IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 7 Issue 7
July 2016

ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI
www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 7 Issue 7 July 2016
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

- **Aamir Shaikh**
- **Abbas Al-Ghaili**
  Mendeley
- **Abbas Karimi**
  Islamic Azad University Arak Branch
- **Abdelghni Lakehal**
  Université Abdelmalek Essaadi Faculté Polydisciplinaire de Larache Route de Rabat, Km 2 - Larache BP. 745 - Larache 92004. Maroc.
- **Abdul Razak**
- **Abdul Karim ABED**
- **Abdur Rashid Khan**
  Gomal Unversity
- **Abeer ELkorany**
  Faculty of computers and information, Cairo
- **ADEMOLA ADESINA**
  University of the Western Cape
- **Aderemi A. Atayero**
  Covenant University
- **Adi Maaita**
  ISRA UNIVERSITY
- **Adnan Ahmad**
- **Adrian Branga**
  Department of Mathematics and Informatics, Lucian Blaga University of Sibiu
- **agana Becejski-Vujaklija**
  University of Belgrade, Faculty of organizational
- **Ahmad Saifan**
  yarmouk university
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**
  Ahlia University
- **Ahmed Nabih Zaki Rashed**
  Menoufia University
- **Ajantha Herath**
  Stockton University Galloway
- **Akbar Hossain**
- **Akram Belghith**
  University Of California, San Diego
- **Albert S**
  Kongu Engineering College
- **Alcinia Zita Sampaio**
  Technical University of Lisbon
- **Alexane Bouënard**
  Sensopia
- **ALI ALWAN**
  International Islamic University Malaysia
- **Ali Ismail Awad**
  Luleå University of Technology
- **Alicia Valdez**
- **Amin Shaqrah**
  Taibah University
- **Amirrudin Kamsin**
- **Amitava Biswas**
  Cisco Systems
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
  Maranatha Christian University
- **Anews Samraj**
  Mahendra Engineering College
- **Anirban Sarkar**
  National Institute of Technology, Durgapur
- **Anthony Isizoh**
  Nnamdi Azikiwe University, Awka, Nigeria
- **Antonio Formisano**
  University of Naples Federico II
- **Anuj Gupta**
  IKG Punjab Technical University
- **Anuranjan misra**
  Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**
  University Technology Malaysia(UTM)
- **Aree Mohammed**
  Directorate of IT/ University of Sulaimani
- **ARINDAM SARKAR**
  University of Kalyani, DST INSPIRE Fellow
- **Aris Skander**
  Constantine 1 University
- **Ashok Matani**
  Government College of Engg, Amravati
- **Ashraf Owis**
  Cairo University
- **Asoke Nath**

CONTENTS

A Vertical Handover Management for Mobile Telemedicine System using Heterogeneous Wireless Networks

Hoe-Tung Yew
School of Engineering,
Universiti Malaysia Sabah,
Sabah,
Malaysia

Eko Supriyanto, M Haikal Satria* and Yuan-Wen Hau
IJN-UTM Cardiovascular Engineering Centre
Faculty of Biosciences and Medical Engineering,
Universiti Teknologi Malaysia
Johor, Malaysia

***Abstract*—Application of existing mobile telemedicine system is restricted by the imperfection of network coverage, network capacity, and mobility. In this paper, a novel telemedicine based handover decision making (THODM) algorithm is proposed for mobile telemedicine system using heterogeneous wireless networks. The proposed algorithm select the best network based on the services requirement to ensure the connected or targeted network candidate has sufficient capacity for supporting the telemedicine services. The simulation results show that the proposed algorithm minimizes the number of unnecessary handover to WLAN in high speed environment. The throughput achieved by the proposed algorithm is up to 75% and 205% higher than Cellular and RSS based schemes, respectively. Moreover, the average data transmission cost of THODM algorithm is 24% and 69.2% lower than the Cellular and RSS schemes. The proposed algorithm minimizes the average transmission cost while maintaining the telemedicine service quality at the highest level in high speed environment.**

***Keywords*—Mobile telemedicine system; vertical handover; heterogeneous networks; unnecessary handover; throughput; cost**

I. INTRODUCTION

The rapid growth of wireless communication technologies has led to the development of telemedicine. Telemedicine provides remote monitoring and diagnosis services via information and communication technologies. Wireless Local Area Network (WLAN), Worldwide Interoperability for Microwave Access (WiMAX) and cellular networks are the three wireless technologies widely applied in telemedicine.

WLAN is the most preferable by users due to the high transmission capacity and low network access cost, however, the small coverage area limits the user's mobility support. WLAN based telemedicine systems are typically for indoor application (home, hospital, clinic, etc.) [1-4]. To tackle the issue in WLAN, researches on cellular network based telemedicine were raised. Authors in [5-8] present a cellular network based telemedicine system. The advantages of cellular based telemedicine system are that it supports high mobility and offers large service coverage. However, the capacity of cellular network is insufficient for high quality images and continuous video transmission as the channel bandwidth is limited. The high bandwidth Fourth Generation Long Term Evolution (4G-LTE) system is still under deployment. The coverage is imperfect in rural and suburban areas. Consequently, the use of 4G-LTE in telemedicine is limited.

The performance of telemedicine service is dependent on the network quality. Poor network quality will disrupt the health data in transmission. Authors in [9, 10] proposed WiMAX based telemedicine application to provide higher bandwidth than cellular network with extended network coverage than WLAN. Authors in [11] integrate both WLAN and WiMAX networks where WLAN is for indoor application and WiMAX is for outdoor environment. However, handover scheme between WLAN and WiMAX is not discussed by authors. WiMAX technology overcomes the issues of small coverage and insufficient bandwidth encountered by WLAN and cellular network, respectively. Unfortunately, most of the network service providers are ceasing development of WiMAX [12]. As a result, the mobile telemedicine system that relies on WiMAX technology cannot guarantee the users connect continuously to the telemedicine services provider at anywhere due to the imperfection of network coverage.

Each wireless technology has its own advantages and disadvantages. None of them can fully support the telemedicine services in terms of data transmission rate and mobility. Therefore, application of existing mobile telemedicine system is sometimes restricted by mobility, coverage and constraints of data transmission rate issues. The continuous service connection and guarantee of data transmission rate are the two main factors to maintain the quality of telemedicine services. For this purpose, a mobile telemedicine system that has capability of accessing multiple wireless networks is needed so that the system has wider service coverage and guarantee the service quality by connecting to the best network based on the services requirement.

The rest of the paper is organized as follow. Section II reviews the existing handover algorithms in heterogeneous networks. In Section III, the framework of mobile telemedicine is introduced. Section IV descries the proposed THODM handover algorithm for mobile telemedicine system. The performance of the proposed algorithm is discussed in Section V. Finally, Section VI concludes the paper.

This research was supported by Ministry of Education Malaysia and Universiti Teknologi Malaysia, Grant Q.J130000.2745.02K17.

II. VERTICAL HANDOVER IN HETEROGENEOUS NETWORKS

Vertical handover in heterogeneous network is a process of the mobile terminal (MT) migrating network connection from one network technology to another. The vertical handover process can be divided into three phases which are handover initiation, decision and execution [13]. The handover initiation phase discovers and obtains available network information via Media Independent Handover Function (MIHF) [14]. The handover decision is a process of selecting the most suitable network based on the calculated network information and triggering handover at the right time. Executive phase is establishing the connection with targeted network and releasing the old network.

Numerous handover decision making algorithms have been previously proposed. A received signal strength (RSS) based handover algorithm introduced by [15, 16] reduces handover failure rate from WLAN to cellular network based on the MT's speed and handover latency. Authors in [17, 18] proposed a prediction based handover decision scheme to estimate MT dwelling time in WLAN coverage. Mobile terminal triggers handover to WLAN cell if and only if the estimated dwelling time is greater than the time threshold. However, high handover delay is observed because these methods need to take two RSS sample points (P_1 and P_2 in Fig. 1) within WLAN coverage for dwelling time estimation process. This processing time will reduce the dwelling time within WLAN as soon as the handover process is done at P_2.

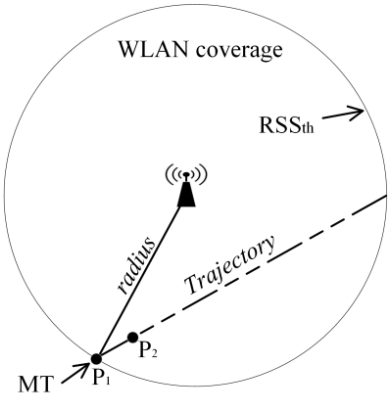


Fig. 1. Scenario of prediction process in [17, 18]

Cost function based handover algorithm for wireless wide area network (WWAN) and WLAN integrated networks has been proposed by [19]. The results showed that WLAN is more preferred than WWAN due to the low network access cost. Also, authors in [20] proposed a cost per signal-to-noise-plus-interference-ratio (SINR) function to improve the throughput and reduce the cost for accessing the integrated wireless networks (WLAN and WWAN). In this approach, authors assumed total cost equal to packet transmission cost plus handover processing cost. However, MT's velocity is not considered by these schemes. The small coverage of WLAN cell will lead to high number of unnecessary handovers in high mobility.

Authors in [21-23] proposed a multiple attribute decision making (MADM) based handover algorithms. In these schemes, a weighting system is given to the handover criteria based on user preference. The network candidate that scores the highest weight sum is selected as a handover target. Recently, intelligent based MADM handover algorithm is presented to improve the performance of handover. A Neuro-Fuzzy based MADM handover algorithm is proposed by [24]. Authors in [25] presented Fuzzy based MADM handover algorithm. Furthermore, Artificial Neural Network (ANN) based handover decision making in heterogeneous networks is presented in [26]. The disadvantage of using intelligent based handover algorithm is high handover latency caused by ANN learning/training, and Fuzzy Logic fuzzification or defuzzification processes. Moreover, handover latency further increases while more handover criteria are taken into account.

In high speed environment, most of the handover algorithms optimize their performance by predefining a speed threshold for WLAN. MT triggers handover to WLAN if and only if MT's speed is below the predefined speed threshold to avoid the handover failure and unnecessary handover to WLAN. The application of WLAN is restricted to static or pedestrian navigation environment [27]. For example, authors in [21, 28-30] predefined the speed threshold for WLAN at 10m/s and below (depending on the radius of WLAN). In addition, Fuzzy MADM based handover algorithm presented by [25] and [31] set the fuzzy if-else rule, "if MT velocity is low then the probability of rejecting WLAN is low; else the probability of rejecting WLAN is high". In this paper, Telemedicine based vertical handover decision making (THODM) algorithm is proposed for mobile telemedicine system aims to maintain the quality of telemedicine service at the highest level with minimum cost in high speed environment.

III. FRAMEWORK OF MOBILE TELEMEDICINE SYSTEM IN HETEROGNEOUS NETWORKS

Fig. 2 shows the proposed mobile telemedicine system framework. The telemedicine device is integrated with various type of electronic health sensors such as pulse oximetry, body temperature and Electrocardiogram (ECG) sensors. The signals or data collected by these sensors will be analyzed by an embedded self-interpretation algorithm [7]. In case an abnormal health signal is detected, the system will set patient health condition (PHC) to "LOW" or "0" and give priority to "Critical" buffer to transmit the abnormal health signal to hospital to let the patient get treatment promptly. The PHC is set to "HIGH" or "1" when the patient is in normal health condition. The health data is stored in the "Non-critical" buffer queue for transmission via WLAN or cellular network.

Telemedicine based handover decision making (THODM) algorithm assists the device to select the best wireless network to transmit the patient's health data to hospital based on the inputs from accelerometer, MIHF, user setting (e.g. video, audio, signal, image, etc.) and predefined database. The function of these modules is illustrated in TABLE I.

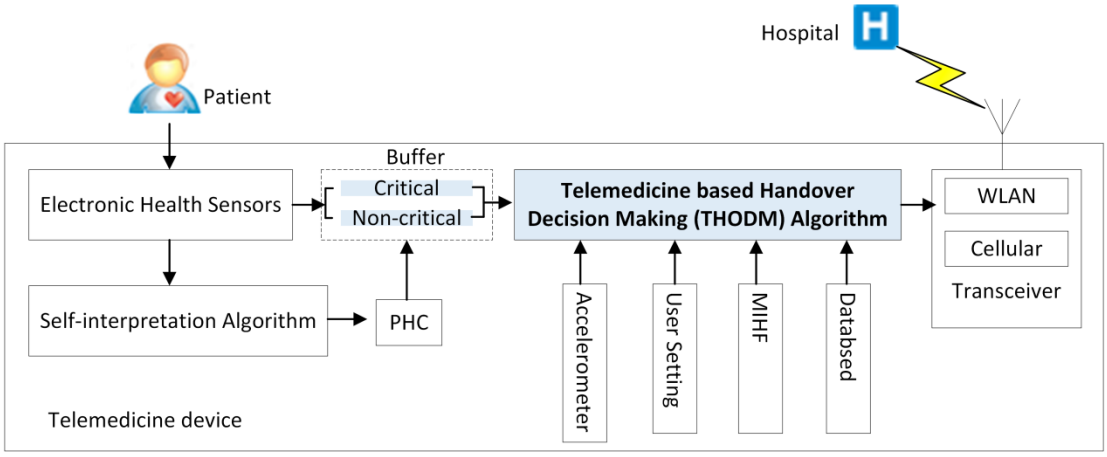


Fig. 2. Mobile Telemedicine Framework

TABLE I. INPUT MODULES OF THODM ALGORITHM

Modules	Function
Accelerometer	Measure the MT traveling speed.
MIHF	Discover neighbouring networks and measure the network quality.
User setting	Monitor type of telemedicine services applied by user.
Database	Contains predefined values such as RSS threshold, network tariff rate, and minimum data rate required by each type of telemedicine service. These values will be the reference values for THODM algorithm during handover decision making.

IV. THODM ALGORITHM

THODM algorithm consists three structured phases of self-inspection, pre-handover filtering and network selection. The handover parameters required by this algorithm are RSS, Signal-to-Noise ratio (*SNR*), data rate requirement (DR_{REQ}), cost (*C*), and velocity (*v*). The RSS and SNR parameters can be obtained via MIHF. The MT's velocity can be measured using an accelerometer. The network access cost and DR_{REQ} are predefined and stored in the database or memory of the telemedicine device.

RSS measurement is used to discover the neighboring wireless networks. The SNR parameter is for algorithm to evaluate the capacity of the available networks. In order to guarantee the quality of telemedicine services, the connected or targeted network candidate must has sufficient capacity to support the data rate required by telemedicine services. The priority level and DR_{REQ} of telemedicine services is shown in Table II.

THODM algorithm also takes network tariff rate into consideration with the purpose of reducing the data transmission cost of telemedicine services. The parameter velocity is used to estimate the dwelling time within WLAN coverage. The energy saving issue is not considered in this work because THODM algorithm is mainly designed for high speed environment such as ambulance or vehicle based telemedicine system. The system could be powered by the power source from ambulance or vehicle.

TABLE II. DATA RATE REQUIRED BY DIFFERENT TYPE OF TELEMEDICINE SERVICES [32, 33]

Data type	Telemedicine service	Data rate (DR_{REQ})	Priority Level
Biosignal data	ECG (12 channels)	24 kbps	1
	Heart rate	2-5kbps	
	Blood pressure	2-5kbps	
Audio	Voice	4-25kbps	2
File transfer	Uncompressed image	30-40Mbytes	3
	Region-of-interest (ROI) JPEG image	15-19Mbytes	
Video	Diagnostic video	768kbps-10Mbps	4

A. Self-inspection Phase

In self-inspection phase, THODM algorithm monitors the RSS of current connected network (RSS_{CCN}), SNR of current connected network (SNR_{CCN}), and DR_{REQ} periodically to ensure the capacity of current connected network fulfills the services requirements. The current connected network (*CCN*) can be WLAN or cellular network. Assumed the cellular technology that integrated in the telemedicine device (Fig. 2) is Universal Mobile Telecommunication System (UMTS) network. The proposed algorithm is continuously searching for WLAN if the current connected network is not a WLAN. THODM algorithm gives priority to WLAN because WLAN provides high capacity with lower cost. The transmission cost can be reduced by optimizing the connection to WLAN. The quality of current connected network (Q_{CCN}) is determined by:

$$Q_{CCN} = F(RSS_{CCN}) * (SNR_{CCN} - SNR_{REQ\_CCN}) \quad (1)$$

where $F(RSS_{CCN})$ is a unit step function as shown in (2) where the output is equal to 1 if RSS_{CCN} is greater than RSS_{CCN} threshold (RSS_{thCCN}), otherwise 0.

$$F(RSS) = F(RSS_{CCN} - RSS_{thCCN}) \begin{cases} 0, & RSS_{CCN} \leq RSS_{thCCN} \\ 1, & RSS_{CCN} > RSS_{thCCN} \end{cases} \quad (2)$$

SNR_{REQ_CCN} is a dynamic SNR threshold defined based on the sum of the data rate of the telemedicine services that

applied by user, DR_{REQ}. It can be calculated by using Shannon-Hartley theorem and given as

$$DR_{REQ} = W_{CCN}\, log_2(1 + SNR_{REQ\_CCN})$$
$$SNR_{REQ\_CCN} = 2^{\frac{DR_{REQ}}{W_{CCN}}} - 1 \quad (3)$$

where W_{CCN} is bandwidth of current connected network. Since SNR of the WLAN and cellular network cannot be compared directly, we standardize the SNR threshold (SNR_{REQ}) for both WLAN and UMTS networks by selecting SNR_{REQ} of WLAN (SNR_{REQ_WLAN}) as a reference value. According to Shannon-Hartley theorem, the DR_{REQ} of WLAN and UMTS channel can be calculated by

$$DR_{REQ} = W_{WLAN}\, log_2(1 + SNR_{REQ\_WLAN}) \quad (4)$$
$$DR_{REQ} = W_{UMTS}\, log_2(1 + SNR_{REQ\_UMTS}) \quad (5)$$

Where SNR_{REQ_UMTS} is SNR_{REQ} of UMTS and W_{WLAN} and W_{UMTS} represent channel bandwidth (Hz) of WLAN and UMTS networks, respectively. Assuming DR_{REQ} for both (4) and (5) are identical, we substitute (4) into (5). The relationship between SNR_{REQ_WLAN} and SNR_{REQ_UMTS} is given as:

$$log_2(1 + SNR_{REQ\_WLAN}) = \frac{W_{UMTS}}{W_{WLAN}} log_2(1 + SNR_{REQ\_UMTS})$$
$$SNR_{REQ\_WLAN} = (1 + SNR_{REQ\_UMTS})^{\frac{W_{UMTS}}{W_{WLAN}}} - 1 \quad (6)$$

By replacing SNR_{REQ_UMTS} in (6) with measured UMTS SNR value (SNR_{UMTS}), we can obtain an equivalent SNR value in WLAN (E_SNR_{UMTS}).

$$E\_SNR_{UMTS} = (1 + SNR_{UMTS})^{\frac{W_{UMTS}}{W_{WLAN}}} - 1 \quad (7)$$

Therefore, set of SNR values for both WLAN and UMTS networks is given by:

$$SNR = SNR_{WLAN} \cup E\_SNR_{UMTS} \quad (8)$$

Rewrite (1),

$$Q_{CCN} = F(RSS_{CCN}) * (SNR - SNR_{REQ\_WLAN}) \quad (9)$$

In the case of $Q_{CCN} \leq 0$ or CCN ≠ WLAN, THODM algorithm scans for neighboring wireless network to find a better network candidate to support the telemedicine services. The detected available network candidates will proceed to the pre-handover filtering phase. Conversely, if no available network is detected, the proposed algorithm will deactivate the lowest priority service systematically in order to adapt to the current connect network [34]. The higher priority services that supportable by current connect network are remaining active with guarantee of service quality.

B. Pre-Handover Filtering Phase

Pre-handover filtering phase consists of dwelling time prediction and network quality evaluation processes. The dwelling time prediction process applies to WLAN cell only whereas the network quality evaluation process is applied to all the available networks. The aim of dwelling time prediction process is to avoid unnecessary handover to WLAN in high speed environment. The proposed dwelling time prediction process predefined two RSS thresholds: RSS boundary ($RSS_{boundary}$) and RSS threshold (RSS_{th}). $RSS_{boundary}$ represents the edge of the WLAN coverage and RSS_{th} denotes minimum RSS for reliable packet delivery. MT initiates prediction process once the measured RSS (RSS_{WLAN}) is greater than $RSS_{boundary}$.

Fig. 3 shows the scenario of MT traveling within the WLAN coverage. MT enters the WLAN coverage at point P_{entry} and exits at point P_{exit}. R is the radius of WLAN cell, r is distance between P_{In_RSSth} and WLAN access point (AP), l is distance between P_{In_RSSth} and P_{Out_RSSth}, and d denotes MT traveling distance from P_{entry} to P_{In_RSSth}. The value of d is varying according to the MT's direction of motion from P_{entry}.

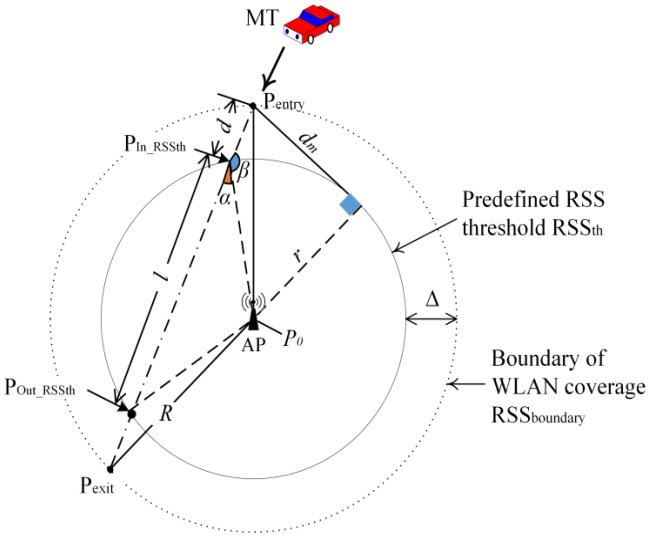


Fig. 3. MT trajectory in WLAN coverage

The distance d is determined by

$$d = t_d * v, \qquad \Delta < d < d_m \quad (10)$$

where t_d is MT's traveling time from P_{entry} to P_{In_RSSth}, and v is MT's velocity. The Δ and d_m are given as

$$\Delta = R - r \quad (11)$$
$$d_m = \sqrt{R^2 - r^2} \quad (12)$$

The R and r values can be calculated by using the Log-distance path loss model [35], expressed as

$$\overline{RSS}_{boundary} = P_{TX} - PL_0 - 10nlog\frac{R}{d_0} + \varepsilon$$
$$R = d_0 * 10^{\left(\frac{P_{TX} - PL_0 - \overline{RSS}_{boundary} + \varepsilon}{10n}\right)} \quad (13)$$

where P_{TX} is AP transmit power, d_0 is the distance from AP to reference point (P_0) usually 1 m, PL_0 is the power loss at P_0, $\overline{RSS}_{boundary}$ is mean of $RSS_{boundary}$, n is path loss exponent, and ε is a zero-mean Gaussian random variable caused by shadow fading. Similarly, the r value can be measured by replacing $\overline{RSS}_{boundary}$ in (13) with $\overline{RSS}_{th}$. The $\overline{RSS}_{boundary}$ can be calculated by

$$\overline{RSS}_{boundary} = \frac{1}{N}\sum_{i=0}^{i=N} RSS_i \quad (14)$$

where N is number of samples. N is adjusted dynamically to the MT's velocity. It given as

$$N = \left\lceil \varphi * \frac{L}{v * T_s} \right\rceil \quad \varphi \in \{0.1, 0.2, \ldots, 0.9\} \quad (15)$$

where T_s is RSS sampling time (5 ms [36]) and L is a fixed distance of 1 m [37]. In this work, MT monitors RSS_{WLAN} periodically in every meter. The higher the MT's velocity is, the smaller the sample size. The maximum N is limited to 20 to prevent over sampling when MT is at low mobility. The impact of Doppler shift can be mitigated by using the Doppler frequency offset estimation and compensation algorithms that presented in [38, 39].

The distance l can be determined by using a trigonometric function as follows:

$$cos\alpha = \frac{l}{2r}$$
$$l = 2rcos\alpha, \quad (16)$$

where $\alpha = \pi - \beta$ (Fig. 3) and r can be calculated by using (13). Rewrite (16),

$$l = -2rcos\beta \quad (17)$$

By using Law of cosine, angle β can be calculated by

$$R^2 = r^2 + d^2 - 2rdcos\beta$$
$$\beta = cos^{-1}\left(-\frac{R^2-r^2-d^2}{2dr}\right) \quad (18)$$

Substitute (18) into (17), estimated traveling distance l is given as

$$l = \frac{R^2-r^2-d^2}{d} \quad (19)$$

The estimated beneficial time to MT within the WLAN coverage (T_{WLAN}) can be determined by

$$T_{WLAN} = \frac{l}{v}$$
$$T_{WLAN} = \frac{R^2-r^2-d^2}{dv} \quad (20)$$

The duration of T_{WLAN} is depending on the R, r, d and v. The WLAN cell which estimated T_{WLAN} is greater than the threshold time (T_{WLAN_th}) proceeds to network quality evaluation process, otherwise rejected. The T_{WLAN_th} for unnecessary handover is 2 seconds [17, 18]. The dwelling time in proposed method will be predicted as soon as MT detected RSS_{th}. This improves the previous method in [17, 18], where the dwelling time prediction is initiated after MT detected RSS_{th}. By using two predefined thresholds, the proposed method reduces the prediction processing time within the dwelling time.

The network quality evaluation process evaluates the quality of all available network candidates except the WLAN cell which estimated T_{WLAN} is less than T_{WLAN_th}. The quality of each network candidate (Q) is evaluated based on the measured RSS, SNR and C values. Q is given as

$$Q_k = \frac{F(RSS_k)(SNR - SNR_{REQ\_WLAN})}{C_k}, \; k = \{1,2,\dots n\} \quad (21)$$

where each network candidate is represented by k of n candidates *and* C_k is predefined cost per Mb of network candidate k.

In this process, the network candidate which scores less than or equal to zero ($Q_k \leq 0$) will be rejected. Only the qualify network candidate that is greater than zero ($Q_k > 0$) will be added to a qualify network list (QNL) for network selection phase.

C. Network Selection Phase

The network selection phase usually falls into three possible conditions. The first condition is that the number of network candidates in QNL (U_{QNL}) is equal to zero. In such case, MT will adjust the services requirement DR_{REQ} by removing the lowest priority service and back to the self-inspection phase. Next condition is only one network candidate in QNL ($U_{QNL}=1$). MT triggers handover to the sole network directly. The last condition is that $U_{QNL} > 1$. Typically, the network candidate which has the highest score will be selected as a handover target or the best network. The best network ($\mathscr{B}$) is given as

$$\mathscr{B} = max(Q_k) \quad (22)$$

V. Results And Discussion

In this section, performance of the THODM algorithm is evaluated by its number of unnecessary handovers, throughput, and cost of transmission. The evaluation is done by simulating the proposed algorithm at high speed environment (50km/h to 120km/h). Fig. 4 shows the simulation scenario where six WLAN cells are covered by a UMTS cellular network. The scenario involves an ambulance or an MT traveling from point A to point B crossing WLAN_1, WLAN_2, WLAN_3 and WLAN_4. The actual traveling distance l within WLAN_1, WLAN_2, WLAN_3 and WLAN_4 is 71.4m, 43.6m, 34.1m and 19.9m, respectively.

The performance of THODM algorithm is compared with the RSS threshold based handover (RSS) algorithm, Cellular based scheme and ideal solution. The RSS threshold based handover algorithm triggers handover to WLAN whenever RSS_{WLAN} is greater than RSS_{th}. Cellular based scheme always connect to the UMTS network. It represents the existing handover algorithms which set the speed threshold for WLAN and only select the WWAN at high mobility. Ideal solution is a theoretical result of MT connection to WLAN and UMTS network while traveling from point A to point B without any unnecessary handover.

It is assumed network providers reserve certain network channels at each UMTS base station and WLAN access point for telemedicine purpose [11]. These reserved channels have average throughput of 1 Mbps [24] and 6 Mbps [19] for UMTS and WLAN, respectively. Therefore, telemedicine user does not have to worry about the network channel availability. The average transmission cost of WLAN and UMTS is 1 and 5 units cost per Mb, respectively [19].

The experiment is simulated for 100 loops. The starting point A is set randomly within the range of $\mathscr{O}$ (as shown in Fig. 4) so that the ambulance or MT has different starting point A in every loop. The experiment parameter settings are shown in Table III.

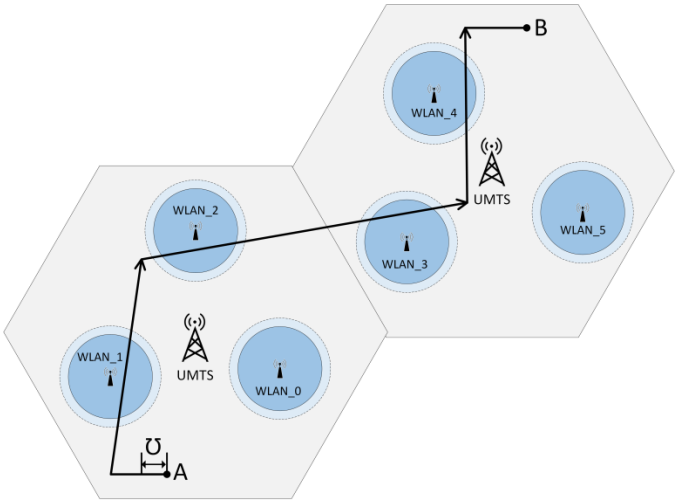


Fig. 4. Simulation scenario in heterogeneous wireless networks (MT travels from point A to point B)

TABLE III. SIMULATION METRICS

Parameter	Value
DR_{REQ} (kbps)	Video + ECG + Voice + Heart rate ≈ 800kbps
P_{TX}	100 mW [20]
n	3.5 [36]
ε	4.3 dB [36]
RSS at boundary, $RSS_{boundary}$	-76.61 dBm
RSS threshold, RSS_{th}	-75.16 dBm
WLAN data rate (Mbps)	6[40]
UMTS data rate (Mbps)	1 [24]
WLAN cost per Mb (unit)	1 [19]
Cellular cost per Mb (unit)	5 [19]
R (m)	55
r (m)	50 [18]
v	50 to 120 km/h
Monitoring time interval (s)	1m / v
T_{WLAN_th} (s)	2 [18]
$\vec{\sigma}$ (m)	Random [0~5]
φ	0.2

A. Unnecessary Handover

Unnecessary handover is defined as a handover that does not benefit the user. It occurs when user failed to establish connection with targeted network due to an abnormal call release or inappropriate handover and the reestablishment of a connection with previous network is required. The number of unnecessary handover (*NUHO*) can be determined by

$$NUHO_a = NHO_a - NHO_{ideal} \quad , \quad a \in \{\text{THODM}, \text{RSS}\} \quad (23)$$

where NHO_{ideal} is total number of handover achieved by ideal solution. In ideal solution, any handovers to WLAN_4 will be considered as unnecessary at the speed of 35km/h and above because T_{WLAN} in WLAN_4 (T_{WLAN_4}) is less than T_{WLAN_th}. Similar to WLAN_3, WLAN_2 and WLAN_1, an unnecessary handover occurs when the velocity of MT is higher than 61, 78 and 128 km/h, respectively. Fig. 5 and Fig. 6 show the total number of handover and unnecessary handover occurred in THODM, RSS and ideal solution schemes.

The result in Fig. 6 shows that THODM algorithm has less number of unnecessary compared to RSS scheme. This is contribution of dwelling time prediction process which rejects all the WLAN cells that estimated T_{WLAN} is less than T_{WLAN_th} even though these WLAN cells have better network quality than UMTS.

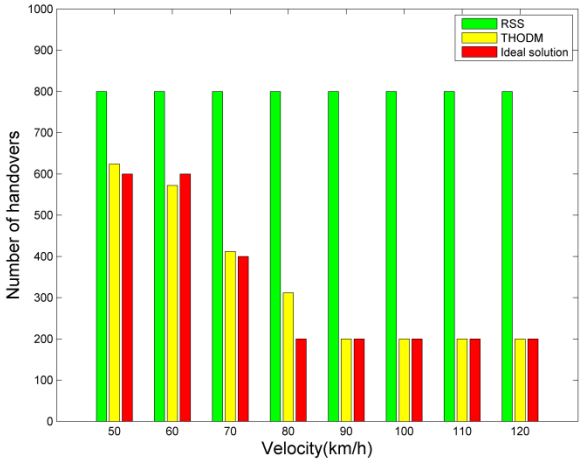


Fig. 5. Total number of handovers

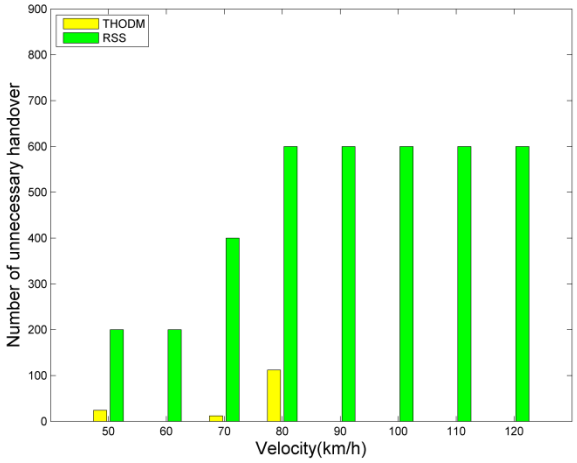


Fig. 6. Number of unnecessary handovers

B. Throughput

The total throughput achieved by MT is dependent on the connection time with UMTS and WLAN, respectively. However, it is affected by the number of unnecessary handovers. The total throughput ($T_{Throughput}$) achieved by MT can be determined by [24]

$$T_{Throughput} = (R_{cellular}t_{cellular} + R_{wlan}t_{wlan}) - \frac{(R_{cellular}+R_{wlan})*NUHO*T_{UHO}}{2} \quad (24)$$

where T_{UHO} is time consumed by each unnecessary handover (2 seconds) [18], $R_{cellular}$ and R_{wlan} represent average data rate of UMTS and WLAN, $t_{cellular}$ and t_{wlan} denotes total time connected to UMTS and WLAN, respectively.

Fig. 7 shows the average throughput achieved by MT based on different approaches (Cellular, RSS, THODM, and ideal solution) in single loop (from point A to point B) at speed of 50km/h to 120km/h. The average throughput decreases when the MT's velocity increases. This is because MT takes less time to travel from point A to B. It can be seen that the total throughput obtained by the proposed THODM algorithm is higher than RSS and Cellular schemes. Furthermore, the throughput achieved by the THODM algorithm is proximate to ideal solution.

The percentage of throughput gain (G) can be determined by

$$G = \frac{TP_{THODM} - TP_x}{TP_x} \times 100\% \ , \qquad x \in \{\text{Cellular}, \text{RSS}\} \quad (25$$

where TP_{THODM} is total throughput achieved by THODM and TP_x represents total throughput of RSS or Cellular based scheme. As depicted in Fig. 8, the throughput of THODM is up to 75% and 205% higher than Cellular and RSS schemes, respectively.

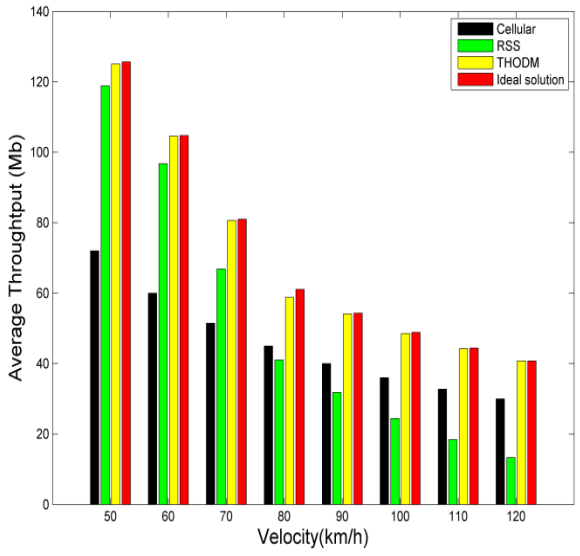


Fig. 7. Average throughput achieved by MT in single loop at the speed of 50km/h to 120km/h.

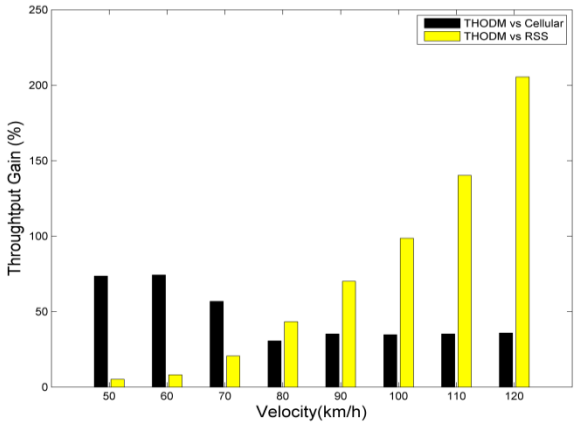


Fig. 8. Percentage of throughput gain

C. Transmission Cost

Assumed the transmission cost of UMTS is five times higher than WLAN [19]. The transmission cost per Mb (C) can be calculated by

$$C = \frac{R_{cellular}t_{cellular}C_{cellular} + R_{wlan}t_{wlan}C_{wlan} + N_{HO}C_{HO}}{T_{Throughput}} \quad (26)$$

where N_{HO} is number of handover, C_{HO} represents handover cost (predefined C_{HO} = 3 units), C_{wlan} and $C_{cellular}$ denote average cost per Mb offered by WLAN and cellular network. As can be seen in Fig. 9, THODM algorithm has the lowest average cost per Mb compared to RSS and Cellular schemes. At speed of 120 km/h, the average cost of THODM is 24% and 69.2% lower than Cellular and RSS schemes, respectively.

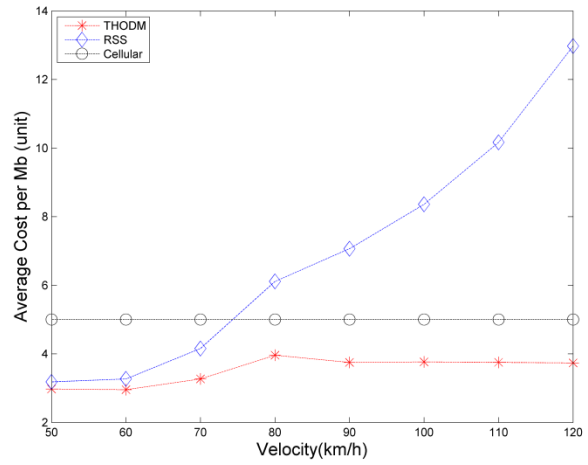


Fig. 9. Average cost per Mb

VI. CONCLUSION

This paper presented a THODM algorithm to support telemedicine service in high speed heterogeneous environment. The proposed algorithm resolves the problems such as limited coverage and mobility issue faced by the existing mobile telemedicine system by selecting the best network according to the services requirement. The dwelling time prediction process in THODM algorithm has successfully reduced the number of unnecessary handovers while optimizing the usage of WLAN in high speed environment. The simulation results show that the proposed algorithm has higher throughput and more cost effective than RSS and Cellular schemes. The proposed THODM algorithm is suitable for ambulance based mobile telemedicine system.

The limitation of this work is that we assume the MT moves at constant speed when crossing the WLAN coverage. For future work, we will enhance the dwelling time prediction method in THODM algorithm so that it can accurately estimate the MT's dwelling time in WLAN coverage even though MT moves in dynamic speed.

REFERENCES

[1] J. C. Tejero-Calado, C. Lopez-Casado, A. Bernal-Martin, M. A. Lopez-Gomez, M. A. Romero-Romero, G. Quesada, et al., "IEEE 802.11 ECG monitoring system," in Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, 2005, pp. 7139-7142.

[2] A. Moein and M. Pouladian, "WIH-Based IEEE 802.11 ECG Monitoring Implementation," in Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007, pp. 3677-3680.

[3] P. A. Sakthi and R. Sukanesh, "A reliable and fast routing data transmission protocol for Wi-Fi based real-time patient monitoring system," in Electronics and Communication Systems (ICECS), 2014 International Conference on, 2014, pp. 1-5.

[4] K. Cai and X. Liang, "Development of WI-FI Based Telecardiology Monitoring System," in Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on, 2010, pp. 1-4.

[5] S. Mitra, M. Mitra, and B. B. Chaudhuri, "Rural cardiac healthcare system-A scheme for developing countries," in TENCON 2008 - 2008 IEEE Region 10 Conference, 2008, pp. 1-5.

[6] M. Abo-Zahhad, S. M. Ahmed, and O. Elnahas, "A wireless emergency telemedicine system for patients monitoring and diagnosis," Int J Telemed Appl, vol. 2014, p. 380787, 2014.

[7] H. Anpeng, C. Chao, B. Kaigui, D. Xiaohui, C. Min, G. Hongqiao, et al., "WE-CARE: An Intelligent Mobile Telecardiology System to Enable mHealth Applications," Biomedical and Health Informatics, IEEE Journal of, vol. 18, pp. 693-702, 2014.

[8] H. Mateev, I. Simova, T. Katova, and N. Dimitrov, "Clinical Evaluation of a Mobile Heart Rhythm Telemonitoring System," ISRN Cardiology, vol. 2012, p. 8, 2012.

[9] I. Chorbev and M. Mihajlov, "Wireless telemedicine services as part of an integrated system for e-medicine," in Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean, 2008, pp. 264-269.

[10] D. Niyato, E. Hossain, and J. Diamond, "IEEE 802.16/WiMAX-based broadband wireless access and its application for telemedicine/e-health services," Wireless Communications, IEEE, vol. 14, pp. 72-83, 2007.

[11] D. Niyato, E. Hossain, and S. Camorlinga, "Remote patient monitoring service using heterogeneous wireless access networks: architecture and optimization," Selected Areas in Communications, IEEE Journal on, vol. 27, pp. 412-423, 2009.

[12] D. Pareit, B. Lannoo, I. Moerman, and P. Demeester, "The History of WiMAX: A Complete Survey of the Evolution in Certification and Standardization for IEEE 802.16 and WiMAX," Communications Surveys & Tutorials, IEEE, vol. 14, pp. 1183-1211, 2012.

[13] I. F. Akyildiz, J. McNair, J. S. M. Ho, H. Uzunalioglu, and W. Wenye, "Mobility management in next-generation wireless systems," Proceedings of the IEEE, vol. 87, pp. 1347-1384, 1999.

[14] M. Aguado, J. Astorga, J. Matias, and M. Huarte, "The MIH (Media Independent Handover) Contribution to Mobility Management in a Heterogeneous Railway Communication Context: A IEEE802.11/802.16 Case Study," in Communication Technologies for Vehicles. vol. 6596, T. Strang, A. Festag, A. Vinel, R. Mehmood, C. Rico Garcia, and M. Röckl, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 69-82.

[15] S. Mohanty and I. F. Akyildiz, "A Cross-Layer (Layer 2 + 3) Handoff Management Protocol for Next-Generation Wireless Systems," Mobile Computing, IEEE Transactions on, vol. 5, pp. 1347-1360, 2006.

[16] S. Mohanty, "A new architecture for 3G and WLAN integration and inter-system handover management," Wireless Networks, vol. 12, pp. 733-745, 2006/12/01 2006.

[17] R. Hussain, S. Malik, S. Abrar, R. Riaz, H. Ahmed, and S. Khan, "Vertical Handover Necessity Estimation Based on a New Dwell Time Prediction Model for Minimizing Unnecessary Handovers to a WLAN Cell," Wireless Personal Communications, vol. 71, pp. 1217-1230, 2013/07/01 2013.

[18] Y. Xiaohuan, N. Mani, and Y. A. Sekercioglu, "A Traveling Distance Prediction Based Method to Minimize Unnecessary Handovers from Cellular Networks to WLANs," Communications Letters, IEEE, vol. 12, pp. 14-16, 2008.

[19] K. Hong, S. Lee, L. Kim, and P. Song, "Cost-based vertical handover decision algorithm for WWAN/WLAN integrated networks," EURASIP J. Wirel. Commun. Netw., vol. 2009, pp. 1-11, 2009.

[20] Y. Kemeng, I. Gondal, and Q. Bin, "Multi-Dimensional Adaptive SINR Based Vertical Handoff for Heterogeneous Wireless Networks," Communications Letters, IEEE, vol. 12, pp. 438-440, 2008.

[21] E. M. Malathy and V. Muthuswamy, "Knapsack - TOPSIS Technique for Vertical Handover in Heterogeneous Wireless Network," PLoS ONE, vol. 10, p. e0134232, 2015.

[22] R. Tawil, G. Pujolle, and O. Salazar, "A Vertical Handoff Decision Scheme in Heterogeneous Wireless Systems," in Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE, 2008, pp. 2626-2630.

[23] S.-m. Liu, S. Pan, Z.-k. Mi, Q.-m. Meng, and M.-h. Xu, "A Simple Additive Weighting Vertical Handoff Algorithm Based on SINR and AHP for Heterogeneous Wireless Networks," in Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on, 2010, pp. 347-350.

[24] A. Singhrova and N. Prakash, "Vertical handoff decision algorithm for improved quality of service in heterogeneous wireless networks," Communications, IET, vol. 6, pp. 211-223, 2012.

[25] F. Kaleem, A. Mehbodniya, K. K. Yen, and F. Adachi, "A Fuzzy Preprocessing Module for Optimizing the Access Network Selection in Wireless Networks," Advances in Fuzzy Systems, vol. 2013, p. 9, 2013.

[26] N. Nasser, S. Guizani, and E. Al-Masri, "Middleware Vertical Handoff Manager: A Neural Network-Based Solution," in Communications, 2007. ICC '07. IEEE International Conference on, 2007, pp. 5671-5676.

[27] W. Song and W. Zhuang, "Introduction on Cellular/WLAN Interworking," in Interworking of Wireless LANs and Cellular Networks, ed: Springer New York, 2012, pp. 1-10.

[28] M. Khan and K. Han, "An Optimized Network Selection and Handover Triggering Scheme for Heterogeneous Self-Organized Wireless Networks," Mathematical Problems in Engineering, vol. 2014, p. 11, 2014.

[29] H. T. Yew, E. Supriyanto, M. Haikal Satria, and Y. W. Hau, "User-centric based vertical handover decision algorithm for telecardiology application in heterogeneous networks," Jurnal Teknologi, vol. 77, pp. 79-83, 2015.

[30] T. Janevski and K. Jakimoski, "Mobility sensitive algorithm for vertical handovers from WiMAX to WLAN," in Telecommunications Forum (TELFOR), 2012 20th, 2012, pp. 91-94.

[31] H. T. Yew, Y. Aditya, H. Satrial, E. Supriyanto, and Y. W. Hau, "Telecardiology system for fourth generation heterogeneous wireless

networks," ARPN Journal of Engineering and Applied Sciences, vol. 10, pp. 600-607, 2015.

[32] J. R. Gallego, A. Hernandez-Solana, M. Canales, J. Lafuente, A. Valdovinos, and J. Fernandez-Navajas, "Performance analysis of multiplexed medical data transmission for mobile emergency care over the UMTS channel," IEEE Trans Inf Technol Biomed, vol. 9, pp. 13-22, Mar 2005.

[33] C. Yuechun and A. Ganz, "A mobile teletrauma system using 3G networks," Information Technology in Biomedicine, IEEE Transactions on, vol. 8, pp. 456-462, 2004.

[34] H. T. Yew, E. Supriyanto, M. H. Satria, and Y. W. Hau, "Adaptive network selection mechanism for telecardiology system in developing countries," in 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2016, pp. 94-97.

[35] P. Santi, "Modeling Next Generation Wireless Networks," in Mobility Models for Next Generation Wireless Networks, ed: John Wiley & Sons, Ltd, 2012, pp. 19-32.

[36] Y. Xiaohuan, Y. A. Sekercioglu, and N. Mani, "A method for minimizing unnecessary handovers in heterogeneous wireless networks," in World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a, 2008, pp. 1-5.

[37] B. Singh, "An improved handover algorithm based on signal strength plus distance for interoperability in mobile cellular networks," Wireless Personal Communications, vol. 43, pp. 879-887, 2007.

[38] Y. Yang, P. Fan, and Y. Huang, "Doppler frequency offsets estimation and diversity reception scheme of high speed railway with multiple antennas on separated carriages," in Wireless Communications & Signal Processing (WCSP), 2012 International Conference on, 2012, pp. 1-6.

[39] E. A. Feukeu, K. Djouani, and A. Kurien, "Compensating the effect of Doppler shift in a vehicular network," in AFRICON, 2013, 2013, pp. 1-7.

[40] Z. Yan, N. Ansari, and H. Tsunoda, "Wireless telemedicine services over integrated IEEE 802.11/WLAN and IEEE 802.16/WiMAX networks," Wireless Communications, IEEE, vol. 17, pp. 30-36, 2010.

A New Approach for Improvement Security against DoS Attacks in Vehicular Ad-hoc Network

Reza Fotohi
Young Researchers and Elite Club
Germi Branch, Islamic Azad University
Germi, Iran

Yaser Ebazadeh
Department Of Computer Engineering
Germi Branch, Islamic Azad University
Germi, Iran

Mohammad Seyyar Geshlag
Department Of Computer Engineering
Shabestar Branch, Islamic Azad University
Shabestar, Iran

***Abstract*—Vehicular Ad-Hoc Networks (VANET) are a proper subset of mobile wireless networks, where nodes are revulsive, the vehicles are armed with special electronic devices on the motherboard OBU (On Board Unit) which enables them to trasmit and receive messages from other vehicles in the VANET. Furthermore the communication between the vehicles, the VANET interface is donated by the contact points with road infrastructure. VANET is a subgroup of MANETs. Unlike the MANETs nodes, VANET nodes are moving very fast. Impound a permanent route for the dissemination of emergency messages and alerts from a danger zone is a very challenging task. Therefore, routing plays a significant duty in VANETs. decreasing network overhead, avoiding network congestion, increasing traffic congestion and packet delivery ratio are the most important issues associated with routing in VANETs. In addition, VANET network is subject to various security attacks. In base VANET systems, an algorithm is used to dicover attacks at the time of confirmation in which overhead delay occurs. This paper proposes (P-Secure) approach which is used for the detection of DoS attacks before the confirmation time. This reduces the overhead delays for processing and increasing the security in VANETs. Simulation results show that the P-Secure approach, is more efficient than OBUmodelVaNET approach in terms of PDR, e2e_delay, throughput and drop packet rate.**

Keywords*—component; VANET; P-Secure Protocol; DoS Attack; detection; OBUmodelVaNET; security

I. INTRODUCTION

VANETs is particular, MANETs by which where vehicles and fixed location at the roadside can keep in touch speak with each. It can self-structure, spread comfortably and cost low with open structures. VANET can pleasure an increasingly significant role in multiple regions: when an event an episode occurs, it sends the procedure message speed to other cars besides the procedure regions that actually help to prevent crashes again; cars catch vehicle velocity, density status of several roads, and they can they are able to take actions ahead of time which facilitates traffic congestion; also, cars can surf the internet via the fixed stations at the roadside. As a result of more and more apps, VANET has turned into an into the focus of research institutes and scientists worldwide [1-5]. Each year there are more and more traffic jams on the roads. This is large due to every year there are more and more cars on the streets so that in 2013 there will be1410 a million vehicle in the world. It is feasible to find conditions where communications between cars can help to prevent accidents. In this approach, cars can help with these questions every week. This can prevent great wastes of time, money and of oil reserves, in addition, governments spend lots of money and destroy the landscape when creating more roads because existing roads do not support the generated traffic. The resultant, a restructuring of traffic can prevent some of the aforementioned dilemmas. There are many points to consider in any wireless networks in overall. The Figure following gives an illustration of VANETs.

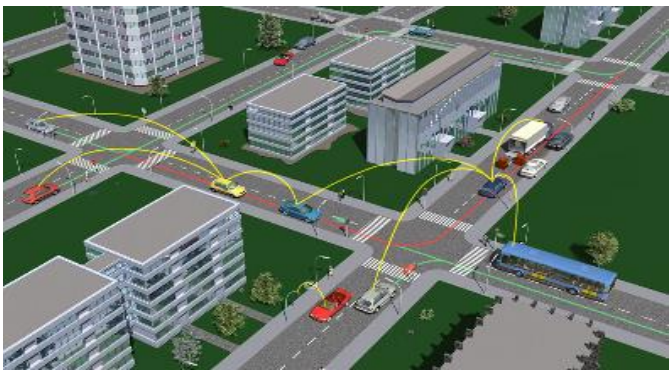


Fig. 1. An example of VANETs senario

In VANET's safety is a significant issue that needs to be taken into account when any wireless network is designed. VANETs have weaknesses to various kinds of DoS attacks [6], Blackhole, gray hole and wormhole portion of the region of the safety problems existent in this kind of networks. Nevertheless, an another approach is presented in this paper. The main purpose of this work is to define a proposed approach, a scalable free system for VANETs where users can cooperate via their mobile technology and obtain updated information of interest about the traffic and attacks area in order to choose the best-refreshed path to their goals. in this paper, we proposed P-Secure approach detection algorithm is that attacks, for detecting DOS attacks used to commit time. This decreases the overhead for processing and securing the VANET is delayed. This approach has better special depending criteria removal rate, throughput, PDR and latency to develop e2e_delay, Therefore, this work proposes a self-managed VANET without any infrastructure, which will serve as an introduction to a more complex VANET, all this with better levels of security. In this paper, we implement the proposed P- Secure approach

as the solution in NS-2 simulator to test its performance it. In the second section of paper, related works are presented. Afterward, section 3, Dos Attack in Vehicular Ad-hoc Network (VANETS), section 4, The proposed method, Section 5 Experimental Data and Analysis, Finally, in section 6 conclusion.

II. Related Work

Security in the network is of specific problems due to man lives are permanently at the condition as in traditional networks the major security concerns include confidentiality, integrity, and availability none of which involves primely with life security. Essential information can't be either changed nor deleted by a malicious node. Yet, security in the network also contains the ability to specify the driver responsibility while maintaining driver privacy. Information about the car and their drivers within must be swapped securely and more importantly, timely in that the delay of message exchange may cause catastrophic consequences such as the collision of vehicles. The spread of a general security model for the network is very challenging in practice.

With its dynamic characteristics and high mobility, the usage of wireless technology also makes VANET vulnerable to DoS attacks that exploit the open and broadcast characteristics of wireless networks. [7] Cryptographic attacks in VANET are classified in the next section. Further common networks security problem, unique security challenges arise because of the unique characteristics of VANET such as high mobility, dynamic topology, short connection duration and frequent disconnections. These novel characteristics bring safety concerns such as trust group formation, location detection, and security as well as certificate management. Corresponding preview work will be given in following sections based on the characteristics of the protection issue in similar work. The clustering system has been well thought in wireless technology in recent years [8]. Nevertheless, considering the natures of VANETs, such as sufficient energy, high speed, the clustering methods proposed for conventional wireless networks are not proper for VANETs. Hence, the clustering approach for VANETs should be designed exactly. The lowest ID clustering algorithm [9] is one of the easiest methods to cluster mobile nodes for VANETs. Using this method, all of the nodes broadcast becomes stages in which the node IDs are encapsulated. Further, these nodes IDs are assigned uniquely. In [10], authors proposed a clustering approach using affinity propagation for VANETs. Affinity propagation is first proposed to solve data clustering problem and it is show that this algorithm can generate clusters more efficiently compared with traditional solutions. The node which has the lowest ID in its neighborhood is selected as the cluster head node, and other nodes are selected as the cluster member nodes. The lowest ID algorithm proposed a basic idea to cluster mobile nodes. First, we need to define a metric to model the property of wireless nodes; and then we can use the metric to group nodes based on some rules. The follow- ing clustering schemes are all based on this basic idea. The difference of various clustering scheme is the metrics used for modeling. Density Based Clustering (DBC) algorithm is proposed in [12]. Using DBC, connectivity level, link quality and traffic conditions are taken into account completely to cluster vehicle nodes. The mobile network is divided into the dense part and sparse part. A node which has links more than a predefined value is considered as in the dense part; otherwise, it is in the sparse part. During the clustering process, link quality is estimated to make a re-clustering decision. According to the experiment results, the cluster head change ratio is less than the lowest ID algorithm [13]. So, there are some works devoted to design and develop specific simulators of VANET. Groove Net [14] is a hybrid simulator which enables communication between simulated vehicles, real vehicles and between real and simulated vehicles, and it models inter-vehicular communication within a real street map-based topology which is based TIGER map data. MOVE and Translated [15] can rapidly generate realistic movement model which can be directly used in network simulators such as NS2, SUMO. VG Sim [16] combines movement model of vehicles and network simulation and transforms vehicular moves and applications to events for further processing of network simulators. NCTUns [17][18] Different kinds of attacks have been analyzed in MANET and their effect on the network. Attack such as grayhole, where the attacker node behaves maliciously for the time until the packets are dropped and then switch to their normal behavior [19]. MANET's routing protocols are also being exploited by the attackers in the form of flooding attack, which is done by the attacker either by using RREQ2 or data flooding [20]. Design and presentation of different security obstacles and attacks in mobile ad hoc networks as well as finding appropriate solutions to them is a challenging research area for researchers. Black hole attack is one of the famous related attacks. In [11], the idea of affinity propagation is used to cluster vehicle nodes in a distributed manner. The vehicle nodes exchange messages with their neighbor nodes to transmit availability and responsibility and make the decision based on the availability and responsibility values for constructing clusters. The simulation results demonstrate that the performance of the clustering scheme using affinity propagation is better than MOBIC in terms of stability.

III. DoS Attack in Vehicular Ad-Hoc Networks

In a VANETs, usually the attacker attacks the communication medium to cause the channel jam or to create count obstacles for the nodes from approachability the network. The essential purpose is to forbid the authentic nodes from approachability the network services and from using the network sources. The attack would result in failure of the nodes and VANET sources. Finally, the VANETs are no longer available to legitimate nodes. In VANETs, DoS should’nt be permissible to happen, where seamless life critical information must reach its intended destination securely and timely. In short, there are 3 routes the attackers could get DoS attacks, scilicet communication channel, network overloading, and dropping the packets [8]. In calculating, a DoS attack is an attempt to make a system or VANET source unavailable to its intended users, like to temporarily cut off or suspend tasks of a host connected to the network. A DDoS is where the attack source is more than one, often hundreds of, unique IP. It is similar to a set of people crowding the entry door and not letting legitimate parties enter into the store, disrupting ordinary tasks. sinful perpetrators of DoS attacks mostly target services hosted on high-profile web servers such as banks,

credit card payment gateways; but motives of revenge, blackmail or activism can be behind other attacks. in Dos Attackers diffuse false information to affect the behavior of other drivers In Figure 2 is shown an example of DOS attacks.

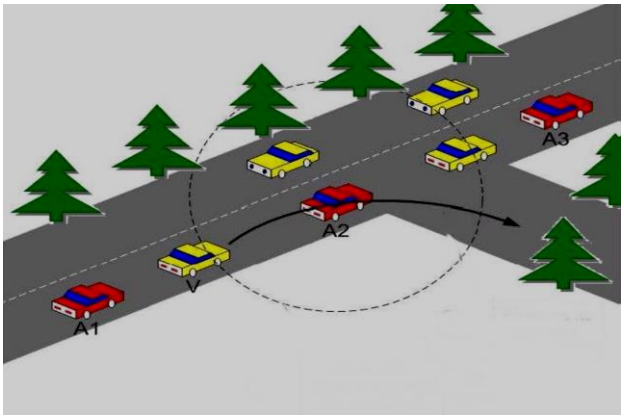


Fig. 2. An example of the problems in the Dos attacks in VANET

In Figure 2, A2 sends a false message with traffic info, and V changes its route and frees the road, in conclusion, Attackers tracks vehicles to obtain those drivers' private information ,and routing disrupts DOS efforts of to influence the sentences are as follows:

A.	Types of outbursts network include TCP, UDP, and ICMP that disrupt legitimate traffic site
B.	Trying to disconnect the machine and thus not being able to use their service
C.	Trying to prevent a particular individual from accessing a service.
D.	Trying to sabotage the service to a specific system or person
E.	Trying to disrupt the hair routing network

IV. The Proposed Method (details)

The essential aim of the VANETs is to supply security and welfare for the travelers. To achieve this object, a specific electronic device is embedded in any vehicle that makes it possible to establish communication between the passengers. Such a network must be implemented without client-server network limits of communication structures. Each vehicle armed with a VANET machinery is a node in a mobile wireless network and is able to receive and send others' messages via the wireless network. Traffic alerts, road signs and traffic observation for a moment that can be transmitted through such a network, provide the necessary tools to make decisions about the best path for the driver. In this paper we proposed a approach in a vehicle network improves road safety, transport efficiency, but also reduces the impact of transport on the environment; all three of these applications are not perfectly perpendicular to each other. For example, reducing the number of accidents, in turn, can reduce the traffic congestion and this can lead to a reduction of the environmental impact.

A. The proposed approach

At first, we must have a system model to express, the proposed method based on which we know the position of the vehicles and roadside equipment as well as antennas etc. We introduce our proposed method with the name P -Secure Protocol that has been named as P-SP in the Figure.

B. The proposed system model

Figure (3) indicates a road that has radio transmissions and vehicles that have onboard radio. RSRU (roadside radio unit) decides to depend on its transmission range and set up in the area where the vehicles can form a network. In fact, we consider it as a threshold. The vehicles can send messages to RSRU through the proposed P-Secure Protocol mechanism. In this way, detection of the exact position of the vehicle that has sent the message is conducted. After discovering the situation, the vehicle's data are saved in some RSRU. Any vehicle with OBU and anti-tampering sensors is (Tamper PROOF). These devices have the responsibility of storing the accurate information about the vehicle like speed, location and more. The position of the vehicles is obtained through the frequency and speed of vehicles and the use of OBU. Vehicles can use the P-Secure Protocol mechanism to request from RSRU. In fact, RSRU conducts the approval of vehicles and maintains the database. Location, time, and etc. along with the data package are provided to RSRU. Traffic devices use the requests using another database and provide the service responses only to the radio transmitter which has been approved, therefore causes the reduction of DOS attacks which could be created due to Flooding.

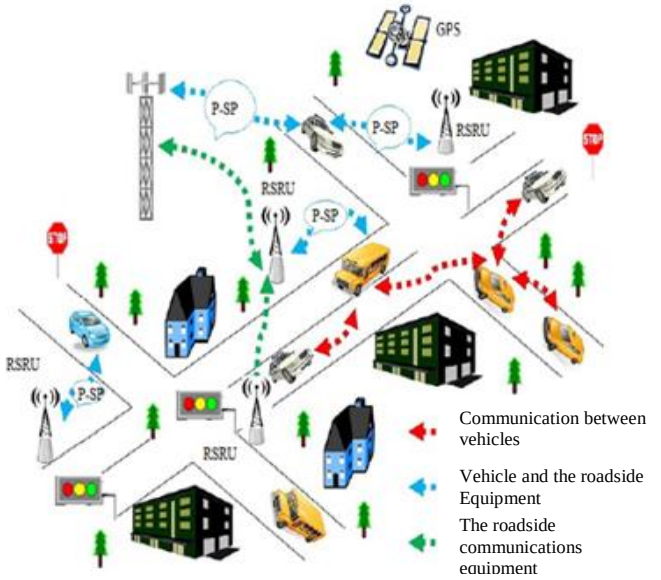


Fig. 3. The proposed system model (RSRU range of vehicles and a process in which the request is sent)

C. The phase 1 of the P-Secure approach

Proposed P-Secure, discovers the vehicle location and information packets that the vehicles have sent and if the packet was not related to the attack, the vehicle is not detected. To conduct the algorithm, it is required to consider a number of variables:

Definition 1: We consider a maximum packet capacity and display it with M where we consider the value 20 for our work.

Definition 2: p parameter is the number of packets sent to RSRU per second.

Definition 3: α is the coefficient which is determined by the characteristics of the road.

Definition 4: V_m is the maximum speed of the vehicle.

Definition 5: V is the speed of the vehicle.

In addition to the maximum speed, a minimum speed is also required that it demonstrated by low. To obtain the number of packets that are sent by the RSRU vehicles per second (P) the Equation (1) is used:

$$P = \alpha * \left| \frac{V - vm}{2} \right| \quad (1)$$

Since the number of packets and the maximum speed is higher than the nod speed, the position of the vehicle is changing rapidly. We consider this as an attack and if the speed is too low, the vehicle's position will not change much and we consider this as an attack. For the attack detection algorithm, we act as the following steps:

1. First we get the required information from the vehicles
2. We set the threshold for RSRU
3. The confirmation requests are sent to RSRUs by the vehicles.
4. A time stamp is expressed for each device.
5. If (the time stamp of the transmitter – time stamp of the receiver) is greater than the threshold value, the packet is discarded, otherwise the package is acceptable.
6. Have the value of M equal to 20 and find the p value for all packets less than or equal to 20.
7. If the p value is greater than or equal to 20 and the v value is greater than or equal to the maximum speed (v_m), the packet is then discarded and the packet is detected as an attack.
8. The next mode of attack is that: if the p-value is less than or equal to 20 and the v value is less than or equal to the minimum speed (low), then the packet is discarded and the packet is detected as an attack.

D. phase 2 of the P-Secure approach

In phase 2 of the P-Secure approach, we detected attacks the packet capacity and/or their speed has been more or less than the minimum whereas the probability of DOS attack also exists in the delivery of any vehicle. To detect these attacks, we improve the proposed algorithm and add the following steps to it:

P-Secure approach discovers DoS attack before the confirmation stage. Location, time stamp, speed and etc. of the vehicle are considered to find out whether it is within the radar range or not. The proposed algorithm is also used in detecting the false alerts. If the count of packets and the MAX speed are above the node's speed, it is considered as an attack on the position of the vehicle is changing quickly. Likely, if speed is too low, the vehicle's position will not change much and this is also considered as an attack. After completing the process of valid vehicles, they are stored in the RSRU database.

TABLE I. REQUAREED PARAMETERS

p	The number of packets per second
M	The maximum capacity packet
α	Coefficient
v_m	Maximum speed
V	Speed
low	Minimum speed

1. The confirmed vehicles are buffered at previous stages in the RSRU
2. Any vehicle wishing to enter the VANET network must submit its application RSRU
3. RSRU updates its' counter for that vehicle.
4. Allocating time slot to the devices is approved by RSRU
5. RSRU counts the number of steps of the vehicles; if the number of steps is equal to the number of counting times, the counter then updates the vehicle's next step and declares the vehicle as valid.
6. The new requests are assessed.
7. If the vehicle has a new request in the same time slot and/or the request number of this vehicle is more than the other vehicles, it is identified as malicious.
8. Indicates the number of the damaging vehicle as the malicious.

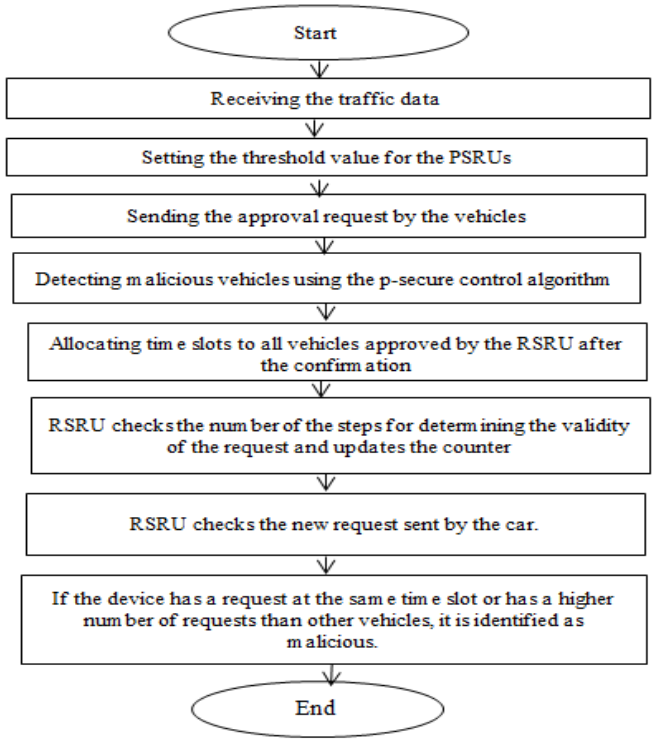


Fig. 4. Flowchart of the P-Secure approach

The phase 2 of the approach is to confirm the new request that wants to join the network. This algorithm compares the previous valid database with new requests and reduces false requests through allowing valid nodes.

The proposed P -Secure Protocol algorithm reduces attacks by limiting the counter and also not allowing the fake vehicles by the attackers. The flowchart of the proposed method is fully displayed in Figure (4).

V. Simulation Result And Analysis

This part contains evaluation of P-Secure approach compared with OBU model VANET approach. With the help of the NS-2 we were able to prove, ns is simulator project, start 1989 as a different of real. We run two senario, one P-Secure approach, and OBUmodelVaNET approach, we have repeated the experiments by changing the several times. To, 200, 400, 800, 1000, and 1200. The simulation parameter are shown in table 2 the parametess used to implementation the performance are given follows.

TABLE II. Simulation Parameters

Simulator	**NS2.34**
Area	1000m X1000m
Density	150-20
Transmission Range	250m
Antenna	Omni Antenna
Simulation duration	200,400,600,800,1000,1200
MAC Layer	802_11
Traffic Type	CBR (UDP)
Buffer Size	150 Packet
Node placement	Random
Simulation methods	*OBUmodelVaNET* AND *P-Secure approach*

As we saw in the previous section, DOS attack causes weakness in making the network unsecure and weakening the ordinary approach of the network, in this article we deal with providing a solution for security improvement in car networks against denial-of-service attacks which improves the standard end-to-end delay, packet delivery ratio, the number of drop packets and throughput.

A. PDR

PDR is define as shown in $Equation$ (2).

$$PDR = \frac{\text{Number of sent Packets}}{\text{Number of received Packets}} * 100 \quad (2)$$

Where PDR is the package delivery rate, SendPacketNo is the number of sent packages, and RecievePacketNo denotes the number of received packages.

B. Packet Drop Rate

It is the measure of the number packets dropped due to malicious node (DoS attack). Thus, we can define $Thro$ as shown in $Equation$ (3).

$$Drop = \frac{\text{Send Packet} - \text{Received Packet}}{\text{Send Packet}} \quad (3)$$

C. End to End Delay

End-to-end delay refers to the time taken for a packet to be transmitted Around the Network from source to destination.

D. Throughput

Throughput is the number of data packets transmited from source node to destination node [21]. Thus, we can define $Thro$ as shown in $Equation$ (4).

$$Thro = \frac{Request}{Time} \quad (4)$$

Where **Thro variable** is the throughput, **requests** is the number of requests that are accomplished by the system, and **time** shows the total time of system observation.

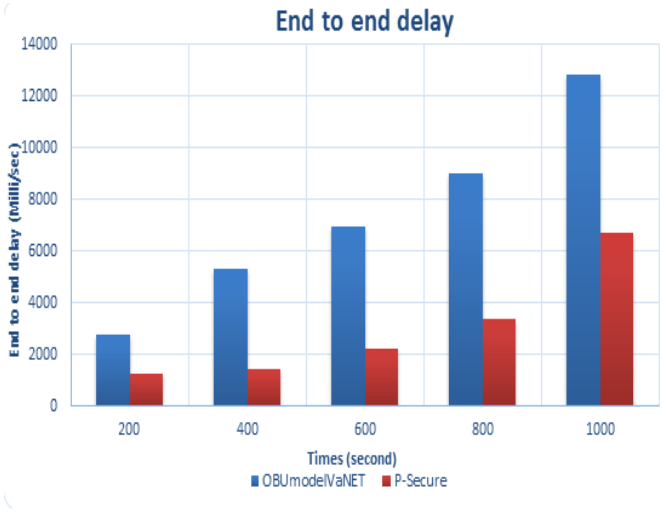


Fig. 5. E2E_delay vs time

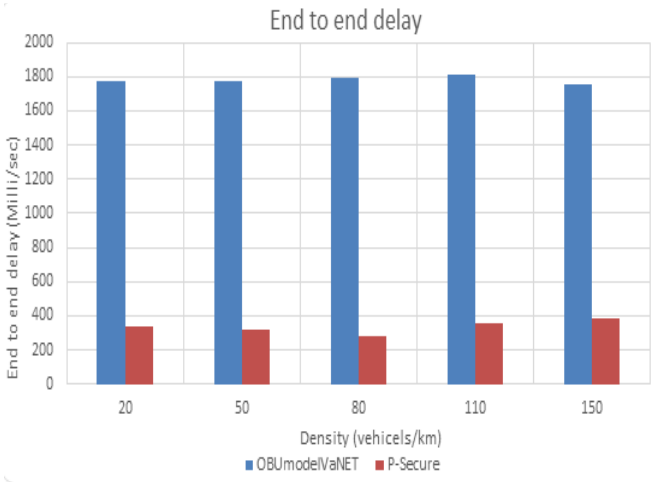


Fig. 6. End-to-end delay vs density

Fig (5) and (6), shows e2e_delay against the time and dencity. P-Secure approach is significantly lower compared to OBUmodelVaNET aproach. The reason is that using the proposed algorithms in the P-Secure approach, the suspicious nodes are not allowed to be sent and the attack was also diagnosed quickly and the security messages are delivered with a very little delay to the vehicles that are at risk.

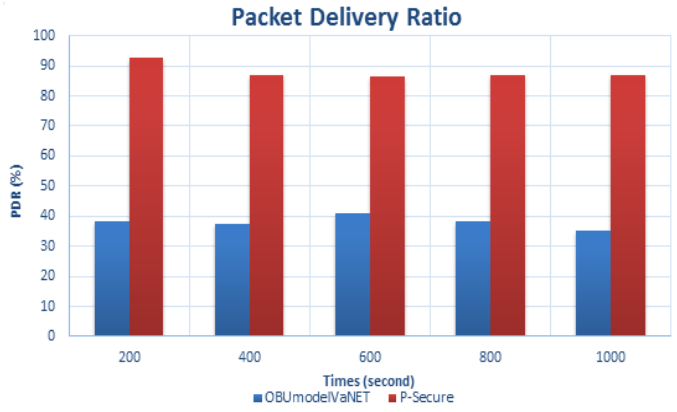


Fig. 7. PDR vs time

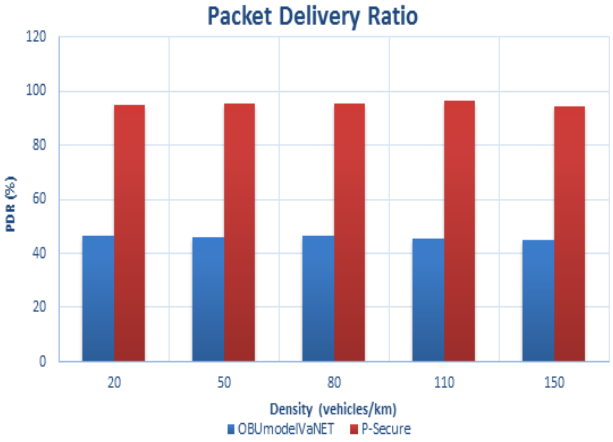


Fig. 8. Packet Delivery Ratio vs density

Fig (7) and (8), shows PDR against the time and dencity. From following graph, we can say that value of PDR is decreasing for OBUmodelVaNET under attack. When we vary pause time from 200 to 1000 as well PDR values for proposed approach is high. Thus, we can say that PDR of proposed method is improved than OBUmodelVaNET under attack which degraded due to DoS attack.

Fig (9) and (10), shows drop ratio against the time and dencity. It shows that, between the pause times 200 to 1000, the OBUmodelVaNET under attack had a high packet drop, while the packet drops of proposed approach in these times, has decreased.

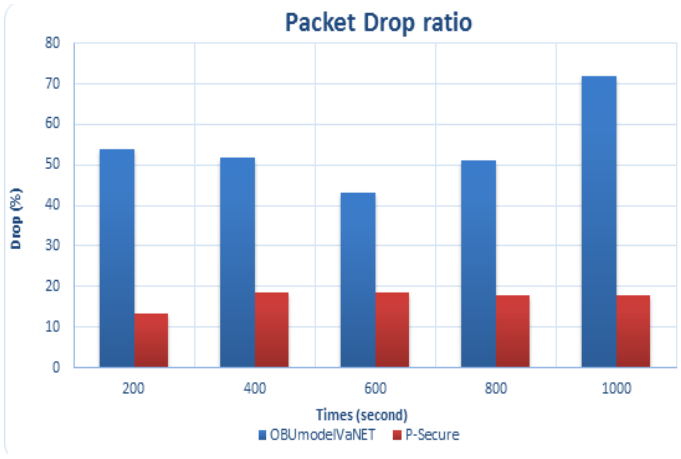


Fig. 9. Dropped Packet Rate vs time

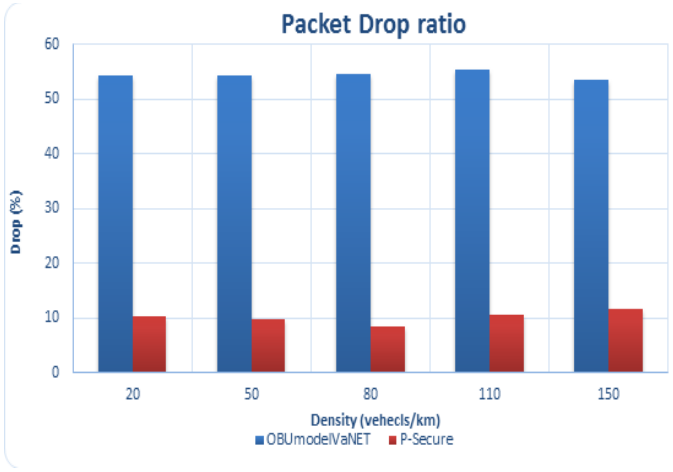


Fig. 10. Dropped Packet Ratio vs density

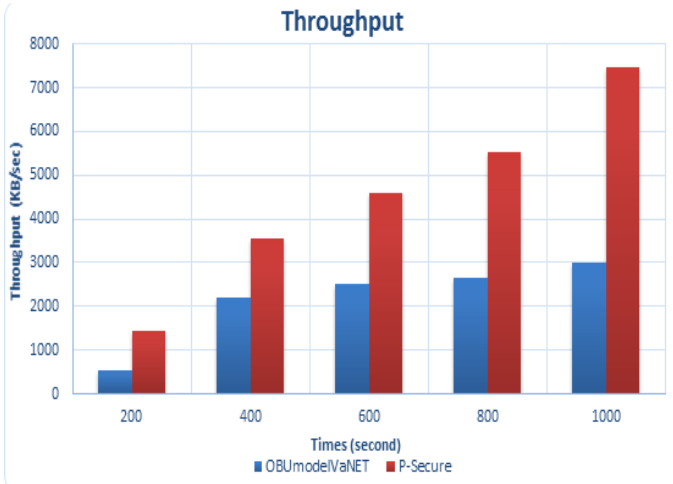


Fig. 11. Throughput vs time

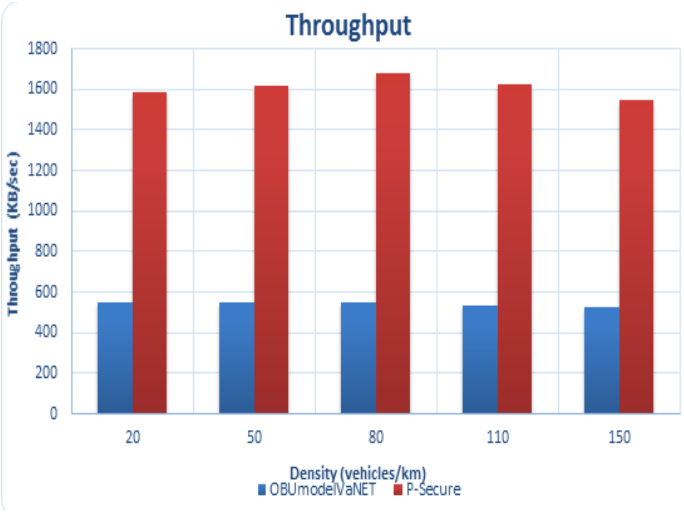


Fig. 12. Throughput vs density

Fig (11) and (12), shows throughput against the time and dencity. So we can claim that the P-Secure approach has better performance in the field of throughput rather than OBUmodelVaNET approach in MANET. This is due to the proposed mechanism that shown in fig 4.

VI. Conclusion

VANETs are a kind of wireless networks which have been created to communication between the adjacent vehicles as well as adjacent vehicles with constant equipment which are usually roadside equipment. One of the main concerns in Vehicular ad hoc networks (VANETs) is the attacks and influence to the system and of course having safety from the key concerns for many road users. In this paper, we introduce the (P-SECURE PROTOCOL) attack detection algorithm that is applied for detecting DOS attacks before the confirmation time. In the proposed RSRU method depending on its transmission range and using the proposed algorithm decides on what is the secure message, and sets the area where the vehicles can form a network. Traffic devices use the requests using another database and provide the service responses only to the radio transmitters which are approved therefore lead to the reduction in DOS attacks. The P-Secure approach leads to the reduction in processing delays and improving safety in VANET. To demonstrate the performance of P-Secure approach using the NS-2 simulator, the proposed system is compared to OBUmodelVaNET approach. The simulation results shows that P-Secure approach outperform than OBUmodelVaNET approach in terms of e2e_delay, PDR, packet drop rate and throughput.

References

[1] Al-Sultan, Saif, et al. "A comprehensive survey on vehicular Ad Hoc network." Journal of network and computer applications 37 (2014): 380-392.

[2] Huo Meimei, Zheng Zengwei, Zhou Xiaowei, “Research overview of simulation of vehicular ad hoc networks,” Application Research of Computers, 2013, Vol.27, No.5, pp.1614-1620.

[3] Jani P. 2002.Security within Ad-Hoc Networks, Seminar on Network Security Position Paper,pp:16-17

[4] Jose Maria de Fuentes, Ana Isabel Gonzalez-Tablas, Arturo Ribagorda, "Overview of security issues in Vehicular Ad-hoc Networks", Handbook of Research on Mobility and Computing 2010.

[5] Deng H, Li W.2002. Routing security in wireless ad hoc networks. Communications Magazine, IEEE,pp: 70-75.

[6] Sumra, I.A.; Hasbullah, H.; Manan, J.A., "VANET security research and development ecosystem," National Postgraduate Conference (NPC), 2011, vol., no., pp.1,4, 1920 Sept. 2011. doi: 10.1109/NatPC.2011.6136344 URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?&arnumber=6136344[3] Jing Zhao, Guohong Gao, “VADD: Vehicle-Assisted Data Delivery in Vehicular Ad Hoc Networks,” IEEE Transactions on Vehicular Technology, 2008, Vol.57, No.3, pp.1910-1922.

[7] J. Yu and P. Chong. A survey of clustering schemes for mobile ad hoc networks. Communications Surveys Tutorials, IEEE, 7(1):32–48, May 2005.

[8] M. Gerla and J. Tzu-Chieh Tsai. Multicluster, mobile, multimedia radio network. Wireless Networks, 1:255–265, 1995. 10.1007/BF01200845.

[9] C. Shea, B. Hassanabadi, and S. Valaee. Mobility-based clustering in VANETs using affinity propagation. In Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE, pages 1–6, December 2009.

[10] C. Shea, B. Hassanabadi, and S. Valaee. Mobility-based clustering in VANETs using affinity propagation. In Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE, pages 1–6, December 2009.

[11] S. Kuklinski and G. Wolny. Density based clustering algorithm for vanets. In Testbeds and Research Infrastructures for the Development of Networks Communities and Workshops, 2009. TridentCom 2009. 5th International Conference on, pages 1–6, April 2009.

[12] M. Gerla and J. Tzu-Chieh Tsai. Multicluster, mobile, multimedia radio network. Wireless Networks, 1:255–265, 1995. 10.1007/BF01200845.

[13] R. Mangharam, D. S. Weller, R. Rajkumar, “GrooveNet: A Hybrid Simulator for Vehicle-to-Vehicle Networks,” Proceedings of Second International Workshop Vehicle-to-Vehicle Communications (V2VCOM) Invited Paper. San Jose, USA. July 2006.

[14] R. Mangharam, D. S. Weller, D. D. Stancil, R. Rajkumar, J.S. Parikh. “Groovesim: A Topography-Accurate Simulator for Geographic Routing in Vehicular Networks. Proceedings of Second ACM International Workshop on Vehicular Ad Hoc Networks (Mobicom/VANET 2005) Cologne, Germany. September 2005.

[15] Karnad I F K, Mo Zhihai, Lan Kun-chan, “Rapid generation of realistic mobility models for VANET,” Proceedings of Wireless Communications and Networking Conference, 2007,pp.2506-2511.

[16] TraNSlite. http://trans.epfl.ch, 2009.

[17] Lu Bojin, Khorashadib, Du Haininf et al. “VGSim: an integrated networking and microscopic vehicular mobility simulation platform,” Communications Magazine, 2009, Vol.47,No.5, pp.134-141.

[18] Wang S Y, Chou C L, “NCTUns 5.0 network simulator for advanced wireless vehicular network researches,” Proceedings of the 10th International Conference on Mobile Data Management System, Services and Middleware, 2009, pp.375-376.[5] S.Marti, T.J.Giuli, K.Lai, M.Baker, “Mitigating Routing Misbehavior in Mobile Ad-Hoc Networks”.

[19] Khabbazian, M., Mercier, H., & Bhargava, V. K. (2009). Severity analysis and countermeasure for the wormhole attack in wireless ad hoc networks. Wireless Communications, IEEE Transactions on, 8(2), 736-745.

[20] M.T.Refaei, V.Srivastava, L.Dasilva, M.Eltoweissy, “A Reputation-Based Mechanism for Isolating Selfish nodes in Ad-Hoc Networks,” Second Annual International Conference on Mobile and Ubiquitous Systems, Networking and Services, pp.3-11, July, 2005

[21] Shahram Behzad, Reza Fotohi, Shahram Jamali,"Improvement over the OLSR Routing Protocol in Mobile Ad Hoc Networks by Eliminating the Unnecessary Loops", IJITCS, vol.5, no.6, pp.16-22, 2013. DOI: 10.5815/ijitcs.2013.06.03

Performance Evaluation of Routing Protocol (RPL) for Internet of Things

Qusai Q. Abuein, Muneer Bani Yassein, Mohammed Q. Shatnawi, Laith Bani-Yaseen, Omar Al-Omari, Moutaz Mehdawi and Hussien Altawssi
Faculty of Computer and Information Technology
Jordan University of Science and Technology
Irbid, Jordan

***Abstract*—Recently, Internet Engineering Task Force (IETF) standardized a powerful and flexible routing protocol for Low Power and Lossy Networks (RPL). RPL is a routing protocol for low power and lossy networks in the Internet of Things. It is an extensible distance vector protocol, which has been proposed for low power and lossy networks in the global realm of IPv6 networks, so it selects the routes from a source to a destination node based on certain metrics injected into the objective function (OF). There has been an investigation of the performance of RPL in the lighter density network. This study investigates the performance of RPL in medium density using of two objective function in various topologies (e.g. grid, random). The performance of RPL is studied using various metrics. For example, Packet Delivery Ratio (PDR), Power Consumption and Packet Reception Ratio (RX) using a fixed Packet Reception Ratio (RX) values.**

Keywords—density network; objective function; zero grid; packet delivery; power consumption; Internet of Things

I. INTRODUCTION

Internet of Things (IoT) is a technology in which everyday objects form an Internet network through where they can communicate with each other. The Internet of Things is a huge network of things or objects that can be embedded with a unique ID which then allows it to be connected to the internet, this huge network allows the devices to exchange data simultaneously for its specific purpose. The IoT allows the object to sense and collects data in the existing network infrastructure, which then will create opportunities for the real-time integration between the Machines and the physical world, this will result in economic benefit, improved accuracy and efficiency. The Internet Engineering Task Force (IETF) STANDARDIZED a powerful and flexible Routing Protocol for Low Power and Lossy Networks (RPL). It selects the ideal routes from a source to a destination node based on certain metrics injected into the Objective Function (OF). Previous studies. Many previous studies have investigated the performance of the OF0 and MRHOF objective functions in the light density network. This study will investigate the performance of the two OFs using various metrics like Packet Delivery Ratio, Energy Consumption in the medium density network. In this study, the performance of RPL will be investigated in terms of two Objective Functions under two topologies (grid, random) which make this work distinctive. To study the RPL performance, various metrics are considered Packet Delivery Ratio (PDR), Power Consumption and RX. The evaluation will be conducted based on these parameters (RX, topology) and compared for both OFs within a medium density network. In [1], Objective function Zero is the default Objective function in the Routing Protocol for Low-Power and Lossy Networks, OF0 is simple, it selects its parent depending on the minimum ranks of the neighbors. The node rank is usually an integer, it represents the nodes location. The most common objective function in RPL is Of0, This objective function permits the upward traffic to be routed through the selected parent (preferred parent) without performing any load balancing. In [2], Minimum Rank with hysteresis Objective Function (MRHOF) was proposed, it is commonly used for metrics as it is based on metric-containers, the container is used to determine the features and the nature of routing objects, this objective function the path cost is equal to the cost of the selected metric, from a child node to the sink node through its neighbors. The route cost is calculated by the node by adding the two components, the cost of the nominated measurement and the selected measurement for the connection to a nominee neighbor.

II. RELATED WORK

The growing attention of the research and industrial communities towards RPL is sworn from the amount of the recently published research, where RPL performance has been studied under the umbrella of different contexts and platforms. The authors of [3-5] show the effectiveness of RPL pertaining to exiguous delay, quick configuration, and self-healing. RPL is a Distance Vector IPv6 routing protocol designed for Low Sensor Networks, it is specifically designed represents the building of Destination Oriented Directed Acyclic Graph (DODAG) using OF0 or MRHOF with a set of constraints/metrics, the purpose is to calculate the best path, the node can operate with multiple OFS concurrently because the distribution varies greatly in different network topologies and different objectives may need to transmit traffic with different necessities of route superiority. The objective function does not require the metric and restrictions however does impose some rules to form the DODAG. One of the responsibilities of the network layer is delivering packets to the destination nodes via multiple hops separating the source node from the destination node. The routing table allows the packet to gain knowledge of the next hop neighbor node, the routing tables is populated by routing protocols. RPL builds a logical routing topology graph which is constructed ended a physical network to come across a assured measures and the network supervisor decides to have

multiple direction-finding topologies operating at the same time used to transmit the stream of traffic with multiple set of requests. Any node in the network can join one or more graphs, in this case they are called RPL occurrences and label the traffic tolerating to the graph characteristics.

The authors [9] provide the comparison for both OFs performances in a light density network under two different topologies (grid, random). RPL supports peer-to-peer communication which means any node in the graph in communicate with any other node in the same graph. When a node communicates with another node in the LLN network, the packet moves 'upwards' to a parent and 'downwards' to the destination.

III. PERFORMANCE EVALUATION

The main point of this study is to investigate the performance of the RPL in terms of OFs under two different topologies. A comparative study of broadcast mechanisms of RPL in IoT in conducted. Broadcast mechanisms using the rooted DAG-like logical structure maintained by the unicast routing protocol in RPL will be introduced and their performance will be studied in order to create a new broadcast mechanism with self-pruning to make the RPL OFs performance more efficient.

A. Results and Deduction

The experiments are conducted under the medium density network which consists of (50, 65, 75 and 85) nodes using random and grid topologies and with a Fixed RX=60.The RPL behavior in terms of power consumption and packet delivery ratio and is investigated. The OF0 was installed and results were obtained.

Figure 2 shows the behavior of the PDR based on a fixed value of RX=60 and a various number of nodes for the grid and the random topology using the objective function OF0. In the Grid topology, the PDR increased between the 50-65 nodes but it decreased between the 75-85 nodes, this shows that the PDR is more efficient when using the OF0 in the grid topology when the density is between 50-65 nodes. In the Random topology, the PDR increased between the 50-65 nodes, the PDR also increased at 75 nodes and above, this shows that the PDR is more efficient when using the OF0 in the random topology when the density is between 50-65 nodes and above 75 nodes.

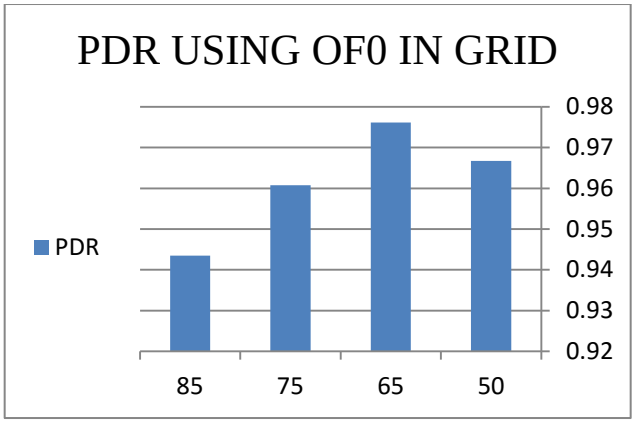


Fig. 1. Values of PDR in GRID Topology using 0FO

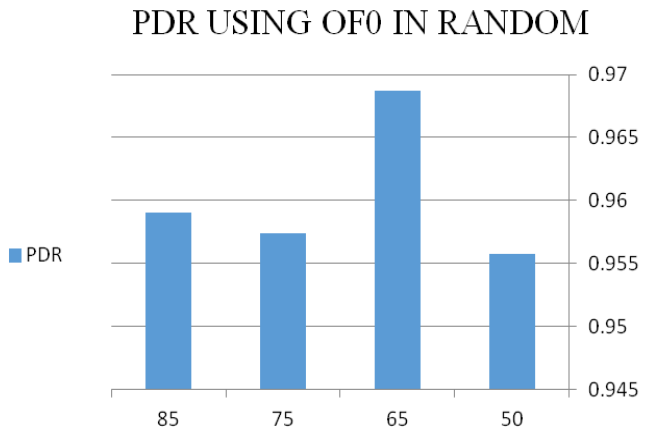


Fig. 2. Values of PDR in random Topology using 0FO

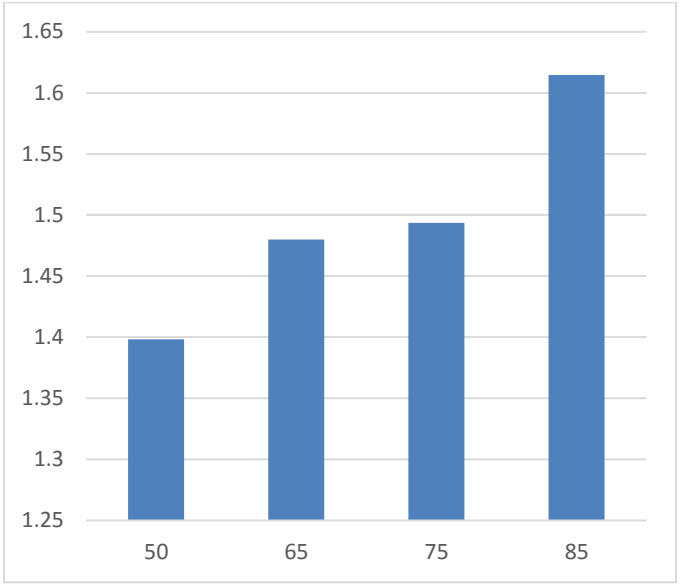


Fig. 3. Power Consumption Using OF0 in Grid

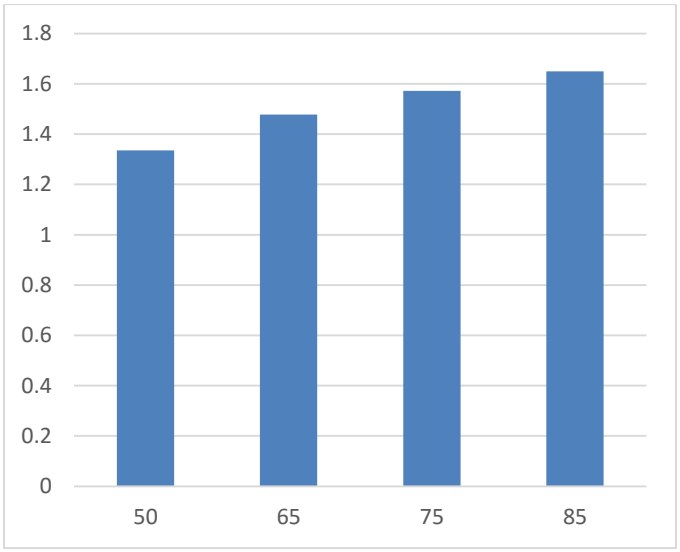


Fig. 4. Power Consumption Using Of0 in Random Topology

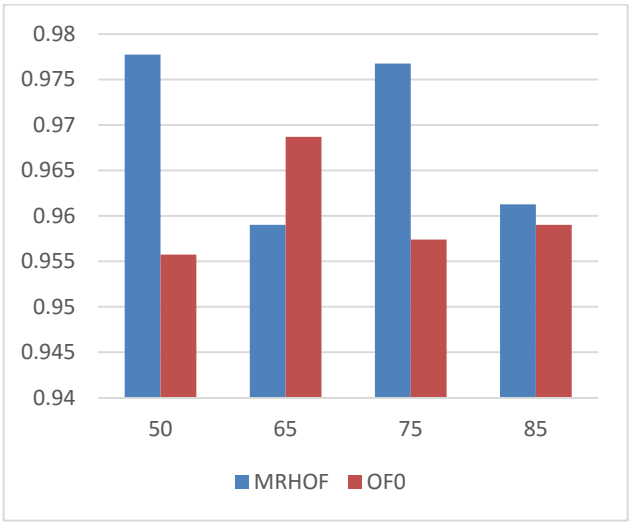


Fig. 5. Values of PDR with OF0 and MRHOF in random topology

Figure 5 shows the presentation of the power depletion based on a fixed value of RX=60 and a various number of nodes for the grid and the random topology using the objective function MRHOF. In the Grid topology, the Power Consumption increased gradually as the number of nodes increased, this shows that the Power Consumption is not efficient when using the MRHOF in the grid topology when the density is between 50-85 nodes. In the Random topology, the Power Consumption increased gradually as the number of nodes increased, the power consumption was almost stable between 65-85 nodes. This shows that the Power Consumption is more efficient to use MRHOF in the random topology when the density is above 50 nodes.

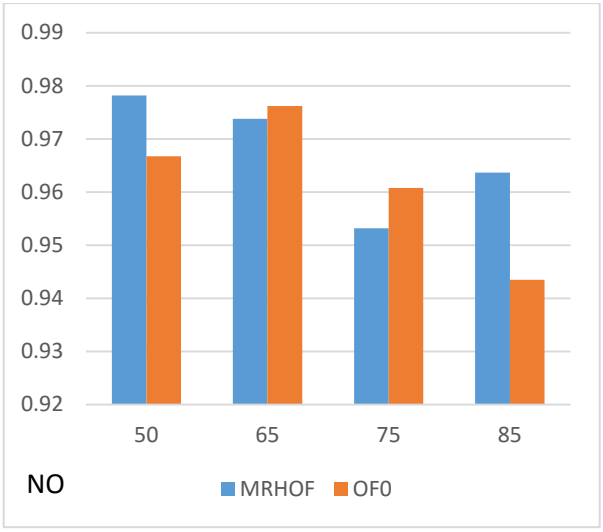


Fig. 6. Values of PDR with OF0 and MRHOF in grid topology

Figure 6 shows that the PDR of Objective Function Zero is roughly 0.956%, and that the PDR of MRHOF is around 0.97%.

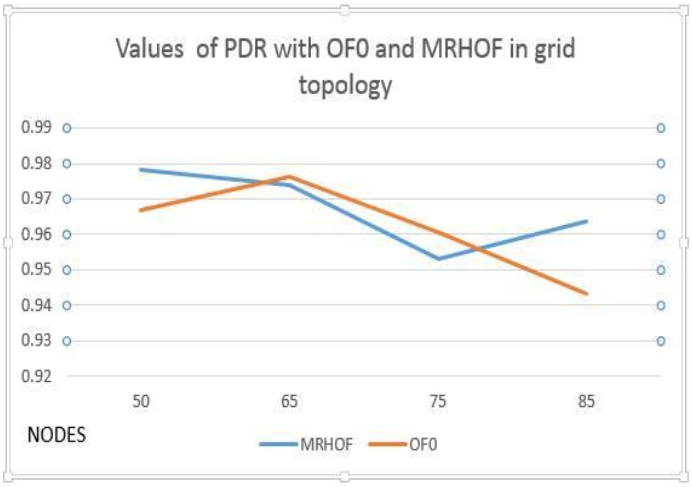


Fig. 7. Values of PDR with Of0 and MRHOF in Grid Topology

In Figure 7, the values of the PDR of MRHOF is approximately 0.97%. The average Packet Delivery Ratio of OF0 decreases as the number of nodes increases.

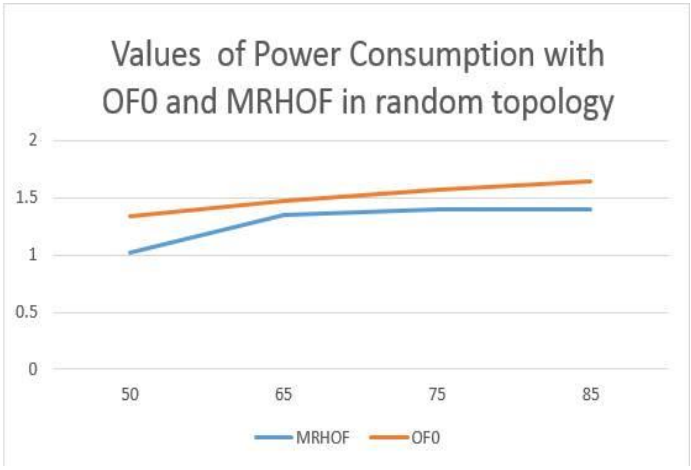


Fig. 8. Values of Power Consumption with OF0 and MRHOF in the Random Topology

In Figure 8, shows the appearance of power depletion with OF0 and MRHOF in the Random Topology, power depletion of MRHOF which is around 1.29% and the power depletion of OF0 is approximately 1.50% using Random Topology. Figure 9 shows the behavior of the Power Consumption has almost stable figures for both MRHOF and OF0 when the Packet Reception Ratio=60% and in a Medium Density Network.

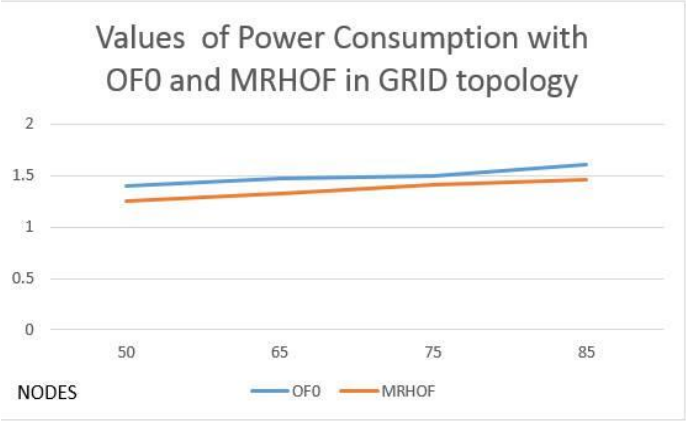


Fig. 9. Values of Power Consumption with OF0 and MRHOF in GRID topology

IV. CONCLUSION

This research proved that the Routing Protocol for Low power and Lossy networks is extremely demanding upon using the Objective function Zero and Minimum Rank hysteresis Objective Functions in terms of Packet Delivery Ratio and Power Consumption in the Medium Density Network. It has been revealed that the Packet Reception Ratio is best when it is equal to 60% for both Objective functions in the relation to Packet delivery ratio and Power Consumption. The best performance of The Routing Protocol for Low power and Lossy networks performances is at its best when the network density is between 50-65 motes for the RX=60% in the Grid and Random Topologies.

REFERENCES

[1] Thubert, "Objective Function Zero," RFC 6552, March 2012.

[2] O. Gnawali and P. Levis, "The ETX Objective Function for RPL," IETF Internet Draft: draft-gnawali-roll-etxof-00, 2010.

[3] N. Accettura, L. Grieco, G. Boggia, and P. Camarda, "Performance Analysis of the RPL Routing Protocol," in Proc. of 2011 IEEE International Conference on Mechatronics, April 2011.

[4] E. Ancillotti, R. Bruno, and M. Conti, "The role of the RPL routing protocol for smart grid communications," IEEE Communications Magazine, vol. 51, no. 1, pp. 75–83.

[5] J. Tripathi, J. de Oliveira, and J. Vasseur, "Applicability Study of RPL with Local Repair in Smart Grid Substation Networks," in 2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 262–267. 2010

[6] O. Gnawali and P. Levis, "The Minimum Rank with Hysteresis Objective Function", RFC 6719 (Proposed Standard), Internet Engineering Task Force, Sep, 2012.

[7] O. Gaddour and A. Koubaa. RPL in a nutshell: A survey. Comput. Netw. pp. 3163–3178, 2012.

[8] Gaddour, O., Koubaa, A. Chaudhry, S. Tezeghdanti, M., Chaari, R., Abid, M.: Simulation and performance evaluation of DAG construction with RPL. In: 2012 Third International Conference on Communications and Networking (ComNet), pp. 1–8, 2012.

[9] Mamoun Qasem, Hussien Altawssi, Muneer BaniYassien, Ahmed Al-Dubai. Performance Evaluation of RPL Objective Functions,2015.

[10] Reactive Discovery of Point-to-Point Routes in Low Power and Lossy Networks - http://tools.ietf.org/html/draft-ietf-roll-p2p-rpl.

[11] Gaddour, O., Koubaa, A., Chaudhry, S., Tezeghdanti, M., Chaari, R., Abid, M.: Simulation and performance evaluation of DAG construction with RPL. In: 2012 Third International Conference on Communications and Networking (ComNet),pp. 1–8 (2012).

[12] M. Vucinic, B. Tourancheau, and A. Duda "Performance comparison of the rpl and loadng routing protocols in a home automation scenario." Proceedings of IEEE WCNC, 2013.

[13] L.Songhua, W. Muqing, C. Chuanfeng, L. Bo and L. Simu. "A high-throughput routing metric for multi-hop Ad hoc networks based on real time testbed." TENCON 2013,Oct. Mar,2013,pp.1-4.

[14] P. Gonizzi, R. Monica, and G. Ferrari. "Design and evaluation of a delay-efficient RPL routing metric." Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International. IEEE,2013.

[15] N. Tsiftes, J. Eriksson, N. Finne, O. Fredrik, J. Höglund, A. Dunkels, A Framework for Low-Power IPv6 Routing Simulation, Experimentation, and Evaluation, SIGCOMM'10, New Delhi, India, November 2010, pp. 479–480.

[16] M. Nuvolone, "Stability analysis of the delays of the routing protocol over low power and lossy networks," Master's thesis, KTH RIT, 2010.

[17] W. Xie, M. Goyal, H. Hosseini, J. Martocci, Y. Bashir, E. Baccelli, A. Durresi, "A performance analysis of point-to-point routing along a directed acyclic graph in low power and lossy networks", 2010.

A New Method to Build NLP Knowledge for Improving Term Disambiguation

E. MD. Abdelrahim[1], El-Sayed Atlam[2], R. F. Mansour[3]
[1,3] Computer science department, Faculty of science, Northern Border University, KSA)
[1,2] Dept. of Mathematics, Computer Science Division, Faculty of Science, Tanta, Egypt
[2]Dept. of Information Science and Intelligent Systems, University of Tokushima, Tokushima, Japan 770-8506

***Abstract*—Term sense disambiguation is very essential for different approaches of NLP, including Internet search engines, information retrieval, Data mining, classification etc. However, the old methods using case frames and semantic primitives are not qualify for solving term ambiguities which needs a lot of information with sentences. This new approach introduces a building structure system of natural language knowledge. In this paper all surface case patterns is classified in advance with the consideration of the meaning of noun. Moreover, this paper introduces an efficient data structure using a trie which define the linkage among leaves and multi-attribute relations. By using this linkage multi-attribute relations, we can get a high frequent access among verbs and noun with an automatic generation of hierarchical relationships. In our experiment a large tagged corpus (Pan Treebank) is used to extract data. In our approach around 11,000 verbs and nouns is used for verifying the new method and made a hierarchy group of its noun. Moreover, the achievement of term disambiguating using our trie structure method and linking trie among leaves is 6% higher than old method.**

Keywords—Information Retrieval; NLP Knowledge; Disambiguation; Word Semantics; trie structure

I. INTRODUCTION

Natural language processing (NLP) systems use many dictionaries. In this paper, we discus two types of information. The first is morphological information about morphemes, or words, and their fundamental attributes such as a part of speech [11], and the second is semantic primitive [16][17][28], and so on.

The understanding of implicit in events is of great interest in recent years. Nouns in NL is assumed as real-world entities due to the implicit with nouns in most of work. The lexical nouns name classes of entities, some of which are kinds and some of which are not. This is compatible with the view of compositional semantic in which nouns are viewed as one-place predicate. They are argument-taking functions which take individuals into truth values. On the other hand, verbs are viewed as n-place predicates, functions which take n-tuples into truth values. The (extensional) meaning of any sentence is composed by recursively combining functional terms with quantifiers, operators, and logical connectives.

So first, generic knowledge of events consist of implication is very essential for understanding. For example a person buys something because he wants it. Such knowledge incorporates the implications that buying is enabled by having enough money, and that asking implies that subject of the asking wants something. The second type of knowledge classifies verbs with subject, object, and place into groups which are considered relation. It is important to design an Implicit Inference of nouns and linking group that can efficiently integrate multi-attribute relation.

Implicit inference iinformation is defined by knowing the verbs deepest meaning, determining a deep knowledge about nouns. Also, multi-attribute relation information is defined by a pair of basic words and its record includes the attribute of relation . Consequently, the problems, are a very large space cost for storing all pairs and a high frequent access of pairs and their attribute in the record. A trie, or a digital search structure, must be introduced to the basic scheme since a word is basically a string. Relational information such compound words is formed significantly, and occupies a large spaces in the morphological dictionary. Artificial intelligence (AI) basic knowledge IS-A also depend on term relationships.

A case frame [25][27] is an important technique to solve ambiguity in syntax and semantic analysis [20][21]. Japanese to English, machine translation systems in both direction [25] requires using case frame to build translation dictionaries.

Aoe et al.[1][2][3][4] and Morita et al.[5] introduced a two-trie structure for storing compound words into the compact structure. Morita et al. [6] presented a link trie.

This paper present an implicit inference of nouns, and collect all knowledge about the sentence and make it groups of linking and high frequent access between verbs with subject, object and place. Moreover, by introducing a trie that can define the linkage among leaves, this paper present an efficient data structure. Therefore, the proposed structure defines, multi-attribute relationships between words which can be merged into the same record.

Section II of this paper describes relational information as multi-relations among terms with a case frame of the basic knowledge. The link trie and an integrating morphological is presented in Section III. The proposed method is verified by simulation results in Section IV. In Section V, we discuss conclusions and potential future work.

II. MULTI – RELATIONSHIPS AMONG WORDS

A. Information Of Multi-Attribute Relation

MOR(x) is the morphological information for word x. here we will discuss relational information, call a multi-attribute relation, for a finite of relational attributes briefly.

Multi-attribute relation's information can be defined as a triplet (x, y, Alpha), where x and y are interrelated, and the attribute is Alpha. In natural language processing one can get a variety of attributes, and clearest meaning by using relationships among words as follows.

B. Case frame

To cope with this complexity we have to use the services of some syntactic and semantic information at the same time for the analysis of a sentential structure. The best grammatical framework for this purpose is the case grammar (C. Fillmore in 1968). the semantic primitive shown in Table 2 is utilized to determine which kind of noun can be in which case slot. For instance, the verb eat load a noun connected with one of the semantic primitive animal as the cause of the verb, and noun of semantic code eatable stuff as an object. This case slot determination is specified for each handling of all verbs in a dictionary.

The information to be inserted in the dictionary record differs depending on each part of speech, but in general include this kind of information: head word, number of character of words end, alternate, root word, correlated words, morphological piece of spoken language, conjugation, prefix information, area code, grammatical part of speech, sub-categorization of piece of speech, patterns case, feature, model, option, semantic primitives, co-occurrence information (adverb, predicative modifier), idiomatic expressions, degrees, degrees of nominality and so on.

Here, Verbs and nouns case pattern is one of the important information. We have renowned over 30 instances, see Table 1. Each case slot in a pattern of verb use include semantic information about the noun, which could be seen in the slot. The noun has the matching semantic code in an entry. We have renowned over 50 semantic primitives (codes) in Table 2.

TABLE I. CASE RELATIONS USED IN THE ENGLISH DEPENDENCY STRUCTURE [M.NAGAO,ET.AL[13]]

(1) Subject	(17) Attribute
(2) Object	(18) Cause
(3) Recipient	(19) Tool
(4) Origin	(20) Material
(5) Partner	(21) Component
(6) Opponent	(22) Manner
(7) Time	(23) Condition
(8) Time-From	(24) Purpose
(9) Time-to	(25) Role
(10) Duration	(26) Content
(11) Space	(27) Range
(12) Space-From	(28) Topic
(13) Space-To	(29) Viewpoint
(14) Space-Through	(30) Comparison
(15) Source	(31) Accompany
(16) Goal	(32) Degree
	(33) Predicative

TABLE II. SYSTEM OF SEMANTIC PRIMITIVES FOR NOUNS (NAGAO ET AL., 1986)

NATION & ORGNAZATION			
ANIMATE	1-HUMAN.PROFFSION	PHENOMENON	1-TURAL PHENOMENON
	2-ANIMAL		2-PHYSCAL PHENOMENON
	3-PLANT		3-POWER&ENERGY
	4-OTHERS		4-PHYSIOLOGICAL PHENOMENON
			5-SOCIAL PHENOMENON
INANIMATE	1-NATURAL SUBSTANCE		6-SOCIAL SYSTEM
	2-PARTS MATERIALS		7-OTHERS
	3-ARTIFICAL PRODUCT		
	4- SYSTEM	FEELING	1-FEELING MENTAL
	5-OTHERS		2-THINKING
			3-OTHERS
ABSTRACT	1-INTERLLECTUAL PRODUCT		
PRODUCT	2-INTERLLECTUAL TOOL	ACTION	1-DOING
			2-MOVING
	3-INTERLLECTUAL MATERIALS		
			3-OTHERS
	4-INTERLLECTUAL GOODS		
	5-OTHERS	MESURMENT	1-NUMERIC
			2-MEASURABLE PROPERTY
PART	1-PARTS ELEMNET		3-STANDARD
			4-UNIT
	2-ORGANS OF HUMAN OR ANIMAL		
			5-OTHERS
	3-OTHERS		
		PLACE LOCATION	
ATTRIBUTE	1-NAME OF ATTRIBUTE		
	2-RELATION	TIME	1-TIME POINT
	3-SHAPE		2-TIME DURATION
	4-STATE		3-TIME PROPERTY
	5-PROPERTY		4-OTHERS
	6-OTHERS		
		OTHERS	

In view of the Machine Translation (MT) example by [M. Nagao, et. al. [26],[27] as in Table 1, the semantic primitive is employed to determine which kind of noun can be in which case slot. For instance, the verb eat requires a noun linked with one of the semantic primitive animal as the cause of the verb, and noun of semantic code eatable substance as an object. That case slot determination is specified for each use of all verbs in dictionary.

The VERB and NOUN (OBJECT, PLACE) relation relations are defined as follows:

Ahmed reside in EGYPT	*<VERB –PLACE>*
Ahmed speak Arabic	*<SUB. – VERB>*
Cat eat food	*<VERB – OBJ.>*
Fish live in water	*<VERB – OBJ.>*
Ibrahim treat sickness	*<SUB. – VERB>*

In the following sub section, more detailed study can be carried out with examples to have an implicit meaning of term disambiguation.

Example [1]. SEMANTIC(“Chocolate”) [PLACE \ MOUNTAIN]

Sentence: Jhon will climb the Chocolate in the next winter holiday.

(ACTOR: Jhon, HUMAN)
(OBJECT: Chocolate)

As in Table 2 of semantic primitives, we will find that “Chocolate” is an OBJECT, and by the information of verb “climb”, then the noun “Chocolate” is a PLACE where HUMAN will climb on it. Therefore, SEMANTIC(“Chocolate”) in the previous sentence is [PLACE \ MOUNTAI][9-15].

Example [2]. SEMANTIC(“Chocolate”) [FOOD \ EAT]

Sentence: Hala eats Chocolate.

(ACTOR: Hala, HUMAN)
(OBJECT: Chocolate)

As in Table 2 of semantic primitives, we will find that “Chocolate” is an OBJECT, and by the information of verb “eats”, then the noun “Chocolate” is an EATABLE MATERIAL that HUMAN will eat. Therefore, SEMANTIC(“Chocolate”) in this previous sentence is [FOOD \ EAT].

Example [3]. SEMANTIC(“Chocolate”) [PRODUCT MOBILE PHONE]

Sentence: Data is organized in Chocolate.

(ACTOR: Data, INTELGENT PRODUCTs)
(OBJECT: Chocolate)

As in Table 2 of semantic primitives, we will find that “Chocolate” is an OBJECT, and by the information of verb “organized”, then the noun “Chocolate” is an INTELGENT PRODUCTs that can be organized data on it. Therefore, SEMANTIC(“Chocolate”) in this sentence is [PRODUCT \ MOBILE PHONE].

Examples show in sentence ambiguities, WSD can be carried out based on the clear semantic primitives in sentences. However, in the case of context ambiguities, although the sentence includes semantic primitives, context ambiguities are still hard to be solved, Appendix A.

C. Implicit Inference of a Noun

It is very essential to have systematic study on the verbs and nouns, to have a deep knowledge. Due to implicitly in the events, a generic knowledge is necessary. By creating a verb semantic representation in the case frame, we can get more information about noun. A detailed study has been carried out with many examples to get nouns implicit inference as follows:

Example [1]: Mr. Atlam eat fried fish in a restaurant.

For a case frame of this sentence;

(ACTOR: Mr. Atlam)
(OBJECT: fried fish)
(LOCATION: Restaurant)

We notice that, a noun has just semantic primitive. For example in this example, we find that fried fish is one kind of food. This means that by using Table 2 (semantic primitive) that food one PLANTS and plants have no knowledge about ‘eatable material.’ Also, a restaurant by Table 2 just a LOCATION, and has no knowledge about ‘eating place.’

By using implicit inference (deep information of a verb), we find a SEMANTIC_REFER of slot, which indicates that the frame may possibly refer to the semantic depiction of that slot. The knowledge of the verb eat in this example refers to knowledge of fried fish and a restaurant, then an object fried fish is referred to ‘eatable material’ and a restaurant is referred to ‘eating place.’

Example 2: Mr. Samouda swims in a river.

For another case frame of the sentence;

(ACTOR: Mr. Samouda)
(LOCATION: River)

By using the semantic primitive of noun Table 2, we find in this example that a river refers to only LOCATION, and has no knowledge about ‘swimming place,’ and Mr. Samouda refers to the HUMAN.

But by employing the deep information of the verb we see a slot of SEMANTIC_REFER indicating that the frame may refer to the semantic description of that slot. The knowledge of the verb swim in this example is refers to the knowledge of a river, then the place a river is referred to ‘swimmable place’ where the HUMAN swim, and if there is relation that LOCATION has OBJECT, then the dynamics knowledge that a river has water.

Example 3]. Mr. Atlam wants to buy a computer from a store.

For this case frame of this sentence;

(ACTOR: Mr. Atlam)
(OBJECT: Computer)
(LOCATION: Store)
(TOOL: $10,000)

By using the semantic primitive for nouns Table 2, this refers to a computer which is an ARTIFICIAL PRODUCT, a store refers to the place location, and $10,000 just TOOL (has no information about the price of computer).

But by employing implicit inference, we find the slot of SEMANTIC_REFER indicating that the frame may possibly refer to the semantic depiction of that slot. The knowledge of the verb buy in this example is refers to knowledge of a computer, a store, and $10,000, then the knowledge refers that Mr. Atlam is enabled, so by having money and the cost is $10,000, that Mr. Atlam intends to use what he buys, also store is referred to 'buyable place', if there is a relation that LOCATION has OBJECT, then a store has a computer.

By using verb information in the case structure, implicit knowledge of nouns can be derived. By extending this knowledge, we can build some linkage groups between a subject with a verb, a verb with an object, and a verb with a place. We can write the same typical sentence, as follows:

1) My wife eat some meat in a restaurant.
2) My father eat some rice in a restaurant.
3) Mr. Mohammad swims in the sea.
4) My son swims in a pool.
5) My daughter buys a toy from the store.
6) Mr. Mathew buys a television from the store.
7) My Mother buys some fruits from the market.
8) The students drink a glass of juice in his house.
9) Math teacher drinks some coffee in the school, and so *on.*

By collecting a large number of this examples, we could build the following groups: the linking between nouns and verbs is shown in the figure 1, and this group is arranged from down to up depending on the strong and has the weak relation. This means that the relation between nouns and high-leaky verbs, such as talk, think, speak, and so on. They are verbs of a higher animal action, and strong verbs. But another is general verbs, such as eat, drink, and so on, or a weak verb, as follows:

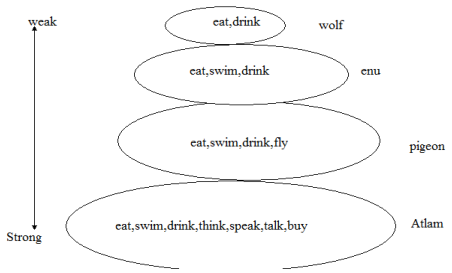


Fig. 1. Group of Link between a Subject and Verbs

Although each knowledge dictionary in primitive systems is built separately, almost all modern natural language applications become more complicated combining the above relationships. For this reason, it become necessary to design a fast and compact structure to be efficiently integrated with any of multi-attribute relation.

Since information about multi-attribute relation is defined by a pair of basic words and its record as well as the attribute of relation, the problems become a very large space cost for storing all pairs and a high frequent access of pairs and their attributes in the record. Since a word is basically a string, a trie, must be added to the basic scheme representation

D. Compound Word

The triple <x, y, α> is called Compound Word relations which indicates that x composite with y to give new information. By using case frame relation <tool> + <Verb> → of computer processing, and <Subject> + <Verb> → of language processing are called compound word relation. By using this case frame relation the clearest meaning and information about word can be extract rather than the single one, another example as follow:

Information Retrieval

Natural language.

III. LINK TRIE (LT) FUNCTION

A. Tries and Efficient Representation of Verb and Noun Linkage

Trie is an n-array tree [2], [10], [11], [15] having n-place vectors as nodes with components corresponding to digits or characters. For confusion avoidance between keys like the and then, let us insert a special end marker; # to the end of all keys, so no prefix of a key can be a key itself [1]. Let K be a keys set. Each path in the trie starting from the initial node (root) to a leaf corresponds to a key in K. Therefore, the nodes of the trie correspond the prefixes of keys in K. A trie definition is as follows [3], [4], [5].

1) S is a limited set of nodes, represented as a positive integer.
2) I is a limited set of input characters, or symbols.
3) g is a goto function from S □ I to S ∪{fail}.

This means that, a node r is in F if and only if there is a path from 1 to r reads some string x in K. A move titled with a (in I) from r to t means g(r, a) = t. The nonexistence of a move means stoppage (failure). Figure 2 shows a trie example for eleven words with '#', where enclosed in a square nodes will be later discussed. The key 'Atlam#' retrieving can be done by applying the transitions g(1, 'a') = 3, g(3, 't') = 22, g(22, 'l') = 14 , g(14, 'a') = 32, g(32,'m')=15, and g(15,'#')=2, sequentially.

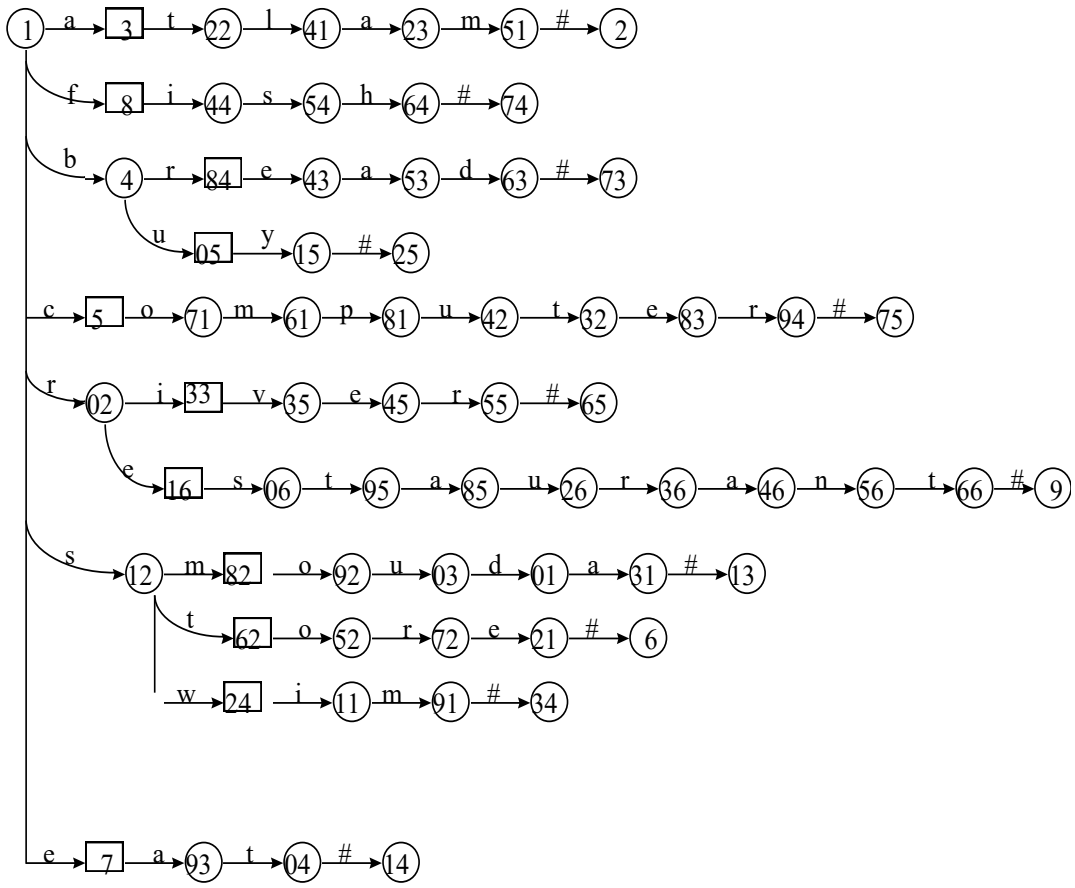


Fig. 2. Example of Trie Structure

B. *Link Trie(LT) Function [K. Morita, 6]*

Term Relationships Definition

Assume (X, Y, R) is the relation R between termss X and Y. With tries, there is one-to-one correspondence between leaves and keys, so we can define its link trie by linking leaf s for X and leaf t for Y. In such case, the definition of function LINK is t ∈ LINK(s) and the relation by the record R ∈ CONTENTS(s, t). Link trie is the trie including the function LINK and CONTENTS. Link information for figure 2 is shown in Table 3.

We can see the relationship between Atlam as a subject and buy as a verb by the trie and there exist one-to-one correspondence, the leaf 2 correspondence key Atlam and leaf 52 correspondence key buy, and link function is defined by 52 ∈ LINK(2) and the record (<subject>, <verb>) ∈ CONTENTS(2, 52). We can see the relationship between words (<verb>, <object>) and (<verb>, <place>), as follows:

Retrieval Algorithm

For the relationship (X, Y, R), the proposed retrieval algorithm (i): retrieve Y and R from X, (ii): retrieve R from X and Y.

For LT and for key X, the function GET_LEAF(LT, X) gives the leaf for X# and gives fail if LT has no X#. The function GET_LEAF (LT, "store") gives leaf 6 in Figure 2.

For the relationship (X, Y, R), the following ALGORITHM returns leaves s for X# and t for Y# if they are recorded in the trie. s and t could be processed to recover CONTENTS(s, t) including relationship R. If any of s or t is not recorded in the trie, then ALGORITHM outputs s = t = 0.

[ALGORITHM]

start
s← GET_LEAF (LT, X);
t← GET_LEAF (LT, Y);
if (s = fail or t = fail) then output s = t = 0;
if ((t ∈ LINK(s) and R ∈ CONTENTS(s, t)) then output s and t;
end;
(Algorithm End)

C. *System Frame work*

Figure 3, shows the frame work of our approach by Searching for Some English Textbook &Papers, concerning with Cross Language Information, Classification Summarization, and Noun Extraction from the Penn

Treebank• Extract compound noun after stemming and use stop word dictionary , from large Corpus. Moreover, Extract the linkage between verb with noun, verb with place , and verb with place , by using part of speech dictionary, and make linkage group and high leaky relation between them. By using this frequent and high leaky relation we can make disambiguate for word, where the surrounding words frequently associated with a sense are used to disambiguate a word.

TABLE III. EXAMPLES OF INFORMATION LINK

X	S	LINKS	CONTENTS(s, t)
Atlam	2	{41,52}	CONTENTS(2,41)={<subject>,<verb>}
			CONTENTS(2,52)={<subject>,<verb>}
Smouda	31	{2,43}	CONTENTS(31,2)={<subject>,<subject>}
			CONTENTS(31,43)={<subject>,<verb>}
eat	41	{37,47,9}	CONTENTS(41,37)={<verb>,<object>}
			CONTENTS(41,47)={<verb>,<object>}
			CONTENTS(41,9)={<verb>,<object>}
buy	52	{57,6}	CONTENTS(52,6)={<verb>,<place>}
			CONTENTS(52,57)={<verb>,<object>}

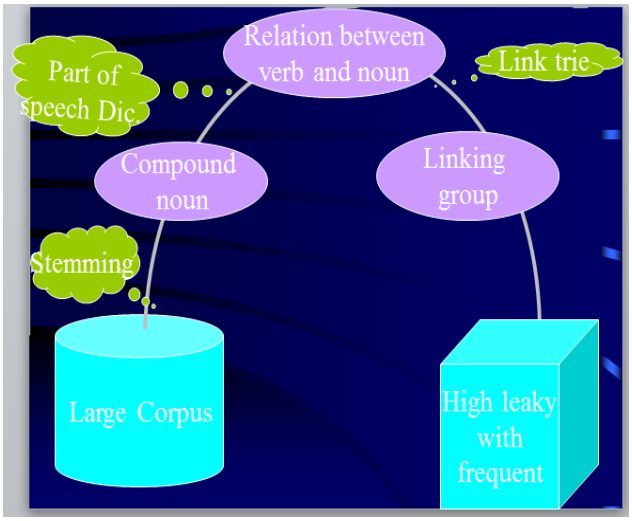


Fig. 3. System Framework

D. Semantic Field Information

As Section 2.1 discussed that some words have many semantic meaning. Therefore, various semantic(x) usually appears in various branches. Table 4 shows that [PLACE \ MOUNTAIN] is in fields <TRIP \ AMERICA>, and <SPORTS \ MOUNTAIN CLIMBING>. [FOOD \ EAT] is in <FOOD \ SUPERMARKET>. [PRODUCT \ MOBILE PHONE] is in <COMPANY \ TELPHONE SHOPE>. Therefore, words with various fields in the context could be utilized to discriminate the semantic(x).

TABLE IV. EXAMPLES OF RELATIONSHIPS BETWEEN SEMANTICS AND FIELDS

SEMANTIC("Chocolate")	semantics Ambiguities	Field
[PLACE \ MOUNTAIN]	After dinner, our manager eat some snacks. Both Hala and Jhon usually eat Chocolate, because they use Chocolate.	<TRIP \ AMERICA>, <SPORTS \ MOUNTAIN CLIMBING>,
[FOOD \ EAT]		<FOOD \ SUPERMARKET>
[PRODUCT \ MOBILE PHONE]		< COMPANY \ TELPHONE SHOPE >

IV. AUTOMATIC KNOWLEDGE GENERATION FOR AN UNKNOWN WORD

This section describe how to get more information & new knowledge from case-frame storing by using trie structure and linking between leaves, perhaps by keeping links between them to reflect some relationships. e.g. Jhon * is unknown word

Context (case frame)

Level1: Jhon * eats apple, Jhon IS – A animal?, Jhon is similar to dog, or human

Level2: Jhon * buys computer, Jhon IS – A human.

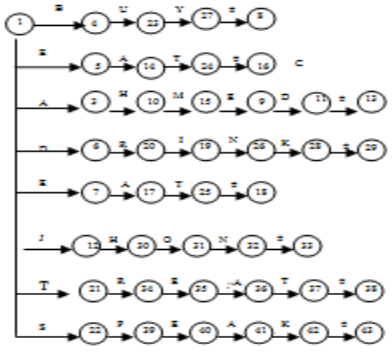


Fig. 4. Trie structure

By using this information as common knowledge, we will show later by using trie structure and link trie, how we can know new automatic & variety relation from this common knowledge. This new knowledge are very useful in NLP, because it make a text more readable and understandable for human, this new knowledge can be combined to provide additional useful <IS –A> hierarchical information, as follow:

In this examples with <SUB. – VERB> relation using the information:

1- Ahmed treat the illness *2- Ahmed cure the sick*
2- Ahmed eat food *4- Ahmed speak with the nurse*
3- Ahmed drink milk *6- Jhon eat orange*
4- Jhon drink tea *8- Jhon speak with his teacher*
5- Cat eat food *10- Cat drink water*

and create the structure of trie as Figure 4. and links between leafs in this trie, we can build the linking as in table 5 from the given information as follows:

TABLE V. TRIE LINK FOR <SUB. –VERB> RELATIONSHIP

X	s	RELATION(s)	ATTRIBUTE(s, t)
Jhon	33	{2,18,8,29}	ATTRIBUTE(33,2)= <SUB. -- VERB > ATTRIBUTE(33,18)= <SUB. -- VERB> ATTRIBUTE(33,8)= <SUB. -- VERB> ATTRIBUTE(33,29)= <SUB. -- VERB>
Ahmed	13	{2,8,29,18,38}	ATTRIBUTE(13,2)= <SUB. -- VERB> ATTRIBUTE(13,8)= <SUB. -- VERB> ATTRIBUTE(13,18)= <SUB. -- VERB> ATTRIBUTE(13,38)= <SUB. -- VERB> ATTRIBUTE(13,29) = <SUB. -- VERB>
Cat	16	{2,29}	ATTRIBUTE(16,2) =<SUB. -- VERB> ATTRIBUTE(16,29) = <SUB. -- VERB>

Next using this automated linking information, one can understand from this linkage that things which can eat and drink only and cannot speak and buy (i.e. eatable & drinkable only) is Animals, also things which can eat ,buy, drink, and speak and cannot treat sickness (i.e. buyable, speakable , eatable, drinkable only) is a provoke (normal) human, and the man who can eat food , drink drinks , buy goods , speak languages and have the ability to care for sickness (treatable) is a doctor. And we can create also this group as in Figure 5.

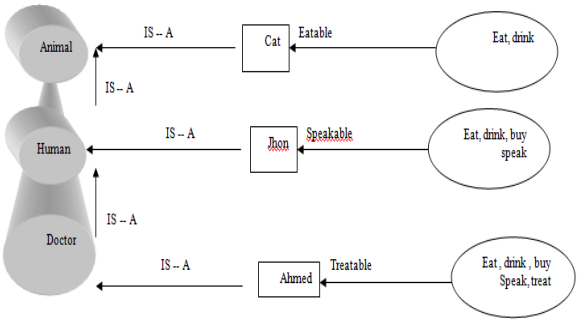


Fig. 5. Hierarchy and clear knowledge extract from link trie

From this link trie, we can get the < IS -- A> hierarchy relationship that *Doctor is a human, and human is an animal.*

V. SIMULATION RESULTS

A. Experimental data and information

99,714 statements from tagged corpus (Pan TreeBank), having diverse of features, is implicated in this experiment.

Data Set 1:

About 11,970 subject-verb case relationship and about 2,514 of verb-object relationship, and 679 verb-places are used. Due to high frequent access of pairs, we could not take them up. See Table 6.

TABLE VI. HIGH FREQUENT ACCESS OF PAIRS & LINKING

Verb	*Subject Frequent*	*Object Frequent*	*Place Frequent*
buy	holders (41) manufacturers (41) consumers (19) worldbank (13) people (12) company(10)	food (28) computer (27) recorder (18) clothes (14)	company(24) institutions (21) market (14) farm (10)
change	Jhon (80) farmers (35) rules (25) money (30)	money (70) shares (30) rate (43) dollar (50)	bank (31) hotel (22) company (13)
sell	traders (24) tourists (19) farmars (15) jhon (13) foreign (10)	car (61) currencies (41) computer (31) bus (50)	institutions (51) company (23)
read	Jhon (80) bank (30) company (48)	money (70) program (30) role (20) shirt (7)	bank (31) company (45) office (34)
eat	Jhon (100) people (98) farmars (70) tourists (19)	apple(150) orange(140) Duck (50)	office (34) resturant (70) company(12)
draw	foreign (100) farmars(67) jhon(50)	book(100) shirt(50) duck (40)	bank (31) institutions(20)

Data Set 2:

Utilizing case frame with trie structure to present a lot of relationships between words as shown in section II.

Data Set 3:

Employing trie structure with linking trie among leaves for additional information as shown in Figure 4, and Table 3.

Data Set4:

Restrict 10 group of typical verbs and objects from Data 1, as in Table 5.

Result [1].

By employing Data 1,2,3. We can establish an automated generation of hierarchy of relations among words as shown in Figure 5 and Algorithm 2.

Result [2].

By gathering this data and establishing relationships among verbs and other kinds of keys with link trie, we can see the hierarchy group. Figure 6 for example shows the subject and verb linkage.

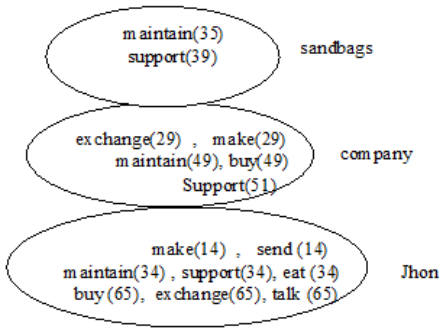


Fig. 6. Subject and verb Linkage group

Utilizing high-leaky this indicates that ***sandbags*** are maintainable and supportable but not exchangeable or buyable, ***company*** exchangeable , buyable supportable, maintainable , but not eatable ,talkable, but ***Jhon*** can , maintain, support, , buy, exchange, eat, talk.

Result [3]. Disambiguation

Figure 7 shows how the algorithm fare with sentences containing ambiguous element, be able to handle many such cases, as will be illustrated here. Consider the pair of sentences below:

1) Investment company support the bank.
2) The sandbags support the bank.

By this three sentence we show the semantic meaning of the word bank have two meaning financial house & edge of rive and by use more information about the word bank by another verbs, we can change the case to disambiguation case. As follow: The first approach: semantic meaning of bank is financial house in the first sentence, this by using another verbs to declare this meaning as in these sentence say : Jhon exchange from bank. and for more information about bank we can say that : Bank buy money . By this more information we find all sentence speak about money this implies more disambiguate for word bank and now the clear semantic is financial institution .the second approach: semantic meaning of bank is edge of river in the second sentence, this by using another verb to declare this meaning as in these sentence: Sandbags maintain bank, and for more information about bank we can say: Bank maintain river i.e. By this more information we find all sentence speak about hold up physically this implies more disambiguate for word bank and now the clear semantic is edge of river.

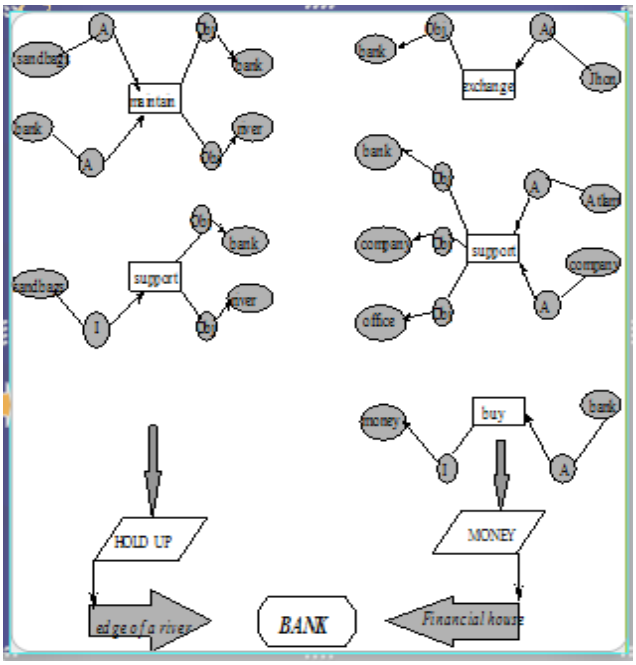


Fig. 7. Example of disambiguation

Result [4]. The accuracy of experimental results is defined as:

$$\text{Accuracy} = \alpha/\beta$$

Where α is the number of words disambiguated correctly and β is total number of ambiguous words.

Table 7 summarizes the experimental observation of using the old method (TM) and new method using the trie structure and linking trie between leaves. All English terms of our experimental are in Appendix A.

In Table 7, both performed with accuracies 73% for TM and 79% by new method using the link trie between leaves which is significant than the value that can be obtained by TM one. This means that, our new approach is viable in solving term ambiguities.

TABLE VII. EXPERIMENTAL RESULTS OF THE WHOLE WORDS

		TM*	NM*
		EN*	EN
Number of terms used in the Experiments			38 372
Sentence Ambiguities	No. of Ambiguous Words	520	520
	No. of Unambiguous Words	380	410
	Accuracy	73%	79%

* EN = English TM*= Traditional method NM*= New method (Using Linkage)

VI. CONCLUSION

In this paper, we proposed a new approach for building structure system of natural language knowledge. In this paper all surface case patterns are classified in advance with the consideration of the meaning of noun. Moreover, this paper introduces an efficient data structure using a trie which define the linkage among leaves and multi-attribute relations. By using this linkage multi-attribute relations, we can get a high frequent access among verbs and noun with an automated generation of hierarchical relationships. In our experiment a large tagged corpus (Pan Treebank) is used

to extract data. In our approach around 11,000 verbs and nouns is used for verifying the new method and made a hierarchy group of its noun. Moreover, the achievement of term disambiguating using our trie structure method and linking trie among leaves is 6% higher than old method. The preliminary result of our method shows a good promise, because the extracted information structures of a special database, can be extended by a more large input of data and more general relations from a large information corpus. The results of disambiguating the word ambiguities are much better than that of case frames. Experimental results also show that enough distinctive terms can help determine the semantic sense of a word in a specific context. The preliminary syntactic analysis can be achieved by many natural language processing system, we will be able to obtain more precise semantic information from the syntactic resource. Moreover, the accuracy of disambiguating words by our method using trie structure and linking trie between leaves is 6% higher than traditional method. Future work could focus in using context analysis to improve disambiguates of words. Extract Arabic keyword By using stop word Dictionary and stemming rule, from large Arabic Corpus with Classification for Arabic text by using Classification engine.

ACKNOWLEDGMENTS

The authors wish to acknowledge the approval and the support of this research study by the grant no. 5-9-1436-5 from the Deanship of Scientific Research in Northern Border University, Arar, KSA.

REFERENCES

[1] A.V. Aho, J. E. Hopcroft, and J. D. Ullman, “Date Structure and Algorithm,” Addison-Wesley, Reading, Mass., pp. 163-169, 1983

[2] M. Ai-Suwaiyel and E. Horowitz, “Algorith for Trie Compaction,” ACM Trans. IEICE, Vol. J76, D-II, No. 11, pp. 243-263, 1984

[3] J. Aoe, “An Efficient Digital Search Algorithm by Using a Double-array Structure”, IEEE Trans. Software Eng., Vol. 15, No. 9, pp. 1066-1077, 1989

[4] J. Aoe, K. Morimoto and T. Sato, “An Efficient Implementation of Trie Structure,” Software-Pract. & Expr. Vol. 22, No. 9, pp. 695-721, 1992

[5] J. Aoe, K. Morimoto, M. Shishibori, and K. Park, “A

[6] Trie Compaction Algorithm for Larg Set Keys”, IEEE. Trans. on Knowledge and Date Eng., Vol. 8, No. 3, 1996

[7] J. Aoe, K. Morita, H. Mochizuki, and Y. Yamakawa, “An Efficient Retrieval Algorithm of Collocational Information Using Trie Structures" (in Japanese), Transactions of the IPSJ, Vol. 39, No. 9, pp. 2563-2571, 1998

[8] J. Aoe, String Pattern Matching strategies, 1994.

[9] E. Brill, “A Case Study in A Part of Speech Tagging”, Computational Linguistics, Vol.21, No. 4, pp. 1-37, 1995.

[10] Abdunabi Ubul, El-Sayed Atlam, Hiroya Kitagawa, Masao Fuketa Kazuhiro Morita and Jun-ichi, Aoe An Efficient Method of Summarizing Documents Using Impression Measurements, COMPUTING AND INFORMATICS Journal, Volume 32, No. 2, 2013. Atlam E.-S, , Morita, K.., Fuketa M, Aoe, & J.. A new for selecting English compound terms and its knowledge representation. Information Processing & Management Journal, 38(6), 807–821. (2002

[11] Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. Document similarity measurement using field association terms. Information Processing & Managemen Journal, 39(6), 809–824. (2003).

[12] Atlam, E.-S, Elmarhomy, G., U. M. Sharif, Fuketa, M., Morita, K., & Aoe, J. Improvement of building field association term dictionary using passage retrieval. Information Processing & Management Journal, 43, 1793–1807. (2007).

[13] M. E. Abd El-Monsef, El-Sayed Atlam and O. El-Barbary, Combining FA Words with Vector Space Models for Arabic Text Categorization, An International Journal of INFORMATION, Vol. 6, No.(6A), pp.3517-3528, 2013.

[14] Atlam El-S. and El-Barbary O., Arabic Document Summarization using FA Fuzzy Ontology, , International Journal of Innovative Computing, Information and Control, 2014.

[15] Atlam El-S., Improving the Quality of FA Word Dictionary based on Co-occurrence Word Information and its Hierarchically Classification, , International Journal of INFORMATION Vol.17, No.2, February, 2014.

[16] K. Dahlgren, Naive semantics for Natural Language Understanding, 1982.

[17] W. B. Frakes, Information Retrieval Data Structure &Algorithms, 1992.

[18] E. Fredkin, “Trie Memory”, Commun. ACM., Vol. 9, No. 2, pp. 490-500, 1960

[19] D. E. Knuth, “The Art of Computer Programming”, Vol. 3, Sorting and Search, pp. 481-505, 1973

[20] F. Fukumoto, “Disambiguating preposition phrase attachment using statistical information” , NLPRS., Vol. 34, No. 2, pp. 752-757, 1995.

[21] Y. Jin, and Y. Tackkim, “Noun-sense Disambiguation from the Concept Base in MT” , NLPRS., Vol. 32, No. 2, pp. 357-362, 1995.

[22] J. Kupiec “A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia, In proceedings of 16th ACM SIGIR international conference, pp. 181-190, 1993.

[23] H. Li., and N. Abe, “ Clustering Words With the MDL Principle”, Journal of Natural Language Processing, Vol. 4, No. 2, pp. 71-88, April 1997.

[24] K. Lim and M. Song, “Morphological Analysis with Adjacency Attributes and Word Dictionary”, In proceedings of the international conference on computer processing of oriental language, pp. 263-268, 1994.

[25] D. W. Loveland, Natural Language Parsing system, 1987

[26] M. Nagao, J. Tsujii, and J. Nakamura, “Machine Translation from Japanese to English”, Vol. 74, No. 7, pp. 993-1012, 1986

[27] A. Oishi, and Y. Matsumoto, “ A Method for Deep Case Acquisition Based on surface Case Pattern Analysis”, NLPRS., Vol. 34, No.2, pp. 678-684, 1995.

[28] T. A. Standish, Data Structure Techniques, 1981

[29] R. E. Tarjan and A. C. Yao, ”Sorting a Sparse Table”, Commun. ACM., Vol. 22, No. 11, pp. 606-611, 1979.

[30] T.Takenobu and I.Makoto, “Text Categorization Based on Weight Inverse Document Frequency”, SIG-IPSJ, pp. 33-39, 1994.

[31] T. Satomi, Atlam El-S., Morita K., Fuketa M. and Jun-ichi Aoe Context Analysis Scheme of Detecting Personal and Confidential Information, , International Journal of Innovative Computing, Information and Control, Vol.8, 5(A), pp.3115-3134, 2012

[32] A. Utsumi, K. Hori, and S. Ohsuga “ An Affective- Similarity-Based Method for Comprehending Attributional Metaphors” , Journal of Natural Language Processing, Vol. 5, No. 3, pp. 3-30 July 1998.

APPENDIX A: ENGLISH WORDS IN EXPERIMENTS

Number	English words	Semantics
1	Kilimanjaro	[Place \ Mountain] [Food \ Drink] [Product \ Software]
2	puma	[Cougar] [Car] [Sportswear]
3	beetle	[Animal \ Insect] [Car]
4	polo	[Sports] [Clothing]
5	hawk	[Animal] [Aircraft in military] [Sports club]
6	queen	[Person] [Animal] [Games]
7	jaguar	[Animal] [Car]
8	apple	[Fruit] [IT company]
9	panda	[Animal] [Chinese car] [Fast-Food] [Software]
10	fish	[Animal] [Food]
11	blackberry	[Fruit] [Mobile phone]
12	Barcelona	[Place \ City] [Sports club]
13	office	[Working place] [Software]
14	thunderbird	[Animal] [Software] [Food \ Drink] [Aircraft in military]
15	tree	[Plant] [Structure in computer science]
16	tiger	[Animal] [A name of golf player] [Car] [Beer]
17	dove	[Animal] [Aircraft] [Food \ Chocolate] [Toiletry]
18	Mont Blanc	[Place \ Mountain] [Writing instruments and accessories]
19	chocolate	[Food] [Place \ Mountain] [Mobile phone]
20	window	[Object] [Software]
21	match	[Tool \ Fire] [Game] [Japanese car]
22	branch	[Plant] [Structure in computer science]
23	Liverpool	[Place \ City] [Sports club]
24	phoenix	[Film] [Sports club] [Television , Broadcasters]
25	Oracle	[Ancient text] [Software company]
26	cobra	[Snakes] [Aircraft] [IT products]
27	rocket	[Vehicle, missile, aircraft] [Sports club]
28	maverick	[Animal] [Sports club] [Car]
29	mustang	[Animal] [Aircraft in military] [Car]
30	QQ	[Messaging program] [Chinese car]
31	lotus	[Plant] [Car]
32	Amazon	[Geography, river] [IT company]
33	crocodile	[Animal] [Aircraft in military]
34	penguin	[Animal] [Clothing] [Sports club]
35	virus	[Program] [Infectious agent]
36	bridge	[Architecture] [Sports game] [Hardware]
37	line	[Object \ Product] [Formation] [Calling] [Cord]
38	bass	[Musical sense] [Animal \ Fish]
39	cone	[Part of tree] [Sharp of object] [Part of eye]
40	interest	[Curiosity, attraction] [Advantage] [Financial] [Share]
41	taste	[Preference] [Flavor]
42	sentence	[Punishment] [Set of words]
43	train	[Object \ Series] [Movement \ Prepare]
44	book	[Object \ Published document] [Movement]
45	bank	[Institution] [Architecture \ Ground]
46	serve	[Movement \ Ball game] [Movement \ Food]
47	dish	[Food \ Meal] [Receptacle]

Combination of Neural Networks and Fuzzy Clustering Algorithm to Evalution Training Simulation-Based Training

Lida Pourjafar
Department of Computer, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran
Department of Computer, Khou Zestan Science and Research Branch, Islamic Azad University, Ahvaz Iran

Mehdi Sadeghzadeh
Department of Computer Engineering, Mahshahr Branch Islamic Azad University, Mahshahr, Iran
Department of Computer, Khouzestan Science and Research Branch, Islamic Azad University, Ahvaz, Iran

Marjan Abdeyazdan
Department of Computer Science, College of Electricity and Computer, Mahshahr Branch Islamic Azad University, Iran
Department of Computer, Khouzestan Science and Research Branch, Islamic Azad University, Ahvaz, Iran

***Abstract*—With the advancement of computer technology, computer simulation in the field of education are more realistic and more effective. The definition of simulation is to create a virtual environment that accurately and real experiences to improve the individual. So Simulation Based Training is the ability to improve, replace, create or manage a real experience and training in a virtual mode. Simulation Based Training also provides large amounts of information to learn, so use data mining techniques to process information in the case of education can be very useful. So here we used data mining to examine the impact of simulation-based training. The database created in cooperation with relevant institutions, including 17 features. To study the effect of selected features, LDA method and Pearson's correlation coefficient was used along with genetic algorithm. Then we use fuzzy clustering to produce fuzzy system and improved it using Neural Networks. The results showed that the proposed method with reduced dimensions have 3% better than other methods.**

Keywords*—Educational Data Mining; Simulation-Based Training; Dimensions Reduction; ANFIS

I. Introduction

Data mining is a kind of computer-based information system (CBIS), which can be used for big data warehouse, peer review, production information, and knowledge discovery. The traditional term of mining is affected the foundations of data mining. But instead of searching for minerals, here discover the knowledge. The purpose of data mining is to identify data patterns, hidden links with organized information, communication rules are structured, the unknowns are estimated to be classified topics, create homogeneous clusters of issues and a wide variety of findings that do not come easily obtained by classical CBIS be uncovered. By the way, the results of data mining are invaluable the basis for decision-making.

Education, a new basin for the use of data mining to discover knowledge, decision-making and provide recommendations. The use of data mining in education is early stages and created the field of "educational data mining". The first decade of this century marks the beginning of educational data mining.

Educational data mining can be an example for the design, assignments, methods and algorithms for data discovery learning environments. The purpose of data mining is to find patterns and make predictions of the behavior and development of trained people, content knowledge application environment, assessments, training and application functions to define. With the emergence development of computer technology, simulation systems closer to reality and is one of the most important and has become efficient tool in education. As mentioned, a virtual environment that simulates real conditions creates absolutely arises [1-2].

II. Literature Review

Tian et al in an article work on the evaluation of simulation-based training for pilots. In this article, they have used cubic learn and krikpatrick model. In fact, to review and evaluate of results is used of simplified krikpatrick model. The results show that simulation-based training to 26% better than the education based on booklet [3].

Pamela et al to evaluate simulation-based training for teaching assistants midwife at the birth of babies. Simulation-based training was conducted on 111 persons and 14 persons were trained as usual. This research was conducted at medical centers in Ghana. The analysis was performed using 4 surfaces krikpatrick model. The results showed that better results are Simulation-based training [4].

Natassia et al in an article work on driver training base on simulation. In this paper, they do the impact of simulation training for drivers of vehicles [5].

Sophia et al in an article work on eye surgery Simulation-based training. This article is used a review of five different database. In this paper, expressed validity of the model and learned the ability to transfer to the operating room has been measured [6].

Shu-Hsien et al [7] Recent advances in data mining involves the collection of the works of the last decade have offered.

Collection approaches began Narli, Ozgan and Alcan [8], the theory of multiple intelligences unaccounted for identifying the relationship between people and their learning styles used. By using multiple intelligences the data collected to teach the learning style scales and tests for prospective teachers.

Gonzalez et al [9], the hidden conditional random field applied to complete the reading assignment predict. They are classified as self-paced dialogue as a matter of classification to classify sequences considered dialogue to assess.

Barbell et al [10], they seek to solve some issues: How can learning processes using process models and control rule based, the optimization? How can process models created based on the concept of learning styles? Therefore, they are a method for modeling a student using a combination of learning styles and approaches proposed mining process, and it gave way to model pupil.

Terry, Pardos, Sarkozy and Heffernan [11], a clustering strategy with common building envelope created to predict the results of students' self-learning function.

Yu et al [12] features expanded with redundancy and discrete optimization techniques. Series of features were learned by logistic regression L1. Then, features were dense to help statistical techniques and random trees. Finally, the results were combined together with adjusted linear regression.

Approaches complete set with Levy and Wilensky [13] began in the behavior of query students studied in complex models, while their goal is to create an equation linking the physical parameters of the system. They looked at how students adapt their systems with different behaviors.

III. PROPOSED METHOD

In the framework presented in Figure 1 feature selection is shown. In this paper we proposed a method based on genetic algorithms with two different fitness function based on linear discriminant analysis and Pearson coefficient. A genetic algorithm is an effective method for solving optimization problems.

Here problem space is multi-dimensional, discrete and complex. Genetic algorithm represents each chromosome is a binary vector and each number represents a feature. If ***I*** show the collection features so that it is composed of N member, in this case a subset X of Y ($X \subseteq Y$) represents the i-th chromosome ***I*** with ***N*** gene so that it is equal to one if the i-th feature was selected and the otherwise is zero (Figure 1).

In addition to, simply coding solutions genetic algorithm for solving such problems is very appropriate because the search space of exponentially in a very difficult, complex and non-linear services. algorithm starts with a ***P*** random populations. The fitness of each chromosome created using appropriate fitness function is calculated as described.

After calculating the fitness, 80% chromosomes are selected using the roulette wheel. Combine the two point method is carried out. Then mutations in chromosomes or individuals carried a distinct possibility.

A. Fitness function based on linear discriminant analysis

We use a fitness function with two statements, a J index called the separability index. The second term represents the number of members of set and if it is desirable to be less. Separability index J is derived from linear discriminant analysis.

In this paper, it is assumed that each feature can be modeled as a random variable. Separability index can calculate uses the covariance matrix of the features.

Since the variance of the random variable dispersion around the mean to express a certain value the covariance matrix of n variables, probability distribution around the mean vector in n-dimensional space. If we have n random variables $\{x_1, x_2, ..., x_n\}$, so that each variable is m samples (dimensions m × n stored in the matrix D), the covariance matrix Σ, an n × n is the matrix of the elements in row i and column j indicates the covariance between the variables xi and yj is the (equation 1):

$$\Sigma[i, j] = Cov(x_i, y_j)$$
$$Cov(x_i, y_j) = \frac{1}{m}\sum_{l=1}^{m}(D(l,i) - \mu_i)(D(l, j) - \mu_j) \quad (1)$$

Here μ_i and μ_j are mean's variables in matrix D.

Two classes of observations by a specified means and variances to consider. Fisher separation between the two distributions for the variance between two classes to the variance within the class definition:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(\vec{w}^T.\vec{\mu}_{y=1} - \vec{w}^T.\vec{\mu}_{y=0})^2}{\vec{w}^T\sum_{y=1}\vec{w} + \vec{w}^T\sum_{y=0}\vec{w}} = \frac{(\vec{w}.(\vec{\mu}_{y=1} - \vec{\mu}_{y=0}))^2}{\vec{w}^T(\sum_{y=0} + \sum_{y=1})\vec{w}} \quad (2)$$

$$S_W = w^T S_W w \quad (3)$$
$$S_B = w^T S_B w \quad (4)$$

According to the above definitions can define separability index using eq 5:

$$J(\mathrm{I}) = tr(\frac{\mathrm{W}^T \sum_B \mathrm{W}}{\mathrm{W}^T \sum_W \mathrm{W}}) \quad (5)$$

Here, W is the transition matrix is defined as follows:

Available features four-dimensional space, so we assume N = 4. If the chromosome I = (0,1,0,1) is, in this case on chromosome 2 and 4 are traits I so:

$$W = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \quad (6)$$

tr (.) is a trace matrix. High levels of separability index J(I) show that the subsidiary created by chromosomes ***I*** have been

at the center of the well separation On the basis of the proposed fitness function and genetic algorithms can be defined relationship 7:

$$F(I) = \frac{J(I)}{N} + K\frac{N - N_I}{N} \quad (7)$$

Here, N number of available features, N_I number of features on chromosomes ***I*** and K constant to determine the effect of the second sentence in eq 7.

B. Pearson correlation-based fitness function

A feature largely is correlated with response variable and with other features in sub-features is at a lower correlation. Correlation is a measure of the strength of the relationship between the two variables in bi-directional.

$$Fitness_S = \frac{N\bar{r}_{cf}}{\sqrt{N + N(N-1)\bar{r}_{ff}}} \quad (8)$$

S subset of N feature, $\bar{r}_{cf}$ is that the average correlation feature-class and $\bar{r}_{ff}$ is average correlation features - features. Eq 8 is Pearson correlation relationship in which all variables have been homogeneous.

C. Neural Networks

Back propagation method is the iterative process that runs on a set of training samples and each of the outputs obtained by comparing the output target and this process continues until a specified condition stems. Target values can be labels for training (for classification issues) and continuous values (numerical calculations). For all samples by repeating the correction process is weights modified to minimize the mean square error between the outputs of the network and target values.

This modification is done in the weights in the reverse direction, so that the beginning of the last layer (output layer) started in the hidden layer continues to be the first hidden layer, hence it is called back propagation. There is also no guarantee the convergence of weight. The algorithm are shown in Table 1.

D. Fuzzy Inference System

At first, a fuzzy inference system based on the existing database defined using Takagi-Sugeno-Kang. Using FCM, fuzzy set of rules for modeling the behavior of the database is extracted. The number of fuzzy inputs used here is the number of features of the database and the number of outputs equal to the number of classes (clusters). Since the four clusters (scores), so we considered 4 levels of output. To define the initial fuzzy system of fuzzy membership functions are used. Figure 2 shows an example of fuzzy rules that used for clustering with 8 features in 4 clusters.

E. Neuro-Fuzzy Systems

Figure 3 shows the structure of neuro-fuzzy system for 8 features. As shown any features connected to four membership functions and the fuzzy rules have been used here. Combined method is used for training.

F. Datasets Used

Preparing the database in an educational institution in Mahshahr shooting took place. A total of 200 patients were used for this database. At first tab contains personal information was prepared and then people were trained based simulation. The following characteristics were obtained from each individual 17 are introduced.

The first feature points was shot from a distance of 175 meters. This feature was extracted from the training and expertise in the institute's rate.

The second feature shooting accuracy in the training. Shooting accuracy of the diversity of positions x and y on Targets for each beam i is obtained. Since each person was 6 shooting at this stage of the relationship 9 was used to calculate accuracy:

$$\alpha = \frac{\sum_{i=1, j=i+1}^{6} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{6} \quad (9)$$

Smaller numbers indicate greater accuracy in high regard. The third feature indicates the shooting accuracy in the training. Shooting accuracy of the calculated difference between each beam to the center of target position is calculated from equation 10.

$$\alpha = \frac{\sum_{i=1}^{6} \sqrt{(x_t - x_j)^2 + (y_t - y_j)^2}}{6} \quad (10)$$

x_t and Y_t the center of target location is here considered to be the origin. x_i and y_i the location of each shot is on target.

The fourth feature is in testing phase accuracy that can be obtained from the above equation.

The fifth feature of the profile of the individual units had been achieved.

The sixth feature of confidence in the firing position is that professionals in the Institute's rate.

The seventh Feature is grade in school was that of profiles of individuals.

The eighth Feature is Confidence in eight shooting rule. The ninth feature of confidence in the usefulness of the use of guns.

The tenth features is the accuracy of the test step. The eleventh feature is the body mass index was calculated from height and weight. Relationship between body mass and BMI of 11 is as follows.

$$BMI = \frac{mass_{Kg}}{height_m^2} \quad (11)$$

The twelfth degree of confidence about the performance characteristics shooting.

The thirteenth features is the left eye and 14th feature is the right eye sight.

The 15th Feature is the location of weapons.

The 16th feature is the months of military service status. The seventeenth feature is belief in focus.

IV. Evaluation of the Proposed Method

For this study we used a computer with certain properties, Including:

Processor: Intel Pentium(R) CPU G620, 2.60 GHz 2.60GHz

Installed memory (RAM): 4.00 GB

For modeling and simulation software program MATLAB version R2014a (8.3.0) 64-bit was used.

A. Neural Networks Standard

For non-reducing features the network includes 17 input that used the same database features. The number of neurons in the hidden layer to all states is constant and equal to 20. Since each class is shown with binary code so the output layer of the Neural Networks includes four neuron.

As described in previous, two different methods have been used to reduce the dimensions by using a genetic algorithm Table 2 shows characteristics of each method with the code number. The purpose of coding here, references in the text to the new method is better.

Table 3 shows recognition rates for the standard Neural Networks for database without reducing the size of the feature. As shown in table recognition rate for data in LDA1 and LDA2 is 94.5%.

B. Fuzzy Inference System

Table 3 shows recognition rates for fuzzy inference system for database no decrease in size. Here confusion matrix for all the data is shown. Here is a detection rate of 94.5% for the entire data. Here's detection rate is slightly less than the Neural Networks. Recognition rate for fuzzy inference system for databases with size reduction for LDA2 is better than other. Here confusion matrix for all the data is shown. Here 95% detection rate for all data.

C. neuro-fuzzy system

Figure 4 confusion matrix for neuro-fuzzy systems for database shows no decrease in size. Here confusion matrix for all the data is shown. Here is a detection rate of 94.5% for the entire data. Detection rate here is like Neural Networks.

Figure 5 confusion matrix for neuro-fuzzy system for databases with size reduction method shows LDA3. Here confusion only for all data matrix is shown. Here 95% detection rate for all data. Such as Neural Networks and fuzzy detection rate here is better than PC.

In the case of the Neural Networks with 13 features we have the highest detection rate is 94.5 and the amount is lower than the other two methods. While this happened to fuzzy inference systems also feature 13 to achieve the highest detection rate. But the combination of neuro-fuzzy system achieved the highest rate of diagnosis for nine features with this mode with less computing power to achieve better accuracy.

V. Conclusion

Here a genetic algorithm with binary encoded with two fitness function includes linear discriminant analysis and Pearson coefficient was used. Here the results of three simulations in various environments, including neural networks, fuzzy systems based on fuzzy clustering and neuro-fuzzy were compared. The highest detection rates in the neural networks, fuzzy systems and neuro-fuzzy inference system was shown. As has been stated that the Neural Networks is used in a state where the number is 13 features we have the highest detection rate is 94.5 and the amount is lower than the other two methods. While this happened to fuzzy inference systems also feature 13 is required to achieve the highest detection rates. But the combination of neuro-fuzzy system achieved the highest rate of diagnosis for 9 Features can achieve better accuracy.

References

[1] Bell, B.S. & Kozlowski, S.W.J. (2002). Adaptive guidance: enhancing self-regulation, knowledge, and performance in technology-based training. Personnel Psychology, 55, 267-307.

[2] Bell, B.S., Kanar, A.M. & Kozlowski, S.W.J. (2008). Current issues and future directions in simulation-based training in North America. International Journal of Human Resource Management, 19(8), 1416-1434.

[3] Yongliang Tian, Hu Liu, Jiao Yin, Mingqiang Luo, Guanghui Wu, (2015), Evaluation of simulation-based training for aircraft carrier marshalling with learning cubic and Kirkpatrick's models, Chinese Journal of Aeronautics, Volume 28, Issue 1, 152–163

[4] Pamela Andreatta, Florence Gans-Larty, Domitilla Debpuur, Anthony Ofosu, Joseph Perosky, (2011), Evaluation of simulation-based training on the ability of birth attendants to correctly perform bimanual compression as obstetric first aid, International Journal of Nursing Studies, 48, 10, 1275–1280

[5] Natassia Goode, Paul M. Salmon, Michael G. Lenné, (2013), Simulation-based driver and vehicle crew training: Applications, efficacy and future directions, Applied Ergonomics, 44, 3, 435–444

[6] Ann Sofia S. Thomsen, Yousif Subhi, Jens Folke Kiilgaard, Morten la Cour, Lars Konge, (2015), Update on Simulation-Based Surgical Training and Assessment in Ophthalmology : A Systematic Review, Ophthalmology, Available online.

[7] Shu-Hsien, L., Pei-Hui, C., & Pei-Yuan, H. (2012). Data mining techniques and applications – a decade review from 2000 to 2011. Expert Systems with Applications, 39(12), 11303–11311.

[8] Narli, S., Ozgen, K., & Alkan, H. (2011). In the context of multiple intelligences theory, intelligent data analysis of learning styles was based on rough set theory. Learning and Individual Differences, 21(5), 613–618.

[9] Gonzalez-Brenes, J. P., Duan, W., & Mostow, J. (2011). How to classify tutorial dialogue? comparing feature vectors vs. Sequences. In Proceedings of the 4th international conference on educational data mining, 169–178

[10] Holzhuter, M., Frosch-Wilke, D., & Klein, U. (2012). Exploiting learner models using data mining for e-learning: a rule based approach. In A. Pena-Ayala (Ed.), Intelligent and adaptive educational- learning systems: achievements and trends, smart innovation, systems and technologies, 77–105.

[11] Trivedi, S., Pardos, Z. A., Sarkozy, G. N., & Heffernan, N. T. (2012). Co-clustering by bipartite spectral graph partitioning for out-of-tutor prediction. In Proceedings of the 5th international conference on educational data mining, 33–40.

[12] Yu, H. F., Lo, H. Y., Hsieh, H. P., Lou, J. K., G.McKenzie, T., Chou, J.W., Chung, P. H., Ho, C. H., Chang, C. F., Wei, Y. H., Weng, J. Y., Yan, E. S., Chang, C. W., Kuo, T. T., Lo, Y. C., Chang, P. T., Po, C.,Wang, C. Y., Huang, Y. H., Hung, C.W., Ruan, Y. X., Lin, Y. S., Lin, S. Lin, H. T., & Lin, C. J. (2010). Feature engineering and classifier ensemble for KDD cup 2010. In Proceedings of the KDD 2010 cup 2010 workshop knowledge discovery in educational data, 1–12.

[13] Levy, S. T., & Wilensky, U. (2011). Mining students' inquiry actions for understanding of complex systems. Computers & Education, 56(3), 556–573.

[14] Jiawei Han, (2012), Data Mining: Concepts and Techniques, third edition, elsevire, USA, 398

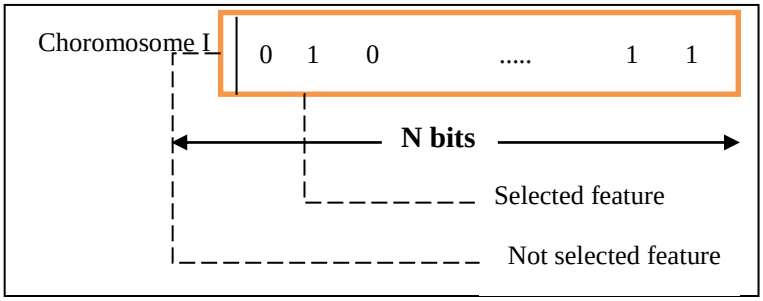


Fig. 1. Is an example of coding set by a series of two-bit Features

TABLE I. BACK PROPAGATION ERROR METHOD BASED [14]

Input:

- D, a data set consisting of the training tuples and their associated target values;
- l, the learning rate;
- *network*, a multilayer feed-forward network.

Output: A trained neural network.
Method:

(1) Initialize all weights and biases in *network*;
(2) **while** terminating condition is not satisfied {
(3) **for** each training tuple $\boldsymbol{X}$ in D {
(4) // Propagate the inputs forward:
(5) **for** each input layer unit j {
(6) $O_j = I_j$; // output of an input unit is its actual input value
(7) **for** each hidden or output layer unit j {
(8) $I_j = \sum_i w_{ij} O_i + \theta_j$; //compute the net input of unit j with respect to the previous layer, i
(9) $O_j = \frac{1}{1+e^{-I_j}}$; } // compute the output of each unit j
(10) // Backpropagate the errors:
(11) **for** each unit j in the output layer
(12) $Err_j = O_j(1 - O_j)(T_j - O_j)$; // compute the error
(13) **for** each unit j in the hidden layers, from the last to the first hidden layer
(14) $Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, k
(15) **for** each weight w_{ij} in *network* {
(16) $\Delta w_{ij} = (l) Err_j O_i$; // weight increment
(17) $w_{ij} = w_{ij} + \Delta w_{ij}$; } // weight update
(18) **for** each bias θ_j in *network* {
(19) $\Delta \theta_j = (l) Err_j$; // bias increment
(20) $\theta_j = \theta_j + \Delta \theta_j$; } // bias update
(21) } }

1. If (in1 is in1cluster1) and (in2 is in2cluster1) and (in3 is in3cluster1) and (in4 is in4cluster1) and (in5 is in5cluster1) and (in6 is in6cluster1) and (in7 is in7cluster1) and (in8 is in8cluster1) then (out1 is out1cluster1) (1)
2. If (in1 is in1cluster2) and (in2 is in2cluster2) and (in3 is in3cluster2) and (in4 is in4cluster2) and (in5 is in5cluster2) and (in6 is in6cluster2) and (in7 is in7cluster2) and (in8 is in8cluster2) then (out1 is out1cluster2) (1)
3. If (in1 is in1cluster3) and (in2 is in2cluster3) and (in3 is in3cluster3) and (in4 is in4cluster3) and (in5 is in5cluster3) and (in6 is in6cluster3) and (in7 is in7cluster3) and (in8 is in8cluster3) then (out1 is out1cluster3) (1)
4. If (in1 is in1cluster4) and (in2 is in2cluster4) and (in3 is in3cluster4) and (in4 is in4cluster4) and (in5 is in5cluster4) and (in6 is in6cluster4) and (in7 is in7cluster4) and (in8 is in8cluster4) then (out1 is out1cluster4) (1)

Fig. 2. Proposed four Fuzzy rules

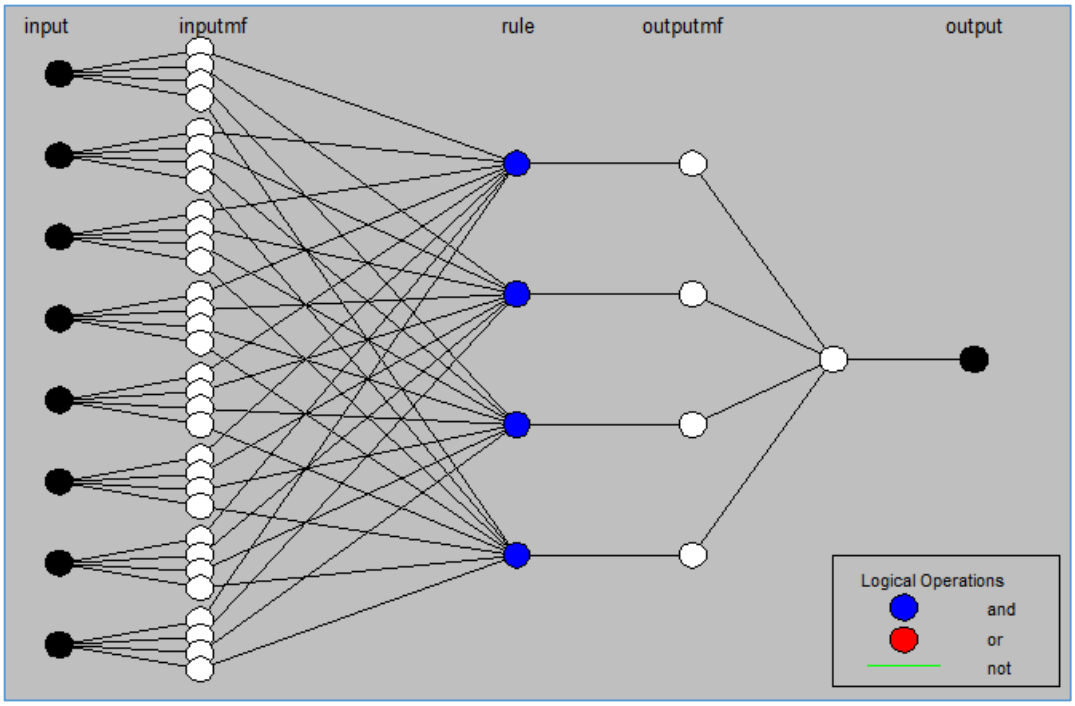


Fig. 3. Neuro-fuzzy inference system for 8 Features

TABLE II. NUMBER OF PROPERTIES WITH EACH PROCEDURE CODE

Dimension reduction method	Pierson coefficient	LDA K=0.8	LDA K=4	LDA K=8	LDA K=10
Number of feature	5	15	13	9	8
Method code	PC	LDA1	LDA2	LDA3	LDA4

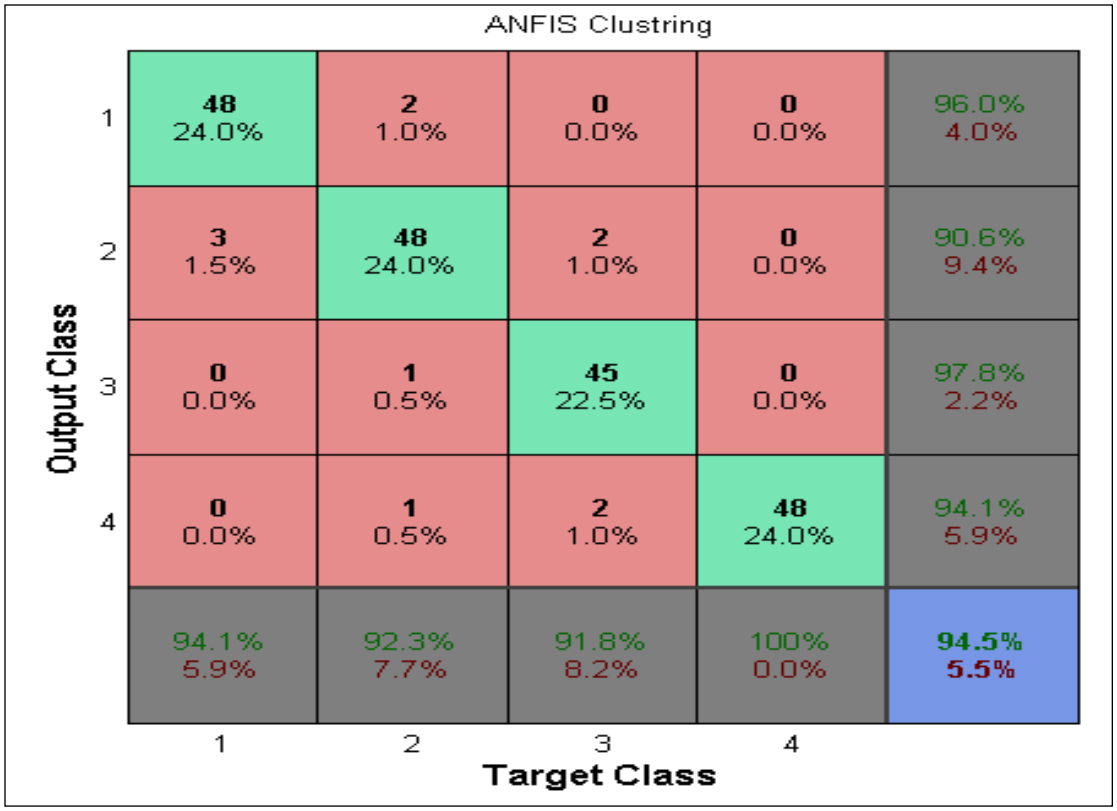


Fig. 4. Shows the confusion matrix for neuro-fuzzy system with no loss of features

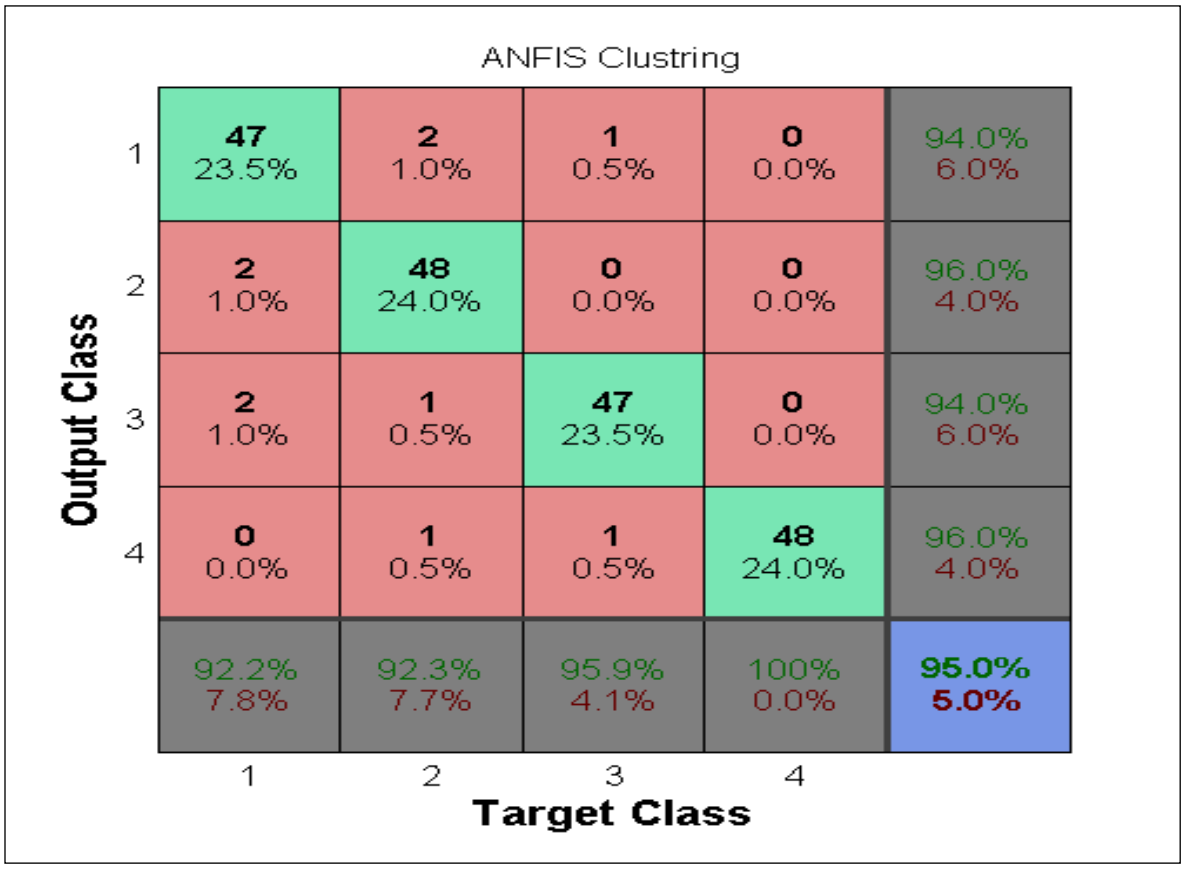


Fig. 5. Confusion matrix for neuro-fuzzy system with reduced features using LDA3

TABLE III. RECOGNITION RATE IN ALL CLASSIFIERS

Method	Dimension reduction	Recognition rate
Neural networks	17 Features	93.50
	PC	78.50
	LDA1	94.50
	LDA2	94.50
	LDA3	93.00
	LDA4	92.00
Fuzzy Inference System	17 Features	94.50
	PC	74.00
	LDA1	95.00
	LDA2	95.00
	LDA3	94.50
	LDA4	94.50
Neuro fuzzy	17 Features	94.50
	PC	83.00
	LDA1	94.50
	LDA2	94.00
	LDA3	95.00
	LDA4	94.50

A Proposed Quantitative Conceptual Model for the Assessment of Patient Clinical Outcome

Mou'ath Hourani
Software Engineering Department, Faculty of Information Technology
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—The assessment of patient clinical outcome focuses on measuring various aspects of the patient's health status after medical treatments and interventions. Patient clinical outcome assessment is a major concern in the clinical field as the current measures are not well developed and, as a result, they may be used without sufficient understanding of their characteristics. This issue retards the development in the clinical field. This paper proposes a general pure quantitative conceptual model for the assessment of patient clinical outcome. The proposed model contains five WHO's International Classification of Functioning, Disability, and Health (ICF) measurable components: body functions impairment, clinical elegancy distortion, pain, death, and shortening of life expectancy. Total patient clinical outcome is measured by summing the five WHO components. Five validity types are used to validate the proposed model: content, construct, criterion, descriptive, and predictive validities.**

Keywords—quantitative; conceptual model; assessment; patient clinical outcome

I. INTRODUCTION

The assessment of patient clinical outcome is an endpoint health care patient assessment. It is a measuring criteria that focuses on measuring various aspects of the patient's health status after medical treatments and interventions. Outcome assessments monitor and evaluate the quality of patient care by recording clinical outcomes resultant from treatment programs and interventions to observe their effectiveness [1, 2].

Recently, the assessment of clinical outcome in health care has been emphasized, in which the effectiveness of care is used to be determined based on purely clinician-centered outcomes to be based on more patient-centered outcomes. Patient-centered outcomes assess the patient's experiences and perceptions of his/her health status. Patient-oriented evidence may be obtained via patient self-report scales that collect a broad range of data that are significant to the patient to measure functional limitations and disability. The use of self-report scales facilitates the assessment of the net effects of both the health care services on the patient's health-related quality of life (HRQOL) and the condition on the patient. According to the means in which the data are gathered, clinical outcomes measures are classified as patient-based outcomes or clinician-based outcomes [1, 3, 4].

Patient-based outcomes provided by the patient via self-report questionnaires or surveys to identify his/her perspective regarding impairments, function, health, and HRQOL. In general, patients have concerns about the effect on their lifestyle due to their condition, including the capability to accomplish common activities such as dressing, attending events, joining in social occasions, and playing sports. The assessment tools of patient self-report outcomes, that capture patient's experiences and perceptions, are the best to evaluate variables such as disability, societal limitations, and quality of life. Patient self-report measures can be used during any point of the patient care to evaluate the status of a patient and to examine changes in the patient status which resulted from treatment of a medical condition. Patient-based outcomes cover up a broad range of health status factors, including symptoms, global judgments of health, physical function, cognitive functioning, psychological, personal constructs, role activities, and sensitivity to care [1, 3, 5].

Clinician-based outcomes are measures that clinicians do, on the patient's response to a treatment intervention, to stress the assessment of illness and impairment. Even though the clinician-based outcomes assessment is needed to assess the illness and the impairment, it can be misinterpreted and improperly used to deduce functional status. Some clinician-based measures assess functional limitations as patient goals should be directed toward enhancing function and disability rather than overcoming impairments. Such measures of functional limitations can then be used to direct treatment to improve activities that are difficult to be performed by the patient and are most important to the patient. Besides, third-party payers are acquiring an enhancement evidence of the patient's functional outcomes after a treatment program. In addition, self-report outcome tools are the best method to find out the disease effect on the patient's capability to carry out activities on a daily basis and to complete wanted or required roles and responsibilities.[1, 3, 6].

The research of clinical outcomes is the basis for evidence-based practice, that has set the standard for modern health care practice, in which it provides the finest research evidence that facilitates clinical decisions. Particularly, the most significant measures in assessing outcome are the patient-based outcome measures due to the importance of the patient's perception of health status and change in health status. High-quality evidence to verify the effectiveness of interventions can be derived from studies of clinical outcomes. Besides, patient-centered data can be obtained from studies of clinical outcomes that integrate patient-based outcomes. For example, there are two major reasons to justify the importance of clinical outcomes assessment and research studies of clinical outcomes within the athletic training profession. First, a clinical outcomes

assessment practice and research can present necessary knowledge to athletic trainers for providing best possible patient care. Second, the integration of clinical outcomes measures will allow the assessment of what is important to the patient and to provide patient-centered health care that concentrates on improving patients' HRQOL. Thus, the clinical outcomes assessment tools can be used to help in enhancing patient care by offering patient-oriented evidence for clinicians concerning the effectiveness of their interventions [1].

A number of clinical outcome assessment methods in some clinical fields have been employed without sufficient understanding of their characteristics [7]. This deficient in understanding the characteristics of outcome assessment methods together with the lack of high-quality tools for outcome measurement of illness manifestations restrain the development of therapy. Novel and enhanced methods of assessing the patient clinical outcome can accelerate the process of therapy development [2]. Accordingly, this paper is proposing a quantitative conceptual model for the assessment of patient clinical outcome. The proposed model contains five WHO's International Classification of Functioning, Disability, and Health (ICF) components measurable components: body functions impairment, clinical elegancy distortion, pain, death, and shortening of life expectancy. Total patient clinical outcome is measured by summing the five WHO components. Five validity types are used to validate the proposed model: content, construct, criterion, descriptive, and predictive validities. The remainder of this paper is organized as follows. Section 2 presents the components of the proposed model. Section 3 discusses the validation of the proposed model. In Section 4, we conclude this work and outline potential research directions

II. Ease of Use the Proposed Patient Clinical Outcome Assessment Model

A. General Definitions of Terms in the Proposed Model

- Harm: it includes any damage in the body function or structure including disease, injury, suffering and death [8, 9].
- Tolerable risk: it is the accepted risk that can be managed. [8].
- Causality assessment: it is concerned with the probability of adverse effect arise from using medicine. [10].
- Strength of clinical evidence: it is concerned with evidence quality [11].
- Patient characteristics: it the related attributes under focus [8].

B. The Mathematical Equations of Components of the Proposed Model

1) The Body Function Impairments Mathematical Equation

According to the WHO classification, there are eleven body functions that are used to calculate the multiple body function impairments score for (uncertain) diseases and adverse events [12]. The following formula us used to calculate the body impairments score:

$$BFI_S = \sum_{BF=1}^{N} IMPR_{BF} \times D_{BF} \times IR_{BF} \times OP_{BF} \times CR_{BF} \times SR_{BF} \quad (1)$$

Where:

BFI_S: (Uncertain) total body function impairments score.

$IMPR_{BF}$: Severity ratio of body function impairment.

D_{BF}: Duration of body function impairment.

IR_{BF}: Intolerability ratio of body function impairment.

OP_{BF}: Occurrence probability of body function impairment.

CR_{BF}: Causality ratio of body function impairment.

SR_{BF}: Strength of clinical evidence ratio of body function impairment.

BF: Body function impairment.

N: No. of body function impairment.

2) The Clinical Elegancy Distortions Mathematical Equation

Clinical elegancy represents the elegant components of the human body, which could be affected by disease or adverse event, and handled in the clinical setting. Clinical elegancy has mainly the following components: Physical appearance change, Undesired odor, Undesired taste, and Undesired audible. The following formula us used to calculate the multiple body size distortions score:

$$BSD_S = \sum_{BS=1}^{M} \left(\frac{|BSNS_P - BSDS_P|}{|BSNS_S - BSDS_S|} \right)_{BS} \times D_{BS} \times IR_{BS} \times OP_{BS} \times CR_{BS} \times SR_{BS} \quad (2)$$

Where:

BSD_S: (Uncertain) total body size distortions score.

$BSNS_P$: Body size at normal state for patient.

$BSDS_P$: Body size at distorted state for patient.

$BSNS_S$: Body size at normal state for the most severe clinical body size distortion case.

$BSDS_S$: Body size at distorted state for the most severe clinical body size distortion case.

D_{BS}: Duration of body size distortion.

IR_{BS}: Intolerability ratio of body size distortion.

OP_{BS}: Occurrence probability of body size distortion.

CR_{BS}: Causality ratio of body size distortion.

SR_{BS}: Strength of clinical evidence ratio of body size distortion.

BS: Body size distortion.

M: No. of body size distortions.

The following formula is used to calculate the multiple skin disclorations score:

$$SDD_S = \sum_{SD=1}^{Q} \left(\frac{DI_P \times BA_P}{DI_S \times BA_S} \right)_{SD} \times D_{SD} \times IR_{SD} \times OP_{SD} \times CR_{SD} \times SR_{SD} \quad (3)$$

Where:

SDD_S: (Uncertain) total skin discoloration distortions score.

DI_P: Skin discoloration intensity for patient.

BA_P: Body area affected with discoloration for patient.

DI_S: Skin discoloration intensity for the most severe clinical skin discoloration case.

BA_S: Body area affected with discoloration for the most severe clinical skin discoloration case.

D_{SD}: Duration of skin discoloration distortion.

IR_{SD}: Intolerability ratio of skin discoloration distortion.

OP_{SD}: Occurrence probability of skin discoloration distortion.

CR_{SD}: Causality ratio of skin discoloration distortion.

SR_{SD}: Strength of clinical evidence ratio of skin discoloration distortion.

SD: Skin discoloration distortion.

Q: No. of skin discoloration distortions.

The following formula is used to calculate the multiple skin hardness distortions score:

$$SHD_S = \sum_{SH=1}^{R} \left(\frac{HD_P \times BA_P}{HD_S \times BA_S} \right)_{SH} \times D_{SH} \times IR_{SH} \times OP_{SH} \times CR_{SH} \times SR_{SH} \quad (4)$$

Where:

SHD_S: (Uncertain) total skin hardness distortions score.

HD_P: Degree of skin hardness for patient.

BA_P: Body area affected with hardness for patient.

HD_S: Degree of skin hardness for the most severe clinical skin hardness case.

BA_S: Body area affected with hardness for the most severe clinical skin hardness case.

D_{SH}: Duration of skin hardness distortion.

IR_{SH}: Intolerability ratio of skin hardness distortion.

OP_{SH}: Occurrence probability of skin hardness distortion.

CR_{SH}: Causality ratio of skin hardness distortion.

SR_{SH}: Strength of clinical evidence ratio of skin hardness distortion.

SH: Skin hardness distortion.

R: No. of skin hardness distortions.

The following formula is used to calculate the multiple undesired odor distortions score:

$$UOD_S = \sum_{UO=1}^{T} \left(\frac{UOS_P}{UOS_S} \right)_{UO} \times D_{UO} \times IR_{UO} \times OP_{UO} \times CR_{UO} \times SR_{UO} \quad (5)$$

Where:

UOD_S: (Uncertain) total undesired odor distortions score.

UOS_P: Severity of undesired odor for patient.

UOS_S: Severity of undesired odor for most severe clinical undesired odor case.

D_{UO}: Duration of undesired odor.

IR_{UO}: Intolerability ratio of undesired odor.

OP_{UO}: Occurrence probability of undesired odor.

CR_{UO}: Causality ratio of undesired odor.

SR_{UO}: Strength of clinical evidence ratio of undesired odor.

UO: Undesired odor distortion.

T: No. of undesired odor distortions.

The following formula is used to calculate the multiple undesired taste distortions score:

$$UTD_S = \sum_{UT=1}^{U} \left(\frac{UTS_P}{UTS_S} \right)_{UT} \times D_{UT} \times IR_{UT} \times OP_{UT} \times CR_{UT} \times SR_{UT} \quad (6)$$

Where:

UTD_S: (Uncertain) total undesired taste distortions score.

UTS_P: Severity of undesired taste for patient.

UTS_S: Severity of undesired taste for most severe clinical undesired taste case.

D_{UT}: Duration of undesired taste.

IR_{UT}: Intolerability ratio of undesired taste.

OP_{UT}: Occurrence probability of undesired taste.

CR_{UT}: Causality ratio of undesired taste.

SR_{UT}: Strength of clinical evidence ratio of undesired taste.

UT: Undesired taste distortion.

U: No. of undesired taste distortions.

The following formula is used to calculate the multiple undesired audible distortions score:

$$UAD_S = \sum_{UA=1}^{V} \left(\frac{UAS_P}{UAS_S} \right)_{UA} \times D_{UA} \times IR_{UA} \times OP_{UA} \times CR_{UA} \times SR_{UA} \quad (7)$$

Where:

UAD_S: (Uncertain) total undesired audible distortions score.

UAS_P: Severity of undesired audible for patient.

UAS_S: Severity of undesired audible for most severe clinical undesired audible case.

D_{UA}: Duration of undesired audible.

IR_{UA}: Intolerability ratio of undesired audible.

OP_{UA}: Occurrence probability of undesired audible.

CR_{UA}: Causality ratio of undesired audible.

SR_{UA}: Strength of clinical evidence ratio of undesired audible.

UA: Undesired audible distortion.

V: No. of undesired audible distortions.

Finally, the following formula is used to calculate the Clinical elegancy distortions score:

$$CED_S = BSD_S + SDD_S + SHD_S + UOD_S + UTD_S + UAD_S \quad (8)$$

Where:

CED_S: (Uncertain) total clinical elegancy distortions score.

BSD_S: (Uncertain) total body size distortions score.

SDD_S: (Uncertain) total skin discoloration distortions score.

SHD_S: (Uncertain) total skin hardness distortions score.

UOD_S: (Uncertain) total undesired odor distortions score.

UTD_S: (Uncertain) total undesired taste distortions score.

UAD_S: (Uncertain) total undesired audible distortions score.

3) Physical and Non Physical Pains Mathematical Equation

Pain is defined as an extremely unlikable physical feeling due to illness or injury. Pain is associated by tissue damage and emotional experience. [13]. Depression, anger, frustration, fear, and anxiety feelings are examples of emotional pain [14]. The following formula is used to calculate the multiple pain types score:

$$PT_S = \sum_{PT=1}^{W} \left(\frac{PTI_P}{PTI_S} \right)_{PT} \times D_{PT} \times IR_{PT} \times OP_{PT} \times CR_{PT} \times SR_{PT} \quad (9)$$

Where:

PT_S: (Uncertain) total pain types score.

PTI_P: Pain type intensity or severity for patient.

PTI_S: Pain type intensity or severity for the most severe clinical case of the same pain type.

D_{PT}: Duration of pain type.

IR_{PT}: Intolerability ratio of pain type.

OP_{PT}: Occurrence probability of pain type.

CR_{PT}: Causality ratio of pain type.

SR_{PT}: Strength of clinical evidence ratio of pain type.

PT: Pain type.

W: No. of pain types.

Pain severity can be quanitifed and scale standardized to enable comparability between differnet cases [15].

4) Death Mathematical Equation

Death is a body functionalities and all clinical elegancy component distortion. Death is associated with the presence of all extreme pain types, therefore, the sum of the highest clinical values for body function impairments, clinical elegancy distortions, and different pain types is used to calculate the death clinical score. Constant value is added to the first three components model to differentiate severity of different clinical death cases. The following formula is used calculate death clinical score:

$$DCS = (HBFI_S + HCED_S + HPT_S + MCV) \times OP_D \times CR_D \times SR_D \quad (10)$$

Where:

DCS: (Uncertain) death clinical score.

$HBFI_S$: Highest clinical value recorded for body function impairments.

$HCED_S$: Highest clinical value recorded for clinical elegancy distortions.

HPT_S: Highest clinical value recorded for different pain types.

MCV: Smallest clinical value recorded for body function impairments, clinical elegancy distortions, or different pain types.

OP_D: Occurrence probability of death.

CR_D: Causality ratio of death.

SR_D: Strength of clinical evidence ratio of death.

C. Total Value of Patient Clinical Outcome Mathematical Equation

By summing all model components, the following formula is used to calculate the patient clinical outcome:

$$PCO = BFI_S + CED_S + PT_S + DCS \quad (11)$$

Where:

PCO: Patient clinical outcome.

BFI_S: (Uncertain) total body function impairments score.

CED_S: (Uncertain) total clinical elegancy distortions score.

PT_S: (Uncertain) total pain types score.

DCS: (Uncertain) death clinical score.

III. The Validation of the Proposed Model

Validity measures the reliability of the results obtained from experiment [16-18]. It also verifies the freedom of results from errors [19]. The proposed model is measured against three types of validity: content, construct, and predictive validities.

- Content validity: it concerns with experiment domain construction. It also signifies the domain clarity,

completeness and confirmation [17, 20-26]. In our model, the domain criteria that are included for validation are all ICF health items.

- Construct validity: it tests and assesses the logical relationships between related concepts [17, 20, 22]. In our proposed model, the logical relationships between all health dimensions are represented and precisely defined.
- Descriptive validity: it is the expression of proposed decisions in any situations [27]. In our proposed model, the output from the model is verified by explaining and describing the obtained decisions in any clinical environment.

To sum up, there is no standard validity test for the results. Using the described validity test described previously does not certain the obtained results [27-29]. Therefore, before validating the results, a threshold should be established.

IV. CONCLUSIONS

This paper proposes a general pure quantitative conceptual model for the assessment of patient clinical outcome. The model contains five major measurable components which are mainly based on the WHO's International Classification of Functioning, Disability, and Health (ICF) components. Those components are body functions impairment, clinical elegancy distortion, pain, death, and shortening of life expectancy. Patient clinical outcome is calculated as the summation of model component values for the patient with specific characteristics and status. The proposed model is verified against five types of validity checks: content, construct, criterion, descriptive, and predictive validities. Future study will focus on identifying facilitators and barriers to the successful implementation of the proposed model in clinical practice.

REFERENCES

[1] T. C. V. McLeod, A. R. Snyder, J. T. Parsons, R. C. Bay, L. A. Michener, and E. L. Sauers, "Using disablement models and clinical outcomes assessment to enable evidence-based athletic training practice, part II: clinical outcomes assessment," Journal of athletic training, vol. 43, pp. 437–445, 2008.

[2] M. K. Walton, J. H. Powers, J. Hobart, D. Patrick, P. Marquis, S. Vamvakas, M. Isaac, E. Molsen, S. Cano, and L. B. Burke, "Clinical Outcome Assessments: Conceptual Foundation—Report of the ISPOR Clinical Outcomes Assessment–Emerging Good Practices for Outcomes Research Task Force," Value in Health, vol. 18, pp. 741-752, 2015.

[3] C. M. Clancy and J. M. Eisenberg, "Outcomes research: measuring the end results of health care," Science, vol. 282, pp. 245-246, 1998.

[4] D. T. Wade, "Outcome measures for clinical rehabilitation trials: impairment, function, quality of life, or value?," American journal of physical medicine & rehabilitation, vol. 82, pp. S26-S31, 2003.

[5] A. Kirkley and S. Griffin, "Development of disease-specific quality of life measurement tools," Arthroscopy: The Journal of Arthroscopic & Related Surgery, vol. 19, pp. 1121-1128, 2003.

[6] J. Binkley, "Measurement of functional status, progress and outcome in orthopaedic clinical practice," Orthop Phys Ther Pract, vol. 11, pp. 14-21, 1999.

[7] P. I. Spuls, L. L. Lecluse, M.-L. N. Poulsen, J. D. Bos, R. S. Stern, and T. Nijsten, "How Good Are Clinical Severity and Outcome Measures for Psoriasis?: Quantitative Evaluation in a Systematic Review," Journal of Investigative Dermatology, vol. 130, pp. 933-943, 2010.

[8] WHO, "The conceptual framework for the international classification for patient safety version 1.1: Final technical report," World Health Organization2009.

[9] W. Runciman, "Shared meanings: Preferred terms and definitions for safety and quality concepts," The Medical Journal of Australia, vol. 184, pp. S41-S43, 2006.

[10] K. Holloway, T. Green, E. Carandang, H. Hogerzeil, R. Laing, and D. Lee, Drug and therapeutics committees: A practical guide. France: World Health Organization, 2003.

[11] J. L. Brożek, E. A. Akl, P. Alonso-Coello, D. L. Lang, R. Jaeschke, J. W. Williams, B. Phillips, M. Lelgemann, A. Lethaby, J. Bousquet, G. H. Guyatt, H. J. Schünemann, and G. W. Group, "Grading quality of evidence and strength of recommendations in clinical practice guidelines," Allergy, vol. 64, pp. 669-677, 2009.

[12] WHO, International Classification of Functioning, Disability and Health (ICF) short version. Geneva: World Health Organization, 2001.

[13] IASP, "Pain terms: A list with definitions and notes on usage," Pain, vol. 6, pp. 249-252, 1979.

[14] J. B. Wade, D. D. Price, R. M. Hamer, S. M. Schwartz, and R. P. Hart, "An emotional component analysis of chronic pain," Pain, vol. 40, pp. 303-310, 1990.

[15] R. L. Kane, B. Bershadsky, T. Rockwood, K. Saleh, and N. C. Islam, "Visual Analog Scale pain reporting was standardized," Journal of Clinical Epidemiology, vol. 58, pp. 618-623, 2005.

[16] D. Sornette, A. B. Davis, K. Ide, K. R. Vixie, V. Pisarenko, and J. R. Kamm, "Algorithm for model validation: Theory and applications," Proceedings of the National Academy of Sciences, vol. 104, pp. 6562-6567, April 17, 2007.

[17] L. Mokkink, C. B. Terwee, D. L. Patrick, J. Alonso, P. W. Stratford, D. L. Knol, L. M. Bouter, and H. C. W. de Vet, "The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes," Journal of Clinical Epidemiology, vol. 63, pp. 737-745, 2010.

[18] D. Bryant and N. Fernandes, "Measuring patient outcomes: A primer," Injury, vol. 42, pp. 232-235, 2011.

[19] E. J. Thomas and L. A. Petersen, "Measuring errors and adverse events in health care," Journal of General Internal Medicine, vol. 18, pp. 61-67, 2003.

[20] K. N. Lohr, N. K. Aaronson, J. Alonso, M. Audrey Burnam, D. L. Patrick, E. B. Perrin, and J. S. Roberts, "Evaluating quality-of-life and health status instruments: Development of scientific review criteria," Clinical Therapeutics, vol. 18, pp. 979-992, 1996.

[21] S. J. Coons, S. Rao, D. L. Keininger, and R. D. Hays, "A Comparative review of generic quality-of-life instruments," PharmacoEconomics, vol. 17, pp. 13-35, 2000.

[22] Scientific Advisory Committee of the Medical Outcomes Trust, "Assessing health status and quality-of-life instruments: Attributes and review criteria," Quality of Life Research, vol. 11, pp. 193-205, 2002.

[23] M. Ryan, D. Scott, C. Reeves, A. Bate, E. v. Teijlingen, E. Russell, M. Napper, and C. Robb, "Eliciting public preferences for healthcare: A systematic review of techniques," Health Technology Assessment vol. 5, 2001.

[24] J. H. Duffus, M. Nordberg, and D. M. Templeton, "Glossary of terms used in toxicology, 2nd edition (IUPAC Recommendations 2007)," Pure & Applied Chemistry, vol. 79, pp. 1153-1344, 2007.

[25] G. H. Guyatt, D. H. Feeny, and D. L. Patrick, "Measuring health-related quality of life," Annals of Internal Medicine, vol. 118, pp. 622-629, April 15, 1993.

[26] A. Y. Chen and A. S. Whigham, "Validation of Health Status Instruments," Journal for Oto - Rhino - Laryngology and Its Related Specialties, vol. 66, pp. 167-172, 2004.

[27] C. McCabe and S. Dixon, "Testing the validity of cost-effectiveness models," PharmacoEconomics, vol. 17, pp. 501-513, 2000.

[28] J. Hay, J. Jackson, B. Luce, J. Avorn, and T. Ashraf, "Panel 2: Methodological issues in conducting pharmacoeconomic evaluations–modeling studies," Value in Health, vol. 2, pp. 78-81, 1999.

[29] C. B. Terwee, F. W. Dekker, W. M. Wiersinga, M. F. Prummel, and P. M. M. Bossuyt, "On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation," Quality of Life Research, vol. 12, pp. 349-362, 2003.

Identifying and Prioritizing Evaluation Criteria for User-Centric Digital Identity Management Systems

Sepideh Banihashemi / Master Candidate
Department of Computer Science
Ryerson University
350 Victoria Street Toronto, Ontario, Canada
Shahid Beheshti University
Tehran, 1983963113, Iran

Alireza Talebpour/Assistant professor
Cyberspace Research
Institute
Shahid Beheshti University
Tehran, 1983963113,
Iran

Elaheh Homayounvala/Assistant Professor
Cyberspace Research Institute
Shahid Beheshti
University
Tehran, 1983963113, Iran

Abdolreza Abhari / Professor
Department of Computer Science
Ryerson University
350 Victoria Street
Toronto, Ontario, Canada

***Abstract*—Identity Management systems are used for securing digital identity of users in reliable, automated and compatible way. Service providers employ identity management systems which are cost effective and scalable but cause poor usability for users. Identity management systems are user-centric applications which should be designed by considering users' perspective. User centricity is a remarkable concept in identity management systems as it provides more powerful user control and privacy. This approach has been evolved from amending past paradigms. Thus, evaluation of digital identity management systems based on users' point of view, is really important. The main objective of this paper is to identify the appropriateness of the criteria used in evaluation of user-centric digital identity management systems. These criteria are gathered from the literature and then categorized into four groups for the first time in this work to examine the importance of each parameter. In this approach, several interviews were performed as a qualitative research method and two questionnaires have been filled out by forty six users who were involved with identity management systems. Since the answers are perception based data the most important criteria in each category are assessed by using fuzzy method. This research found that the most important criteria are related to security category. The results of this research can provide valuable information for managers and decision makers of hosting companies as well as system designers to adapt and develop appropriate user-centric digital identity management systems.**

Keywords—management of information technology; digital identity management systems; evaluation criteria; fuzzy analytical hierarchy process (FAHP); user-centricity

I. INTRODUCTION

In today's Information Systems, users have various compounds of login-name and password for every online service or even distinct credentials for different roles inside the services which are available for them. This can result in privacy risk for end-users and jeopardize service providers by security threats. Applying identity management systems is therefore the solution. These systems issue a digital identity for each user and users can control the full life cycle of their identities, from creation to termination [30]. In federated identity management model, identities from various service provider, particularly identity domains, are identified across all domains [3]. The major goals of these systems are to increase user convenience and privacy, and to decentralize user management tasks inside or across the trust circle [1]. It is complicated for user to choose his/her identity provider along with username and password in federated identity management systems which can be considered as a drawback. In addition, the user should remember which federations he/she belongs to or can utilize [33]. In Single Sign-On (SSO) solution, the user authenticates him or herself only once and it is very similar to the federated identity scenario as the same identifier of the user is automatically used by each service provider when the user logged into. Capturing the information of authentication and identification by system and giving the user access to services is the functionality of single sign-on systems [29].

In order to overcome the complicated, unintuitive difficulties which affect the actual users' needs, the latest approach in identity management systems that is user-centric identity management systems has been emerged [32]. These systems support user control and privacy and designed from the users' perspective [7]. A beneficial control of the use and management of Personal Identification Information (PII) is considered in these systems [35]. There are two different user-centric identity management concepts: relationship-focused in which a relationship between the user and identity provider must be established, and credential-focused identity management that is offering user a long-term credentials from the identity provider and keeping them locally [6].An instance of a user-centric identity management system is PRIME which is a European government-funded project [8]. Enhancing user control which is accepted by user-centric identity management paradigms is one the objectives of PRIME project along with

considering identity credential misuse and physical stealing of devices [26]. Former studies have presented features and frameworks, and considered numerous metrics for user-centricity paradigm, but most of them have not examined the prioritizing of a comprehensive classification concerning the most important criteria. Since, user-centric digital identity management approach has been emerged in order to conquer the drawbacks of previous identity management models, and concentrates particularly on users' perspective rather than other entities, prioritizing the evaluation criteria which can be aggregated in a universal system is very worthwhile and will assist the design and implementation of these systems beneficially. This paper identifies and prioritizes evaluation criteria for user-centric digital identity management systems. These evaluation criteria have been surveyed in the literature and then evaluated through several interviews with experts in Iran and Canada. In this research, Fuzzy Analytical Hierarchy Process (AHP) has been applied for prioritizing identified criteria, by providing a web-based questionnaire based on the most important criterion in each group.

The rest of this paper is structured as follow. Section two provides literature review. The paper's approach is evaluated in section three. In particular, fuzzy AHP and pairwise comparisons which have been performed in each category are discussed in section four. Finally, we conclude and give an outlook for future work in section five.

II. Literature Review

In order to identify appropriate criteria for evaluating identity management systems, we have conducted a vast search in related literature. In this section, user-centric digital identity management systems are examined based on their characteristics and requirements. Vossaert et al. [40] proposed a user-centric federated identity management approach based on trusted secure modules which meets several requirements, including: 1) Verification to prove that the only information from identity providers for which they gave their consent, is inquired. 2) Performing access restriction to the information by users. 3) Managing the disclosure of personal information. 4) Trustworthiness of service providers in order to request their information. 5) A flexible revocation procedure can be predicted. 6) Scalability property in order to add new identity and service providers. 7) User consent on release of data.

According to Ahn et al. [1] privacy is a major issue as a result of the immense exchange of sensitive information. Pseudonymity is the key principle for protecting user identities and personal information. Furthermore, user-centric models used in the organizations are required to pursue four key principles: 1) Notice: gaining notice about information practice. 2) Choice: Users have the capability of the usage of information type and its purpose. 3) Access: Users should have access to their personal information and be able to modify it whenever is essential. 4) Security: Organizational system must confirm securing users' personal information.

As stated by Ahn et al. [2] an identity metasystem is designed to provide minimal disclosure for a limited usage and consistent experience across contexts in order to improve security and privacy enhanced interoperable architecture, based on the laws of identity.

Poursalidis et al. [31] introduced a multi-pseudonym Identity Management Infrastructure in which users can manage and make an excessive amount of pseudonyms. Their scheme has several advantages. First, users can maintain their anonymity. Next, preventing the existence of a single point that keeps numerous digital identities to preserve the privacy of the user.

According to Ben Ayed et al. [5] the notion of user-centricity has emerged by offering convenience and control to the users over their personal data and fulfilling to their requirements. The attribute management systems are developed to guarantee that any system section can't collect an individual's confidential attributes. From privacy-preserving perspective, keeping track of which digital identity attributes have been revealed and operate by whom, are also considerable issues. In order to prohibit other parties' unpleasant context-spanning linkage and profiling, pseudonyms can be applied.

Claycomb et al. [12] discussed that, the user control over the kind of information being kept, the actual content of the information and the authorizing individuals to view the information are the major motivations in the concept of user-centric identity management systems. Another motivation is privacy and confidentiality, accomplished by offering users the option about what is shared, and with whom it is shared. Furthermore, various service providers such as financial institutions or online merchants must use a centralized repository of user information. Scalability and data authenticity should be taken into account as well.

Jøsang et al. [21] proposed a user-centric identity management approach in a single tamper resistant device in order to improve usability, simplify the user experience, provide mobility by supporting the user in using any hardware platform while obtaining online services and enhance user control. These systems introduce process automation and system support of the identity management at the user side.

According to El Maliki et al. [17] there are some basic rules which have been considered in the new user-centric identity paradigm, specifically: 1) Enhancing the user privacy by providing them full control over their identity information 2) Usability and user experience quality as a result of consistent identity interface and using the same identity for each identity transaction 3) Decreasing identity attacks, including phishing 4) Reducing reachability/disturbances caused by spams 5) Policy specification on both sides, identity providers and service providers 6) Profiting from huge scalability 7) Providing secure conditions at the time of data exchange 8) Separating the digital data from applications.

As stated by Suriadi et al. [35] communication security, minimal data sharing and disclosure, negotiation, user registration, anonymous authentication, data storage, accountability and user control are the requirements for user-centric identity management systems. It also requires that users have an effective control of the use and management of their personal identifiable information, leading to a better privacy.

Some properties have been laid out in Bhargav-

Spantzeletal et al. upon which user-centric federated identity management is based on. The key properties of a user-centric federated identity management system are user control and consent, and numerous system properties help to achieve user control. The properties that are not based on the realization of other properties are basic properties whereas composite properties are composed of basic properties. There are four basic system properties: 1) User chosen identity provider 2) Policy specification and enforcement 3) Auditing 4) Assurance support. Another basic property is transaction property. Transaction properties concern all the transactions which deal with identity-related information that is: 1) Context bound transactions 2) Unlinkability 3) User consent. The final properties in this category are identity information properties which are: 1) Confidentiality 2) Integrity 3) Availability 4) Stealing protection 5) Revocation 6) Portability 7) Sharing prevention 8) Selective release and 9) Conditional release. Several composite properties are defined which build on one or more of the basic properties: 1) Attribute security 2) Service protection 3) Non-repudiation 4) Data minimization 5) Attribute privacy 6) Accountability 7) Privacy policy, obligations, and restrictions 8) Notification 9) Anonymity 10) User in the middle. It is also stated that multi-device management and usability are the unique properties which are essentials for these systems. Usability addresses the relationship between the user-centric tools and their users. Some key aspects are 1) To have consistent user experience, 2) An intuitive and easy UI which may also help required functionality from the user like policy specification, and finally 3) Process automation that is, automating user-side processes of identity management as far as possible through policy and preferences-driven methods [6]. On the other hand, some research projects look at Digital Identity Management as the core of the Internet economy and from public policy concept [14]. Or another research project studies identity through one's whole life. [19].

To sum up, previous research projects have surveyed key principles and properties required in user-centric digital identity management systems. Our work demonstrates taxonomy of criteria in terms of security, user control, system capabilities and cost-effectiveness. These groups of criteria and criteria within each group are first evaluated and then prioritized based on fuzzy Analytical Hierarchy Process.

III. Evaluation Approach

As the first step to evaluate identity management criteria, a thorough list of identified criteria was provided to the specialists in this domain in order to obtain their verification. Then a common decision making tool has been used to prioritize these criteria.

A. Decision Making Models

In recent decades, researchers have paid attention to multi criteria decision making model (MCDM) for complex decision making. In such models, instead of using one optimal evaluation criterion, several evaluation criteria may be used. [22]

These decision making models are categorized into two groups: Multi objective decision making models (MODM) and Multi attribute decision making models (MADM). Multi objective models are used to design the alternatives whereas multi attribute models include the choice of the best option [30]. One of the methods for MADM is Analytical Hierarchy Process (AHP) which is based on pairwise comparison [37].

B. Analytical Hierarchy Process

Analytical Hierarchy Process was developed by Thomas L. Saaty in 1970 which is a tool of decision making that can deal with structured and semi-structured decisions [23]. In AHP, both qualitative and quantitative features of human thoughts are included in decision making process. The analytical hierarchy process deals with the inconsistency because people are more likely to be inconsistent when they are making judgments. Therefore, the pairwise comparison matrix is used which is perfectly consistent [34].

The first step in AHP is creating a multi-level hierarchical structure of objectives, criteria, sub criteria, and alternatives [36]. Then, the priorities for each level of criteria are required which come from pairwise comparison [34]. These comparisons obtain the relative importance of each factor that is defined by their weights [37]. The decision maker has to present his idea about the value of one single pairwise comparison at a time [36]. After obtaining the relative weights, the best alternative can be determined from the aggregation value of them [37]. Relative weights can be evaluated from least square, geometric means, and eigenvalue methods [36]. In order to quantify pairwise comparison which is the most crucial step in decision making process, a scale is used. Since people cannot distinguish between two very close values of importance (e.g., 3.00 and 3.02), Saaty used 9 as the upper limit and 1 as the lower limit in his scale [11] and for the comparison of factors, the available values are the members of this set: {1.9, 1.8, …, 1.2, 1, 2, …, 8, 9} [38].

C. Fuzzy Analytical Hierarchy Process

Although the aim of applying Analytical Hierarchy Process is to obtain the opinions of experts, the typical AHP method does not reflect the human thoughts because the exact numbers are used in pairwise comparisons method. After supplying the graph of hierarchy in FAHP, the decision makers are asked to compare the elements of each level to each other and to express the relative importance of elements by using fuzzy numbers [9].

Van Laahoven *et al.* [38] have introduced the triangular fuzzy numbers based on vector operation to represent the decision maker's opinion for alternatives compared to each criterion.

Chang [9] introduced triangular fuzzy numbers as a new approach in fuzzy AHP. This approach uses triangular fuzzy numbers for pairwise comparisons in FAHP. Noorul Haq *et al.* [28] proposed a model to evaluate and select the supplier based on fuzzy AHP approach. The main advantage of their proposed method was considering qualitative and quantitative criteria in hierarchy structure and problem solving of supplier selection using fuzzy AHP. Lee *et al*. in [25] applied fuzzy AHP method for assessing the importance of effective factors in choosing the supplier. These factors include: cost, performance and number of suppliers. Then based on fuzzy

AHP results, goal planning was used to formulate the constraints. Lee [24] utilized the fuzzy AHP approach in order to analyze and evaluate the relation between the supplier and purchaser.

IV. Analysis and Results

A. Interview

Interviews are among the most familiar strategies for collecting qualitative data. The interview is a method in which, the researcher establishes direct contact with subjects and through this method he/she assesses the perceptions and attitudes. Table I shows the first full list of criteria which have been confirmed and modified by experts. For instance, according to them, confidentiality and user's privacy must be presented as one item, with respect to their definitions. In addition, it was stated that sharing prevention should be a second-order criterion related to security issues in that we only share credentials when we try to obtain services and then we need to invent security mechanisms to avoid identity theft and misuse. As a result of experts' verification and change, second list of criteria, as depicted in Fig 1, was prepared which indeed became an outline for the main questionnaire.

First interview resulted to removing some of the criteria. Security and stealing protection covers features and characteristics of some other criteria. Therefore, these criteria should be removed. In addition, unlinkability criterion should be eliminated since it can't be applied in face-to-face healthcare transactions. Policy specification and enforcement is also not obvious because it should be identified that the policies are related to entities or they are related to privacy policy. Sharing prevention should be considered as a second-level and related to security criteria since sharing the credentials; connection of apps is tempted to share feeds of "data" efficiently so, there is no need to share data by value. Data minimization is an important one but it's very hard to achieve given business model imperatives in most ecosystems. Scalability is one of the most main criterion because without that, no system is likely to succeed any more.

Another interview leads to merging security and stealing protection criteria as they both have the same meaning. In addition, anonymity criterion prevents from revealing identification information of a person and when the conditional release of information exist, this one is fulfilled too. Furthermore, Pseudonymity with anonymity were combined because a person has an identity in the system but he/she has a pseudonym and its anonym.

The outcome of third interview was that Notification should be considered as a second-level criterion related to systems capabilities' criteria. The definition of auditing criterion is that it must support enforcement of responsibility for actions among several loosely coupled identity actors in case of unexpected results. In addition, User Chosen Identity Provider criterion is a hard criterion to achieve but must be considered an important goal to strive for. Many governments are managing to achieve it through contracting with private sector partners.

In the last interview, it is concluded that Confidentiality and Privacy criteria can be considered as one criterion since they have two aspects: Data protection is usually about the service provider's intended security mechanisms, vs. its policies, where it may intend to release sensitive data because it suits the organization's own ends (such as making money). Additionally, Conditional Release seems very second-order criterion related to system and users' security criteria though it's an important one. Verifiability criterion must have remediation abilities in the face of incorrect data.

TABLE I. First List of Evaluation Criteria

Study	Criteria	Study	Criteria
Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al*. [32], Hoffman [20], Mashima *et al*. [26], El Maliki *et al.* [17]	Context Bound Transaction Context—Detection	Ahn *et al.* [1], Suriadi *et al*. [35], Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al*. [32], Mashima *et al.* [26]	Data Minimization Minimal disclosure Minimal data sharing
Ben Ayed *et al*. [4], Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al*. [32], Marx *et al*. [27]	Unlinkability	Suriadi *et al.* [35], Bhargav-Spantzeletal *et al.* [9], Quasthoff *et al.* [32], Marx *et al*. [27]	Accountability
Vossaert *et al.* [40], Claycomb *et al*. [12], Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al.* [32]	Confidentiality Controlling the disclosure of personal information	Ahn *et al*. [2], Bhargav-Spantzeletal et al, [9], Suriadi et al. [35], Quasthoff et al [32], Mashima et al. [26]	Notification Notice user awareness by SMS
Claycomb *et al.* [12], Bhargav-Spantzeletal *et al.* [9], Quasthoff *et al.* [32], Cottrell [13]	Integrity data authenticity Accuracy	Ben Ayed *et al*. [4], Claycomb *et al*. [12], Jøsang *et al*. [21], Suriadi *et al*. [35], Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al*. [32], Mashima *et al*. [26]	User in the middle giving sovereignty to the users over their personal data user control
Vossaert *et al.* [40], Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al*. [32], Mashima *et al*. [26], Marx *et al.* [27]	Verifiability	Ahn *et al*. [1], Jøsang *et al*. [21], El Maliki *et al*. [17]	User experience quality consistent experience simplify the user experience

El Maliki *et al*. [17], Bhargav-Spantzeletal *et al*. [9], Suriadi *et al*. [35], Quasthoff *et al*. [32], Poursalidis *et al*. [31], Mashima *et al*. [26]	Stealing Protection		Vossaert *et al*. [40], Claycomb *et al*. [12], El Maliki *et al*. [17], Marx *et al*. [27]	Scalability
Vossaert *et al*. [40], Bhargav-Spantzeletal *et al*. [6], Quasthoff et al. [32], Poursalidis *et al*. [31], Marx *et al*. [27]	Revocation		Vossaert *et al*. [40], Ahn *et al*. [2], Ahn *et al*. [1], El Maliki *et al*. [17], Poursalidis *et al*. [31]	Security
Vossaert *et al*. [40], Bhargav-Spantzeletal *et al*. [6], Quasthoff *et al*. [32]	Conditional release Access Restriction		Jøsang *et al*. [21], El Maliki *et al*. [17], Bhargav-Spantzeletal *et al*. [9], Mashima *et al*. [26]	Usability
Bhargav-Spantzeletal *et al*. [6], Suriadi *et al*. [35], Quasthoff *et al*. [32], Mashima *et al*. [26]	Sharing Prevention		Vossaert *et al*. [40], Bhargav-Spantzeletal *et al*. [6], Suriadi *et al*. [35], Quasthoff *et al*. [32], Mashima *et al*. [26]	User Consent (Negotiation: users should be allowed to negotiate on the PII that they want to reveal and at what level they are willing to disclose it)
Jøsang *et al*. [21], Quasthoff *et al*. [32], Bhargav-Spantzeletal *et al*. [6], Marx *et al*. [27]	Portability Mobility		Bhargav-Spantzeletal *et al*. [6], Mashima *et al*. [26], Vecchio *et al*. [39[40]	Delegation
Bhargav-Spantzeletal *et al*. [6], Rieger [33], Quasthoff *et al*. [32], Poursalidis *et al*. [31], Mashima *et al*. [26], Choi *et al*. [10]	User chosen Identity Provider		Jøsang *et al*. [21], Bhargav-Spantzeletal *et al*. [6], Ben Ayed [5], Marx *et al*. [27]	Fault Tolerant (tamper resistant)
Vossaert *et al*. [40], Bhargav-Spantzeletal *et al*. [6], El Maliki *et al*. [17], Quasthoff *et al*. [32], Mashima *et al*. [26]	Policy Specification and enforcement Privacy policy, obligation and restriction		Ahn *et al*. [2], Mashima *et al*. [26], Claycomb *at al*. [12]	Availability (Accessibility) (User access)
Vossaert *et al*. [40], Bhargav-Spantzeletal *et al*. [6], Mashima *et al*. [26]	Auditing the log of the transactions activities		Bhargav-Spantzeletal *et al*, [6], Vossaert *et al*. [40], Quasthoff *et al*. [32], Poursalidis *et al*. [31]	Service Protection
Vossaert *et al*. [40], Ben Ayed *et al*. [4], Bhargav-Spantzeletal *et al*. [9], Quasthoff *et al*. [32]	Attribute Security		Vossaert *et al*. [40], Ahn *et al*. [2], Ahn *et al*. [1], Poursalidis *et al*. [31], Ben Ayed *et al*. [4], Claycomb *et al*. [12], El Maliki *et al*. [17], Suriadi *et al*. [35], Poursalidis *et al*. [31]	Privacy
Vossaert *et al*. [40], Bhargav-Spantzeletal *et al*. [6], Poursalidis *et al*. [31], Marx *et al*. [27]	Dependable Trustworthiness Legitimacy of the end-entities Authorized entity Justifiable parties		Vossaert *et al*. [40], Ahn *et al*. [2], Poursalidis *et al*. [31], Suriadi et al. [35], Bhargav-Spantzeletal *et al*. [6], Quasthoff *et al*. [32], Poursalidis *et al*. [31]	Pseudonymity and anonymity

Fig. 1. Second list of criteria-Hierarchical decision tree

B. The Results of Solving Hierarchy Model Using Change Approach

TABLE II. FUZZY SPECTRUM AND THE CORRESPONDING VERBAL EXPRESSIONS

Row	Verbal Expressions	Fuzzy Numbers
1	Equally Important	(1, 1, 1)
2	Weakly Important	(0.75, 1, 1.25)
3	Strongly Important	(1, 1.25, 1.5)
4	Very Strongly Important	(1.25, 1.5, 1.75)
5	Extremely Preferred	(1.5, 1.75, 2)

Step1. Hierarchical decision tree of this project is created as it is shown in Fig. 1. Step2. In order to perform pairwise comparison, the verbal expressions are used, namely: Equally Important to Extremely Preferred, as depicted in Table II.

Results of fuzzy AHP approach for prioritizing the evaluation criteria are presented in this section. In other words, criteria in each category and their arithmetic means are illustrated in details. Forty six experts (20 in Canada and 26 in Iran) have filled the web-based questionnaires, as depicted in Fig. 2.

The experts were both male and female students and university professors with the range of age, 15 to over 45. The questionnaire begins with some inquiries including services in which the users have registered accounts as well as managing and dealing with identity management systems.

Figures 3 to 7 show the arithmetic mean of experts' opinions in which the numbers are separated by comma in Iran and in Canada within each table. Additionally, the bar charts illustrate the preference degrees of both countries.

Final weights of sub-criteria are displayed in Table III.

As mentioned before, identified criteria in level 2 as a result of interviews are categorized into four groups:

1) Criteria related to security
2) Criteria related to system capabilities
3) Criteria related to user control
4) Criteria related to cost effectiveness

As it can be seen in Fig.3, the highest rank is dedicated to criteria related to security, in both countries. Second and third criteria are different in Iran and Canada, but forth criteria in both countries are cost effectiveness.

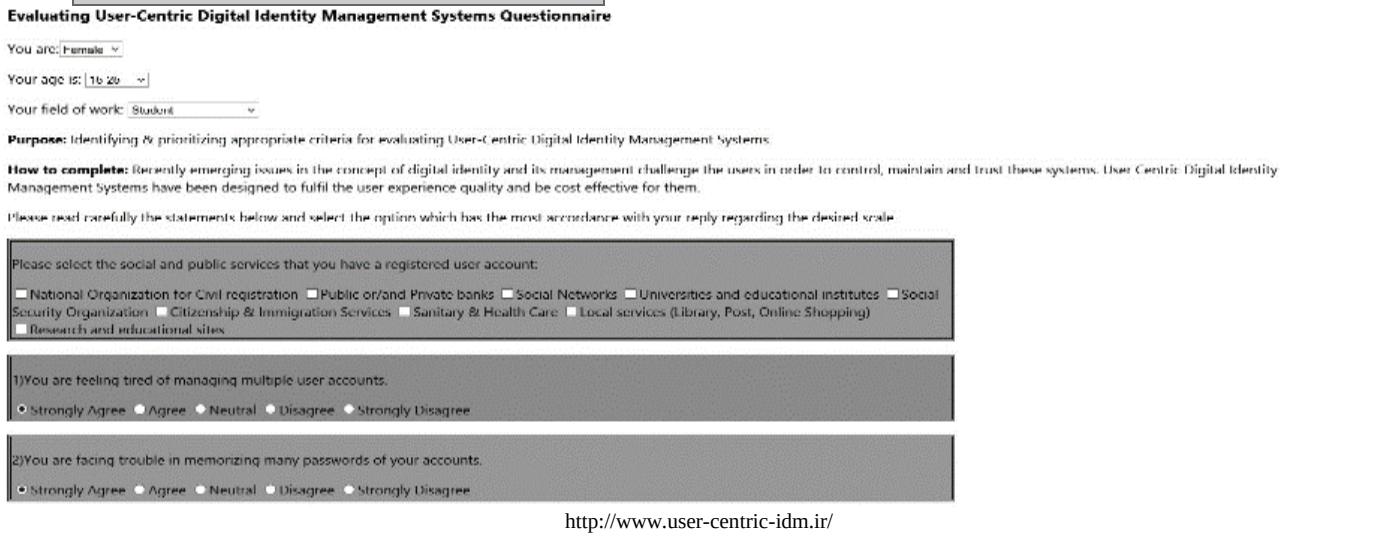
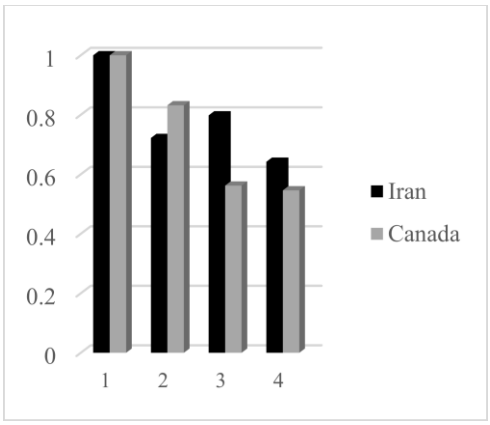


Fig. 2. Web based questionnaire

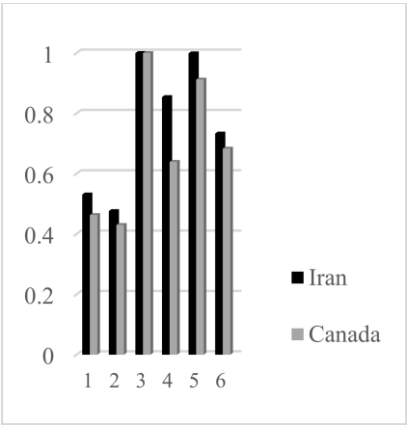
Identifying and Prioritizing The Evaluation Criteria	1	2	3	4
1	-, -	1, 1	1, 1	1, 1
2	0.722, 0.832	-, -	0.916, 1	1, 1
3	0.798, 0.562	1, 0.725	-, -	1, 1
4	0.624, 0.546	0.904, 0.706	0.818, 0.978	-, -



1- Security 2- User control 3- System capabilities 4- Cost-effectiveness

Fig. 3. Arithmetic means and preference degrees of sub-criteria level 2

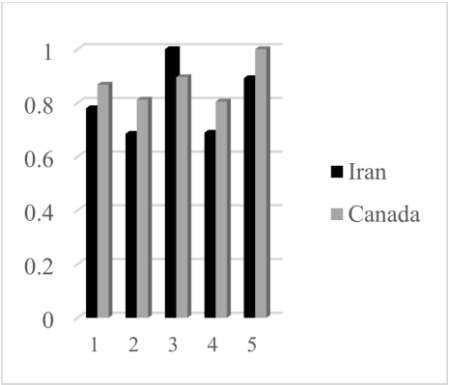
Criteria related to security	1	2	3	4	5	6
1	-, -	1, 1	0.531, 0.463	0.671,0.822	0.541, 0.558	0.803, 0.778
2	0.939, 0.971	-, -	0.476, 0.43	0.614, 0.791	0.487, 0.526	0.745, 0.747
3	1, 1	1, 1	-, -	1, 1	1, 1	1, 1
4	1, 1	1, 1	0.854, 0.639	-, -	0.858, 0.732	1, 0.957
5	1, 1	1, 1	0.999, 0.912	1, 1	-, -	1, 1
6	1, 1	1, 1	0.733, 0.683	0.873, 1	0.739, 0.775	-, -



1- Revocation 2- Conditional Release 3- Confidentiality & user's privacy 4- Sharing Prevention 5-Security & Stealing Protection 6- Integrity

Fig. 4. Arithmetic means and preference degrees of sub criteria level 3 related to *security*

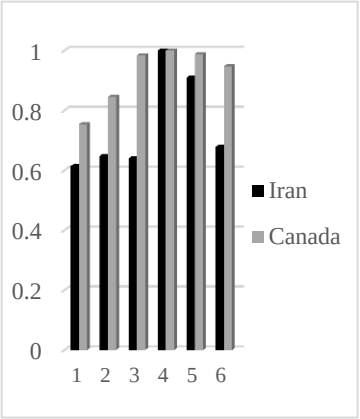
Criteria related to user control	1	2	3	4	5
1	-, -	1, 1	0.78, 0.968	1, 1	0.883, 0.868
2	0.91, 0.934	-, -	0.685, 0.901	1,0.987	0.789,0.812
3	1, 1	1, 1	-, -	1, 1	1, 0.896
4	0.898, 0.947	0.985,1	0.679,0.914	-, -	0.78, 0.805
5	1, 1	1, 1	0.892, 1	1, 1	-, -



1- User chosen identity provider 2- User in the middle 3- User consent 4- Delegation 5- Verifiability

Fig. 5. Arithmetic means and preference degrees of sub criteria level 3 related to *user control*

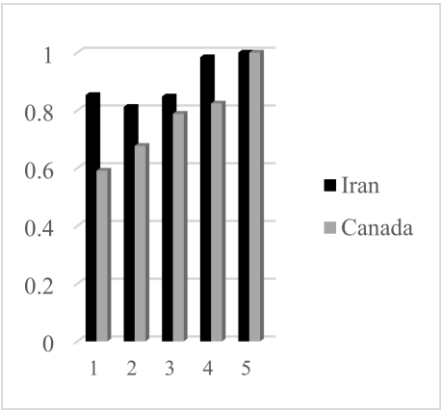
Criteria related to system capabilities	1	2	3	4	5	6
1	-, -	0.969, 0.909	0.976, 0.767	0.614, 0.754	0.701, 0.779	0.932, 0.812
2	1, 1	-, -	1, 0.859	0.648, 0.845	0.734, 0.868	0.964, 0.901
3	1, 1	0.993, 1	-, -	0.64, 0.984	0.726, 1	0.957, 1
4	1, 1	1, 1	1, 1	-, -	1, 1	1, 1
5	1, 1	1, 1	1, 0.993	0.91, 0.988	-, -	1, 1
6	1, 1	1, 1	1, 0.962	0.678, 0.947	0.765, 0.969	-, -



1- Scalability 2- Portability 3- Notification 4- Dependable 5- Accountability 6- Auditing

Fig. 6. Arithmetic means and preference degrees of sub criteria level 3 related to *system capabilities*

Criteria related to cost-effectiveness	1	2	3	4	5
1	-, -	1, 0.905	1, 0.805	0.87, 0.763	0.853, 0.591
2	0.961, 1	-, -	0.97, 0.895	0.828, 0.853	0.812, 0.677
3	0.992, 1	1, 1	-, -	0.864, 0.961	0.848, 0.788
4	1, 1	1, 1	1, 1	-, -	0.984, 0.824
5	1, 1	1, 1	1, 1	1, 1	-, -



1-Data Minimization 2- Availability 3- Fault tolerant 4- Usability 5- Service protection

Fig. 7. Arithmetic means and preference degrees of sub criteria level 3 related to *cost-effectiveness*

TABLE III. FINAL WEIGHT OF SUB CRITERIA

Iran		Canada	
Criterion	**Final absolute weight of criterion**	**Criterion**	**Final absolute weight of criterion**
Revocation	0.037	**Revocation**	0.038
Conditional Release	0.033	**Conditional Release**	0.035
Confidentiality & User's Privacy	0.069	**Confidentiality & User's Privacy**	0.082
Sharing Prevention	0.059	**Sharing Prevention**	0.056
Security & Stealing Protection	0.069	**Security & Stealing Protection**	0.075
Integrity	0.051	**Integrity**	0.058
User Chosen Identity Provider	0.043	**User Chosen Identity Provider**	0.053
User in the Middle	0.039	**User in the Middle**	0.053
User Consent	0.057	**User Consent**	0.056
Delegation	0.039	**Delegation**	0.052
Verifiability	0.051	**Verifiability**	0.065
Scalability	0.036	**Scalability**	0.026
Portability	0.037	**Portability**	0.029
Notification	0.036	**Notification**	0.034
Dependable	0.057	**Dependable**	0.038
Accountability	0.051	**Accountability**	0.034
Auditing	0.037	**Auditing**	0.033

Data Minimization	0.038	**Data Minimization**	0.028
Availability	0.035	**Availability**	0.032
Fault Tolerant	0.038	**Fault Tolerant**	0.035
Usability	0.044	**Usability**	0.039
Service Protection	0.044	**Service Protection**	0.048

V. Conclusion and Future Research

According to literature review, transformation of Identity Management Systems can be in the range of development of silo models to federated user-centric identity management models. User-Centric Identity Management Systems should consider scalability and cost-effectiveness issues from users' perspective. Scalability is important because users register with a growing number of services and deal with complexity of managing more personal credentials which has become an impediment [21]. This paper presented an approach for identifying and prioritizing appropriate criteria in order to evaluate user-centric digital identity management systems. It is believed that no single perfect set of criteria is perceived which can be implemented in all user-centric identity management systems. four categoriesare proposed to place the evaluation criteria for accomplishing the notion of user-centricity. It can be observed that based on pairwise comparison matrix and preference degrees of sub-criteria, the highest rank is dedicated to criteria related to security.This could be due to the fact that security issues enhance the trust to these systems which is very important for the user. In addition, most of the survey participants have had users' account in financial institutions and banks.

The second-best criteria in developing country (e.g. Iran), are system capabilities whereas user control in the developed countries (e.g. Canada) have had this spot as the second best criteria. Perhaps for the reason that, digital identity management systems have been more used in developed countries like Canada than developing countries is because system capabilities are more advanced in the developed countries so users are more concern with user control. Lastly, cost-effectiveness criteria have had the least priority both in developed and developing countries. Furthermore, considering sub-criteria of confidentiality and user's privacy, dependability, user consent and service protection in Iran, whereas Confidentiality and user's privacy, dependability, verifiability and service protection in Canada were the ones with highest preference degrees resulted from prioritizing criteria using fuzzy AHP. Based on literature review, it can be concluded that the future outlook of this research will be further taxonomies of appropriate criteria in which the most predominant one could be specified regarding to assessment of user-centric systems. Interoperability with traditional identity management systems would be an asset for this user-centricity concept as it should incorporate the advantages presented by the previous approaches and focus on adaptability [2]. Another important direction for future work is unifying the corresponding criteria implementable in user-centric devices, applications and solutions that facilitates user control and privacy when accessing increasing amount of online services [35]. Currently, web identity management is a technology centered concept, designed to be profitable for service providers but not for users. The browser must be a user-centered identity layer between the service provider and the user, leading to better control for user over his/her identity attributes [13]. Progress in digital identity management systems will become feasible to deploy user-centric paradigm which operate on a massive scale and control the full life cycle of digital identities from creation to termination, maintaining its major advantage that is, involvement in each transaction and improving its main drawback which is not being able to handle delegation [6] along with focusing on users, controlling what information is shared about them, the content of the information and who is allowed to access it [12].

Acknowledgment

We would like to thank the students of faculty of Management and Accounting in Shahid Beheshti University, John Molson School of Business in Concordia University, University of Montreal, University of Quebec in Montreal and University of Toronto as well as Sebastian Medina Lopez Computer Technician at University of Quebec in Montreal, Prof. Dimitrios Hatzinakos at University of Toronto and director of Identity, privacy and Security Institute at UoT, Dr. Jian Yun Nie, full professor at IT department of University of Montreal, Eve Maler, VP Innovation and emerging technology at ForgeRock, and Andre Boysen, chief identity officer at SecureKey Inc. for their support and encouragement as well as sharing their thoughts and experiences with us.

References

[1] G. J. Ahn, and M. Koo. (2007, Nov). User-centric privacy management for federated identity management. Presented at International Conference on Collaborative Computing: Networking, Applications and Work sharing.

[2] G. J. Ahn, M. Ko, and M, Shehab. (2009, June). Privacy enhanced user-centric identity management. Presented at IEEE International Conference on Communications. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Baldwin, A., Casassa Mont, M., Beres, Y., & Shiu, S. (2010). Assurance for federated identity management. Journal of Computer Security, 18(4), 541-572.

[4] G. Ben Ayed (2014), Architecting user-centric privacy as-a-set-of services, Digital identity related privacy framework, Ph.D. dissertation, Dept. IS, Lausanne. Univ., Lausanne, Switzerland, 2014.

[5] G. Ben Ayed, and S. Ghernaouti-Hélie. (2011, Sept). Digital identity management within network information systems, From Vertical Silos View into Horizontal User-Supremacy Processes Management. Presented at 14th International Conference on Network-Based Information Systems. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] A. BhargavSpantzel, J. Camenisch, T. Gross, and D. Sommer. (2007, Oct). User Centricity: A Taxonomy and Open Issues. Journal of Computer Security. 15 (5), pp.493-527.

[7] P. Bramhall, M. Hansen, k. Rannenbeg, and T. Roessler. (2007, July). User-Centric identity management, New trends in standardization and regulation. IEEE Security & Privacy. [Online]. 5 (4), 84-87.

[8] J. Camenisch, A. Shelat, D. Sommer, S. Fischer-Hubner, M. Hansen, H. Krasemann, G. Lacoste, R. Leenes, and J. Tseng, (November 2005), "Privacy and identity management for everyone," DIM '05., New York., NY, pp. 20-25.

[9] D. Y. Chang. (1996, Dec). Applications on the extent analysis method on fuzzy AHP. European Journal of Operation Research. [Online]. 95(3), 649-655.

[10] D. Choi, S. H. Jin, and H. Yoon. (2007, June). Trust management for user centric identity management on the internet. Presented at IEEE International Symposium on Consumer Electronics.

[11] M. T. Chu, and R. Khosia, (2010), Benchmarking of communities of practice model for R&D organizations, Presented at 6th European Conference on Management Leadership and Governance., London, pp. 76-77.

[12] W. Claycomb, D. Shin, and D. Harelland. (2007, Oct). Towards Privacy in Enterprise Directory Services: A User-Centric Approach to Attribute Management. Presented at 41st Annual IEEE International Carnahan Conference on Security Technology.

[13] R. Cottrell, (2011), User-centered identity management, evaluating the role of the browser, Project Report, Faculty of Life Sciences., University College London, London, 2011.

[14] Digital identity management- Enabling innovation and trust in the internet economy, (2011), OECD, Organization for Economic CO-Operation and Development, Paris, 2011.

[15] O. Duran, and J. Aguilo. (2008, Apr). Computer-aided machine-tool selection based on a Fuzzy-AHP approach. Expert Systems with Applications, 34(3),17871794.

[16] T. M. Eap, M. Hatala, and D. Gašević. (2007, July). Enabling user control with personal identity management. Presented at IEEE International Conference on Services Computing.

[17] T. El Maliki, and J. Seigneur. (2007, Oct). A survey of user-centric identity management technologies. Presented at International Conference on Emerging Security Information, Systems and Technologies.

[18] T. El Maliki, and J. Seigneur, (2013), "Online identity and user management services," Computer and Information Security Handbook, 2nd ed. MA, Morgan Kaufmann Publishers, ch. 25, pp. 459–484.

[19] M. Hansen, A. Pfitzmann, and S. Steinbrecher. (2008, May). Identity management throughout one's whole life. Information Security Technical Report. 13(2), pp. 83-94.

[20] M. Hoffmann, (2005), User-centric identity management in open mobile environment, Privacy, Security and Trust within the Context of Pervasive Computing, 1st ed. City of Publisher, Springer US, pp. 99–104.

[21] A. Jøsang, and S. Pope. (2005, May). User centric identity management. Presented at AusCERT.

[22] C. Kahraman, U. Cebeci, and Z. Ulukan. (2003). Multi criteria supplier selection using fuzzy AHP. Journal of Enterprise Information Management. 16(6), 382-394.

[23] M. Kwiesielewicz, E. V. Uden, (2004, Apr). Inconsistent and contradictory judgment in pairwise comparison method in the AHP. Computers and Operations Research. 31 (5), 713-719.

[24] A. H. I. Lee. (2008, Feb). A fuzzy supplier selection model with the consideration of benefits, opportunities, costs and risks. Expert Systems and Applications. 36(2), 2879-2893.

[25] A. H. I. Lee, H. Y. Kang, C. F. Hsu, and H. C. Hung. (2009, May). A green supplier selection model for high-tech industry. Expert Systems with Applications. [Online]. 36(4), 7917-7927.

[26] D. Mashima, D. Bauer, M. Ahamad, and D. M. Blough, (2011), User-centric management architecture using credential-holding identity agents, Digital Identity and Access Management, PA, IGI Global, ch. 5, pp. 78-96.

[27] R. Marx, H. S. Fhom, D. Scheuermann, K. M. Bayarou, and A. Perez, (2010) Increasing security and privacy in user-centric identity management: the IdM card approach," International Conference on P2P., Washington., DC, 2010, pp. 459-464.

[28] A. Noorul Haq, and G. Kannan. (2006, July). Fuzzy analytical hierarchy process for evaluating and selecting a vendor in a supply chain model. The International Journal of Advanced Manufacturing Technology. [Online].29(7-8),826-835.

[29] A. Pashalidis, and C. J. Mitchell, (2003), A Taxonomy of Single Sign-On Systems, Information Security and Privacy, Springer Berlin Heidelberg, pp. 249-264.

[30] W. Pedrycz, P. Ekel, and R. Parreiras, (2011), Decision-Making in System Project, Planning, Operation, and Control: Motivation, Objectives, and Basic Concepts, Fuzzy Multicriteria Decision-Making: Models, Methods and Applications, 1st ed. New Delhi, India: John Wiley & Sons, ch. 1, sec. 1.3, pp. 9-10.

[31] V. Poursalidis, and C. Nikolaou. (2006, September 4-8) A new user-centric identity management infrastructure for federated systems. Trust, Privacy, and Security in Digital Business. 4083.

[32] M. Quasthoff, and C. Mienel, (2007). User-centricity in healthcare infrastructure. LNI. [Online]. 108, pp. 141-152.

[33] S. Rieger. (2009, May). User-centric identity management in heterogeneous federations. Presented at Fourth International Conference on Internet and Web Applications Services.

[34] T. L. Saaty, and L. G. Vargas, (2012), Why is the principal eigenvector necessary?, Models, Methods, Concepts, and Applications of the Analytical Hierarchy Process, 2nd ed. New York, Springer Science & Business Media, ch. 4, sec. 4.1, pp. 63–64.

[35] S. Suriadi, E. Foo, and A. Jøsang. (2009, March). A user-centric federated single sign-on system. Journal of Network and Computer Applications. 32(2), pp. 388-401.

[36] E. Triantaphyllou, and S. H. Mann. (1995, Jan). Using the analytical hierarchy process in decision making in engineering applications: some challenges. Inter'I Journal of Industrial Engineering: Applications and Practice. 2 (1), pages.

[37] G. H. Tzeng, and J. J. Huang, (2012), Analytical Hierarchy Process?, Multiple Attribute Decision Making: Methods and Applications, FL, CRC Press, ch.2, pp. 15-16.

[38] P. J. M. Van Laarhoven, and W. Pedrycz. (1983). A Fuzzy extension of Saaty's priority theory. Fuzzy Sets and Systems. 11(1-3), 199-227.

[39] D. D. Vecchio, J. Basney, and N. Nagaratman. (2005, July). CredEx: user-centric credential management for grid and web services. IEEE International Conference on Web Services.

[40] J. Vossaert, J. Lapon, B. De Decker, and V. Naessens. (2013, April). User-Centric identity management using trusted modules. Mathemathcal and Computer Modeling. 58(7-8), pp. 15921605.

Direct Torque Control of Saturated Doubly-Fed Induction Generator using High Order Sliding Mode Controllers

Elhadj BOUNADJA
Department of Automatic
Ecole Nationale Polytechnique
Elharrach, Algiers, Algeria

Abdelkader DJAHBAR
Department of Electrical Engineering
University HASSIBABENBOUALI
Chlef, Algeria

Mohand Oulhadj MAHMOUDI
Department of Automatic
Ecole Nationale Polytechnique
Elharrach, Algiers, Algeria

Mohamed MATALLAH
Department of Technology
University DJILALI BOUNAAMA
Khemis Meliana, Ain defla, Algeria

***Abstract*—The present work examines a direct torque control strategy using a high order sliding mode controllers of a doubly-fed induction generator (DFIG) incorporated in a wind energy conversion system and working in saturated state. This research is carried out to reach two main objectives. Firstly, in order to introduce some accuracy for the calculation of DFIG performances, an accurate model considering magnetic saturation effect is developed. The second objective is to achieve a robust control of DFIG based wind turbine. For this purpose, a Direct Torque Control (DTC) combined with a High Order Sliding Mode Control (HOSMC) is applied to the DFIG rotor side converter. Conventionally, the direct torque control having hysteresis comparators possesses major flux and torque ripples at steady-state and moreover the switching frequency varies on a large range. The new DTC method gives a perfect decoupling between the flux and the torque. It also reduces ripples in these grandeurs. Finally, simulated results show, accurate dynamic performances, faster transient responses and more robust control are achieved.**

Keywords—Doubly Fed Induction Generator (DFIG); Magnetic saturation; Direct Torque Control (DTC); High Order Sliding Mode Controller (HOSMC)

I. INTRODUCTION

Recently, worldwide awareness for renewable energy resources has been increasing. In particular, wind energy has been largely considered because of its economy and reliability. Wind turbines contribute a certain amount of the word electricity consumption [1]. They usually use a Doubly-Fed Induction Generator (DFIG) for the electrical energy conversion process. As deduced from literature, many workers investigate the DFIG from diverse aspects. However, in these works, many simplifying hypotheses are considered in the modelling of the DFIG, with the neglect of magnetic saturation being the most important as in [1-10]. However, the phenomenon of saturation is present in all electrical machines. In addition, the exact calculation of the machine dynamic performances depends considerably on the saturation of the mutual and leakage fluxes [11-17]. Because of this reason and in order to realize an accurate representation of the DFIG, saturation must be taken into account in their mathematic modelling. Consequently, an accurate DFIG model taking into account the saturation effect both in mutual flux and in leakage fluxes is used in this paper.

After major advances in power electronics and material technologies, many works have presented the DFIG with different control algorithms. One of the conventional control schemes used actually for the DFIG-based wind turbine is the Direct Torque Control based on switching table and hysteresis comparators [18]. This strategy, however, has a few disadvantages which limit its use, such as variable switching frequency and torque ripple [19, 20]. In many research works on DTC, these disadvantages are reduced by using SVM scheme, but with the price of scarifying the robustness of the control [21].

To incorporate a robust DTC without torque and flux ripples, we propose, in the present work, a direct torque control based on high order sliding mode controllers (DTC-HOSMC) for a DFIG in saturated state. Proposed by Levant in [22] the HOSMC strategy has many attractive features such as its robustness towards parametric uncertainties of the DFIG, and moreover, it reduces the chattering effect.

The present work provides the important features of the DTC-HOSMC and presents simulation results for a DFIG system. We compare the proposed strategy with a conventional DTC. The present paper is organized as follows: we present the modelling of the wind turbine and the DFIG using the saturated model in section II. In section III, the proposed DTC-HOSMC, is applied to control the saturated DFIG. The implementation and the results obtained from the proposed controller are shown in section IV. Finlay, it will be shown that using the developed DFIG model and the proposed controller, the dynamic responses of the system can be determined accurately and more precise robust control is achieved.

II. Modeling of Wind Energy Conversion System

The wind energy conversion system adopted in this work is based on wind turbine driven a DFIG. In such configuration, the stator is directly connected to the network, whereas, the rotor is fed by the grid via two converters (AC/DC) and (DC/AC). In addition, the rotor side converter (DC/AC) is used to control independently the DFIG active and reactive powers.

A. Modeling of the wind turbine

The mechanical power captured from the wind turbine, used in this investigation, is expressed as below:

$$P_t = 0.5 C_P(\lambda,\beta) R^2 \rho\, v^3 \tag{1}$$

With:

R: radius of turbine (m), ρ: density of air (kg/m3), v: speed of wind (m/s) and C_P: the power coefficient.

According to [3], the power coefficient C_p is function of the tip speed ratio λ and the blade pitch angle β (deg), as follows:

$$C_p = 0.5109\left(\frac{116}{\lambda_i} - 0.4\beta - 5\right)\exp\left(-\frac{21}{\lambda_i}\right) + 0.0068\lambda \tag{2}$$

With:

$$\frac{1}{\lambda_i} = \frac{1}{\lambda + 0.08\beta} - \frac{0.035}{\beta^3 + 1} \tag{3}$$

The tip speed ratio λ is given by:

$$\lambda = \frac{\Omega_t R}{v} \tag{4}$$

In (4) Ω_t represent the rotational speed of the wind turbine.

B. Modeling of the DFIG

Below, we develop the conventional model of the DFIG without saturation. According to this model, both mutual flux and leakage fluxes saturation are considered. In these models, the DFIG is considered as a generalized wound rotor induction machine taking the stator resistance into consideration. The latter is neglected in many works e.g. in [1-19].

1) Linear DFIG model

The *d* and *q* equivalents circuits for the DFIG are shown in Fig. 1 [15]. Based in these schemas, the voltages equations of the DFIG in the *d-q* synchronous referential are given by:

$$\begin{cases} v_{sd} = R_s i_{sd} + \dfrac{d}{dt}\psi_{sd} - \omega_s \psi_{sq} \\ v_{sq} = R_s i_{sq} + \dfrac{d}{dt}\psi_{sq} + \omega_s \psi_{sd} \\ v_{rd} = R_r i_{rd} + \dfrac{d}{dt}\psi_{rq} - \omega_r \psi_{rq} \\ v_{rq} = R_r i_{rq} + \dfrac{d}{dt}\psi_{rq} + \omega_r \psi_{rd} \end{cases} \tag{5}$$

Where the rotor frequency ω_r is given by:

$$\omega_r = \omega_s - \omega_m \tag{6}$$

The flux linkages in (5) are obtained from the following equation system:

$$\begin{cases} \psi_{sd} = L_s i_{sd} + L_m i_{rd} \\ \psi_{sq} = L_s i_{sq} + L_m i_{rq} \\ \psi_{rd} = L_r i_{rd} + L_m i_{sd} \\ \psi_{qr} = L_r i_{qr} + L_m i_{qs} \end{cases} \tag{7}$$

The equation system in (7) is used to calculate the *d* and *q* components of stator and rotor currents:

$$\begin{cases} i_{sd} = \dfrac{1}{\sigma L_s L_r}(L_r \psi_{sd} - L_m \psi_{rd}) \\ i_{sq} = \dfrac{1}{\sigma L_s L_r}(L_r \psi_{sq} - L_m \psi_{rq}) \\ i_{rd} = \dfrac{1}{\sigma L_s L_r}(L_s \psi_{rd} - L_m \psi_{sd}) \\ i_{rq} = \dfrac{1}{\sigma L_s L_r}(L_s \psi_{rq} - L_m \psi_{sq}) \end{cases} \tag{8}$$

With:

$$\sigma = 1 - \frac{L_m^2}{L_s L_r},\ L_s = L_{s\sigma} + L_m\ ,\ \ L_r = L_{r\sigma} + L_m \tag{9}$$

The magnetizing current i_m is given as follows [14, 15]:

$$i_m = \sqrt{i_{md}^2 + i_{mq}^2} \tag{10}$$

Where:

$$i_{md} = i_{sd} + i_{rd}\ ,\ i_{mq} = i_{sq} + i_{rq} \tag{11}$$

In steady state and by aligning the q-axis of synchronous rotating reference frame on stator flux vector, the following equations can be written [2, 8]:

$$\psi_{ds} = \psi_s = \frac{V_s}{\omega_s},\ \ \psi_{qs} = 0 \tag{12}$$

$$\psi_r = \sigma L_r i_{rd} + \frac{L_m V_s}{L_s \omega_s} \tag{13}$$

$$\begin{cases} v_{rd} = R_r i_{rd} + L_r \sigma \dfrac{di_{rd}}{dt} - g\omega_s L_r \sigma i_{rq} \\ v_{rq} = R_r i_{rq} + L_r \sigma \dfrac{di_{rq}}{dt} + g\omega_s L_r \sigma i_{rd} + g\omega_s \dfrac{L_m V_s}{L_s \omega_s} \end{cases} \tag{14}$$

$$T_{em} = -p\frac{L_m V_s}{L_s \omega_s} i_{qr} \tag{15}$$

2) DFIG model with saturation

Using the linear model explained in the previous section, we develop a DFIG model taking the mutual flux saturation into consideration. The corresponding saturated value L_{ms} replaces the unsaturated mutual inductance L_m in this approach, (7)-(9). This saturated mutual inductance is calculated by multiplying the corresponding unsaturated value, L_m, with a saturation coefficient K_{sm}, corresponding to the saturation state.

The saturated mutual inductance L_{ms}, which is a function of the magnetizing current i_m, can be written as follows:

$$L_{ms} = \begin{cases} L_m & i_m < I_{msat} \\ K_{sm}(i_m).L_m & i_m \geq I_{msat} \end{cases} \quad (16)$$

The saturation coefficient K_{sm} can be represented by the function [11]:

$$K_{sm}(i_m) = \begin{cases} 1 & i_m < I_{msat} \\ \frac{2}{\pi}\left[\arcsin\left(\frac{I_{msat}}{i_m}\right) + 0{,}5\sin\left(2\arcsin\left(\frac{I_{msat}}{i_m}\right)\right)\right] & i_m \geq I_{msat} \end{cases} \quad (17)$$

Here I_{msat} represents the magnetizing current at which the saturation starts. It is around *0.5 pu*, i.e $0.7 \times I_n$ [14,25].

In addition, in the modelling of DFIG, the representation of saturation includes the variation in the stator and rotor leakage inductances caused by the saturation in the leakage flux paths. The saturation in the leakage flux paths is taken into account in the model developed in previous section by replacing the unsaturated stator and rotor leakage inductances ($L_{s\sigma}$, $L_{r\sigma}$) in (9) by their corresponding saturated values ($L_{s\sigma s}$, $L_{r\sigma s}$). The latter inductances are obtained by multiplying their respective unsaturated values by a saturation coefficient $K_{s\sigma}$. This coefficient depends to the stator current or the rotor current.

The stator and rotor leakage inductances are given as function of their corresponding currents as follows:

$$L_{s\sigma s}(i_s) = \begin{cases} L_{s\sigma} & i_s < I_{sat} \\ K_{s\sigma}(i_s).L_{s\sigma} & i_s \geq I_{sat} \end{cases} \quad (18)$$

$$L_{r\sigma s}(i_r) = \begin{cases} L_{r\sigma} & i_r < I_{sat} \\ K_{s\sigma}(i_r).L_{r\sigma} & i_r \geq I_{sat} \end{cases} \quad (19)$$

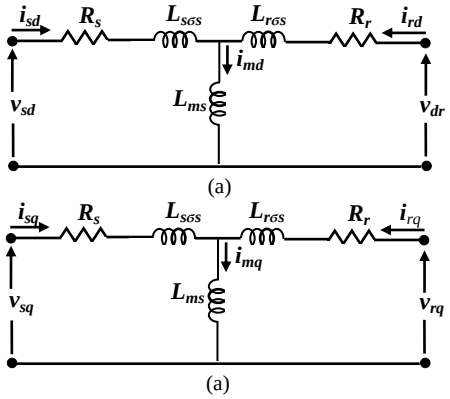


Fig. 1. Equivalent circuits of a DFIG: (a) on d-axis, (b) on q-axis

The saturation coefficient $K_{s\sigma}$ can be represented by the function [11]:

$$K_{s\sigma}(i) = \begin{cases} 1 & i < I_{sat} \\ \frac{2}{\pi}\left[\arcsin\left(\frac{I_{sat}}{i}\right) + 0{,}5\sin\left(2\arcsin\left(\frac{I_{sat}}{i}\right)\right)\right] & i \geq I_{sat} \end{cases} \quad (20)$$

Saturation is taken into account at values of the current I_{sat} in the range *1,3-3 pu*, that is $1{,}8 \times I_n$–$4{,}2 \times I_n$ [14,26,27].

III. Direct Torque Control using High Order Sliding Mode Controllers of DFIG

The goal of DTC-HOSMC is to regulate both the torque and the rotor flux magnitude of the DFIG. The flux is regulated using the direct axis voltage V_{dr}, while the torque is controlled using the quadrature axis voltage V_{qr}. The phenomenon of chattering that represents the important problem of the conventional sliding mode control can be very harmful for the DFIG due the fact that the discontinuous control can cause overheating of the coils and the excitation of unmodelled high frequency dynamics. In [28] some solutions were proposed in order to avoid this disadvantage. The main idea was to adjust the dynamics in a small region of the discontinuity surface so to escape the real discontinuity meanwhile conserving the major characteristics of the entire system. The lately proposed HOSMC generalizes the main sliding mode idea which acts on the high order time derivatives of the system deviation from the constraint in place of influencing the first deviation derivative as in standard sliding modes [6]. In addition to keeping the major advantages of the original technique, they discuss the chattering effect and even represent higher accuracy in a real implementation. In HOSMC algorithm implementations, the main difficulty consists of the increase in the needed information. In fact, the knowledge of $\dot{S}, \ddot{S}, \ldots, S^{(n-1)}$ is required to achieve an n^{th} order controller. As an exception to all the algorithms proposed for the HOSMC, the super-twisting algorithm needs only the information on the sliding surface [6]. As a consequence, this algorithm has been utilized for the proposed control method. As mentioned in [28], the stability can be obtained for all high order sliding mode controllers with this algorithm. Figure 2 shows the proposed DTC-HOSMC, which is used to control both the rotor flux and the electromagnetic torque of the DFIG.

In this study, the errors between reference and measured of the electromagnetic torque and the rotor flux have been chosen as sliding mode surfaces, so the following expression can be written:

$$\begin{cases} S_{\psi_r} = \psi_{r\_ref} - \psi_r \\ S_{T_{em}} = T_{em\_ref} - T_{em} \end{cases} \quad (21)$$

By substituting the rotor flux and the electromagnetic torque in (21) by their expressions given, respectively, by (13) and (15), one obtains:

$$\begin{cases} S_{\psi_r} = \psi_{r-réf} - \sigma L_r i_{rd} - \frac{L_m}{L_s}\Psi_s \\ S_{T_{em}} = T_{em-réf} + p\frac{L_m V_s}{L_s \omega_s}\, i_{qr} \end{cases} \quad (22)$$

The first derivative of (22), gives:

$$\begin{cases} \dot{S}_{\psi_r} = \dot{\psi}_{r-réf} - \sigma L_r \dot{i}_{rd} \\ \dot{S}_{T_{em}} = \dot{T}_{em-réf} + p\dfrac{L_m V_s}{L_s \omega_s} \dot{i}_{rq} \end{cases} \quad (23)$$

If we replace the *d* and *q* rotor currents derivatives in (23) by their expressions given from (14), one obtains:

$$\begin{cases} \dot{S}_{\psi_r} = \dot{\psi}_{r-réf} - \left[v_{rd} - R_r i_{rd} + g\omega_s \sigma L_r i_{rq} \right] \\ \dot{S}_{T_{em}} = \dot{T}_{em-réf} + p\dfrac{V_s L_m}{\sigma L_s L_r} \left[v_{rq} - R_r i_{rq} - g\omega_s \sigma L_r i_{rd} - gV_s \dfrac{L_m}{L_s} \right] \end{cases} \quad (24)$$

We define the functions G_1 and G_2 as follows:

$$\begin{cases} G_1 = \left[R_r i_{rq} - g\omega_s \sigma L_r i_{rq} \right] + \dot{\psi}_{r-réf} \\ G_2 = p\dfrac{V_s L_m}{\sigma L_s L_r} \left[-R_r i_{rq} - g\omega_s \sigma L_r i_{rd} - g\omega_s \dfrac{L_m}{L_s} \psi_s \right] + \dot{T}_{em-réf} \end{cases} \quad (25)$$

After substituting (25) in (24), the derivative of (24) gives:

$$\begin{cases} \ddot{S}_{\psi_r} = \dot{v}_{rd} + \dot{G}_1 \\ \ddot{S}_{T_{em}} = p\dfrac{V_s L_m}{\sigma L_s L_r} \dot{v}_{rq} + \dot{G}_2 \end{cases} \quad (26)$$

Basing on the super-twisting algorithm established by Levant in [22,23], the high order sliding mode controller contains two parts:

$$\begin{cases} v_{rd} = -\alpha_1 \int sign(S_{\psi_r}) dt - \beta_1 \left| S_{\psi_r} \right|^{0.5} sign(S_{\psi_r}) \\ V_{rq} = -\alpha_2 \int sign(S_{T_{em}}) dt - \beta_2 \left| S_{T_{em}} \right|^{0.5} sign(S_{T_{em}}) \end{cases} \quad (27)$$

In order to guarantee the convergence of the sliding manifolds to zero in set time, the constants α_1, β_1, α_2 and β_2 can be chosen as follows [2,6,28]:

$$\begin{cases} \alpha_1 > \mu_1 \\ \beta_1^2 \geq 4\mu_i \dfrac{(\alpha_1 + \mu_1)}{(\alpha_1 - \mu_1)} \\ \mu_1 > |G_2| \end{cases} \quad (28)$$

$$\begin{cases} \alpha_2 > \mu_2 \dfrac{pV_s L_m}{\sigma L_s L_r} \\ \beta_2^2 \geq 4\mu_2 \left(\dfrac{pV_s L_m}{\sigma L_s L_r} \right)^2 \dfrac{(\alpha_2 + \mu_2)}{(\alpha_2 - \mu_2)} \\ \mu_2 > |G_2| \end{cases} \quad (29)$$

IV. Simulation Results

In this section, simulations are realized with a 7.5 KW DFIG coupled to a (311V, 50 Hz) network, utilizing the Matlab/Simulink environment. The parameters of the machine are shown in Table 2.

The consideration of saturation into account for the mutual flux of the investigated DFIG is realized by taking I_{msat} in (16)-(17) to be equal to $0.7 \times I_n = 6\ A$. In this equality, I_n is the rated current given in Table 2. The mutual flux saturation coefficient K_{sm} used in the determination of the saturated value of the mutual inductance L_{ms} is sketched in figure 3. Likewise, the consideration of the leakage flux saturation, I_{sat} in (18)-(20) was assumed to be equal to $1.8 \times I_n = 15.8\ A$, where the leakage flux saturation coefficient $K_{s\sigma}$ is shown in figure 4.

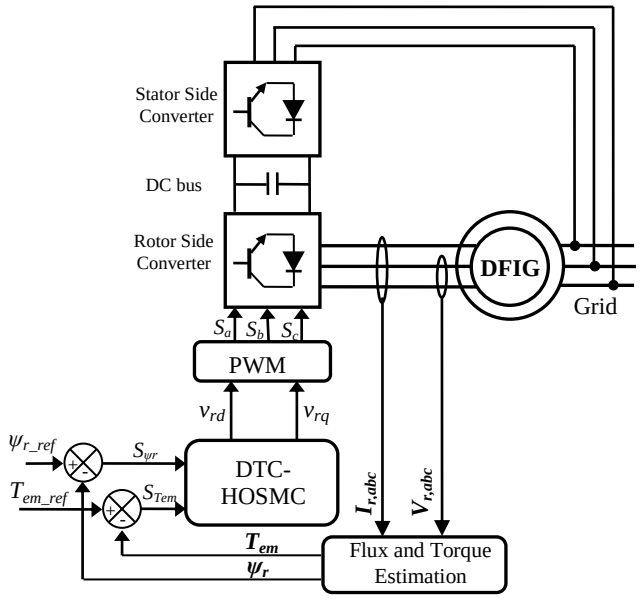


Fig. 2. Bloc diagram of HOSMC-DTC applied to the DFIG

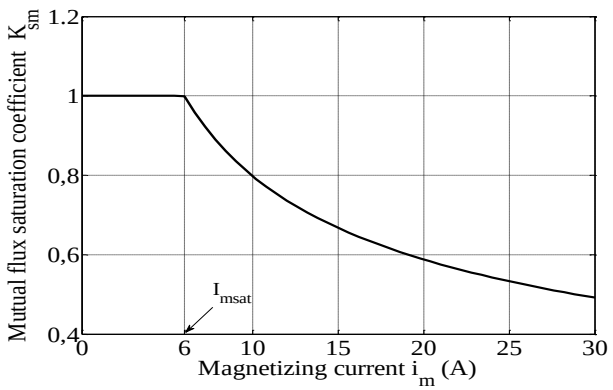


Fig. 3. Mutual flux saturation coefficient K_{sm}

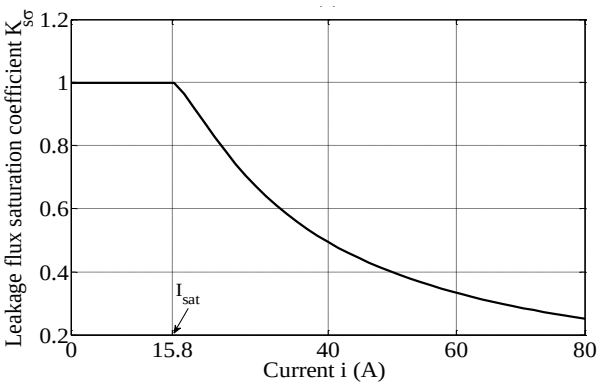


Fig. 4. Leakage flux saturation Coefficient $K_{s\sigma}$

The two DTC control strategies; classical DTC and DTC-HOSMC are simulated and compared in terms of reference tracking and robustness against machine parameter variations.

A. *Test of reference tracking*

The goal of this test is to explore the behaviour of the two DTC control strategies while maintaining the DFIG's speed at its nominal value. The simulation results are shown in figures 5 and 6. From these figures we can see that for the DTC-HOSMC control method, the torque and rotor flux track almost perfectly their references values. In addition, and contrary to the classical DTC strategy in which the coupling effect between the two axes is quite apparent, we remark that, in the present DTC-HOSMC strategy, the decoupling between the axes is guaranteed.

B. *Test of robustness*

In order to check the robustness of the used DTC control strategies, the machine parameters namely the stator and the rotor resistances R_s and R_r have been intentionally doubled. The DFIG is working at its nominal speed and is in state of saturation. Figures 6 and 7 show the simulation results. From these Figures, we see that the parameters variation of the machine increase the time-response of the classical DTC strategy slightly. However the results show that these variations cause a marked effect on the torque and flux variations and this effect is more marked for the classical DTC strategy than that with DTC-HOSMC.

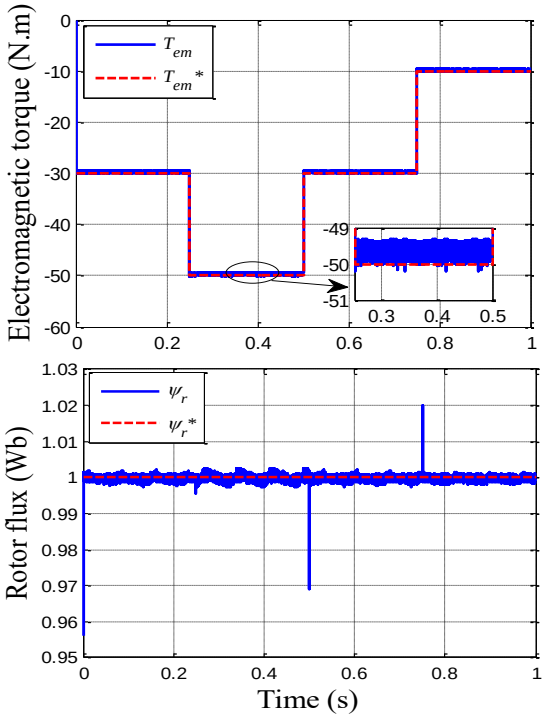


Fig. 5. Classical DTC strategy responses (reference tracking test)

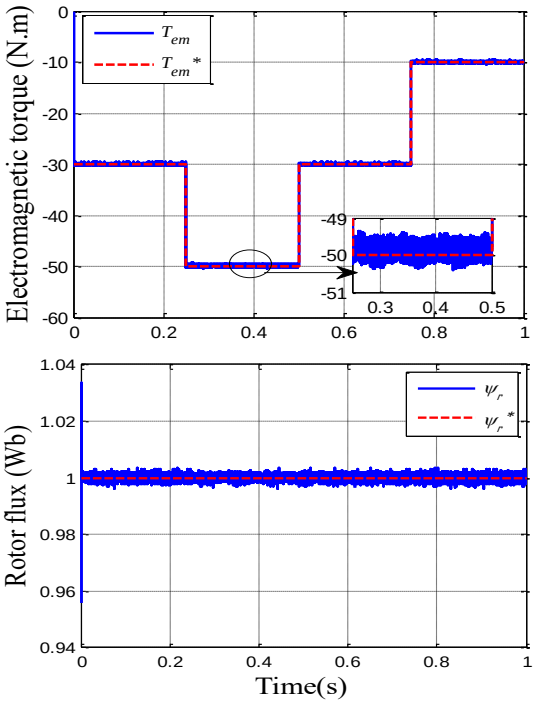


Fig. 6. DTC-HOSMC strategy responses (test of reference tracking)

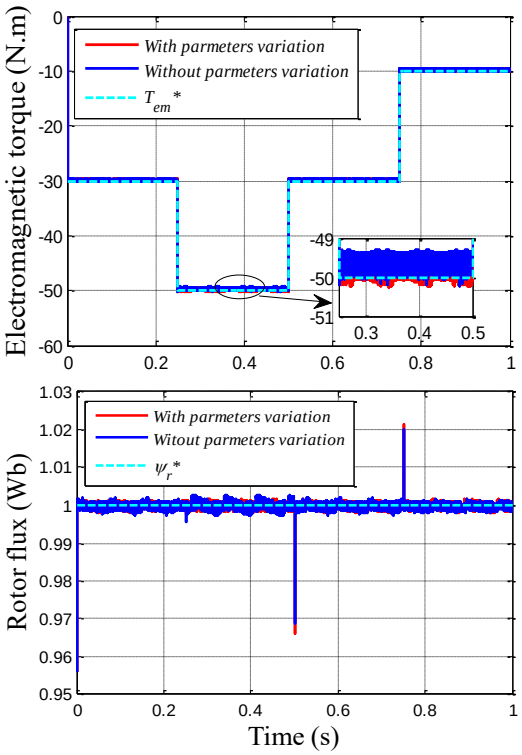


Fig. 7. Classical DTC strategy responses (test of robustness)

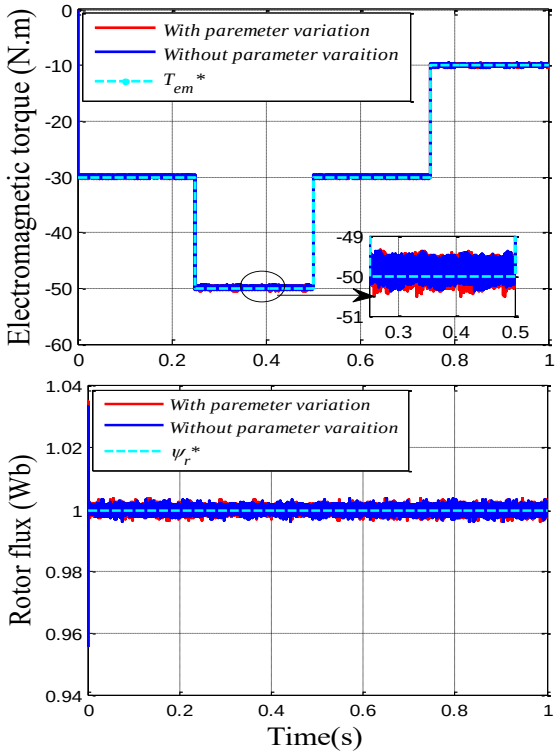


Fig. 8. DTC-SOSMC strategy responses (robustness test)

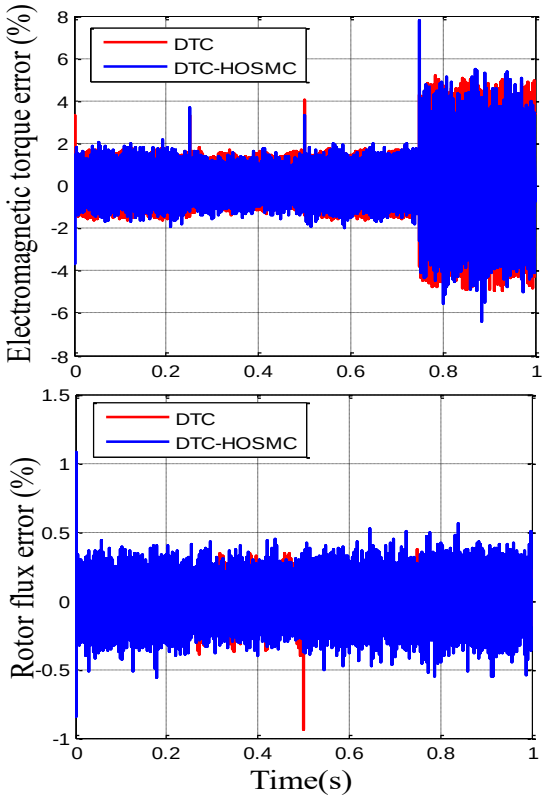


Fig. 9. Error curves (robustness test)

V. CONCLUSION

In this paper, we have investigated a Direct Torque Control using a high order sliding mode controllers. The goal of this control strategy has been to improve the calculation of the dynamic performances of saturated DFIG driven by a wind turbine driven. In the first place, the modelling of the saturated DFIG-based wind turbine has been carried out. The saturation of both the magnetizing flux and of the leakage fluxes have been taken into account in the proposed DFIG model. Then, the synthesis of a new DTC combined with HOSMC has been performed and this DTC-HOSMC has been compared with the conventional DTC in term of reference tracking. The tacking of their references was achieved almost perfectly by the two DTC strategies, however there appeared a coupling effect in the conventional DTC responses, whereas this coupling was eliminated in the DTC-HOSMC. An investigation of robustness test has also been realized in which the parameters of the DFIG have been intentionally modified. Some disturbances on the torque and flux responses have been induced by these changes but with a major effect with the conventional DTC strategy than with the proposed DTC-HOSMC. On the light of these results, we conclude that the robust DTC-HOSMC control method is a very attractive solution for those devices that use the saturated DFIG as happens in wind energy conversion systems.

APPENDIX

TABLE I. LIST OF SYMBOLS

Symbol	Significance
DFIG	Doubly-fed induction generator
DTC	Direct Torque Control
HOSMC	High Order Sliding Mode Control
v_{ds}, v_{qs}, v_{dr}, v_{qr}	*d* and *q* axis stator and rotor voltages,
ψ_{ds}, ψ_{qs}, ψ_{dr}, ψ_{qr}	*d* and *q* axis stator and rotor fluxes,
ψ_{r_ref}	Reference rotor flux
i_{ds}, i_{qs}, i_{dr}, i_{qr}	*d* and *q* axis stator and rotor currents,
R_s, R_r	Stator and rotor resistances,
L_s, L_r	Stator and rotor inductances,
$L_{s\sigma}$, $L_{r\sigma}$	Stator and rotor leakage inductances,
L_{sas}, L_{ras}	Stator and rotor saturated leakage inductances,
σ	Leakage coefficient
I_n	Rated current,
L_m	Mutual inductance,
L_{ms}	Saturated mutual inductance,
p	Number of pole pairs,
s	Generator slip,
ω_s, ω_r	Stator and rotor current frequencies (rd/s),
ω_m	Mechanical rotor frequency (rd/s),
T_{em}, T_{em_ref}	Electromagnetic, Reference electromagnetic torque.

TABLE II. MACHINE PARAMETERS

Parameters	Rated Value	Unit
Nominal power P_n	7.5	KW
Stator voltage V_n	220	V
Stator voltage amplitude V_s	311	V
Stator current I_n	8,6	A
Stator frequency f	50	Hz
Number of pairs poles p	2	
Nominal speed $\boldsymbol{\omega_m}$	144	rad/s
Stator resistance R_s	1.2	Ω
Rotor resistance R_r	1.8	Ω
Mutual inductance L_m	0.15	H
Leakage stator inductance $L_{\sigma s}$	0.0054	H
Leakage rotor inductance $L_{\sigma r}$	0.0068	H

REFERENCES

[1] D. Kairous and B. Belmadani, "Robust Fuzzy-Second Order Sliding Mode based Direct Power Control for Voltage Source Converter", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 8, 2015, pp. 167-175.

[2] S. Abddaim, A. Betka, "Optimal tracking and robust power control of the DFIG wind turbine", Electrical Power and Energy Systems, vol. 49, 2013, pp. 234–242.

[3] A. Davigny, " Participation in the system services of wind farms variable speed integrated storage inertial energy"; PhD thesis, University of Lille, France, 2007.

[4] M. Edrah, K. L. Lo, O. Anaya-Lara, "Impacts of high penetration of DFIG wind turbines on rotor angle stability of power systems", IEEE Trans. Sustain Energ, vol. 6, 2015, pp. 759-766.

[5] N. Bounar, A. Boulkroune, F. Boudjema, M. M'Saad, M. Farza, "Adaptive fuzzy vector control for a doubly-fed induction motor", In: Neurocomputing, vol. 151, no. 2, 2015, pp. 756-769.

[6] B. Beltran, M.E.H. Benbouzid, T. Ahmed-Ali, "Second order sliding mode control of a doubly fed induction generator driven wind turbine, IEEE Trans., Energy Convers., Vol. 27, No 2, 2012, pp. 261–269.

[7] M. Benkahla, R. Taleb and Z. Boudjema, "Comparative Study of Robust Control Strategies for a DFIG-Based Wind Turbine", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 2, 2016, pp. 455-462.

[8] F. Poitiers, T. Bouaouiche, M. Machmoum, "Advanced Control of a Doubly-Fed Induction Generator for Wind Energy Conversion", Electric Power Systems Research, vol. 79, no 7, 2009, pp.1085–1096.

[9] X. Zhu, S. Liu, Y. Wang, "Second-order sliding-mode control of DFIG-based wind turbines", In: IEEE 3rd Renewable Power Generation Conference, 24-25 September 2014; Naples, Italy, pp. 1-6.

[10] S. Z. Chen, N. C. Cheung, K. C. Wong, J. Wu, "Integral sliding-mode direct torque control of doubly-fed induction generators under unbalanced grid voltage", IEEE T Energy Conver, vol. 25, 2010, pp. 356-368.

[11] L. Monjo, F. C´orcoles, J. Pedra, "Saturation Effects on Torque and Current Slip Curves of Squirrel Cage Induction Motors", IEEE transactions on energy conversion, vol.28, No1, 2013, pp. 243-254.

[12] N.C. Kar, H. M. Jabr, "A novel PI gain scheduler for a vector controlled doubly-fed wind driven induction generator, in: Proceedings of 8th IEEE Int. Conf. Electrical Machines and Systems, vol. 2, 2005, pp.948–953.

[13] J. Zhao, W. Zhang, Y. He, J. Hu, "Modeling and Control of a Wind-Turbine-Driven DFIG Incorporating Core Saturation during Grid Voltage Dips", Electronic Machines and Systems (ICEMS), 2008, pp.2438-2442.

[14] H.M. Jabr , N.C. Kar, "Effects of main and leakage flux saturation on the transient performances of doubly-fed wind driven induction generator", Electric Power Systems Research, Vol. No.8, 2007, pp 1019–1027.

[15] H.M. Jabr, N.C. Kar, "Leakage flux saturation effects on the transient performance of wound rotor induction motors", Electric Power Systems Research, vol. 78, 2008, PP.1280-1289.

[16] M. Iordache, L. Dumitriu, "Voiculescu, R. Nicolae, D., Galan, N.: Saturated Induction Machine Steady-state Performance assessment through simulations", IEEE transactions on energy conversion, 2014, pp. 368-374.

[17] H.M. Jabr, N.C. Kar, "Starting performances of saturated induction motors", In: Proc. IEEE. Conf, Power Engineering Society General Meeting, Tampa Florida, USA, 24-28 June, 2007, pp. 1-7.

[18] S. Mefoued, "A second order sliding mode control and a neural network to drive a knee joint actuated orthosis", Neurocomputing 2015, vol. 155, pp. 71-79.

[19] X. Yao, Y. Jing, Z. Xing. "Direct torque control of a doubly-fed wind generator based on grey-fuzzy logic", In: International Conference on Mechatronics and Automation, 2007, Harbin, China. pp. 3587-3592.

[20] G.S. Buja, M.P. Kazmierkowski, "Direct torque control of PWM inverter-fed AC motors - a survey", IEEE T Ind Electron 2004, vol. 51, pp. 744-757.

[21] S.Z. Chen, N.C. Cheung, K.C. Wong, J. Wu, "Integral sliding-mode direct torque control of doubly-fed induction generators under unbalanced grid voltage", IEEE Trans. Energy Conver. 2010, vol. 25, pp. 356-368.

[22] A. Levant, L. Alelishvili, "Integral high-order sliding modes", IEEE T Automat Contr, vol. 52, 2007, pp. 1278-1282.

[23] A. Levant, "Higher-order sliding modes, differentiation and output feedback control", Int J Control, 2003; vol. 76, pp. 924-941.

[24] X. Zhu, S. Liu, Y. Wang, "Second-order sliding-mode control of DFIG-based wind turbines", In: IEEE 2014 3rd Renewable Power Generation Conference; 24-25 September 2014, Naples, Italy, pp. 1-6.

[25] N. C. Kar, H. M. Jabr, "A novel PI gain scheduler for a vector controlled doubly-fed wind driven induction generator", in: Proceedings of 8th IEEE Int. Conf. Electrical Machines and Systems, vol. 2, 2005, pp.948–953.

[26] P. Kundur, "Power Systems Stability and Control", New York, McGraw-Hill, USA, 1994, pp. 296–297.

[27] G. J. Rogers, D. S. Benaragama, An induction motor model with deep-bar effect and leakage inductance saturation", Arch Elektrotech, vol. 60, 1978, pp. 193–201.

[28] B. Beltran, M.E.H. Benbouzid, T. Ahmed-Ali, "High-order sliding mode control of a DFIG-based wind turbine for power maximization and grid fault tolerance", In: IEEE International Electric Machines and Drives Conference, May 2009; Miami, Florida, USA, pp. 183-189.

Energy Dissipation Model for 4G and WLAN Networks in Smart Phones

Shalini Prasad
Research Scholar, Jain University
Dept. of Electronics and Communication Engineering
City Engineering College, Bangalore-560061, India

S. Balaji
Center for Engineering Technologies
Jain Global Campus, Jain University
Jakkasandra Post, Kanakapura Taluk
Ramanagara Dist.-562112, Bangalore, India

***Abstract*—With the modernization of the telecommunication standards, there has been considerable evolution of various technologies to assist cost effective communication. In this regard, the fourth generation communication services or commonly known as 4G mobile networks have penetrated almost every part of the world to offer faster and seamless data connectivity. However, such services come at the cost of energy drained from the smart phone supporting 4G services. This paper presents an algorithm that is capable of evaluating the actual amount of energy being dissipated while using next generation mobile networks. The study also performs a comparative analysis of energy dissipation of 4G networks with other wireless local area networks to understand the networks that cause more energy dissipation.**

***Keywords*—G Wireless Networks; Energy Consumption; Smart phone; Wi-Fi**

I. INTRODUCTION

There has been a remarkable improvement in the cellular technologies right from handheld devices to services in last 5 years. This phenomenon has lead to a new era of mobile computing that has potential supportability of ubiquitous and pervasive computing. At present, the mobile applications are mushrooming in faster pace in the commercial markets and its adoptability seems to be exponentially higher. The users make use of these mobile applications for multiple purposes: i) entertainment, ii) social networking, iii) business utilities, iv) remote monitoring system, v) educational purpose, vi) public service utilities and many more. Although the usage of such mobile apps makes the work easier and saves lots of productive time, it is done at the cost of battery. Normally, mobile apps with extensive threads use the hardware resources (brightness, contrast, sound, touch etc.) leading to high energy dissipation [1]. With advancement in hardware circuitry and mobile operating system, various features have been evolved that are in compliance with Moore's Law [2] [3] as witnessed in the existing evolution of different processors and circuit design in the existing smart phones. But with the size of the hardware circuits scaling down, the energy dissipation is quite difficult to control and obtaining maximal performance is another challenging factor. There are two prospects in this: (i) novel hardware design to support execution of new mobile applications and (ii) unique networking protocols that is anticipated to support seamless application. The existing wireless standards e.g. IEEE 802.11, 802.16, 802.15, etc. are used for Wireless Local Area networks, 3G, 4G networks respectively. This IEEE standard is now extensively used in creating a network of mobile communications. It *should* be known that all these IEEE standards do have distinct protocol stacks and unique networking characteristics. Interestingly, the usage of IEEE standard is not uniform in smart phone devices. The prime reason behind this is that smart phone has multiple forms of antenna e.g. main antenna, WLAN antenna, GPS antenna, FM antenna, and diversity antenna. All these antennas use different IEEE standards and specifications that cause excessive energy dissipation. Hence, if the network has more supportability of multicarrier signals e.g. 3G and 4G networks than it is quite evident that energy dissipation from battery would be increased. However, the users also switch option between the usage of 3G/4G network as well as IEEE 802.11 networks for obtaining faster access to Internet-based resources. Usage of IEEE 802.11 standard for wireless access by the cell phone causes faster access even for heavier applications which cannot be seen much in 3G/4G network. A simple example is Skype calls whose quality is quite poor in 3G and 4G networks while it is superior in Wi-Fi networks. However, there is an unsolved question i.e. which is energy efficient network 3G/4G or IEEE 802.11 standard. The researcher in [4] has shown that usage of any wireless network works by task sharing and cooperation process in order to balance the load and energy. Therefore, multiple network participation permits decreasing the energy utilization of cell phones and allow the use of a few services, like file streaming or downloading, in addition to web scanning. Additionally, the author in [5] depicts another efficient , method for saving energy in portable VoIP. The presented technique has used GSM system in order to solve the energy consumption problem in WLAN network assisting in making voice calls. This paper discusses about a simple technique based on radio resource control to evaluate energy dissipation in wireless networks. It also provides a few results to demonstrate the effect of 4G and WLAN networks on the energy utilization of smart phones concentrating on voice services and data association. Further, this paper exhibits a cost efficient approach for dependable measurements of energy on smart phones and provides a modeling technique for scaling the battery consumption over IEEE 802.11 networks or in 4G networks. Section II briefly explains the background and motivation of the proposed work. It includes cellular power management, Wi-Fi power management, and the measuring methodologies for 4G and Wi-Fi, motivation limitations of mobile devices and finally quality of experience perception factors. Section III reviews the

work being carried out by other researchers. Section IV explains design and implementation of energy consumption model along with an algorithm for the proposed model. Section V presents results and discusses the proposed scheme. Section VI gives concluding remarks and direction for future work is given in Section VII.

II. Background and Motivation

Different vendors differentiate among the mobile devices by the services they offer and utilize completely accessible computational energy to offer new services. As shown in Figure 1, as more and more applications and services a smart phones offer, more the energy utilization will be. A phone, when runs out of battery, cannot provide access to mobile services for its users, thus reducing the revenue generated by the service to the service provider. This leads the manufacturers to focus on developing ways to extend battery life and hence the device operational time.

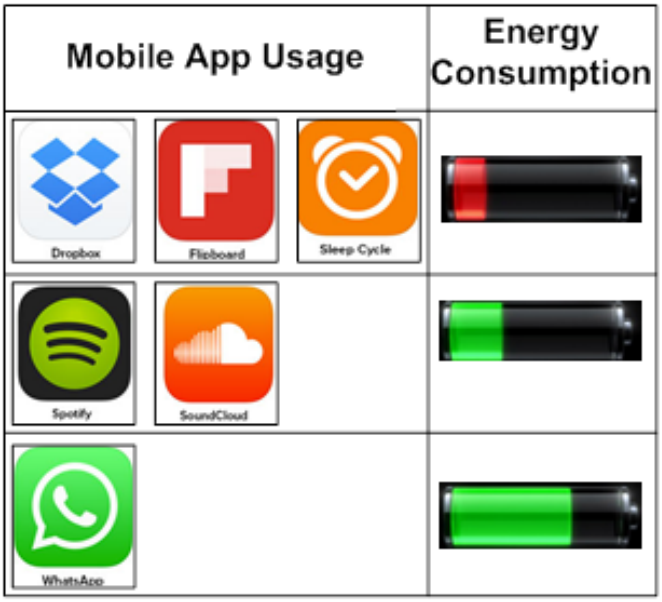


Fig. 1. Energy Consumption by Mobile Applications

A. Background

In order to retain superior service quality, the existing manufacturers have started emphasizing on energy optimization in smart phones along with associated constraints of size and heat dissipation [6][7]. Basically, every smart phone has three units: processing unit, power unit, and radio (or communication) unit. All these units have varying energy requirements. The energy drainage in the smart phone is controlled by two significant parameters: i) radio resource control and ii) transmission energy. The radio resource control is responsible for managing the control plane and also causes establishment of connection, broadcasting, notification of paging etc. Similarly, the transmission energy is responsible for allocation of energy required for forwarding definite bits of data in a defined communication channel. Both the parameters are dominantly used as the first preference to perform energy management in low power communication devices. Figure 2 shows a description of the radio resource control and its associated mechanism when implemented on existing GSM networks as well as on any WCDMA networks with 4G compliance [5]. The mechanism turns the radio in idle phase when there is no significant activity in a network. This state of idleness also consumes some amount of power. The radio immediately switches to state of high power if an active network is sensed. This principle either uses Forward Access Channel (FACH) or Dedicated Channel (DCH). The prime responsibility of DCH is to retain the dedicated channel in order to obtain maximized throughput as well as minimized latency. However, all these are achieved at the cost of energy. On the other hand FACH state is responsible for channel sharing with all available devices. It is preferred option of energy management when there is less availability of traffic for performing transmission. Power consumption capability of FACH is better in comparison to DCH.

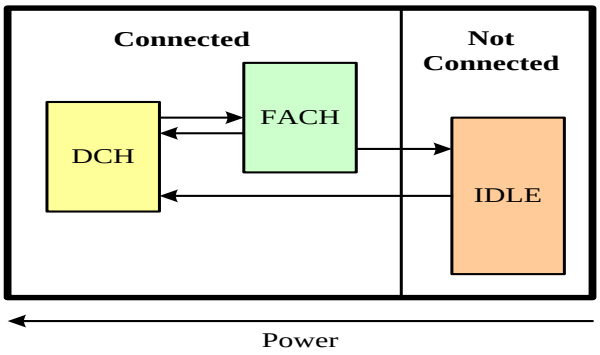


Fig. 2. Radio Resource Control Mechanism

The routers or the access points in the Wi-Fi networks are responsible for controlling the energy dissipation. However, it all depends on i) what type of Wi-Fi network it is? ii) How big it is?, and iii) what protocol is used in routing? An older version of Wi-Fi uses IEEE 802.11 a/b/g family quite frequently. However, now we have IEEE 802.11n family too and much higher versions are also available. Majority of them use power up to 6 watts in 24 Hours. The WLAN router does not have many options to save its energy consumption other than turning it off. However, Wi-Fi features available on smart phones use power saving mode to minimize unnecessary energy depletion. According to the concept of power save mode, the size of the data to be transferred and the transmitted power value are directly proportional to amount of energy being considered for data transmission in Wi-Fi networks. The next part of the background discussion is on cellular networks that support 4G services. The usage of 4G services make use of LTE (Long Term Evolution) which already provides multiple tools to conserve energy during radio access mechanism in 4G networks. There is also availability of various routing mechanisms that have significant energy conservation properties in 4G.

B. Motivation

4G telephones offer good services compared with 3G telephones, particularly with regards to bit rate while downloading or transferring information. Besides, they can support information and voice activity in the meantime permitting video calls, for instance. However, utilization of

data services is just gradually turning out to be more widespread. Numerous customers still utilize their telephone primarily for voice and Short Message Service (SMS) and very little for data broadband services. Also, numerous areas still have constrained 4G scope and the telephone persistently makes hand-offs from 3G to 4G systems and the other way around as the cell telephones move into and out of 4G scope. Due to this switch over from 4G to 3G and vice versa particularly when no information transmission is required, has a high cost as far as energy utilization is concerned.

The need for Wi-Fi based phones is increasing rapidly due to the ubiquitous presence of WLANs. Power consumption is the vital issue in the selection of a mobile phone [9]. Since all these devices make use of rechargeable batteries, the attractive services and features unfortunately drain lot of energy stored in a capacity limited battery of a smart phone. Achieving low power consumption for wireless devices has become a key design issue [10] which motivated us to carry out this research. For any mobile device, the major constraints are battery life, size and weight [11]. Such handheld devices will require enough power to perform processing. However, they attempt to minimize energy consumption during idle networking conditions. Discussion on energy saving schemes on such devices can be seen in [12] [13]. In [14], recommendations for amplifying the battery life of an Android device are given. It is a typical issue among smart phone manufactures in figuring out ways to expand battery life of their device and permit clients to utilize portable administrations for a more drawn out time [12]. This is also another motivation behind carrying out this research. Though mobile devices are making it easier for users to communicate, many users get frustrated while accessing certain applications and services using their mobile devices. The technological advancements of mobile device are new and yet they are improving [15]. Certain limitations of the mobile devices are:

a) Limited memory
b) Limited processing power
c) Battery consumption
d) Simplicity
e) Accessibility

Apart from technical specifications, it is also essential that various factors should be used to scale the experience of user. This perception is termed as QoE which determines the usability of the service or application in subscriber's perspective. The following are some of the critical perception factors of QoE [16]

a) Speed
b) Accessibility
c) Session quality
d) Integrity
e) Flexibility

The way clients perceive the execution of a system or a mobile service is an ultimate method for measuring that specific administration or system. This perception is named as QoE and is a definitive method for measuring that specific administration or system. The next section discusses about the existing techniques that have emphasized on the various techniques for conserving energy with respect to various network-based services.

III. Related Work

A few studies have proposed models for assessing the energy utilization of mobile services. In any case, as far as we are concerned, proposed model is the first outline stage energy utilization estimation model considering the different energy utilization properties like signalling and media exchanges. The related work depicted below clearly explains the advance energy consumption schemes used in the past.

Gross et al. [17] present a study that can compute the cumulative energy being drained in a smart phone which was done on the basis of component-based modeling. The study has shown the capability of the device to do so by approximately 4.7% of error. Another study by Haverinen et al. [18] showcase an impact of certain forms of messages and notifications on the lifetime of WCDMA networks. Usually such forms of networks are capable of faster data delivery with high throughput. The experiment performed by the author shows radio resource control is highly influenced by the energy being consumed. Vegara et al. [19] present a tool called Energy Box which can measure the energy consumption in the devices connected in 3G networks or in Wi-Fi Networks. According to the author, the traffic pattern is responsible for the energy being consumed over such mobile networks. The study shows an accuracy of 99% in energy computation for both the networks.

Balasubramanian et al. [20] present a mechanism to evaluate the energy being consumed in utilizing regular GSM networks as well as Wi-Fi networks. The computation of the energy was carried out with respect to the overhead in tail energy. The significant contribution of the study is to formulate a protocol called Tail Ender for minimizing the energy dissipation in mobile applications. Kelenyi et al. [21] make use of distributed hash tables to analyze the energy being spent by the mobile phones. Lane et al. [22] compute energy in smart phone equipped by sensors. The technique allows aggregation of the sensory data from the cellular phone to minimize the energy overhead for the user. The mechanism allows collection of the active usage data to develop a decisive model for energy conservation. The studies carried out by Damasevicius et al [23] have used the concept of measuring the energy consumed due to running of multiple applications on mobile device. The study also uses 3DMark06 (a benchmarking tool) to measure the effectiveness of the technique. Perala et al. [24] present a tool that can compute the extent of energy utilization on WCDMA networks governed by radio resource control.

Han et al. [25] present a scheme for energy being consumed in smart phones. Study on energy consumption over IEEE 802.11g network by Xiao et al. [26] and nearly similar direction of study was also carried out by Zhang et al. [27]. Study towards energy efficiency in 4G networks as well as a WLAN network was done by Harjula et al. [28]. The author presents a sophisticated model for testifying the impact of protocols residing in application layer on the energy dissipation

of a cellular phone. Miranda et al. [29] investigate the role of transport layer security on the energy consumption factor of the cellular phones. Similar work was also carried out by Abbas et al. [30]. However, the authors have incorporated a machine learning-based algorithm. The technique also uses cross-validation mechanism for evaluating the effectiveness of the mechanism. The study carried by Le et al. [31] presents an investigation of multiple radio-technologies used in advanced mobile networks and studied the energy usage over uplink and downlink transmission. Ravi et al. [32] have shown the possible use of cloudlets to minimize massive energy consumption in 3G and 5G mobile networks. The technique has investigated the handoff mechanism considering multiple parameters e.g. bit rate, signal strength, and number of interaction. Fuzzy logic was applied to further strengthen the decision. Wang et al. [33] and Sun et al. [34] have performed research in similar lines to minimize the energy consumption due to 4G and Wi-Fi networks.

Thus, it can be seen that there are many researchers working on the existing systems to address the energy dissipation problems while accessing the data from the mobile devices using wireless networks like WLAN, 3G, and 4G. However, a closer look into the existing systems will also show that majority of the techniques have used radio resource control but did not develop a mathematical model to quantize the measurements. Hence, to bridge this gap, we propose a simple mathematical model which is illustrated in the next section.

IV. Design and Implementation of Energy Consumption Model

This section discusses about a simple computational model that can perform evaluation of energy on mobile devices due to the usage of mobile networks. This discussion of the proposed system was carried out with respect to two distinct algorithms.

A. Energy Consumption Model

The core objective of this model is to investigate signalling properties of mobile networks and their possible connections with energy dissipation. We take the case study of standard IEEE 802.11 network and 4G network protocols as the communication media whose signalling properties will be assessed. The proposed system does not use any form of offline or pre-stored communication data in order to perform computation of energy being dissipated from devices. We develop a computational model with a backbone design of analytical model considering the signalling properties of both the types of the mobile networks. We also consider radio resource control and its associated features with respect to the mobile networks of Standard IEEE 802.11 and 4G networks. The presented technique also considers the inherent characteristics of data transfer of both the mobile networks. . The simulation of the proposed logic was carried out using Matlab where we developed user interface for both client and server. The design principle of the study considers monitoring the energy usage based on the multiple applications running on the mobile devices. The study considers a cut-off based scheme deployed over an interval of the transmission of the data packets. We anticipate that our mechanism will bring better probability of the minimization of an energy being consumed with better linearity on the energy curves. We check for multiple prioritized applications and their respective threads running on the mobile devices. The study was conducted based on end-to-end monitoring of the energy factor on the mobile devices that is connected by wireless communication media of Standard IEEE 802.11 or 4G networks.

The discussion of the technique in the form of algorithm is showcased here that is developed for evaluating the energy on the mobile device.

Algorithm-1

Input: Fs, Ds, R_Ip, R_Port, Rec_port;
Output: D_matrix, Id_list;

1. ***Start***
2. *Initialize all the input parameters;*
3. *get Pkt_Size;*
4. *compute S^{th} value;*
5. *if (Pkt_Size = Th)*
6. *Pkt++;*
7. *read Id_list and D_matrix;*
8. *End;*
9. *if (Pkt_Size < 0)*
10. *Packet bad size;*
11. *End;*
12. *If (Pkt_Size < Th)*
13. $P_{tra}(k)=X*S_{sig}- Y*k +Z$*;*
14. *End;*
15. *If ($S_{sig}< S_{th}$)*
16. $P_{fach}(k) = \frac{k_1}{k}.[P_{tra}(k_1) - P_{fach}]+P_{fach}$*,;*
17. *End;*
18. *If (Pkt_Size >= Th)*
19. $P_{idl}(k) = \frac{k_3}{k}.[P_{tra}(k_3) - P_{idl}]+P_{idl}$*;*
20. *End;*
21. ***End;***

Algorithm-1 presents a technique based on signalling properties of wireless channels. The overall algorithm is summarized in Algorithm-1 and Algorithm-2. Algorithm-1 gives the algorithm for the upload WLAN system. It uses Fs, Ds, R_Ip, R_Port, Rec_port as inputs and the output is a data matrix and an ID list. Here, the first step is to initialize the input parameters at the server side GUI and select simulation option, WLAN or 4G, and enter the values for remote IP address, remote port number, received port number, average signal frequency, average data size and data transfer rate. If the packet size is equal to the threshold value, then increase the packet size and read the ID list and data matrix. If the packet size is less than zero value, the input is bad packet size. Then discard the packet. If the packet size is less than threshold value, then compute the equation (1) given below:

$$P_{tra}(k)=X*S_{sig}- Y*k +Z, \text{ where } k\leq k_1 \quad (1)$$

then similarly, if the packet size is less than the signal threshold value, then compute the equation (2) given below:

$$P_{fach}(k) = \frac{k_1}{k}.[P_{tra}(k_1) - P_{fach}]+P_{fach}, \text{ where } k_1\leq k\leq k_3 \quad (2)$$

Finally check if packet size is greater than or equal to threshold value; then compute equation (3) given below:

$$P_{idl}(k) = \frac{k_3}{k}.[P_{tra}(k_3) - P_{idl}]+P_{idl}, \text{ where } k\geqslant k_3 \quad (3)$$

where, k, k_1, k_2 and k_3 are the threshold values. In the above equation $P_{tra}(k)$ represents transmission power of the signal, the variable $P_{fach}(k)$ represents FACH power consumption of the system, while the variable $P_{idl}(k)$ represents ideal power dissipation of the signal. Then, X, Y and Z are the scaling factors.

Algorithm2 shows the server part of the proposed system. Here the data is uploaded through the wireless LAN.

Algorithm-2

Input: Fs, Ds, R_Ip, R_Port, Rec_Port;
Output: D_matrix, Id_list;

1. ***Start***
2. *Initialize all the input parameter;*
3. *Get Pkt_Siz;*
4. *Select network;*
5. *Initialize R_Ip, R_Port_No, Recv_Port_No, Avg_Fs, Avg_Pkt_Size, D_Tr;*
6. *Get msg;*
7. *If (WLAN=1)*
8. *Calculate Msg_In;*
9. *Download_WLAN;*
10. *End;*
11. *If (4G=1)*
12. *Calculate Msg_In;*
13. *Download_4G;*
14. *End;*
15. *Data downloaded/received successfully;*
16. ***End;***

Algorithms-2 gives the steps for server side and client side respectively. In this procedure the input is taken as Fs, Ds, R_Ip, R_Port, Rec_Port, msg, D_flg, Msg_In; at the output side is data matrix and ID list. First initialize all parameters and then start at step 2 and 3 of Algorithm-2. Then select the input packet size along with message input, select WLAN option and calculate the message input value. Then upload the WLAN side. If 4G is selected directly upload the 4G message input. Finally, download or receive successfully the uploaded input message.

Table-1 shows the variables used in this algorithm for different input and output side.

TABLE I. NOTATIONS USED

Sl. No.	Variables	Description
1.	Fs	Signal frequency
2.	Ds	Data Size
3.	R_Ip	Remote IP
4.	R_port	Remote Port
5.	Rec_port	Received port
6.	D_matrix	Data matrix
7.	Id_list	Id list
8.	Pkt_Size	Packet size
9.	Sth	$S_{threshold}$
10.	Th	Threshold
11.	S_{sig}	S_{sig}
12..	Avg_Fs	Average Signal Frequency
13.	D_Tr	Data Transfer Rate
14.	Msg_In	Input Message
15.	D_flg	Data flag

V. RESULTS AND DISCUSSION

The results we obtained are discussed in this section.

A. Server Side

The inputs given to server side are:

a) Choose: Simulation Option

b) Choose Device Specific: WLAN Network Option

c) Enter Remote IP: local host (you can also enter System IP)

d) Enter the Remote Port No: 300

e) Enter the Received Port No: 301

f) Average Sign Frequency: 200

g) Average Pkt Size (bytes): 250

h) Data Transfer (kbs): 15,000

B. Client Side

The inputs given to client side are:

a) Choose: Simulation Option

b) Choose Device Specific: WLAN Network Option

c) Enter Remote IP: local host (you can also enter System IP)

d) Enter the Remote Port No: 301

e) Enter the Received Port No: 300

f) Average Sign Frequency: 200

g) Average Pkt Size (bytes): 250

h) Data Transfer (kbs): 15,000

The results analysis is carried out on the basis of an energy being dissipated from the client module. Figure 3 highlights the energy consumption trend for the proposed system with increasing time scale. The assessment of the proposed model is accomplished by transmitting a test data in the distinct wireless channel of standard IEEE 802.11 and 4G networks. The trend of Figure 3 shows that in idle state power consumption is quite low while in an upload condition the power consumption is found increasing linearly and then it maintains a better linear behavior for 20-80 seconds of time limit. The trend is then found to decrease in its overhead at the same time. The trend eventually shows that power consumption lowers down in the state of network.

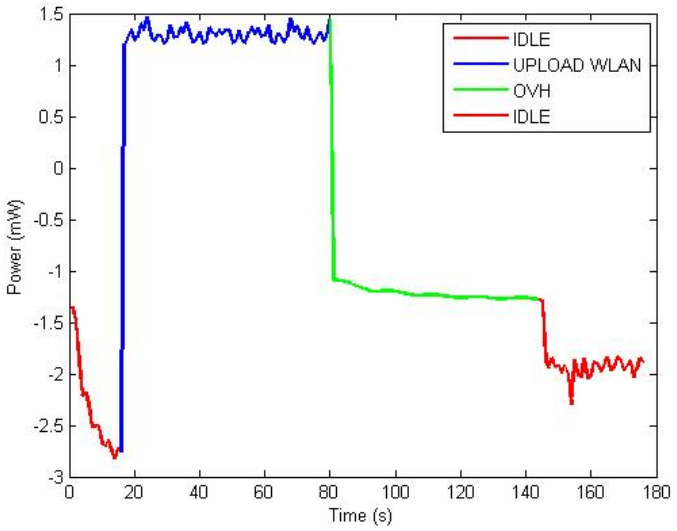


Fig. 3. Power (mW) Vs. Time (s)

We have also performed a comparative analysis of the proposed model in the context of Wi-Fi networks and 4G networks with respect to lookup time in terms of hours as shown in Figure 4. The study shows that energy dissipation is more for 4G networks in contrast to Wi-Fi networks. The same test file is forwarded twice in both networks in order to check the rate of energy dissipation effectively. Both uplink as well as downlink transmission were tested for this purpose and it was found that there are frequent switch overs of FACH to DCH by the 4G network which has resulted in exponential increase of energy consumption in comparison to Wi-Fi networks.

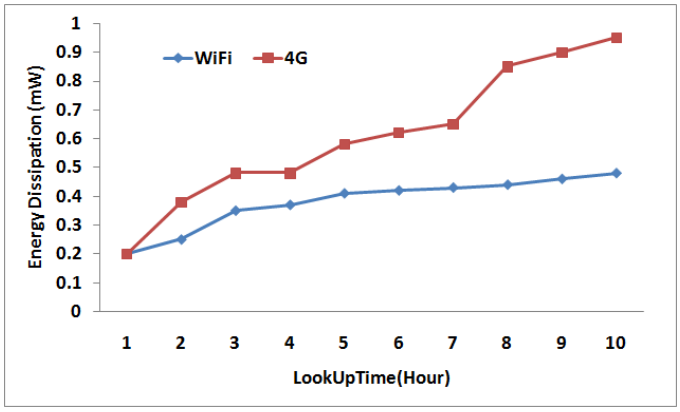


Fig. 4. Comparative Analysis on Different Networks

VI. Conclusion

Energy consumption is one of the critical problems in smart phones and successful execution of multiple mobile applications calls for extensive battery lifetime as well as energy conservation schemes. The biggest problem with the smart phone devices is to understand the importance of various threads running inside the applications. The existing energy conservation schemes call for suppressing some threads leading to temporary minimization of energy which has potential impact on application performance on the mobile device. Therefore, a simple and yet robust framework is required that can study the signaling properties of the usage of mobile networks and then perform investigation on the energy drainage. We argue that a precise detection of energy dissipation will always assist the energy saving schemes. In this research manuscript, we present one such computational model that is capable of measuring accurate energy drainage rate.

VII. Future Work

The study towards the future direction will be to develop a model for compensating the energy that was found to be dissipating. As the presented model is capable of assessing cumulative energy drainage owing to wireless local area network and 4G, the study could now trace the priority by identifying the applications or services consuming more energy and choose to suppress those applications for balancing the necessary power. As 4G services also offer increasing data transfer, we will investigate a better possibility of antenna management techniques for compensating the energy loss.

References

[1] G.P. Perrucci, F. H.P Fitzek, G. Sasso, W. Kellerer, and J. Widmer, "On the impact of 2G and 4G network usage for mobile phones' battery life", In Wireless Conference, European, pp. 255-259, 2009

[2] Y. Borodovsky, "Marching to the beat of Moore's Law." In SPIE 31st International Symposium on Advanced Lithography,International Society for Optics and Photonics, pp. 615301-615301, 2006

[3] K.D. Sattler, "Handbook of Nanophysics: Nanomedicine and Nanorobotics", CRC Press, pp. 887, 2010

[4] F. Fitzek and M. Katz, editors. Cooperation in Wireless Networks: Principles and Applications – Real Egoistic Behavior is to Cooperate, ISBN 1-4020-4710-X. Springer, April 2006.

[5] G. Perrucci, F. Fitzek, G. Sasso, and M. Katz. Energy saving strategies for mobile devices using wake-up signals. In MobiMedia - 4th International Mobile Multimedia Communications Conference, Oulu - Finland, 2008.

[6] G. P. Perrucci, "Energy Saving Strategies on Mobile Devices," Ph. D. Dissertation, Dept. Multimedia information and Signal Processing, Aalborg University, Denmark, 2009.

[7] Miao, Guowang, Himayat, Nageen, Li Ye, Swami, Ananthram, " Cross Layer OPtimization for Energy-E_cient Mobile Networking: A survey," Wireless Communication and Mobile computing, vol. 9, pp. 529-542, Apr. 2009.

[8] G. Perrucci, F. Fitzek, G. Sasso, and M. Katz. Energy saving strategies for mobile devices using wake-up signals. In MobiMedia - 4th International Mobile Multimedia Communications Conference, Oulu, Finland, 2008.

[9] A. Gupta, P. Mohapatra, "Energy Consumption and Conservation in WiFi Based Phones: A Measurement-Based Study," in 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad-Hoc Communications and Networks, 2007, pp.122-131.

[10] Hyun-Ho Choi, Jung-Ryun Lee, Dong-Ho Cho, "On the use of a power-saving mode for mobile VoIP devices and its performance evaluation,"IEEE Transactions on Consumer Electronics, vol. 55, no. 3, pp. 1537-1545, Aug. 2009.

[11] R. Want, "When Cell Phones Become Computers," Pervasive Computing, IEEE , vol. 8, no. 2, pp. 2-5, Apr. 2009.

[12] G. P. Perrucci, F. H. P. Fitzek, G. Sasso, W. Kellerer, J. Widmer, "On the impact of 2G and 4G network usage for mobile phones' battery life," Wireless Conference, Europe, pp. 255-259, May 2009.

[13] Battery Life, (Online: Veri_ed January, 2011). Available: http://android.bigresource.com/Samsung/-Vibrant-Battery-indicator-is-wrong--hxAgM7rn.html. Accessed on 11th July 2016.

[14] Maximising Battery Life, (Online: Veried December, 2010).Available:http://www.howtogeek.com/howto/25319/complete-

guide-to-maximizing-your-android-phones-attery-life/. Accessed on 11th July 2016.

[15] Limitations of Mobile Devices, (Online: Veri_ed April, 2011). Available:http://www.wireless-center.net/Mobile-and-Wireless/2505.html. Accessed on 11th July 2016.

[16] Yu Du, Wenan Zhou, Jian Liu, Junde Song, "A study on the extraction and application of customer perception indexes in mobile network," in IEEE International Symposium on IT in Medicine & Education, 2009, pp.436-440.

[17] C. Gross, F. Kaup, D. Stingl, D. Richerzhagen, D. Hausheer, and R. Steinmetz, "EnerSim: An energy consumption model for large-scale overlay simulators", IEEE-Local Computer Networks, pp. 252-255, 2013.

[18] H. Haverinen, J. Siren, and P. Eronen, "Energy Consumption of Always-On Applications in WCDMA Networks," in, IEEE Vehicular Technology Conference, Dublin, Ireland, 2007, pp. 964-968.

[19] E.J. Vergara, and S.N. Tehrani, "Energybox: A trace-driven tool for data transmission energy consumption studies," Energy Efficiency in Large Scale Distributed Systems. Springer Berlin Heidelberg, pp. 19-34, 2013.

[20] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications", in, ACM Internet Measurement Conference, Chicago, USA, 2009, pp. 280-293.

[21] I. Kelenyi and J.K. Nurminen, "Energy Aspects of Peer Cooperation - Measurements with a Mobile DHT System," in, IEEE International Conference on Communications, Beijing, China, 2008, pp. 164-168.

[22] N. D. Lane, Y. Chon, L. Zhou, Y. Zhang, F. Li, "Piggyback Crowd Sensing (PCS): energy efficient crowd sourcing of mobile sensor data by exploiting smart phone app opportunities", Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, 2013.

[23] R. Damasevicius, V. Stuikys, and J. Toldinas, "Methods for measurement of energy consumption in mobile devices", Metrology and Measurement Systems. No. 3, pp.419-430, 2012.

[24] P.H.J. Perälä, A. Barbuzzi, G. Boggia, K. Pentikousis, "Theory and Practice of RRC State Transitions in UMTS Networks," in, IEEE Broadband Wireless Access Workshop, Hawaii, USA, 2009, pp.1-6.

[25] H. Han, J. Yu, H. Zhu, Y. Chen, Y, "Energy-efficient engine for frame rate adaptation on smart phones,". Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, Vol.15, 2013.

[26] Y. Xiao, P. Savolainen, A. Karppanen, M. Siekkinen, and A. Ylä-Jääski, "Practical power modeling of data transmission over 802.11g for wireless applications" in, International Conference on Energy-Efficient Computing and Networking,Passau, Germany, 2010.

[27] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R.P. Dick, Z.M. Mao, and L. Yang, "Accurate Online Power Estimation and Automatic Battery Behavior Based Power Model Generation for Smart phones," in, International Conference on Hardware-Software Codesign and System Synthesis, Scottsdale, USA, 2010, pp. 105–114.

[28] Harjula, E.; Kassinen, O.; Ylianttila, M., "Energy consumption model for mobile devices in 4G and WLAN networks," in Consumer Communications and Networking Conference (CCNC), 2012 IEEE , vol., no., pp.532-537, 14-17 Jan. 2012.

[29] Miranda, P.; Siekkinen, M.; Waris, H., "TLS and energy consumption on a mobile device: A measurement study," in Computers and Communications (ISCC), 2011 IEEE Symposium on , vol., no., pp.983-989, June 28 2011-July 1 2011.

[30] Abbas, N.; Taleb, S.; Hajj, H.; Dawy, Z., "A learning-based approach for network selection in WLAN/4G heterogeneous network," in Communications and Information Technology (ICCIT), 2013 Third International Conference on , vol., no., pp.309-313, 19-21 June 2013.

[31] Le Wang; Manner, J., "Energy Consumption Analysis of WLAN, 2G and 4G interfaces," in Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom) , vol., no., pp.300-307, 18-20 Dec. 2010.

[32] Ravi, A.; Peddoju, S.K., "Mobility managed energy efficient Android mobile devices using cloudlet," in Students' Technology Symposium (TechSym), 2014 IEEE , vol., no., pp.402-407, Feb. 28 2014-March 2 2014.

[33] Yumei Wang; Lin Zhang; Hongyu An; Bin Xu; Geng Xi, "Power consumption testing and optimization for mobile router based on data aggregation and compression," in Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on , vol., no., pp.1-5, 24-27 June 2013.

[34] L. Sun, R.K. Sheshadri, W. Zheng, and D. Koutsonikolas, "Modeling WiFi Active Power/Energy Consumption in Smart phones," IEEE-34th International Conference on Distributed Computing Systems, pp. 41-51, 2014

Effective Data Mining Technique for Classification Cancers via Mutations in Gene using Neural Network

Ayad Ghany Ismaeel
Information System Engineering Department
Technical Engineering College, Erbil Polytechnic University
Erbil, Iraq

Dina Yousif Mikhail
Information System Engineering Department
Technical Engineering College, Erbil Polytechnic University
Erbil, Iraq

Abstract*—The prediction plays the important role in detecting efficient protection and therapy/treatment of cancer. The prediction of mutations in gene needs a diagnostic and classification, which is based on the whole database (big dataset enough), to reach sufficient accuracy/correct results. Since the tumor suppressor P53 is approximately about fifty percentage of all human tumors because mutations that occur in the TP53 gene into the cells. So, this paper is applied on tumor p53, where the problem is there are several primitive databases (e.g. excel genome and protein database) contain datasets of TP53 gene with its tumor protein p53, these databases are rich datasets that cover all mutations and cause diseases (cancers). But these Data Bases cannot reach to predict and diagnosis cancers, i.e. the big datasets have not efficient Data Mining method, which can predict, diagnosis the mutation, and classify the cancer of patient. The goal of this paper to reach a Data Mining technique, that employs neural network, which bases on the big datasets. Also, offers friendly predictions, flexible, and effective classified cancers, in order to overcome the previous techniques drawbacks. This proposed technique is done by using two approaches, first, bioinformatics techniques by using BLAST, CLUSTALW, etc, in order to know if there are malignant mutations or not. The second, data mining by using neural network; it is selected (12) out of (53) TP53 gene database fields. To clarify, one of these 12 fields (gene location field) did not exists inTP53 gene database; therefore, it is added to the database of TP53 gene in training and testing back propagation algorithm, in order to classify specifically the types of cancers. Feed Forward Back Propagation supports this Data Mining method with data training rate (1) and Mean Square Error (MSE) (0.00000000000001). This effective technique allows in a quick, accurate and easy way to classify the type of cancer.

Keywords*—Detection; Classification; Data Mining; TP53 Gene; Tumor Protein P53; Back Propagation Network (BPN)

I. INTRODUCTION

Cancer is a main cause of death worldwide; it has calculated for 7.4 million deaths in 2004 with an estimated 12 million deaths in 2030 [1]. Tumor protein P53, which is produced by Tumor Protein (TP53) gene, is a sequence-specific transcription factor that acts as a large tumor suppressor in mammals. The disorder in the function of the tumor suppressor p53 is one of the most common genetic changes in human cancer, which is close to 50% of all human tumors carry p53 gene mutations within their cells [2]. Fig. 1 shows the cancers and TP53 mutations on the worldwide.

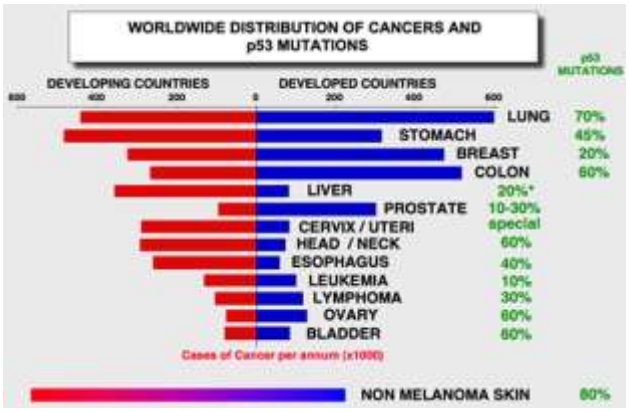


Fig. 1. Shows P53 (TP53 gene) mutations

Nowadays, biologists use a computer system like any other trained professionals but in general function. In addition, they use computers to solve problems that are very specific to them in the specialist tasks. They are taken together, to support the field of bioinformatics. More specifically, bioinformatics' focus is to analyze biological data and to do anticipations about biological systems, in order to provide more knowledge about how living organisms function [3]. Bioinformatics is an emerging discipline that bases upon the strengths of computer sciences, mathematics, and information technology to determine and analyze genetic information [4].

For instance, to predict whether two proteins react or not, it could be used computational biology. If the prediction is correct, then biological data that got from a wet lab experiment, including the proteins, should be analyzed by using computational biology to know how these proteins contribute to the physiology of an organism. Computational biology can be further broken down into molecular modeling and bioinformatics [3].

Data Mining (DM) is defined as the essence of the Knowledge Discovery in Databases (KDD) process. It includes the algorithm conclusions that explore the data, develop the model, and discover previously unknown models. The model is applied to understand phenomena from the data analysis and prognosis. The accessibility and abundance of data today makes knowledge discovery and Data Mining an issue of great necessity and importance [5].

At last, Data Base (DB) related to tumor protein P53 (TP53 gene) contains large amounts of data, these data in the database

are represented as excel sheet file, and regular techniques may not be helpful and impractical in such large volumes of data. So, artificial intelligence techniques such as Data Mining are used to simplify and improve the process of research and education. Data Mining is the method of analyzing the data by linking them with artificial intelligence techniques to examine and search for specific information, in addition, to take the useful data from a large amount of data. Moreover, Data Mining can be done through the process of linking this data analysis and artificial intelligence methods to become the most. This proposed method predicts, diagnose, and Classifies cancer mutations; So, the comparison between a person and standard gene protein sequence is done firstly by using BioEdit package. If differences might be found between the two proteins, then there is a malignant mutation. This stage has been used by bioinformatics techniques. Secondly, Data Mining technique (back propagation algorithm) is trained by using UMD Cell-line-2010 p53 mutation database that must be carefully selected to function correctly.

Artificial Neural Network (ANN) can discover how to solve problems by itself. Later, this trained Back Propagation Neural Network (BPNN) offers an effective and flexible predictions, diagnosis and genetic diagnosis technique of cancers.

II. Related Work

E. Adetiba, J. C. Ekeh, V. O. Matthews, and et al. [2011], this proposed study is aimed at assessing the best back-propagation learning algorithm for a genomic-based on ANN system for NSCLC diagnosis. It used the nucleotide sequences of EGFR's exon 19 of a noncancerous cell to learn ANN. MATLAB R2008a was used to test many BPNN training algorithms to get an optimal algorithm for learning the network. It were examined in the nine different algorithms and achieved the better performance (i.e. the least Mean Square Error MSE) with the minimum epoch (training iterations) and learning time using the Levenberg-Marquardt algorithm (trainlm) [6].

Syed Umar Amin1, Kavita Agarwal, and et al. [2013], introduced a new method to predict heart disease based on the neural network and genetic algorithm. The whole existent systems predicted heart diseases that depended on the clinical dataset, which is collected from complex tests that conducted in pathology labs. There was no method, which predicts heart diseases that depended on risk factors like diabetes, age, family history, high cholesterol, alcohol intake, tobacco smoking, obesity or physical inactivity etc. However, this system gave a patient a warning about a probable existence of heart disease even before he/she makes medical checkups. Two Data Mining tools, genetic algorithms and neural networks were used in this system. In this method, the system may not fall into the local minimum, because the genetic algorithm was applied for optimization of neural networks weights. This system used a multilayered feed-forward network with structure 12 nodes in the input layer, 10 nodes in the hidden layer and 2 nodes in the output layer, where the number of input nodes depends on the final set of risk factors for each patient. In the initial stage, the 'configure' function that available in MATLAB was used to initialize the neural network weights. After that, "these configured weights were passed to the genetic algorithm for optimization according to the fitness function". Once the weights were optimized, the 'trainlm' back propagation algorithm was used for training and learning. The accuracy of the system that predicts heart disease risk is 89%, because the learning process of the derived system was quick, more steady, and accurate as compared to back propagation neural network [7].

Ayad. Ghany Ismaeel, and Raghad. Zuhair Yousif [2015], proposed technique to classify, diagnose mutations' patient, and predict the mutation's position for the patient. TP53 gene (tumor protein P53) datasets were used and (6) fields were selected from UMD_Cell_line_2010 database, in order to train and test Quick back Propagation Network (QPN).The mining method was based on training (QPN), which is an improvement of the back propagation network, since (283-141-1) the number of nodes were used in input, hidden and output layers, by Alyuda NeuroIntelligence package. The training for all datasets (train, test, and validation dataset) led to the following results: the Correlation (0.9993), R-squared (0.9987), and mean of Absolute Relative Error (0.0057) [8].

III. Proposed of Effective Data Mining Method for Classification Cancers

The major functions of suggested Effective Data Mining technique for classifying specific cancer are shown in Fig. 2. The technique of classification specific cancer is done by using two approaches. The first approach predicts whethere the person has mutations that cause cancer or not. The second approach tha classifies the mutations are obtained from the first approach to know which kind of cancer it caused(cancers types) . Those two approaches are:

A. Bioinformatics Techniques:

There are many bioinformatics techniques for analysis and search genome, some of these helpful techniques are explained below:

*1) BLAST: "*Is defined as A powerful tool for searching sequence databases with an implementing sequence. BLAST is Basic Linear Alignment Sequence Tool. An earlier program, BLAST, worked by identifying local regions of similarity without gaps and then combining them together. BLAST includes an iterative process, as the emergent pattern becomes better defined in sequential stages of the search"[9].

2) CLUSTALW: It is the first technique for examining whether the person has a malicious mutation or not, which is based on the idea of "Two proteins can have very different amino acid sequences, it still be biologically similar (Homology)" [10]. The gene mutation is detected by using CLUSTALW. The gene mutation increases the probability of cancer. In CLUSTALW, users must know there are two types of sequence: one of them is the normal sequence of each gene (without mutation), the second is the person's gene sequence. The matching between them is examined [11].

The previous studies [11, 12] supported the following algorithm, which clarifies the major functions of the bioinformatics tools (sequences alignment).

Input: Standard Gene and persons TP53 Gene Sequence
Output: Diagnose are there malicious mutations or not in Person's P53 sequence.
BEGIN
Step1: Make FASTA format of Standard Gene and persons Gene Sequence
Step2: Use ClustalW for Sequences Similarity Check
Step3: If there is matching
Its normal gene
Step4: Else
Convert Gene Sequence from DNA to Protein
Step5: Apply ClustalW for Protein Similarity Check
Step6: If there is matching
Its normal gene
Step7: Else
There is Malignant Mutations.
END

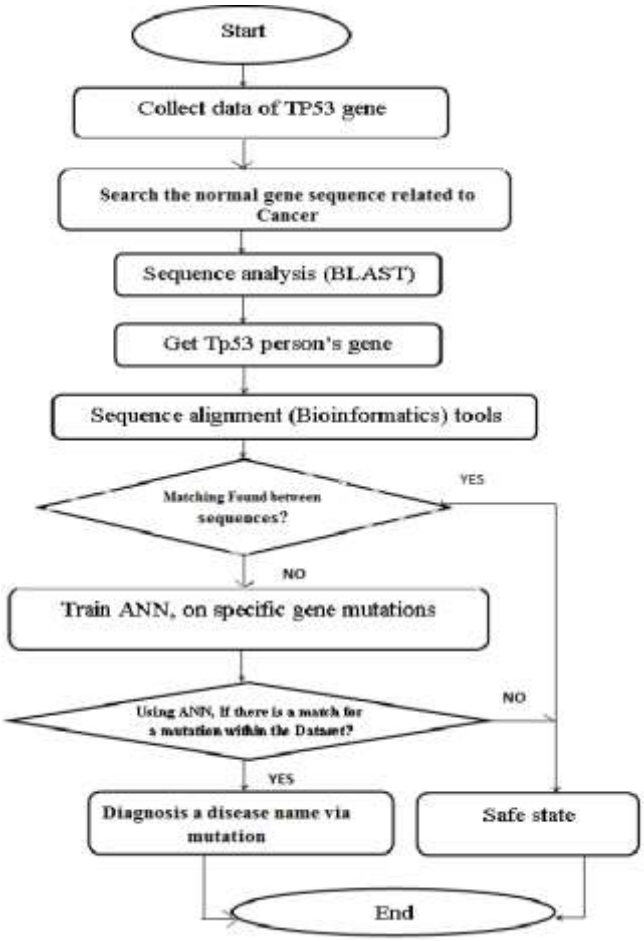


Fig. 2. Flowchart for main tasks of proposed Data Mining method

B. Data Mining technique (Feed Forward Back Propagation Neural Network):

After the first approach predicts there is a malignant mutation or not in the person's genes, then these results which obtained from bioinformatics technique(first approach) are not enough to classify a specific type of cancer. So, this proposed method needs the second approach, which focuses on training back-propagation neural network, because mutations which are related to cancer and mutations are gotten from the first approach need to be classified by neural networks. Learning or training stage is done by applied The Levenberg-Marquardt back propagation algorithm 'trainlm' is a network learning function that updates weight and bias values according to Levenberg-Marquardt optimization. The proposed structure of training BPN has 3 layers (input layer within layers of BPNN).

"The BPNN is a multilayered neural network applies a supervised learning method and feed-forward architecture. It is by far the most extensively used network". Classifying and predicting are done by using BPNN, because it is one of the most frequently used neural network techniques. "The principle of BPNN runs by approximating the non-linear relationship between the input and the output by adjusting the weight values internally". The neural network model is constructed by using the supervised learning algorithm of back propagation. [13].

The feed forward BPNN is a very common model in neural networks. The errors are back propagated during training, because it does not have feedback connections [14]. The Back-propagation learning process includes two stages in all different layers of the network: forward pass and backward pass [15]:

Forward pass: Input vector is entered to the sensory nodes of the network and its effect spreads out through the network, layer by layer. Lastly, a set of outputs is generated as the actual reaction of the network. During the forward pass, the synaptic weights of the network are all steady.

Backward pass: "The synaptic weights are all modified in accordance with an error correction base". An error signal can be computed by subtracted the actual response of the network from the desired response. "Then the error signal is back propagated through the network, against the direction of synaptic connections".

The steps of training BPN are shown below [16]: The terminologies needed in the algorithm are explained below:

xi – Input value

vij – input weight of hidden node

v0j– Weight of bias node from input to hidden

z _ inj – Weight from input to hidden node

Zj– output weight of hidden node

Wjk – bias node weight from hidden to output

Wok– bias node weight from hidden to output

y _ inj – input to hidden node

y j– Final output value

During the forward pass, information passes from the input node to the hidden node, until reaching to the output node. "All input nodes in the input layer are loaded with the values that are given for training. And for each input pattern, a target

output is also supplied. Each hidden node sums up all incoming values and its bias and then is passed to an activation function f(x)".

$$Z_{-in_j}=v_{oj}+\sum_{i=1}^{n} x_i \; v_{ij} \quad (1)$$

$$z_j=f(z_{-in_j}) \quad (2)$$

The output value is passed from the hidden node to each output node. The value from each hidden node is summed up by the output node, and then the output value passes to activation function.

$$y_{-in_k}=w_{ok}+\sum_{i=1}^{p} z_j \; w_{jk} \quad (3)$$

$$y_k=f(y_{-in_k}) \quad (4)$$

Determining the error is the start of the backward pass phase. The difference between the target and actual value represents the error. This error is back propagated to each hidden node. "In order to find that it is passed through the derivative of activation function".

$$\delta_k=(t_k-y_k)f'(y_{-in_k}) \quad (5)$$

Once is found as δ_k, the change in weight can be easily computed

$$\Delta w_{jk}=\alpha\delta_k z_j \quad (6)$$

$$\Delta w_{ok}=\alpha\delta_k \quad (7)$$

Learning rate determines how fast the model learns. If the Learning rate sets to a small value, then the network will need a long time to learn, but if it sets to a high value, then it will make the network inefficient "when there are variations in the input pattern. Updating the weights between the input and hidden layers require more calculations".

$$\delta_{-in_j}=\sum_{k=1}^{m} \delta_k \; w_{jk} \quad (8)$$

$$\delta_j=\delta_{-in_j} f'(z_{-in_j}) \quad (9)$$

$$\Delta v_{ij}=\alpha\delta_j x_i \quad (10)$$

$$\Delta v_{oj}=\alpha\delta_j \quad (11)$$

To obtain the updated weights, the old weights are added with the change.

$$w_{jk}(new)=w_{jk}(old)+\Delta w_{jk} \quad (12)$$

$$w_{ok}(new)=w_{ok}(old)+\Delta w_{ok} \quad (13)$$

The process is repeated until the selected error criterion is satisfied.

IV. Experimental Results

The Implementation of the Effective Data Mining Technique for classifying cancers via mutations in the gene (Tp53) is explained below:

1) First, the normal TP53 gene sequence is obtained from The Catalogue of Somatic Mutations In Cancer (COSMIC) site. It provides normal genes, genes information, and datasets. The search for the normal gene can be done by sending gene's name to the server, then selecting the sequence option which provides access to the normal gene (Deoxyribonucleic acid (DNA) or protein sequence).

2) It uses BioEdit package to get TP53 gene for person by selecting World Wide Web, then selecting BLAST at the National Center for Biotechnology Information (NCBI), after that selecting nucleotide blast. Then the gene sequence is pasted or uploaded from the file of the normal TP53 gene sequence, which is formatted in FASTA file.

3) The Effective Data Mining method uses BioEdit package to complete the first approach for prediction and diagnosis mutations. It applies clustalW to display alignment result between the normal gene and person's gene sequences. A comparison between normal's gene sequences (e.g. Tp53) with the person's gene is executed to find whether there are mutations in person's gene or not, as shown in Fig.3.

4) Step (3) is not enough, because its result cannot determine whether mutation affects in protein function or not. So, normal and person TP53 gene are transformed to tumor protein P53. Then, the same tool ClustalW in BioEdit package is used in order to diagnose whether there is malignant mutation as it is done in step (3) or not (No risk). Fig. 4 shows there is malignant mutation (ACC → CCC), i.e. the alignment finds the codon 155 converted from T (at Normal P53) to P (at Person's P53 gene).

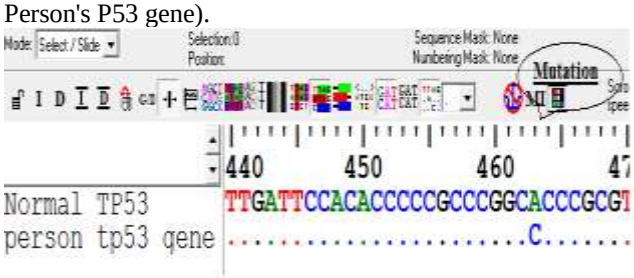


Fig. 3. Shows there is a malignant mutation in TP53 gene

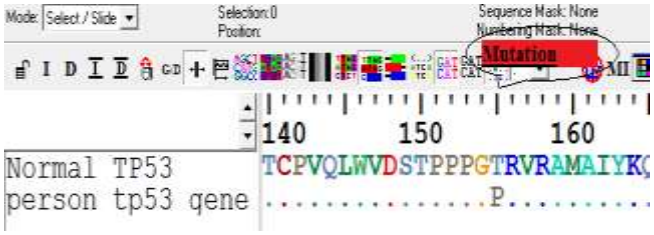


Fig. 4. Shows there is a malignant mutation in P53 protien

5) The previous step is used only to detect and predict malignant mutations. Also it does not give the cancer classification results, because the malicious mutations, which are discovered, are general. These malicious mutations related to Tp53 gene database. The common database (UMD_Cell_line_2010) is used to train BPN , which consists of (53) fields and 1448 records. The database (UMD_Cell_line_2010) from TP53 website which is modern and comprehensive database under URL:http://p53.free.fr/Database/p53_MUT_MA T. html[2]. But in Effective Data Mining method, (12) fields are selected for training and testing BPNN; (11) fields are selected from UMD_Cell_line_2010 database. The remaining field is a new field called (gene location field), which is added to the (11) fields selected, in order to get accurate and efficient results in cancer classification. The sample of this database is shown in Fig. 5.

6) Matlab R2015a is used on PC type core i5 for neural network toolbox, because it contains several tasks. The classification of malicious mutations for cancer is created successfully by using the structure of feed-forward BPNN and (trainlm) algorithm to obtain an optimal classifier for classification cancer with MSE (0.1E10-13) as shown in Fig. 6.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Mutation position	Codon	WT codon	Mutant	WT AA	Mutant	Event	Mutatio	Exon	Type	gene location	Cancer
2	94	31	GAG	CAG	Glu	Gln	G>C	B	5/8	Tv	Lung	Lung (NSCLC)
3	94	31	GAG	CAG	Glu	Gln	G>C	B	GBD	Tv	Gastric	Gastric carcinoma
4	94	31	GAG	CAG	Glu	Gln	G>C	B	8/8	Tv	T-cell Lymphobl	T-cell Acute Lymphoblast
5	110	38	CAG	CTG	Gln	Leu	A>T	B	GBD	Tv	Lung	Lung (NSCLC)
6	114	37	GAA	GAT	Glu	Asp	A>T	B	8/8	Tv	choriocarcinoma	choriocarcinoma
7	135	34	AAA	AAT	Lys	Asn	A>T	B	GBD	Tv	choriocarcinoma	choriocarcinoma
8	163	33	CCC	ins1	Pro	Fs.	ins	F	GBD	Fr.	Lung	Lung (NSCLC)
9	163	34	CCC	del1a	Pro	Fs.	Stop at 43	F	5/8	Fr.	Lung	Lung (NSCLC)
10	166	33	TTG	ins1b	Leu	Fs.	Stop at 42	F	5/8	Fr.	Lung	Lung (NSCLC)
11	172	37	TCC	CCC	Ser	Pro	T>C	B	8/8	Ts	T-cell Lymphobl	T-cell Acute Lymphoblast
12	199	46	TCC	del1b	Ser	Fs.	Stop at 122	F	GBD	Fr.	bronchus	Lung (SCLC)
13	203	47	CCG	CTG	Pro	Leu	C>T	B	5/8	Ts	Lung	Lung (NSCLC)
14	206	48	GAC	GGT	Asp	Gly	A>G/C>T	T	5/8	Ts/Ts	Biliary tract	Biliary tract carcinoma
15	208	49	GAT	CAT	Asp	His	G>C	B	GBD	Tv	bone marrow	Acute Myelogenous Leuk
16	214	51	GAA	TAA	Glu	Stop	G>T	B	5/8	Tv	Colorectal	Colorectal carcinoma
17	217	52	CAA	TAA	Gln	Stop	C>T	B	5/8	Ts	Colorectal	Colorectal carcinoma
18	217	52	CAA	ins3a	Gln	InF	In frame ins	I	5/8	Fr.	head SCC cell	Head and Neck SCC
19	222	53	TGG	TGA	Trp	Stop	G>A	B	5/8	Ts	bronchus	Lung (SCLC)
20	224	54	TTC	TAC	Phe	Tyr	T>A	B	5/8	Tv	head SCC cell	Head and Neck SCC
21	226	55	ACT	ins1c	Thr	Fs.	Ins	F	5/8	Fr.	Astrocytoma	Astrocytoma
22	232	57	GAC	del17	Asp	Fs.	Stop at 142 ?	D	5/8	Fr.	Gastric	Gastric carcinoma
23	232	57	GAC	del1c	Asp	Fs.	Stop at 122	F	5/8	Fr.	lymph nodes	Mantle Cell Lymphoma

Fig. 5. Shows sample of dataset (database) which used in training BPN

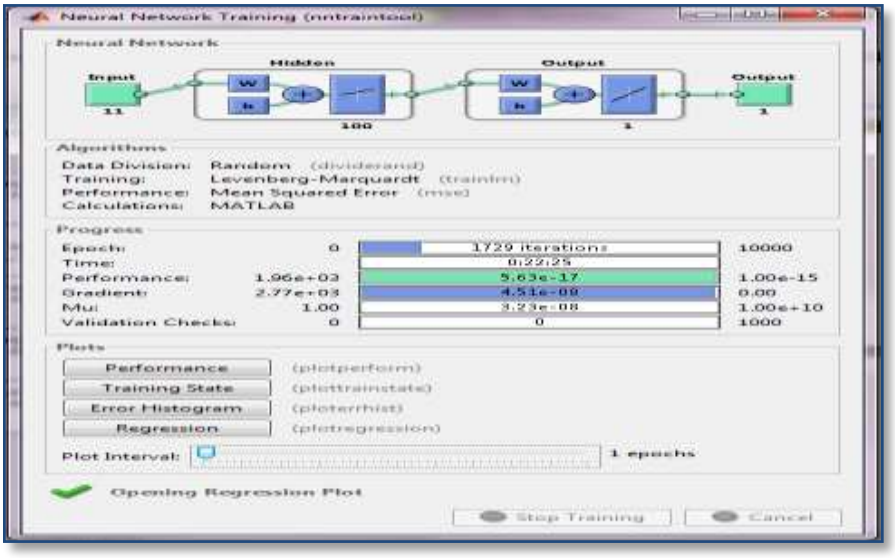


Fig. 6. Shows training result of BP algorithm

Fig. 7 shows plots and the elements of this learning process (Fig. 7; A reveals performance, Fig. 7; B shows regression and Fig 7; C reveals training state).

7) The trainer BPNN with malignant mutations of TP53 is completed. Then designed GUI for the doctors , biologists and other users of proposed method is tested. The Effective Data Mining method allows to classify cancers via mutations of a certain person (by entering each field of data manually in GUI).

8) The malignant mutation at codon 155 (ACC CCC) is obtained from using ClustalW in BioEdit package. Then trainer BPNN can be used to classify cancer type that may occur due to this mutation, for example, the result of classifying the malignant mutation at codon 155 (ACC CCC) is Head and Neck SCC Cancer, as shown in Fig.8.

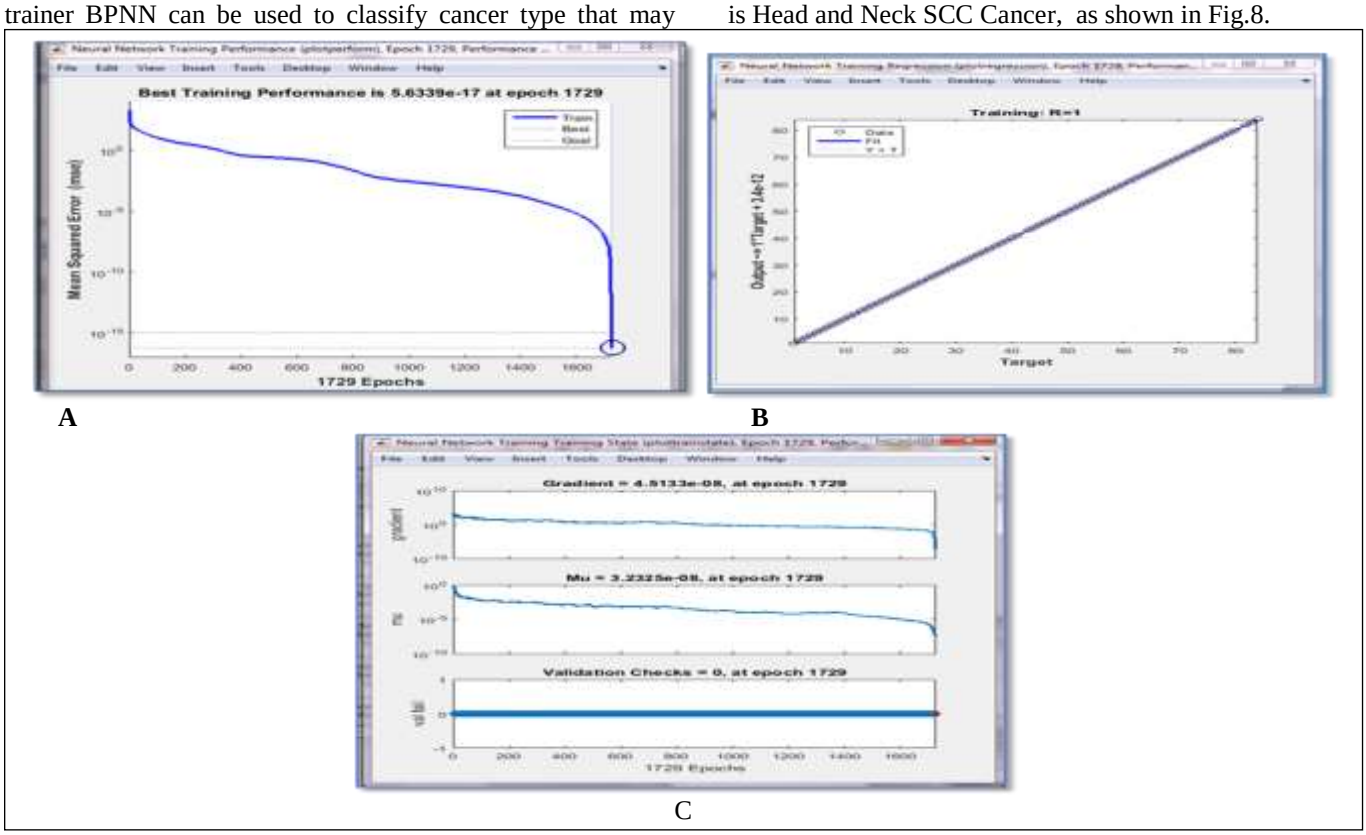


Fig. 7. Shows plots and the three training BPNN elements

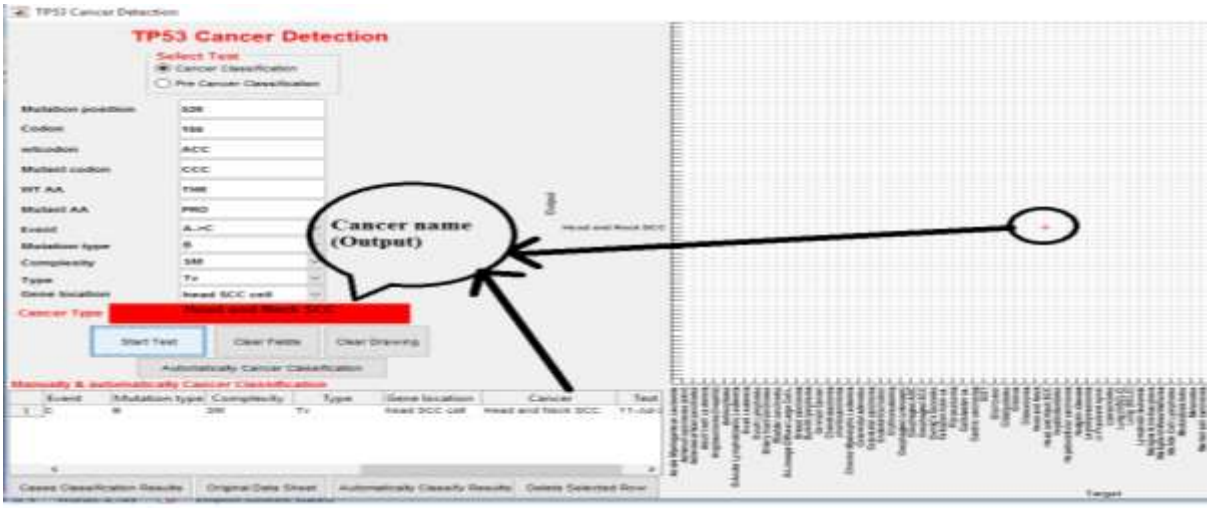


Fig. 8. Classfy cancers via mutations of Person's P53

V. Discussion the Results

The learning process is achieved, and it is highly successful. To meet the performance goal, it only takes 22 minutes to complete the learning process. Then the problem is presented to the trained model to classify the cancers. The DM method is an effective way in the classification cancers via mutation, since BPN is used in training and testing a minimum number of fields, which is (12) out of (53) fields in each record of TP53 database. The data of p53 database was saved in columns and records in Excel sheet file, as shown in Fig. 5. Whereas, (7) out of (53) fields for each record of TP53 database were used in the Novel Mining method, and 14 fields were used in the heart diseases method. These fields depend on the final set of risk factors for each patient. In addition, the proposed DM method adds a new field called Gene Location

field to the UMD TP53 database in order to make the neural network be able to classify specific type of cancer, and give accurate results; While the other related methods base only on the original database. Moreover, the proposed DM method of classifying cancer is compared with other related techniques or methods. The comparison is done in terms of goals, the used database, neural network structure, techniques, and performance; While all the other related methods use neural network but with different structure, training algorithm, performance, and results. More details are shown in Table I.

TABLE I. SHOWS COMPARISON OF EFFECTIVE DATA MINING METHOD WITH TWO OTHER METHODS

Feature	**The proposed method**	**Amin, et. al., [7]**	**Ayad, et. al., [8]**
The goal	Prediction, diagnosis ,classification specific cancer	prediction ,diagnosis heart diseases	Prediction, diagnosis ,classification mutation position
Universal method	Yes with two approaches, and adding new data field to DB	no	yes
Including Tp53 gene	yes	no	yes
DNA and Protein check	yes	no	yes
Sequence similarity check	yes	no	Yes
Data base used	UMD TP53 mutation DB	American Heart Association survey	UMD TP53 mutation DB
Technique used	Bioinformatics and DM tools(BPN algorithm)	DM techneques (neural networks and genetic algorithms)	Bioinformatics tools, quick BPN algorithm
Weight update function	trainlm	trainlm	QBP
ANN topology	(11-100-1)	(12-10-2)	(283-141-1)
Performance	0.1E10-13	0.034683	0.000006
Program	MATLAB R2015a	MATLAB R2012a	Alyuda NeuroIntellig e-nce
Support for	Researchers, bioinformatics doctors, and biomedical Eng.	Doctors	Researchers

VI. CONCLUSIONS

The proposed Data Mining method of cancer classification explains the following conclusions:

1) The proposed Data Mining method provides flexible diagnosis and prediction. Also, it classifies cancers via mutations in tumor protein P53 sequence. BBNN algorithm is used with the best performance (MSE), which reaches to (5.6339E-17), and the training rate(R) equals (1), as shown in Fig.7A&C. While in the Novel Mining method, Quick BPN algorithm was used with performance (0.000006), the training rate(R) equals (0.9987). Whereas in the heart diseases method, BP neural network and genetic algorithms were used with performance (0.034683).

2) The proposed approach shows the classification of cancer via predicting mutated P53 gene, in order to reduce the risk of cancer infection. This is done to keep people away from radiation, exposure to toxins, control themselves at older ages, and arrange their food system. In addition, the earlier diagnosis can predict the therapy for the mutated tumor protein P53.

3) Since cancer is an inherited disease and many different cases have been appeared in many families history around the world, this study is important for a further work to set up a database for a local area (e.g. for Middle East). This database would include the background or history of each family datasets, by taking into consideration the genetic diseases data of the family history. Then a complete system would constructed that is able to predict genetic disease early. Also, this local database would support therapy process by using therapy techniques like replacement, drug discovery, etc.

4) The results are obtained from this method can be forwarded to include the gene therapy by using therapy techniques as it is mentioned in the previous point, where therapy is a field in Biotechnology science.

ACKNOWLEDGMENT

(Mr. HAIDER HASAN HUSSIEN) Thanks for helping me when adding a new field that called gene location field in UMD database and yours advice at the medical side.

REFERENCES

[1] I. CATH Tee, and A. H. Gazala, "A Novel Breast cancer prediction system.", IEEE, pp.621-625, 2011.

[2] France database of TP53 gene, update 2010, can accessed at URL: http://p53.free.fr/Database/p53_MUT_MAT.html.

[3] J.Claverie, and C. Notredame, "Bioinformatics for dummies", Wiley publishing Inc, 2nd Edition., 2007.

[4] G. B. Singh, " Fundamental of bioinformatics and computational biology", Springer .

[5] O. Maimon, and L. Rokach, "Data mining and knoelodge discovery handbook", Springer Science+Business Media, 2005.

[6] E. Adetiba, J. C. Ekeh, V. O. Matthews, S.A. Daramola, and M.E.U. Eleanya, "Estimating an optimal backpropagation algorithm for training An ANN with the EGFR exon 19 _ucleotide sequence: an electronic diagnostic basis for non–small cell Lung cancer(NSCLC)", Journal of Emerging Trends in Engineering and Applied Sciences JETEAS, vol2, issue 1, P74-78, 2011.

[7] S. U. Amin, K. Agarwal, and R. Beg, " Genetic neural network based data mining in prediction of heart disease using risk factors ", Proceedings of IEEE Conference on Information and Communication Technologies (ICT), pp.1227-1231, 2013.

[8] A. Gh. Ismaeel, and R. Zuhair Yousif, " novel mining of cancer via mutation in tumor protein p53 using quick propagation network", International Journal of Computer Science and Electronics Engineering (IJCSEE),vol.3, issue 2, pp.121-126, (2015).

[9] A. M. Lesk, "Introduction to bioinformatics", Oxford University Press Inc., New York, 2002.

[10] Autumn, "Introduction to bioinformatics", Chapter 7: Rapid alignment methods: FASTA and BLAST, pp. 83-116 , 2007

[11] A. Gh. Ismaeel, and A. A. Ablahad, "Novel method for mutational disease prediction using bioinformatics techniques and backpropopagation algorithm", IRACST- Engineering Science and

Technology: An International Journal Vol. 3, pp. 150-156, 2013, (online).

[12] A. Gh. Ismaeel, and A. A. Ablahad, "Enhancement of a novel method for mutational disease prediction using bioinformatics techniques and backpropagation algorithm", international journal of scientific & engineering research, vol. 4, issue 6, 2013.

[13] W. D. Chen, W.Chen, and W. Pei, "Back-propagation neural network based importance-performance analysis for determining critical service attributes", Chung Hua University, pp.1-26, 2006.

[14] K.U. Rrami, " Parallel approach for diagnosis of breast cancer using neural network technique", International Journal of Computer Applications, Vol.10, PP.1-5, 2010.

[15] M. Hajek, "Neural networks", Neural networks doc , pp.(39), 2005.

[16] V.Kalaichelvi, and A. Shamir Ali., "Application of neural networks in character recognition", International Journal of Computer Applications ,Vol. 52, pp.1-6, 2012.

AUTHOR PROFILE

Ayad Ghany Ismaeel is a professor at 7 Aug, 2012 and awarded his Ph.D. computer science in the qualification of the computer from University of Technology, Baghdad- Iraq at 2006. M.Sc. in computer science (applied) from the National Computers Center NCC (currently ICCI) Baghdad-Iraq at 1987, and then B.Sc. in Informatics/ statistics from Al-Mustansiriyah University, Baghdad- Iraq at 1982.

Professor Ismaeel is coordinator and organizer of computer/IT center in Baquba Technical Institute - Technical Education Foundation TEF Baghdad-Iraq at 1990, founder and coordinator the dept. of Computer Systems. in Baquba Technical Institute- TEF Baghdad-Iraq at 2000, then founder and coordinator the department of Information Systems Engineering, Erbil Technical Engineering College - Erbil Polytechnic University, Iraq at 2007. He has 28-year experience in teaching an undergraduate and graduate in computer science, information systems, software engineering and fields related (IT, bioinformatics, biomedical Engineering/Informatics, etc) in many universities from 2007 till now in Kurdistan, Iraq. He is editorial, advisor, reviewer board member (one of them IJACSA of SAI Org:: http://thesai.org/Reviewers/Details/0a1f2c5d-6c63-4232-9fa2-d790812be480) and program committee member of many international journals and conferences. His research interest mobile network, cloud computing, semantic web, distributed system, healthcare systems bioinformatics & biomedical Eng. /Informatics. He has experiences and skills for Advising, Counseling, Teaching, Training, Industrial and Curricula Development using European/Germany standards. More details visit: http://drayadghanyismaeel.wix.com/ayad-ghany-ismaeel-

Dina Yousif Mikhail: obtained B.Sc. (Bachelor of engineering) Medical Instrumentation engineering at 2003/2004 in Technical Collage - University of Mosul-Iraq and Higher Diploma (Software Engineering) at 2009/2010 in Engineer Collage-University of Sallah Al Dien, Erbil-Iraq. She is M.Sc. researcher's student in Information System Engineering in Engineering Technical Collage - Erbil, Polytechnic University Erbil-Iraq. She is currently in the department of Information Systems Engineering, Technical Engineering College, Erbil Polytechnic University, IRAQ. Her research interests in web application, bioinformatics & biomedical Engineering, Artificial intelligent. She has 10-year experience in teaching an undergraduate in information systems engineering.

Firefly Algorithm for Adaptive Emergency Evacuation Center Management

Yuhanis Yusof, Nor Laily Hashim, Noraziah ChePa, Azham Hussain
School of Computing
Universiti Utara Malaysia
Sintok, MALAYSIA

***Abstract*—Flood disaster is among the most devastating natural disasters in the world, claiming more lives and causing property damage. The pattern of floods across all continents has been changing, becoming more frequent, intense and unpredictable for local communities. Due to unforeseen scenarios, some evacuation centers that host the flood victims may also be drowned. Hence, prime decision making is required to relocate the victims and resources to a safer center. This study proposes a Firefly Algorithm (FA) to be employed in an emergency evacuation center management. Experimental analysis of a minimization problem was performed to compare the solutions produced by FA and the ones generated using Tabu Search. Results show that the proposed FA produced solutions with smaller utility value, hence indicating that it is better than the benchmark method.**

Keywords*—Firefly Algorithm; Swarm Intelligence; Flood Management; Evacuation Center Management

I. INTRODUCTION

Disaster management is extremely important in today's world and it focuses on the organization and management of resources and responsibilities for dealing with all humanitarian aid. Despite all activities accomplished by governments in the disaster preparation stage, flood occurs and affected people's daily routines and the economic flow since offices, businesses and schools are closed. In the past decade, Malaysia has experienced a number of major floods. Floods are caused by a combination of natural and human factors. Malaysians are historically riverside people as early settlements grew on the banks of the major rivers in the peninsula. Coupled with natural factors such as heavy monsoon rainfall, intense convection rain storms, poor drainage and other local factors, floods have become a common feature in the lives of a significant number of Malaysians. The vast increasing numbers of the lost due to flood enforces the government to take proactive steps such as setting up supervisory bodies, implementing flood mitigation programmes, implementing non-structural steps with the setting up of flood forecasting and warning systems for the flood prone area. The evacuation and relocation of flood victims involves a lot of capital. As informed by the Minister in the Prime Minister Department in March 2011, almost USD 21.12 million was spend in for 89,000 flood victims in five states effected by the disaster and it was estimated that 53 percent of that amount was spent on relocation of the victims, which also includes food and other daily necessities. At the moment, there are 5,143 evacuation centres (EC) which could accommodate 1.4 million flood victims around the country.

Flood evacuation centres in Malaysia are managed by the Department of Social Welfare (*Jabatan Kebajikan Masyarakat*) and works closely with a number of governmental and non-governmental agencies to provide necessary steps to ensure safety and comfort in every evacuation centre. These individuals are the backbone of the flood evacuation centre and often face difficulty in decision making such as for resource allocation. Various work can be seen in resource allocation pertaining to disaster management such as in flood. In the work by Zhu, Huang, Liu and Han [1] , the researchers propose a resource allocation model that is aimed at determining the location of reserve depots and the amount and type of resources to be stored. It is modelled based on discrete scenarios that is divided into two; local government and national. Their optimization focuses on the commodities inventory holding and transportation cost. On the other hand, a more recent work [2] was discussed in that identifies the optimal number, location and inventory level of warehouses around the world in the occurrence of a disaster. The model considers uncertainties on product quality, availability and production capacity in affected areas.

This study proposes the employment of a swarm intelligence algorithm (i.e Firefly Algorithm) in the Adaptive Emergency Evacuation Center Management (AEECM) that monitors and manage evacuation centers. Similar to existing work on disaster management, the AEECM focuses on resource allocation. However, the study reported in this article is limited to the management of victims located in ECs. The proposed AEECM adapts a recent computing approach known as Swarm Intelligence. Swarm Intelligence is defined as an emergent collective intelligence of groups of simple agents [3]. It is used to find optimal solutions in hard problems, such as Travelling Salesman (TSP) [4], scheduling [5] and nurse rostering [6]. Examples of algorithms that are considered as Swarm Intelligence are the Ant Colony Optimization, Artificial Bee Colony, Fish School, Bat Algorithm and Firefly Algorithm (FA). In this study, a variant of Firefly Algorithm that provides optimal solution to the management of victims in an evacuation center is presented. In particular, the proposed FA determines to where (i.e which available centers) victims in an evacuation center should be re-located and how many of them should be moved to each identified center.

II. Related Work

Meta-heuristic algorithms are defined as optimization algorithms that determines the best solution (optimal or near optimal solution) from a set of available solutions [7]. The identification of best solution is achieved by evaluating a predefined objective function that can be addressed as either minimum or maximum function. The design of an objective function is based on the problem in-hand, that is if the goal of the problem is to obtain minimum cost then a minimum objective function is designed [8] and vice versa.

Existing meta-heuristic algorithms can be categorised into two types as shown in Figure 1; single meta-heuristic and population meta-heuristic [9]. For the first category, (i.e single meta-heuristic), the algorithm generates a single solution and iteratively enhance it. An example of such algorithm is the Tabu Search (TS) which was introduced by Glover in 1986 [10]. On the other hand, the population meta-heuristic algorithms generate a set of solutions and select one of it as the best solution. The swarm algorithms such as Firefly Algorithm is an example of meta heuristic algorithm [3, 11].

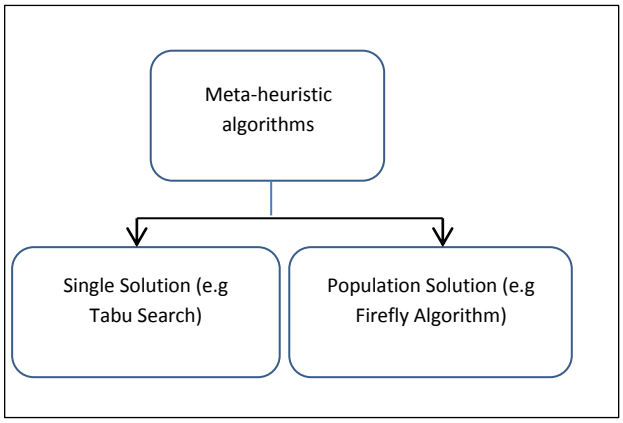


Fig. 1. Categories of Meta-heuristics Algorithm

A. Tabu Search

Tabu Search (TS) generates a single solution by extending the search space beyond local optimality to identify better solution [12, 13]. TS classify a subset of the moves in a neighbourhood as tabu. A neighbourhood constructs to reach adjacent solution from a current solution. The main idea in TS is to avoid recently visited solution space areas and move towards promising area [14]. TS has been adapted into various optimization problems (Glover & Laguna, 2013) such as colour texture histogram [13], scheduling [5], test data software generation [12], cell formation [14], nurse rostering [6], graph colour [15, 16], assignment [16] and max-cut problem [17].

In TS, there are two important factors; tabu moves and tabu condition. The first factor is determined by a function that utilizes information from the search process, while, the second factor is a linear inequality or logical relationships that is used to choose the tabu moves [18]. Figure 2 details the pseudo code of the Tabu Search algorithm [18]. In Step 1, a random solution is selected and assigned as the best solution. A new subset of solutions will be generated based on the identified best solution. Comparison between solutions is performed in order to identify the best one.

Step1: Select $X \in S$ and set $X^* = X$.
Set $K = 0$. Set $T = \emptyset$
Step2: Set $K = K + 1$
Generate a subset of solutions in the neighborhood *N-T* of *X*.
Step3: Choose the best solution s from this neighborhood and set $X = S$.
Step4: if $f(X) = f(X^*)$
Then $X^* = X$.
Step5: if termination criteria
Then stop
Else
Update T.
Go to step 2.
end

Fig. 2. Pseudo Code of Tabu Search [18]

B. Firefly Algorithm

Firefly algorithm (FA) is an example of algorithm that is based on nature inspired computing. It has the ability to identify global optimal solution [19]. The main concept of Firefly algorithm is realized in two factors; light intensity and attractiveness between fireflies. The light intensity of a firefly is more related with the objective function, *f(x),* and can be a maximization or minimization function. On the other hand, the attractiveness, β, between fireflies is associated with the distance between two fireflies, where β is based on the change of distance. Figure 3 shows the steps in Firefly Algorithm [11, 20].

Step1: Generate Initial population of firefly randomly xi (i=1, 2,.., n),Light Intensity, I, at x_i is determined by Objective function f(xi).
Step2: Define light absorption coefficient, γ.
Step3: While (t < Max Generation)
Step4: For i=1 to N (N all fireflies)
Step5: For j=1 to N
Step6: If (Ii<Ij) { $X^i = X^i + \beta_0 exp^{(-\gamma r_{ij}{}^2)} * (X^j - X^i) + \alpha\varepsilon_i$ }
Step7: $\beta = \beta_0 exp^{(-\gamma r_{ij}{}^2)}$
Step8: Evaluate new solutions and update light intensity.
Step9: End For i
Step10: End For j
Step11: Rank the fireflies and find the current global best g*.
Step12: End While

Fig. 3. Pseudo Code of Firefly Algorithm [11, 20]

Firefly Algorithm has been implemented in many optimization problems such as image segmentation [21], traffic forecasting[22], discrete optimization [23], data classification [24], data clustering [25], text clustering [26] and economic dispatch problems [27]. In all of these domains, Firefly Algorithm has proven to be successful in solving the problems and identifying the optimal solution.

III. Methods

The proposed work is realized by performing 3 phases; data collection, design of algorithms and evaluation.

A. Data Collection

The obtained data on evacuation centers is represented as four independent variables as depicted in Table 1. It covers information on the 106 ECs in one of the district in Malaysia.

TABLE I. INDEPENDENT VARIABLES

Data	Parameter
Size of EC	V1
Distance of EC to the closed EC	V2
Water level of nearby river	V3
Distance of EC to nearby river	V4

B. Design of Algorithms

In this section, elaboration on the adaptation of Tabu Search and Firefly Algorithm in determining number of victims to be evacuated is presented. In the Adaptive Emergency Evacuation Center Management, the two algorithms are triggered when a decision on closing a particular EC is made. Assuming that the chosen EC has *n* number of victims, the proposed Tabu Search and Firefly Algorithm provides suggestion on the location (i.e EC) to where the victims should be relocated. Furthermore, the suggestion also includes information on the number of relevant victims. In the AEECM, the proposed Tabu Search is termed as TS_{Flood} while the variant of FA is known as FA_{Flood}. Both of these algorithms employ an objective function as depicted in Eq. 1

Utility function F= Summation of (75% of V1) of available EC + Summation of V2 of available EC + Summation of V3 of available EC – Summation of V4 of available EC

(Eq. 1)

The proposed objective function is of minimum problem as most of the included parameters V1, V2, and V3 prefer small values. For example, an EC with a smaller distance to the closed EC is preferred compared to the EC that has farther distance. On the other hand, the fourth parameter which is V4 is of maximum value as the system needs to avoid EC that is near to a river. In addition, the first variable includes the constraint of 75% usage as we need to ensure that there isn't any EC that is 100% occupied for safety and convenience purposes.

Tabu Search for Optimal Evacuees Management

In Figure 4, pseudo code of the proposed Tabu Search is presented. The TS_{Flood} starts by randomly generating an initial solution, *X*, and denote the solution as the best solution where *X*=X*. The solution, *X*, is represented in binary form, where the length of the representation is based on the number of evacuation centers. Each bit in this solution represents one EC; if the value is 0 it means that the EC is not chosen and if it is 1 it shows that EC is selected. The best solution will then undergo an evaluation using the objective function.

In Step 6 of the TSFlood, if the termination criteria (a termination criterion is based on number of iteration) is reached then the process is stopped, else continues by generating a subset of solutions based on the best solution (as shown in step 7). The solutions are produced by adding or deleting item from the best solution. The number of addition is determined by a random number, while the number of for deletion equals the difference between the number of sub solutions and the addition.

The generated sub solutions are later evaluated using objective function (utility function in equation 1) and the best solution is identified. The aim of the proposed TS_{Flood} is to identify a solution that the smallest utility function value.

Input:
Step 1: Input the dataset that includes four variables [v1, v2, v3, v4].
Step 2: Determine the max capacity for EC.

Process:
Step3: Generate random start solution X, and set the best solution X*=X,
Step4: Evaluate the best solution using objective function (Equation 1)
Step5: Check the generated solution, it must fulfill the constraint of summation of (75% of v1) of all record in one solution exceed the capacity value. If fail, the generated solution will undergo refinement process (one item will be added randomly to the solution).
Step6: If termination criteria is fulfil, then stop (termination criterion is based on number of iteration.
Step7: Generate a subset of solutions based on best solution by adding or deleting item from best solutions. The number of adding process is determined by random number, while the number of deleting process equals the difference between the number of sub solutions and the number of adding process.
Step8: Evaluate the sub solutions using objective function as shown in equation 1 and check the generated solution as in Step3.
Step9: Choose the best solution from this sub solutions
if f(x) < f(X*) then X*= X
Step 10: End

Fig. 4. Pseudo code of proposed Tabu Search algorithm (TS_{Flood})

Firefly Algorithm for Optimal Evacuees Management

The pseudo code of proposed Firefly algorithm (FA_{Flood}) in decision making for AEECM is illustrated in Figure 5. Further, in the input of the proposed FA_{Flood} needs to define some important parameters for operating the algorithm such as the light absorption coefficient γ, where it set to 1 in the algorithm, the value of initial attractiveness β0, where it sets to 1, the number of max generation, the value of capacity which equal to value of the number of vectims that need to be in safe places, the number of fireflies which equal to 10% from the number of records in dataset, and finally is the number of initial solution which is equal to the number of fireflies.

The proposed FA_{Flood} starts to operate by generating initial solutions which are represented in binary form [5], where the dimension of one solution equals the number of records in dataset (i.e evacuation centers). Generation of the solutions are performed using two ways: If variable v1 of a record is higher than the capacity, it will take it as one solution by assigning 1 in the solution, else, it randomly generate a solution that has 2 or more records. These solutions then undergo a verification

process that checks if the summation of v1 for all records in one solution exceed the capacity value. After that, the generated initial solution ned to be evaluated based on objective function (Utility function) as shown in equation 1. Assign the inverse of utility value (fitness) of each solutions to each firefly as initial light (I). Then, the initial position of each firefly are determined, which is represented the solution.

Fireflies compete between them to determine the best solution that has the highest fitness value. Firefly with brighter light attracts the less bright ones and this is based on the distance between two solutions using Hamman distance as shown in equation 2. Then, the attractiveness between two solutions using equation 3 is calculated. The less bright firefly will move to the brighter one using equation (4).

$$r_{ij} = \frac{\textbf{the number of dissimilar bit in i and j solutions}}{\textbf{the number of records in dataset}} \quad (2)$$

$$\boldsymbol{\beta = \beta_0 \exp^{(-\gamma r_{ij}^2)}} \quad (3)$$

$$X^i = X^i + \beta_0 \exp^{(-\gamma r_{ij}^2)} * (X^j - X^i) + \text{rand} - 0.5 \quad (4)$$

After moving, a new solution is generated, If the summation of v1 in new solution greater or equal to capacity, then, the utility function and fitness for new solution are calculated and compared with old solution, if a new solution is better than old solution then replace it. In the situation where the summation of v1 in a new solution is smaller than the capacity, a mutation process is conducted on the new solution by adding one bit randomly until pass the capacity value. Then the utility function of the new solution is calculated and compared against the old solution in order to identify the best solution. Once the predefined number of iteration is reached, the fireflies are sorted based on their brightness that indicates the utility value.

C. Evaluation

The effectiveness of TS_{Flood} and FA_{Flood} is evaluated based on two criteria; utility value and computational time. The AEECM prefers the method that produces solution with the lowest utility value and computational time. In addition, two scenarios were employed; the first scenario investigates solutions for a to-be closed EC with number of victims that is larger than the capacity of any available EC. Meaning that the solution is expected to consist a combination of ECs. On the other hand, the second scenario represents situations where the to-be closed EC has the same or less number of victims as the EC.

Input:
Step 1: Input the dataset that includes four variables [v1, v2, v3, v4].
Step 2: Define light absorption coefficient γ, where γ=1.0
Step 3: Define initial attractiveness $\beta_0 = 1$
Step 4: Determine the Max Generation.
Step 5: Determined the Capacity.
Step 6: Determined the number of fireflies which equal 10% from the number of records in dataset.
Step 7: Determine the number of initial solution which is equal the number of fireflies.
Process:
Step 8: The solutions are represented in binary form (0,1), where the dimension of one solution equal the number of records in dataset.
Step 9: Generated the solutions are undertaken in two ways:
Step 9.1: If the variable v1 in one record pass the capacity, it will take it as one solution by assigning 1 in solution.
Step 9.2: If the v1 in one record less than the capacity, then, the solution will generated randomly that take more than one record in one solution.
Step 9.3: Check the generated solution must pass the constrain which is the summation of v1 of all record in one solution pass the capacity value.
Step 10: Calculated the initial utility function for each solution as in Eq 1

Step 11: Assign the utility value to each firefly as initial light (I)….meaning that fitness = 1/utility
Step 12: Determine position of each firefly, which is the initial solution.
Step 13: While (t < Max Generation): Max_Generation = No of Firefly = 0.1 * number of EC
Step 14: For i=1 to N (N all fireflies)
Step 15: For j=1 to N
Step 16: If fitness_i < fitness_j (Ii<Ij) {
Step 17: Calculated the distance between two solutions using Hamman distance the following equation:
$$r_{ij} = \frac{\textbf{the number of dissimilar bit in i and j solutions}}{\textbf{the number of records in dataset}} \quad (2)$$
Step 18: Calculated the attractiveness between two solutions using the following equation:
$$\beta = \beta_0 \exp^{(-\gamma r_{ij}^2)} \quad (3)$$
Step 19: Move less brighter firefly to high brighter firefly
$$X^i = X^i + \beta_0 \exp^{(-\gamma r_{ij}^2)} * (X^j - X^i) + \text{rand} - 0.5 \quad (4)$$
Step 20: If summation of v1 in new solution >= Capacity
Step 20.1: Calculated the utility function and fitness for new solution.
Step 20.2: Compare with old solution.
Step 20.3: Replace old solution with new solution.
Step 21: Else if summation of v1 in new solution < Capacity
Does mutation for one bit random in new solution until pass Capacity.
Step 21.1: Calculated the utility function for new solution.
Step 21.2: Compare with old solution.
Step 21.3: Replace old solution with new solution.
Step 22. End For i
Step 23. End For j
Step 24. End While
Step 25. Rank the fireflies and find the current global best utility function, and best solution.
Output
Step 26. Sort the best solution based on v1.

Fig. 5. Pseudo code of proposed Firefly Algorithm (FAFlood)

IV. RESULTS

In Table 2, result for the first scenario is presented while Table 3 depicts results for the second scenario. The proposed algorithms were executed for three times. For the first scenario, assuming the closed EC is ECID_76 and the number of victims in this EC is (298) greater than the capacity of any EC. As shown in Table 2, the utility values of FA_{Flood} are (5635.15, 5235.45, and 4068.59) which are smaller than TS_{Flood} which are (5754.95, 5545.95, and 4521.59). However, the execution time for TS_{Flood} (305, 368, and 295) is better than FA_{Flood} which took 676, 641, and 641ms. The graphical illustration is provided in Figure 7.

In Table 3, the closed EC is ECID_35 and the number of people in this EC is (221) smaller or equal input EC. As can observe in Table2, the utility values of proposed FA_{Flood} are (260.1, 265 and 276) for three executions better than (minimum is better) TS_{Flood} which are (260.1, 288, and 276), (260.1, 283.1, and 279.05) and (260.1, 265, and 279.05) as shown in Figure 7.c, however, the time execution of proposed TS_{Flood} algorithm is (277, 282 and 266) better (smaller is better) than proposed FA_{Flood} algorithm which are (594) in three executions. The graphical representation is as shown in Figure 7.d.

V. CONCLUSION

Over the past decades, the pattern of floods across all continents has been changing, becoming more frequent, intense and unpredictable for local communities. Even though various strategies have been implemented in managing flood evacuation centers, unforeseen scenarios may lead to the closure of the center. Hence, in this study, a variant of Firefly Algorithm (FA) that provides optimal solution to the management of victims in an evacuation center is presented. In particular, the proposed FA determines the number of victims to be relocated to a particular center. Evaluation of the proposed FA is undertaken by comparing its result (i.e utility function and computational time) against the one produced by Tabu Search which is also an example of a meta heuristics algorithm. Two types of experiments were performed; number of victims in a to-be closed EC is larger and, smaller or equals the capacity of any available EC. Results of the first experiment show that the average value of utility function produced by FA solutions (i.e 4980) is smaller than the one obtained by Tabu Search (i.e 5274). Similar pattern can also be seen in the utility function values of the second experiment. In this study, as the problem is formulated as a minimization function, FA that produces a smaller utility function is preferred. Nevertheless, in both experiments, it is noted that FA consumes larger computational time compared to Tabu Search. FA requires at least 594ms to produce a solution while Tabu Search only took 266ms.

ACKNOWLEDGMENT

Authors would like to thank the Malaysian Ministry of Higher Education for the financial support given under the Fundamental Research Grant Scheme (S/0 Code 13183).

TABLE II. RESULTS OF SCENARIO 1

Center to-be closed	Victims	FA_{Flood}			TS_{Flood}		
		Solution	Utility value	Time (ms)	Solution	Utility value	Time (ms)
ECID_76	298	ECID_38 = 255 ECID_34 = 43	5635.15	676	ECID_38 = 255 ECID_80 = 43	5754.95	305
		ECID_34 = 245 ECID_64 = 53			ECID_80 = 253 ECID_70 = 45		
		ECID_64 = 242 ECID_60 = 56			ECID_70 = 249 ECID_90 = 49		
ECID_76	298	ECID_70 = 249 ECID_34 = 49	5235.45	641	ECID_70 = 249 ECID_90 = 49	5545.95	368
		ECID_34 = 245 ECID_64 = 53			ECID_90 = 203 ECID_125 = 95		
		ECID_64 = 242 ECID_121 = 56			ECID_125 = 197 ECID_103 = 101		
ECID_76	298	ECID_88 = 255 ECID_70 = 43	4068.59	641	ECID_37 = 189 ECID_46 = 109	4521.59	295
		ECID_70 = 249 ECID_34 = 49			ECID_46 = 188 ECID_57 = 110		
		ECID_34 = 245 ECID_64 = 53			ECID_57 = 179 ECID_105 = 119		

TABLE III. RESULTS OF SCENARIO 2

Center to-be closed	Victims	FAFlood			TSFlood		
		Solution	Utility value	Time (ms)	Ssolution	Utility value	Time
ECID_35	221	ECID_34 = 221	260.1	594	ECID_88 = 221	260.1	277
		ECID_38 = 221	265.0		ECID_70 = 221	288.0	
		ECID_64 = 221	276.0		ECID_34 = 221	276.0	
ECID_35	221	ECID_34 = 221	260.1	594	ECID_88 = 221	260.1	282
		ECID_38 = 221	265.0		ECID_70 = 221	283.1	
		ECID_64 = 221	276.0		ECID_80 = 221	279.05	
ECID_35	221	ECID_34 = 221	260.1	594	ECID_88 = 221	260.1	266
		ECID_38 = 221	265.0		ECID_121 = 221	265.0	
		ECID_64 = 221	276.0		ECID_80 = 221	279.05	

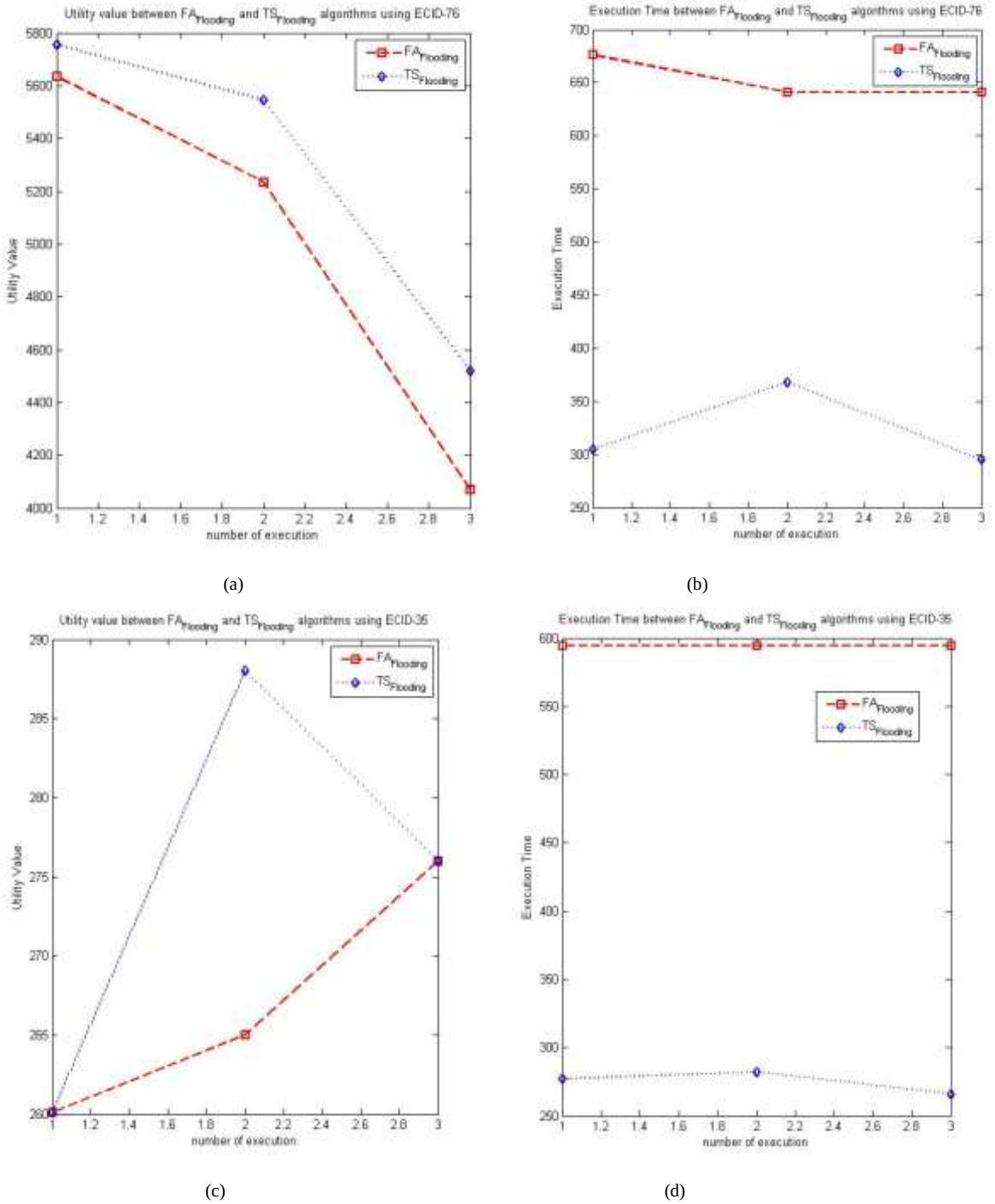


Fig. 6. A graphical representations of utility function and execution time results: FA_{Flood} vs. TS_{Flood}. (a) Utility for ECID-76 (b) Execution time for ECID-76 (c) Utility for ECID-35 (d) Execution time for ECID-35

REFERENCES

[1] J. Zhu, J. Huang, D. Liu, and J. Han, "Resources Allocation Problem for Local Reserve Depots in Disaster Management Based on Scenario Analysis," presented at 7th International Symposium on Operations Research and Its Applications (ISORA'08), Lijiang, China, 2008.

[2] S. Duran, M. A. Gutierrez, and P. Keskinocak, "Pre-Positioning of Emergency Items for CARE," Interfaces, vol. 41, pp. 223-237, 2011.

[3] E. Bonabeau, G. Theraulaz, and M. Dorigo, Swarm Intelligence: From Natural to Artificial Systems: Oxford University Press, 1999.

[4] M. Dorigo and L. M. Gambardella, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem," IEEE Tranasactions on Evolutionary Computation, vol. 1, pp. 53-66, 1997.

[5] A. C. Adamuthe and R. S. Bichkar, "Tabu search for solving personnel scheduling problem," presented at nternational Conference on Communication, Information & Computing Technology (ICCICT), 2012.

[6] S. N. Vu, M. H. N. Nguyen, L. M. Duc, C. Baril, V. Gascon, and T. B. Dinh, "Iterated local search in nurse rostering problem," in Proceedings

of the Fourth Symposium on Information and Communication Technology. Danang, Vietnam: ACM, 2013, pp. 71-80.

[7] S. Das, A. Abraham, and A. Konar, Metaheuristic Clustering, vol. 178: Springer-Verlag Berlin Heidelberg, 2009.

[8] F. Rothlauf, Design of Modern Heuristics: Principles and Application: Springer Publishing Company, Incorporated, 2011.

[9] I. BoussaiD, J. Lepagnot, and P. Siarry, "A survey on optimization metaheuristics," Information Sciences, vol. 237, pp. 82-117, 2013.

[10] F. Glover, "Future paths for integer programming and links to artificial intelligence," Computational Operation Research, vol. 13, pp. 533-549, 1986.

[11] X.-S. Yang, Nature-Inspired Metaheuristic Algorithms: Second Edition: Luniver Press, 2010.

[12] A. Rathore, A. Bohara, R. G. Prashil, T. S. L. Prashanth, and P. R. Srivastava, "Application of genetic algorithm and tabu search in software testing," in Proceedings of the Fourth Annual ACM Bangalore Conference. Bangalore, India: ACM, 2011, pp. 1-4.

[13] B. K. Koora, N. R. Satpute, and A. Adiga, "Tabu search based implementation of object tracking using joint color texture histogram," presented at 2012 IEEE 7th International Conference on Industrial and Information Systems (ICIIS), 2012.

[14] T. Dinh, T. Dinh, and J. Ferland, "A meta-heuristic approach for cell formation problem," in Proceedings of the Second Symposium on Information and Communication Technology. Hanoi, Vietnam: ACM, 2011, pp. 11-18.

[15] D. Chalupa, "Population-based and learning-based metaheuristic algorithms for the graph coloring problem," in Proceedings of the 13th annual conference on Genetic and evolutionary computation. Dublin, Ireland: ACM, 2011, pp. 465-472.

[16] J. Xie, Y. Mei, and A. Song, "Evolving Self-Adaptive Tabu Search Algorithm for Storage Location Assignment Problems," in Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation. Madrid, Spain: ACM, 2015, pp. 779-780.

[17] E. Arraiz and O. Olivo, "Competitive simulated annealing and Tabu Search algorithms for the max-cut problem," in Proceedings of the 11th Annual conference on Genetic and evolutionary computation. Montreal, Quebec, Canada: ACM, 2009, pp. 1797-1798.

[18] C. Grosan and A. Abraham, Intelligent Systems: A Modern Approach, vol. 17: Springer-Verlag Berlin Heidelberg, 2011.

[19] R. Tang, S. Fong, X. S. Yang, and S. Deb, "Integrating nature-inspired optimization algorithms to K-means clustering," presented at International Conference on Digital Information Management (ICDIM), 2012.

[20] X. S. Yang, "Firefly algorithm, stochastic test functions and design optimisation," Int. J. Bio-Inspired Comput., vol. 2, pp. 78-84, 2010.

[21] M.-H. Horng and T.-W. Jiang, "Multilevel Image Thresholding Selection Based on the Firefly Algorithm," presented at 7th International Conference on Autonomic and Trusted Computing (UIC/ATC), 2010.

[22] Y. Yusof, F. K. Ahmad, S. S. Kamaruddin, M. H. Omar, and A. J. Mohamed, "Short Term Traffic Forecasting Based on Hybrid of Firefly Algorithm and Least Squares Support Vector Machine," presented at Proceedings of First International Conference on Soft Computing in Data Science, Putrajaya, 2015.

[23] M. K. Sayadi, A. Hafezalkotobb, and S. G. J. Nainia, "Firefly-inspired algorithm for discrete optimization problems: An application to manufacturing cell formation," Journal of Manufacturing Systems, vol. 32, pp. 78-84, 2013.

[24] N. Sudarshan, S. P. Pratim, and D. Achintya, "Analysis of a Nature Inspired Firefly Algorithm based Back-propagation Neural Network Training," International Journal of Computer Applications, vol. 43, pp. 8-16, 2012.

[25] A. J. Mohammed, Y. Yusof, and H. Husni, "Document Clustering Based on Firefly Algorithm," Journal of Computer Science, vol. 11, pp. 453-465, 2015.

[26] A. J. Mohammed, Y. Yusof, and H. Husni, "GF-CLUST: A Nature Inpsired Algorithm for Automatic Text Clustering," Journal of Information and Communication Technology (JICT), vol. 15, pp. 57-81, 2016.

[27] X.-S. Yang, S. S. S. Hosseini, and A. H. Gandomi, "Firefly Algorithm for solving non-convex economic dispatch problems with valve loading effect," Appl. Soft Comput., vol. 12, pp. 1180-1186, 2012.

A Conversion of Empirical MOS Transistor Model Extracted from 180 nm Technology to EKV3.0 Model using MATLAB

Amine AYED, Mongi LAHIANI, Hamadi GHARIANI
LETI Labortory-ENIS
Sfax, Tunisia

***Abstract*—In this paper, the EKV3.0 model used for RF analog designs was validated in all-inversion regions under bias conditions and geometrical effects. A conversion of empirical data of 180nm CMOS process to EKV model was proposed. A MATLAB developed algorithm for parameter extraction was set up to evaluate the basic EKV model parameters. Respecting the substrate, and as long as the source and drain voltages remain constant, the DC currents and g_m/I_D real transistors ratio can be reconstructed by means of the EKV model with acceptable accuracy even with short channel devices. The results verify that the model takes into account the second order effects such as DIBL and CLM. The sizing of the elementary amplifier was considered in the studied example. The sizing procedure based on g_m/I_D methodology was described considering a semi-empirical model and an EKV model. The two gave close results.**

***Keywords*—EKV model; g_m/I_D methodology; analog design; MATLAB**

I. Introduction

Developing high performance low-voltage analog circuits is required for implantable biomedical devices and portative systems. Several attempts have been made proposing some MOS models and analog circuit design methodologies. The MOS transistors modelling for analog integrated circuit and RF design has to extremely precise to predict correctly the behaviour of a real transistor and cover all the transistor operation regions.

The PSP model [1] is surface-potential-based considered as the most recent advanced MOSFET model. It was even selected by Compact Model Council as the new industry standard MOSFET model aimed to replace the BSIM3/4 for the advanced CMOS designs. It includes all the essential effects in the state of the art MOS transistors from the effects of the reverse short-channel to the long channel degradation. Although the PSP is very instrumental for the understanding of the MOS transistors operation modes, it is not suited for a circuit design: The PSP model relies on an explicit formulation of the potential of surface according to the terminal voltage of the MOS device. The analog circuit setting equations according to a PSP model is difficult. This paper opted, therefore; for a PSP model rather than a dimensioning tool. Enz, Krumenacher and Vittoz [2] [3] as well as others [4] respectively suggested the EKV model and ACM models which were specially developed for this purpose. They were derived from the gradual channel approximation. As for [2] [5], they proposed more advanced versions considering short channel effects and mobility degradation.

The basic notions of the E.K.V3 model were reviewed in this paper. In fact the charge-based compact EKV3 MOSFET model is an analog/RF IC design tool. The first versions of this compact model used an empirical current-voltage relationship [6] to address the moderate inversion successfully. It was pioneer in adopting a substrate instead of source, and exploiting the symmetrical forward-reverse operation of MOS transistors [7]. A design methodology based on the level of inversion (or inversion coefficient, IC) was developed by [8]. The developed EKV3 model [9] included several other specificities for non-quasi static (NQS) operation [10], RF operation [11], NQS thermal noise [12] and handling of short-thermal noise [13]. Further details on EKV3 may be found in [14] [15].

The EKV model led to the development of a ratio-based design technique known the gm/ I_D based methodology intended for low-power analog circuits. In such circuits the moderate-inversion region is often applied as it allows a good compromise between speed and power consumption [5][16][17]. The g_m/I_D sizing methodology was first introduced in [18]. Since then, the concept has been referenced by many publications [19] [20].

The ACM model has also led to the development of the g_m/I_D based Methodology [5] [21]. According to the above mentioned models, the gm/I_D based methodology using the characteristic of g_m/I_D as a function of the normalized current diagram is very useful from the point of view power and speed for the analog circuit design.

II. Motivation and Organisation of the Work

The success of RF design depends heavily on transistor modeling. This requires efficient and compact models for the active and passive circuit elements. Since the MOS transistor is the essential circuit element, great effort has been made to model its DC and AC behavior accurately. Furthermore designing a circuit for electronic systems with reduced power consumption is the ultimate purpose of any circuit designer. For this low power design, it is vital to use low voltage and low current circuits. This means that MOSFETs can operate in the weak or moderate inversion region in the low power circuit.

The motive behind this work was to develop an EKV 3.0 model for 180nm TSMC technology with few numbers of parameters which allow precise designs in all-inversion regions of MOS transistor. This modeling would provide flexibility and optimal sizing for analog RF designers using 180nm TSMC technology.

The aim of this compact model was to obtain simple, fast, and accurate representations of the device behavior. This paper tried to validate the EKV model according to PSP model and real transistor. Table lookup models called empirical models was implemented on *MATLAB* in the form of matrix containing device data for different bias points, were needed to evaluate real transistor. In this paper, a dimensioning of intrinsic gain stage based on g_m/I_d methodology using semi-empirical an EKV model was introduced.

The remainder of the paper was organized as follows. Section III presented the EKV formulation, comparative study with the PSP model was achieved. In Section IV, the validity of EKV model according to real transistor was reviewed. Section V presented the application of g_m/ I_D methodology on intrinsic gain stage dimensioning. Conclusions were drawn in section VI.

III. THE EKV3.0 MODEL

A. Presentation and formulations

The compact EKV 3.0 model was designed to simplify the MOS transistors dimensioning in advanced analog IC designs. It provides analytical, continuous, and physically correct description of weak, moderate and strong inversion including linear and saturation operation. The EKV3.0 MOS transistor has a hierarchical design, built through successive steps considering the major physical effects that may influence the transistor operation.

The EKV model Formulations rely on three basic parameters: The slope factor n, the specific current I_S and the threshold voltage V_{T0}. The latter is defined as the channel voltage for which the inversion charge becomes zero in the assumption of a strong inversion. The main equations constituting the model are given below.

The expression of the specific current is given by:

$$I_S = 2n(U_T)^2\mu C'_{ox}\frac{W}{L} = 2n(U_T)^2\beta \quad (1)$$

where the normalized drain current $i=I_D/I_S$.

The relation between the normalized drain current and the normalized mobile charge density and vice-versa is given by:

$i=q^2+q$ (2.1) $\quad q = 0.5\left(\sqrt{1+4i}-1\right)$ (2.2) (2)

The following expression relates the channel voltage V on the one hand and the normalized mobile charge density and the pinch-off voltage V_P on the other:

$$\frac{V_P - V}{U_T} = \left[2(q-1)+\log(q)\right] \quad (3)$$

Finally, the pinch-off voltage in EKV is computed as:

$$V_P = \frac{V_G - V_{T0}}{n} \quad (4)$$

where V_G and V_{T0} represent respectively the gate voltage and the threshold voltage.

Opposite to most MOSFET models, the EKV model made use the inherent symmetry of the MOSFET by referring all the terminal voltages to the substrate. Thanks to the device symmetry, the normalized drain current boils below to the difference between a forward component i_F and a reverse component i_R representing the drain current of saturated MOS transistors which source voltages are respectively V_S and V_D:

$$i=i_F-i_R \quad (5)$$

The graphical interpretation of EKV model presented by "Fig. 1" illustrates the drain current delivered by a saturated grounded source transistor whose parameters n, $V_{T0}$and I_S are considered respectively equal to 1.2, 0.4 V and 0.7A with three distinct values of gate voltage.

The corresponding pinch-off voltages predicted by "(4)" are marked by circles.

The $V_T(V)$ curves are plotted in a logarithmic scale proceeding by evaluating the non-equilibrium voltage V for every Vp by means "(3)".

The hatched areas identify $2nU_T^2i$ term that represent the drain currents divided by beta owing to the definition of I_S given by "(1)".

The gate voltage can be noticed to be large 0.6 V, the pinch-off voltage is positive, which is typical of a strong inversion. For $V_G<V_{T0}$ the pinch-off voltage V_P shifts left to become negative and the drain current decreases exponentially.

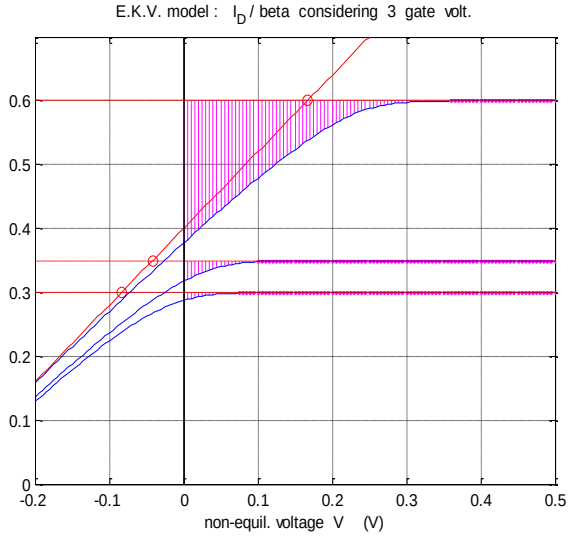


Fig. 1. Graphical illustration of drain current

B. Checking the EKV model against the PSP

In this part, the currents evaluated using the compact model were compared to the currents predicted by the PSP. First, the acquisition algorithm advocated in [3][5] has to be set up by MATLAB to extract n, I_S and V_{T0} from the PSP currents. Second, the currents by means of the E.K.V model have to be reconstructed and have to be compared to the new findings so as to check the validity of the new model.

Taking as a reference the original data, a unary N-type transistor having technological parameters issue from 0.18 µm CMOS process of TSMC technology was considered: An oxide thickness equal to 4.08nm, a substrate impurity concentration of 1.6 10^{17} cm-3 , and a V_{FB} =1 V. The temperature is 300°K.

Two distinct source voltages were selected; one for a weak inversion and the other for a strong one. The gate-to-substrate voltages from 0.6 to 1.8 V in steps 0.2 V was considered to be wide. After running the acquisition algorithm the value of unary specific current is I_{Su}=6.0476 10^{-007}A. The slope factor and the threshold voltage are 1.1227 and 0.0337, respectively.

"Fig.2" compares the reconstructed drain currents by means of the E.K.V model to the original PSP currents. The continuous lines represent the C.S.M. drain current and the circles stand for the strong and weak inversion. From the results illustrated in "Fig.2", the E.K.V compact model is remarked to be a good approximation of the PSP model.

In the following part, g_m/I_D ratios predicted by the compact model and the PSP were compared considering various back-bias voltages. An analytical expression of the g_m/I_D ratio in terms of the EKV compact model is given by [3]:

$$\frac{g_m}{I_D} = \frac{1}{nU_T}\frac{q}{i} = \frac{1}{nU_T}\frac{1}{q+1} \qquad (5)$$

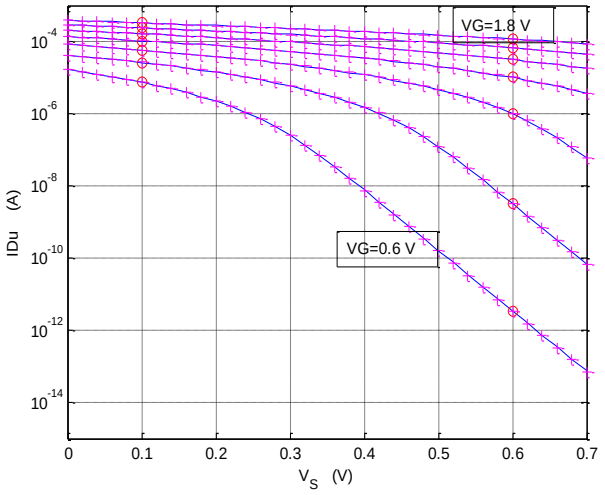


Fig. 2. Comparison between the reconstructed drain currents by Means of the E.K.V model and the original PSP currents

For the PSP, these were evaluated numerically by taking g_m/I_D the derivative of the log of the drain current. In "Fig. 3", the continuous lines represent the g_m/I_D ratios of the Charge Sheet Model. The crosses show the reconstruction based on the EKV model.

"Fig. 3" illustrates that the correspondence is satisfactory except for deep in weak inversion and low back-bias voltages. This might be due to the fact that the compact model does not consider the slight decrease of the subthreshold slope in a weak inversion.

The basic EKV model considered in the previous part was not suitable for real transistors, mobility degradation and short channel effects were ignored.

IV. The Real Transistor

In this part, we showed the impact of the gate length on the EKV model basic parameters in order to predict the drain currents and g_m/I_D ratios of real transistors. The only drawback was the introduction of look-up tables that contain a huge quantity of values extracted from the empirical model.

A. The inflence of the gate length on the model parameters

The gate length brings up the issue of some well-known effects, such as threshold voltage roll-off, reverse short channel effect, DIBL and CML.

Fig. 4" illustrates the impact of the gate length on the slope factors of N- and P-channel transistors, the threshold voltage and the specific current I_S.

Below 1µm, the threshold voltage starts to increase progressively at short gate lengths. The global increase, called the reverse short channel effect. In addition, "Fig. 4" illustrates that the specific currents increase slightly when the drain voltage increases. The effect is commonly designated by the acronym CLM for Channel Length Modulation.

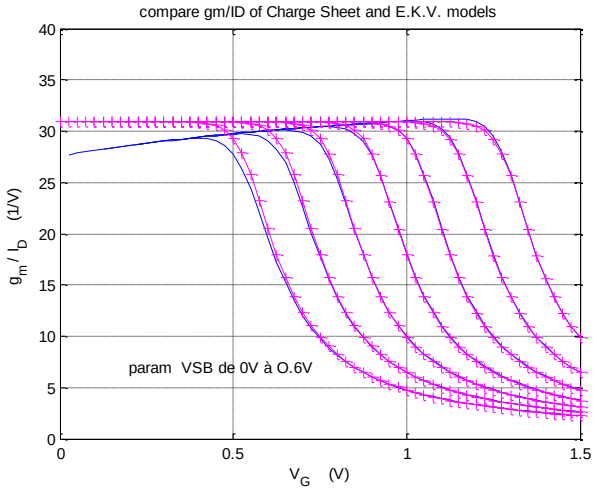
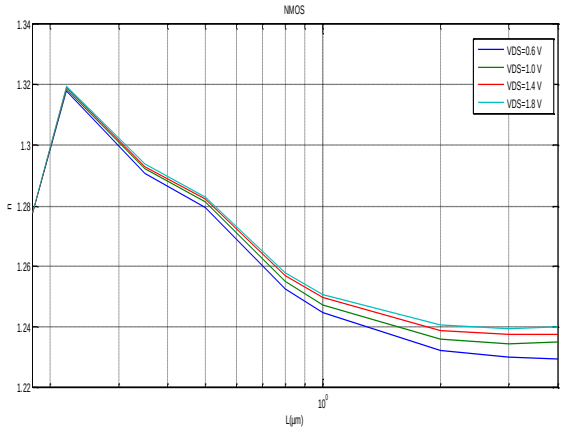
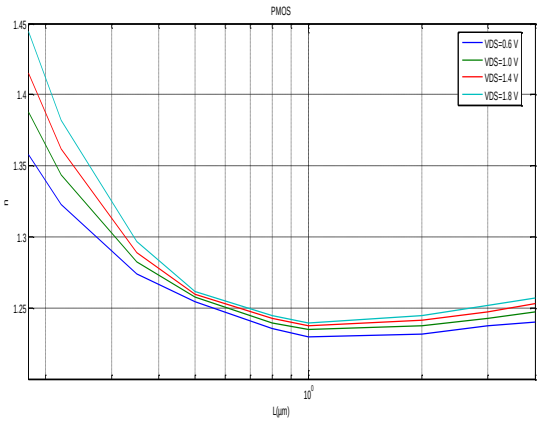


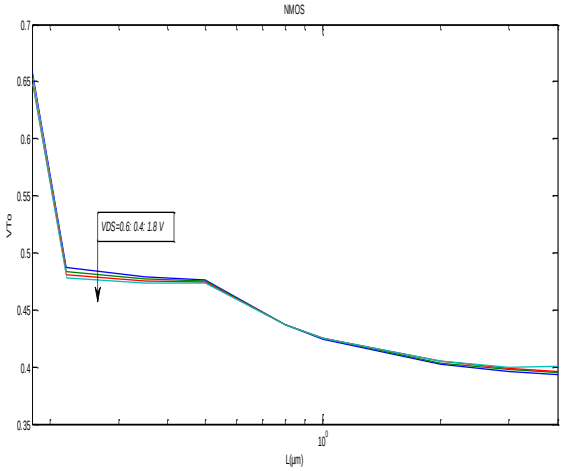
Fig. 3. Comparison between the reconstructed the gm/ID ratio by means of the EKV model to the original PSP considering various back-bias voltages



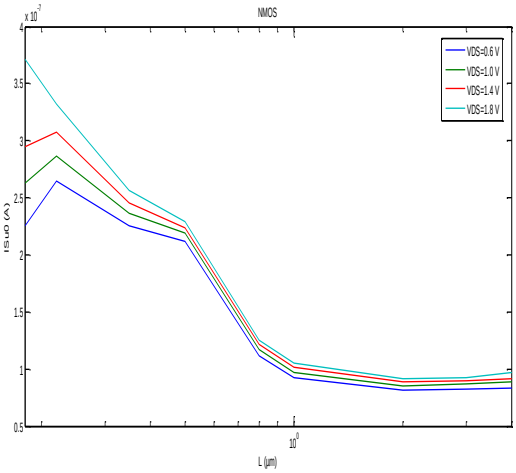
(a) n versus L for N-channel transistor



(b) n versus L for P-channel transistor



(c) V_{T0} versus L for NMOS transistor



(d) I_{SU0} versus L for NMOS transistor

Fig. 4. Plot of the slope factors of N- and P-channel transistors, threshold voltage and the specific current I_S versus the gate length considering four equally spaced drain voltage comprised between 0.6 and 1.8V

B. *Checking the validity of EKV model when its parameters vary with the source and drain voltages*

"Fig.5" displays a sample that shows the drain currents of a 10 µm wide N-channel MOS transistor whose drain-to-source voltage varies from 0.6 to 1.8 V, considering two gate lengths (0.18 µm and 1µm). The device belongs to a 180 nm technology developed by TSMC and consists of look-up tables listing the empirical data and implemented with MATLAB on an organized cell.

In this part, the EKV model was used to reconstruct I_D versus V_{DS} characteristic benefiting from the parameters that depend on the source and drain voltages including hort channel devices impact discussed previously.

Finally, the drain currents predicted by the model were compared to real I_{DS} (V_{GS}) characteristics. To this end, the identification algorithm presented by [5] is needed in order to extract the basic EKV parameters from empirical data achieved on real physical transistors.

"Fig 6" shows the reconstructed drain currents obtained by means of the EKV model.

The drain currents (dots) are compared to those of "Fig 5" (plain lines). As for the dashed lines, they relate to the model when the mobility is supposed to be invariant.

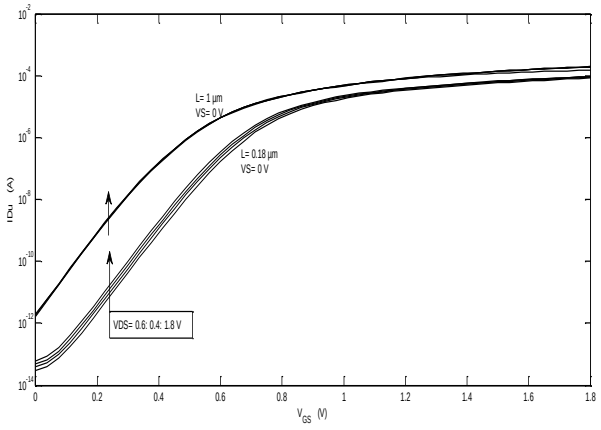


Fig. 5. Drain currents of an N-channel unary transistor. The device belongs to a 180 nm technology developed by TSMC

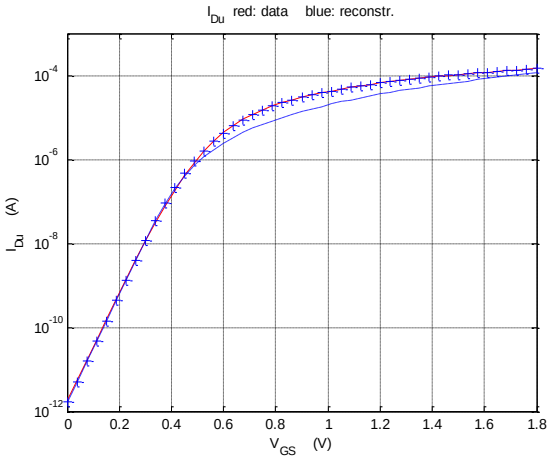


Fig. 6. Comparison between reconstructed drain currents (dots) by means of EKV model to the currents of "Fig. 5" (plain lines)

The assumption that the reconstructed currents agree fairly well with the physical currents is accepted implicitly. The model reproduces reasonably well real I_{DS} versus V_{GS} characteristics.

The extension of the E.K.V model to short channel devices considered in previous part lays down the foundation for the sizing of elementary amplifier.

V. Sizing the Elementary Amplifier

A. The elementary amplifier

The circuit of elementary amplifier called currently the 'Intrinsic Gain Stage' (IGS), shown in "Fig. 7", consists of a saturated common source transistor loaded by a capacitor.

We therefore consider the small signal equivalent circuit shown in "Fig. 8".

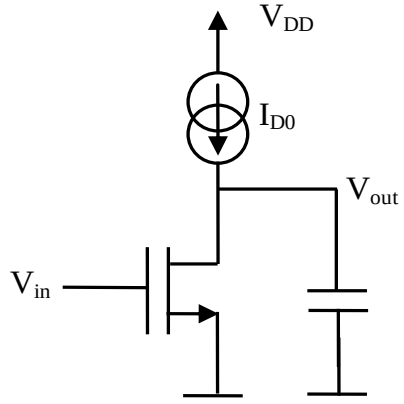


Fig. 7. Elementary amplifier

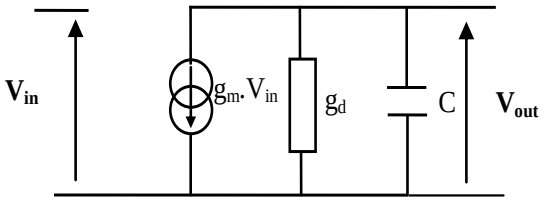


Fig. 8. Small signal equivalent circuit of elementary amplifier

In this section the sizing method of the elementary amplifier based on EKV model was reviewed by means of the g_m/I_D methodology. Our aim was to calculate gate width and drain current optimum values to control the circuit performance and achieve a desired gain-bandwidth product.

The DC gain is given by:

$$|A| = \frac{g_m}{g_d} = \frac{g_m}{I_D} \cdot \frac{I_D}{g_d} = \frac{g_m}{I_D} \cdot V_A \quad (6)$$

where V_A represent the early voltage

The relation between transconductance g_m and transition frequency f_T is given by:

$$g_m = 2\pi \cdot f_T \cdot C \quad (7)$$

The g_m/I_D methodology benefits from the variation of the transconductances and drain currents with the gate width where the key term g_m/I_D ratio is independent of the gate width and offers the possibility to achieve the transconductance derived from the expression below and deduce the gain bandwidth product.

The g_m/I_D ratio can be set up using two strategies. The first makes use of experimental $I_D(V_{GS})$ characteristics carried from measurements on real transistors. This is called the semi-empirical g_m/I_D sizing method. The other method refers to the analytical expressions for g_m/I_D founded in EKV model formulations.

Before applying the semi-empirical g_m/I_D method to size the elementary amplifier, let us look at the dependence of g_m/I_D on the gate-to-source and drain-to-source illustrated in "fig. 9".

For a desired transition frequency fixed at 100MHz, a MATALB computation is developed illustrating a contour plot of intrinsinc gain shown in "fig.10".

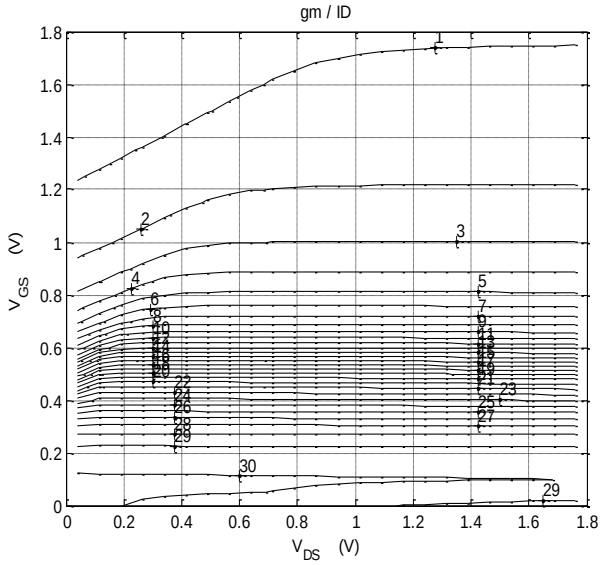


Fig. 9. gm/I_D contours versus drain an gate voltages for 0.18 µm gate length of NOMS transistor

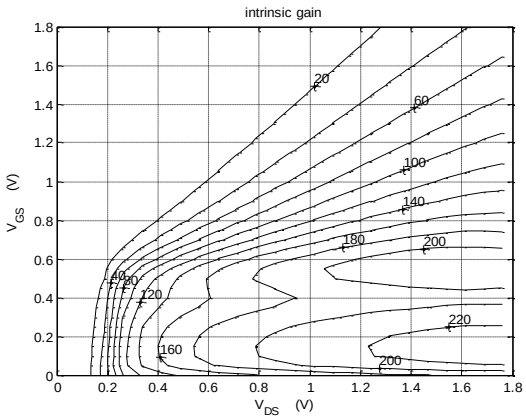


Fig. 10. Intrinsic gain contours versus drain and gate voltages

A series of gate widths, gate voltages V_{GS} and gains achieving the desired gain-bandwidth product is displayed in "Fig. 11" considering four drain voltages V from 0.25 to 1 V.

"Fig. 12" shows the impact of the gate length on the gate width, gate-to-source voltage and gain when L takes the following values 0.18, 0.5 and 0.22 µm.

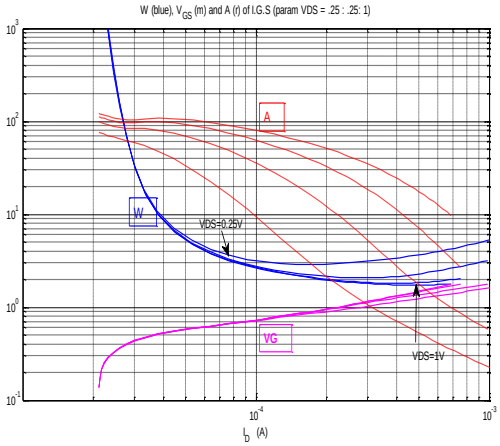


Fig. 11. Plot of the gate widths W, V_G(V) and gain A versus drain current for transistor frequency equal to 100 MHz

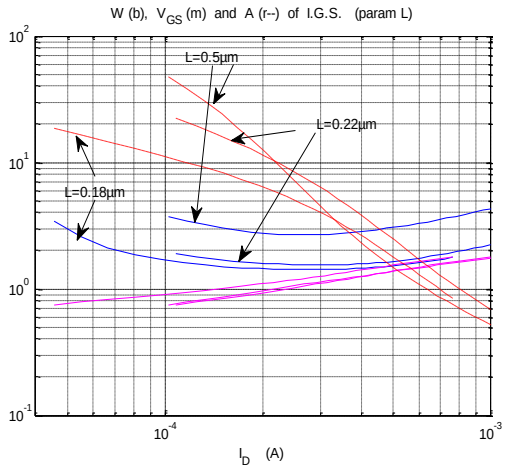


Fig. 12. Illustration of the influence of the gate length

B. The sizing procedure

In this subsection, sizing was undertaken considering the compact model instead of 'semi- empirical' data. It is divided into two parts: first, W and I_D were evaluated and then the Intrinsic Gain A was estimated.

Implemented on MATLAB, the sizing algorithm begins with the extraction of the model parameters from the empirical model. A q_F logspace vector that encompasses the moderate inversion region was then defined. This leads to the estimation and evaluation of the pinch-off voltage and the normalized reverse mobile charge density vector q_R. In a last step, the normalized drain current i was extracted.

In "Fig. 13" the width, gate-to-source voltage and intrinsic gain predicted by the model (continuous lines) are compared to their semi-empirical counterparts (crosses). The gain-bandwidth product is equal to 1 GHz and the output capacitor to 1 pF.

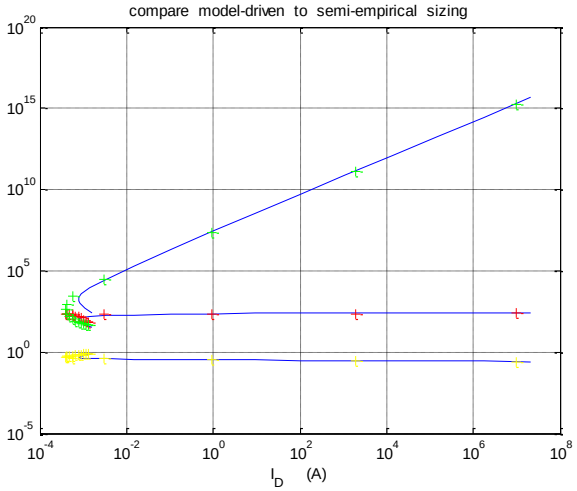


Fig. 13. Comaraison between W, A and V_{GS} predicted by EKV model (continuous lines) and its semi-empirical counterparts (crosses)

Considering the compact and the semi-empirical models, the sizing results are similar in most of the operation regions. Therefore, one of the important benefits of EKV model procedure is that the sizing can be achieved in well defined regions.

VI. Conclusion

This paper proved the consistency of the EKV3 model when describing the real transistor of 180 nm MOSFET technology. Extraction of EKV parameters algorithm was done using MATLAB. The results bring about a number of interesting observations highlighting the impact of the short channel effects on the parameters of the compact model. For a long and short channel transistor, the observed modeling result in weak, moderate and strong inversion cover qualitatively well all the aspects of the MOS transistor.The simplicity of the model has allowed us to reach a performing sizing of real Intrinsic Gain Stage. This underlines the suitability of the EKV3.0 model to be usefully used in analog circuit design for several applications such as OTA circuit and Ring VCO. Such a study is the topic of our future potential perspective.

REFERENCES

[1] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G.D.J. Smit, A.J. Scholten and D.B.M. Klaassen, "PSP: An Advanced Surface-Potential-Based MOSFET Model for Circuit Simulation", IEEE TED, Vol. 53, No. 9, pp. 1979-1993, September 2006.

[2] CC. Enz,"short story of the EKV MOS transistor model", IEEE Solid State Circuits, News13 (3), pp 24–30, 2008.(www.ieee.org/SSCS-news)

[3] CC. Enz, EA. Vittoz, "Charge-based MOS Transistor Modeling.The EKV model for low-power RF IC design", Wiley, Chichester, 2006.

[4] AIA. Cunha, MC . Scheider, C. Galup Montoro,"An MOS transistor model for analog circuit design", IEEE JSCC33(10),pp 1510–1519,1998.

[5] P. Jespers, "The gm/ID Methodology, a sizing tool for low-voltage analog CMOS Circuits: The semi-empirical and compact model approaches", ACSP Springer, 2010.

[6] C. C. Enz, F. Krummenacher, E. A. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications", J. Analog Int. Circ. Signal Processing, Vol. 8, pp. 83-114, 1995.

[7] E. Vittoz, C. Enz, F. Krummenacher, "A Basic Property of MOS Transistors and its Circuit Implications", Workshop on Compact Models, 6th Int. Conf. on Modeling and Simulation of Microsystems, San Francisco, California, USA, February, pp. 23-27, 2003.

[8] D. M. Binkley, C. E. Hopper, S. D. Tucker, B. C. Moss, J. M. Rochelle, D. P. Foty, "A CAD Methodology for Optimizing Transistor Current and Sizing in Analog CMOS Design", IEEE Trans. Computer-Aided Design of Int. Circ. & Syst., pp. 225-237, Vol. 22, N° 2, February 2003.

[9] M. Bucher, C. Enz, F. Krummenacher, J.-M. Sallese, C. Lallement, A.-S. Porret, "The EKV 3.0 MOS Transistor Compact Model: Accounting for Deep Submicron Aspects" (Invited Paper), WCM, 5th Int. Conf. on Modeling and Simulation of Microsystems, pp. 670-673, San Juan, Puerto Rico, USA, April 2002.

[10] A-S. Porret, J.-M. Sallese, C. Enz, "A Compact Non Quasi-Static Extension of a Charge-Based MOS Model", IEEE TED, Vol. 48, N° 8, pp. 1647-1654, 2001.

[11] C. Enz, "An MOS Transistor Model for RF IC Design Valid in All Regions of Operation", IEEE Trans. Microwave Theory and Tech., Vol. 50, N° 1, pp. 342-359, 2002.

[12] A.-S. Porret, C. C. Enz, "Non-Quasi-Static (NQS)Thermal Noise Modeling of the MOS Transistor", IEE Proc. Circuits, Devices and Syst., Vol. 151, N° 2, pp. 155-166, 2004.

[13] A. S. Roy, C. C. Enz, "Compact Modeling of Thermal Noise in the MOS Transistor", IEEE TED, Vol. 52, N° 4, pp. 611-614, April 2005.

[14] M. Bucher, A. Bazigos, F. Krummenacher, J.-M. Sallese, C. Enz, "EKV3.0: An Advanced Charge Based MOS Transistor Model", in W. Grabinski, B. Nauwelaers, D. Schreurs (Eds.), Transistor Level Modeling for Analog/RF IC Design, pp. 67-95, ISBN 1-4020-4555-7, Springer, 2006.

[15] C. C. Enz, E. A. Vittoz, "Charge-based MOS Transistor Modeling", John Wiley & Sons, ISBN 0-470-85541-X, 2006.

[16] T. Eimori , K. Anami , N. Yoshimatsu, T. Hasebe , and K. Murakami,"Analog design optimization methodology for ultralow-power circuits using intuitive inversion-level and saturation-level parameters", Japanese Journal of Applied Physics 53, 04EE23, 2014.

[17] D.Colombo, Fyomi, Nabki, L.F.Ferreira, G.Wirth and Bampi, "A design methodology using the inversion coefficient for low-voltage, low-power CMOS voltage references", Int.Circuits.Syst.6, 7, 2011.

[18] F.Silveira, D.Flandre, P.Jespers, "A gm/ID based methodology for the design of CMOS analog circuits and its application to the synthesis of a silcon-on-isulator micropower OTA", IEEE J Solid State Circuits 31(9), pp. 1314–1319, Sept1996.

[19] Binkley, "Trade offs and optimization in analog CMOS design", Wiley,Chichester,England,ISBN978–0–470–03136–0, 2007.

[20] A. Girardi,FP.Cortes, S.Bampi, "A tool for nautomatic design of analog circuits based on g_m/I_Dmethodology" IEEEISCAS, 2006.

[21] R.Fiorelli, A.Villegas, E.Peralías, D.Vázquez, andA.Rueda."2.4-GHz single-ended inputlow-power low-voltage active front-end for ZigBee applications in 90nm CMOS",In Proceedings of 20th European Conference on Circuit Theory and Design,ECCTD,pp.858–861,Aug.2011.

PSO Algorithm based Adaptive Median Filter for Noise Removal in Image Processing Application

Ruby Verma
M.E Student (ECE Deptt.)
NITTTR,
Sec -26, Chandigarh, India

Rajesh Mehra
Assoc. Professor (ECE Deptt.)
NITTTR,
Sec -26, Chandigarh, India

***Abstract*—A adaptive Switching median filter for salt and pepper noise removal based on genetic algorithm is presented. Proposed filter consist of two stages, a noise detector stage and a noise filtering stage. Particle swarm optimization seems to be effective for single objective problem. Noise Dictation stage works on it. In contrast to the standard median filter, the proposed algorithm generates the noise map of corrupted Image. Noise map gives information about the corrupted and non-corrupted pixels of Image. In filtering, filter calculates the median of uncorrupted neighbouring pixels and replaces the corrupted pixels. Extensive simulations are performed to validate the proposed filter. Simulated results show refinement both in Peak signal to noise ratio (PSNR) and Image Quality Index value (IQI). Experimental results shown that proposed method is more effective than existing methods.**

Keywords—Switching median filter; Particle Swarm algorithm; Noise removal; salt and pepper noise

I. INTRODUCTION

Image is a source of information but due to false capturing process, recorded images are degraded form of original image. Image noise is undesirable random fluctuations in color information or brightness of image. In digital cameras Noise depends on exposure time and amount of light. Long exposure time (slow shutter speed) mainly cause salt and pepper noise due to photodiode leakage currents. Image noise is of course inaudible. Different area of applications like medical imaging, remote sensing, robotics, computer vision and astronomical imaging needs good quality of images.

Digital images are prone to a variety of types of noise. Noisy pixels can take only maximum and minimum value in dynamic range in case of salt and pepper noise. In case of impulse noise, negative impulse appears as black (pepper) points and positive impulse appear as white (salt) noises. As a result, ‘‘noisy’’ input image gives degraded version of original image and carry inaccurate information. This is because input ‘‘noise’’ may be treated as valid information and transferred to output image, significantly degraded system performance [1]. Image filtering can be classified into two main categories: linear and nonlinear filtering. In a group of nonlinear filter, median filter gives good performance on impulse noise. A new adaptive switching median filter (SWM) is better than switching median filter in terms of PSNR [2]. But adaptive SWM filter handle noise up to 60%. Above 60% performance will decrease. Switching median filter with detector using max-min window is proposed [3]. Better than ASWM but it can handle noise only up to 70%. A new algorithm works on both impulse as well as Gaussian noise is known as universal noise removal algorithm. As compare to SD-ROM filter it gives better result in terms of PSNR [4]. If noise is more than 25% algorithm does not work. A noise adaptive soft-Switching median (NASM) filter preserves signal details across a wide range of noise densities and it is ranging from 10% to 50% [5]. If Noise density greater than 50% performance significantly degraded. Recently proposed switching median filter gives better result on salt and pepper noise. It handles noise up to 70%, known as new switching based median filter (NSWM) [6]. But it always considers pixel value 0 and 255 as a corrupted pixel. However practically it may not always true. Fuzzy impulse noise detector works on image corrupted with Gaussian as well as impulse noise if an image is corrupted with random impulse noise, filtering is applied on different part separately [7]. A novel switching median filter with impulse noise detection method, called boundary discriminative noise detection (BDND) works on monochrome as well as color images, but handles noise up-to 70% [8]. Two step filter (FIDRM) has been developed for reducing all kinds of impulse noise [9]. Fuzzy filter based on interval-valued fuzzy sets (IVFS) [10] and Predictive based adaptive switching median filter (PASMF) [11] are neural network based two stage switching median filters. Performance of these filters are better in terms of Peak signal to noise ratio (PSNR) and Image Quality index (IQI) with low noise level.

A new adaptive median filter based on PSO detection technique has been proposed in this paper. PSO algorithm and its features are discussed in 3rd section. 4th section contains block diagram of proposed filter and it’s working. 5th section contains Images and graphs simulated on MATLAB. The conclusion is given in the last section.

II. SALT AND PEPPER NOISE

It appears as randomly scattering white and black pixels over the image. Noisy pixels take either minimum or maximum value in the dynamic range. In non-linear filters, median filter is most popular to remove salt & pepper noise. However when noise level is above 50%, edge details and other information of image are smeared. The Probability density function of bipolar impulse noise is given by

$$P(x) = \begin{cases} P_a & \text{for } x=a \\ P_b & \text{for } x=b \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

If a > b, intensity "b" will appear as white dot in the image and "a" as a dark dot. If P_a or P_b is zero then impulse noise becomes uni-polar. Four different impulse noise models are discussed here [11,12].

A. Noise model 1

Impulse noise is modeled as salt and pepper noise, where pixels are corrupted by two fixed Intensity values, 0 and 255 randomly (for gray-level image), generated with the same probability. Every image pixel have intensity value $S_{i,j}$,where (i,j) is the location of pixel. In noisy image $X_{i,j}$ is the corresponding pixel having probability density function given by,

$$P(x) = \begin{cases} P/2 & \text{for } x=0 \\ P/2 & \text{for } x=225 \\ 1-P & \text{for } x = S_{i,j} \end{cases} \quad (2)$$

B. Noise Model 2

In this model Intensities of two noises are fixed similar to noise model 1, but Image pixels are corrupted by salt and pepper noise with unequal probabilities. That is,

$$P(x) = \begin{cases} P_1/2 & \text{for } x= 0 \\ P_2/2 & \text{for } x= 225 \\ 1-P & \text{for } x= S_{i,j} \end{cases} \quad (3)$$

Where p = *p*1 + *p*2 *and p1* ≠ *p2*.

C. Noise model 3

Instead of two fixed values, impulse noise modeled more realistically by two fixed ranges. It ranges from [0, m] or [255-m, 255], with a length of "m" appears at both extreme ends with equal probability. The probability density function of $X_{i,j}$ will be,

$$P(x) = \begin{cases} P/2m & \text{for } 0 \le x < m \\ P/2m & \text{for } x= (225-m) < x \le 255 \\ 1-P & \text{for } x = S_{i,j} \end{cases} \quad (4)$$

D. Noise Model 4

Model 4 also have two ranges of noise intensities similar to Noise Model 3, except that the intensities of low-density impulse noise and high density impulse noise are unequal. That is,

$$P(x) = \begin{cases} P_1/2m & \text{for } 0 \le x < m \\ P_2/2m & \text{for } x= (225-m) < x \le 255 \\ 1-P & \text{for } x = S_{i,j} \end{cases} \quad (5)$$

where *p* = *p*1 + *p*2 and *p*1 ≠ *p*2.

III. Particle Swarm Optimization

Particle swarm optimization is introduced by Dr. Kennedy and Dr. Eberhart in 1995, is a population based stochastic optimization. It is inspired by social behaviour of bird flocking or fish schooling. Evolutionary techniques such as Genetic Algorithms shares many similarities with PSO[13]. For solving complex problems, the system is initialized with a population of random solutions and searches for optima by updating generations. Each particle is initialized with a random position and a random initial velocity in the search space. The velocity and position of each particle is updated based on its own intelligence and on the experience of its neighbour [14].

A. The Particle Swarm Algorithm

In Particle swarm optimization each particle is refining its knowledge by interacting with one another. Each particle has arbitrarily small mass and volume & and also feels velocities and accelerations [15]. Each Particle updated its coordinates which are associated with the best solution, it has achieved so far. It is local best or pbest. Another 'best' is tracked by the particle taking all the population as its topological neighbors, it is global best or gbest.

Its position is x_i and velocity v_i, each particle stores the best position in the search space it has found thus far in a vector p_i. The velocity of the particle is adjusted stochastically toward its previous best position, and the best position found by any member of its neighborhood:

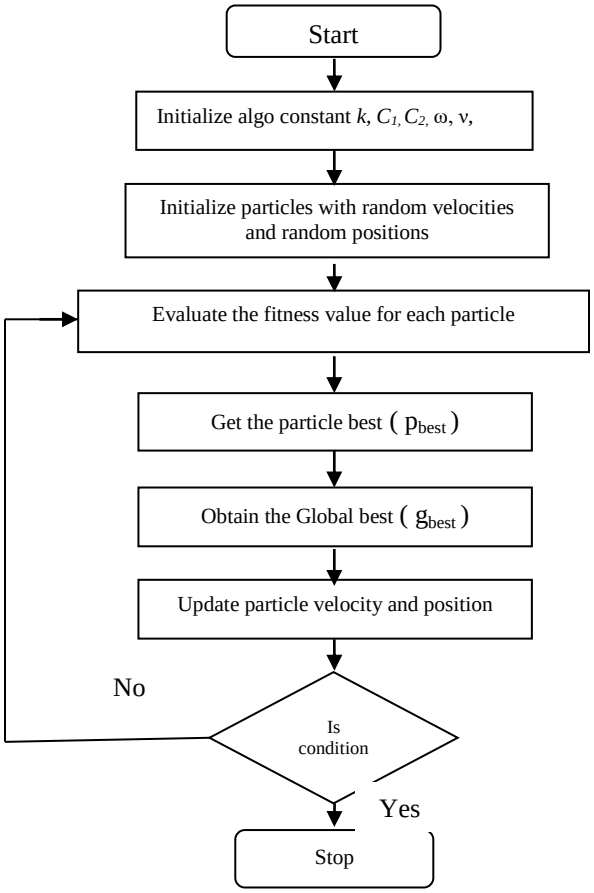


Fig. 1. Block diagram of Particle Swarm optimization

IV. Proposed Median Filter

A proposed median filter works in two steps. PSO optimizer works as Decision maker. It generates the noise map of Image. Noise map gives information about the corrupted and non

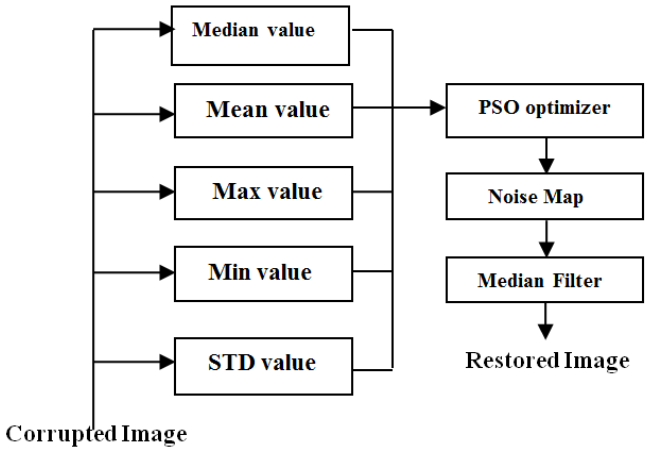


Fig. 2. Block diagram of Proposed Filter

corrupted pixels. If contaminated, a median fitter is applied. Median value is calculated only through the non-corrupted pixels of window.

A. PSO optimizer

The algorithm for PSO based decision maker is as follows:

Step 1: For FVT, taking 3x3 window from image and calculate the mean, median, max, min, std. deviation values of this window.

Step 2: Now taking the difference of centre pixel by these five values.

Step 3: For generating Feature vector Table taking 5000 pixels in which half is corrupted and half uncorrupted.

Step 4: Initialize population p=10. Taking small value of initial position and initial velocity.

Step 5: Multiply the FVT with Particles value.

Step 6: Compared the fitness value with threshold value and get fitness value.

Step 7: Updating the values for getting better fitness value. These values are local best values of particles, called "pbest".

Step 8: After 1000 Iterations "pbest" values becomes "gbest" value.

Step 9: Now training is completed. Detector uses these best value particles for generating the noise map of Image.

$$\text{Noise Map } (i, j) = \begin{cases} 0 & \text{if output of FVT} < 1 \\ 1 & \text{if output of FVT} \geq 1 \end{cases} \quad (6)$$

B. Filtering Stage

Filter uses a 3 x 3 sliding window W, corrupted pixel ($\mathbf{X_{i,j}}$) is located in its centre. Adaptive median filter locally calculates the median value of uncorrupted neighboring pixels of 3 x 3 sliding window. It replaces the value of corrupted pixels by the calculating median value, uncorrupted pixel retains as is it.

$X_{i-1, j-1}$	$X_{i, j-1}$	$X_{i+1, j-1}$
$X_{i-1, j}$	$X_{i, j}$	$X_{i+1, j}$
$X_{i-1, j+1}$	$X_{i, j+1}$	$X_{i+1, j+1}$

Fig. 3. Elements of 3X 3 sliding window W

V. Performance Measures

Four different noise models are introduced to check the performance of filtering process. All the possible combination of noise densities are covered under experiments. Performance are measured in terms of PSNR and IQI.

$$\mathbf{PSNR = 10 \log_{10} (255^2/MSE)} \quad (7)$$

Where MSE = $\sum m \sum n [O(m,n)-R(m,n)]^2/(MN)$

Where MSE is mean squared error, *O* is a original image, R is a restored image and MN is the dimensions of the image. Image quality Index is a Integration of three different factors: loss of correlation, luminance distortion and contrast distortion [16].

$$\mathbf{IQIw = Corr}(Ow, Rw) \times \mathbf{Lum}(Ow, Rw) \times \mathbf{Cont}(Ow, Rw) \quad (8)$$

IQI of an image is an average value. The Image quality index IQI*w* is computed locally within a particular sliding window W. Here *Ow* represents the original and *Rw* represents the sliding window of restored images. IQI can vary from -1 to 1. 1 is the best value represents the best restored image. Image quality map of restored image is appears as a black dots on white background. Black dots in Image quality map shows dissimilarity in original and restored image while white dots shows similarity. Light colour map shows excellent result.

The proposed filter was compared with standard median filter (MED), adaptive median filter (AMED) [2], MNASM filter with BDND detector (MNASM) [8,13], Kaliraj et al.[17] and predictive based adaptive switching median filter (PASMAF) [11]. It is justified by the simulated result that proposed filter gives better results with different noise levels in terms of PSNR and IQI values. In Fig.4 to 11 results are shown is the form of graph, generated by using noise models 1, 2, 3, 4 respectively.

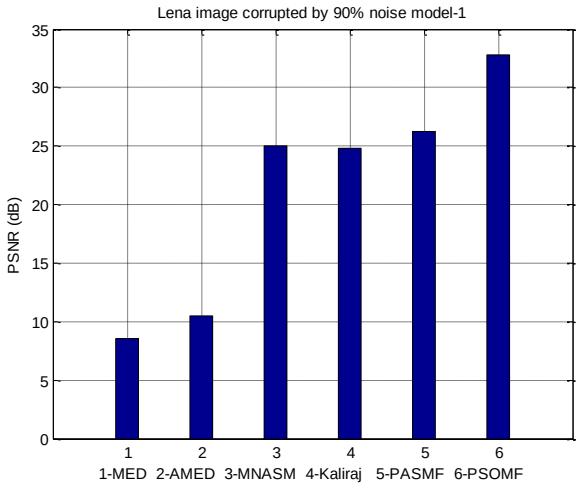


Fig. 4. Comparison in terms of PSNR value Based on noise model 1

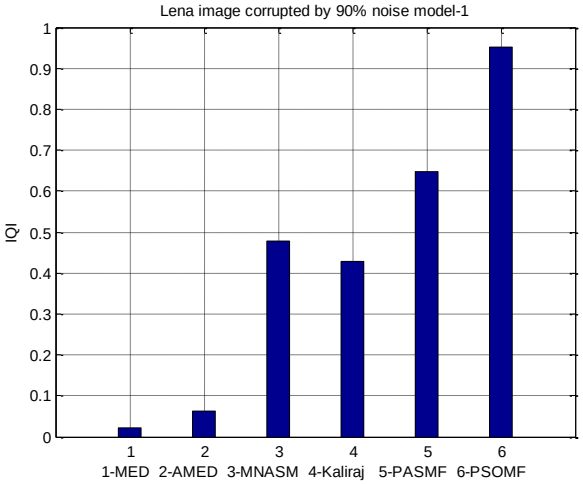


Fig. 5. Comparison in terms of IQI value Based on noise model 1

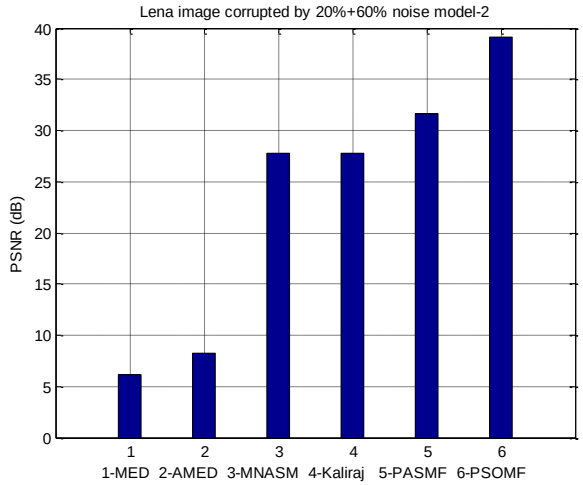


Fig. 6. Comparison in terms of PSNR value Based on noise model 2

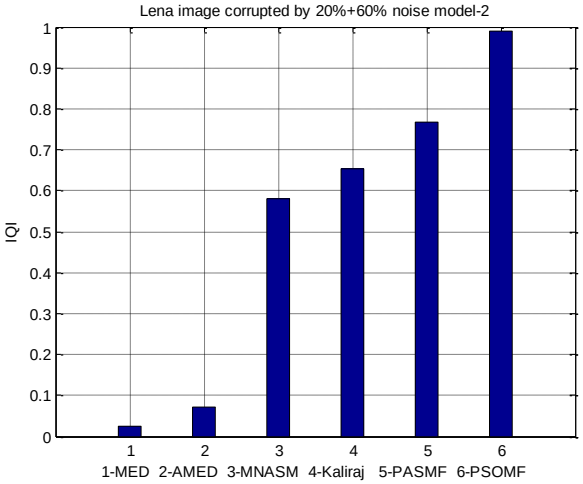


Fig. 7. Comparison in terms of IQI value based on noise model 2

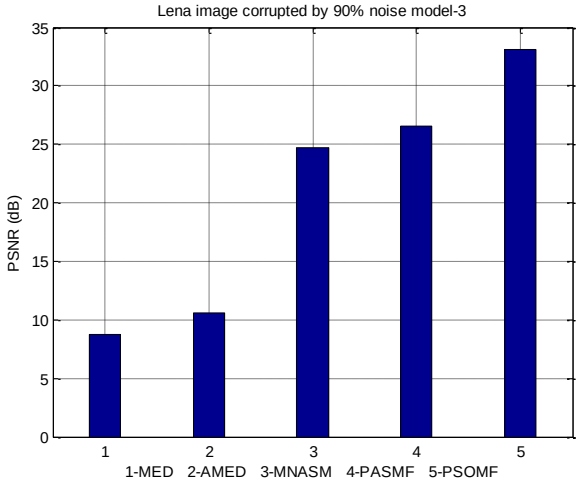


Fig. 8. Comparison in terms of PSNR value based on noise model 3

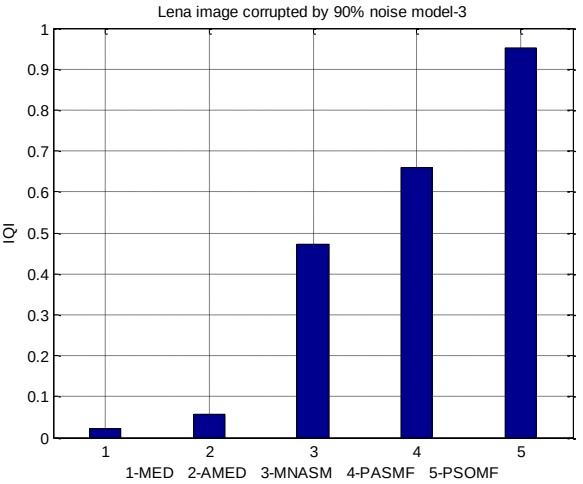


Fig. 9. Comparison in terms of IQI value Based on noise model 3

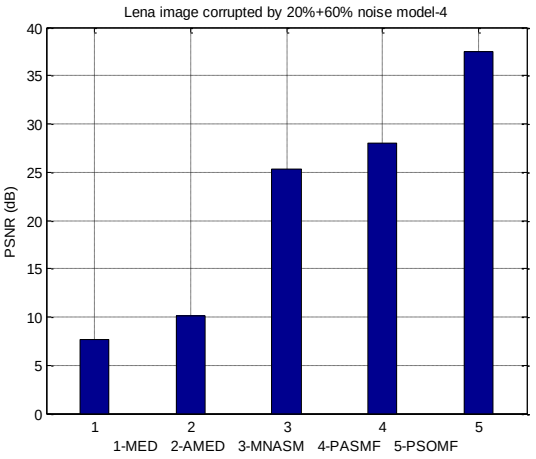


Fig. 10. Comparison in terms of PSNR value Based on noise model 4

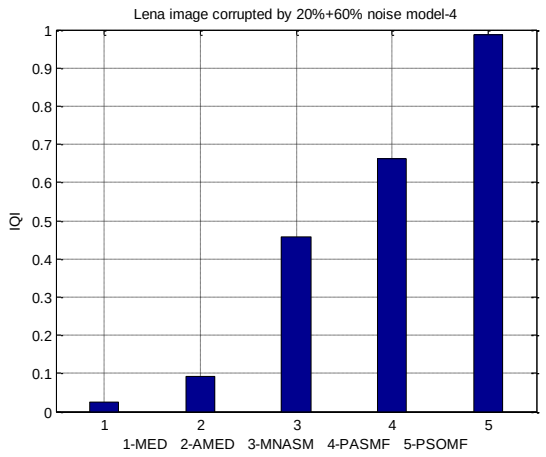


Fig. 11. Comparison in terms of IQI value based on noise model 4

To validate and compare the results of proposed algorithm a gray scale Lena, circuit , Goldhill images having 512 x 512 are being taken. Figure 12, 13, 14, 15 all having result images based on noise model 1, 2, 3, 4 respectively. Each figure consist of original image and corrupted image with edge map of original image and corrupted image produced by applying canny operator. To make it more clear, here one dimensional signal of original image and corrupted image are also shown. It is a histogram of a row intensity of image. The last image shows the image quality map of restored image. It's look like a black dots on white background. Black dots represent the mismatching between original image and restores image. Light quality map shows better restoration. All the simulations are done on MATLAB R2014a software. Noise intensities for different models are different. Every possible combination of noise have covered in experiments. The results have shows enhanced performance in terms of PSNR and IQI.

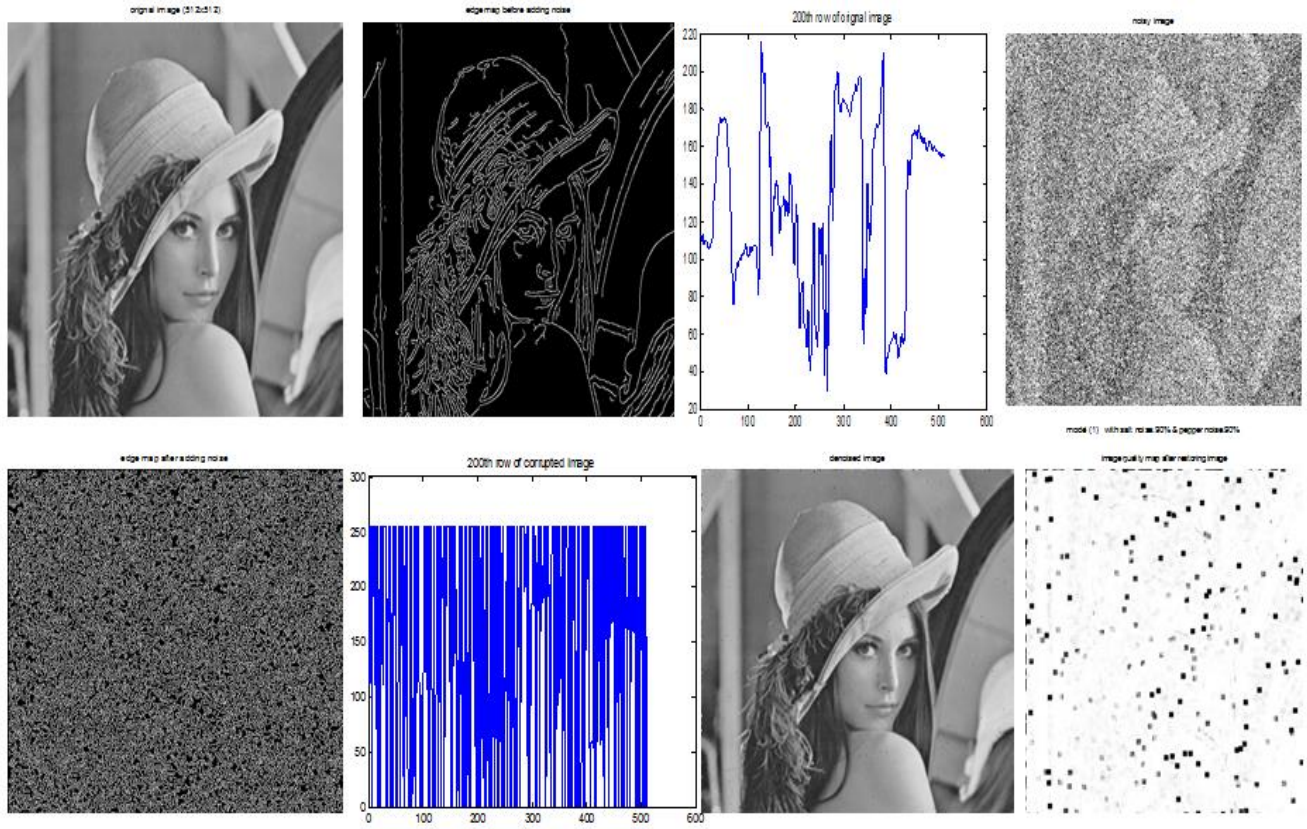


Fig. 12. Lena image corrupted by 90% noise based on noise model 1

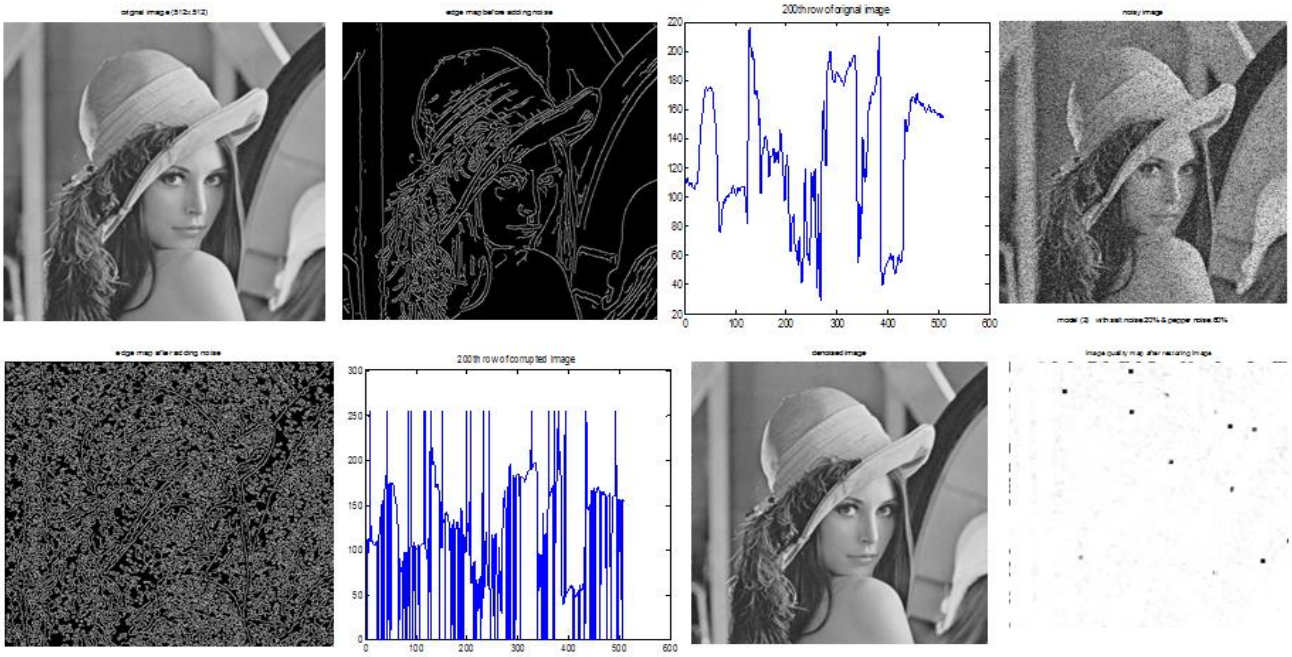


Fig. 13. Lena image corrupted by 20% salt and 60% Pepper noise based on noise model 2

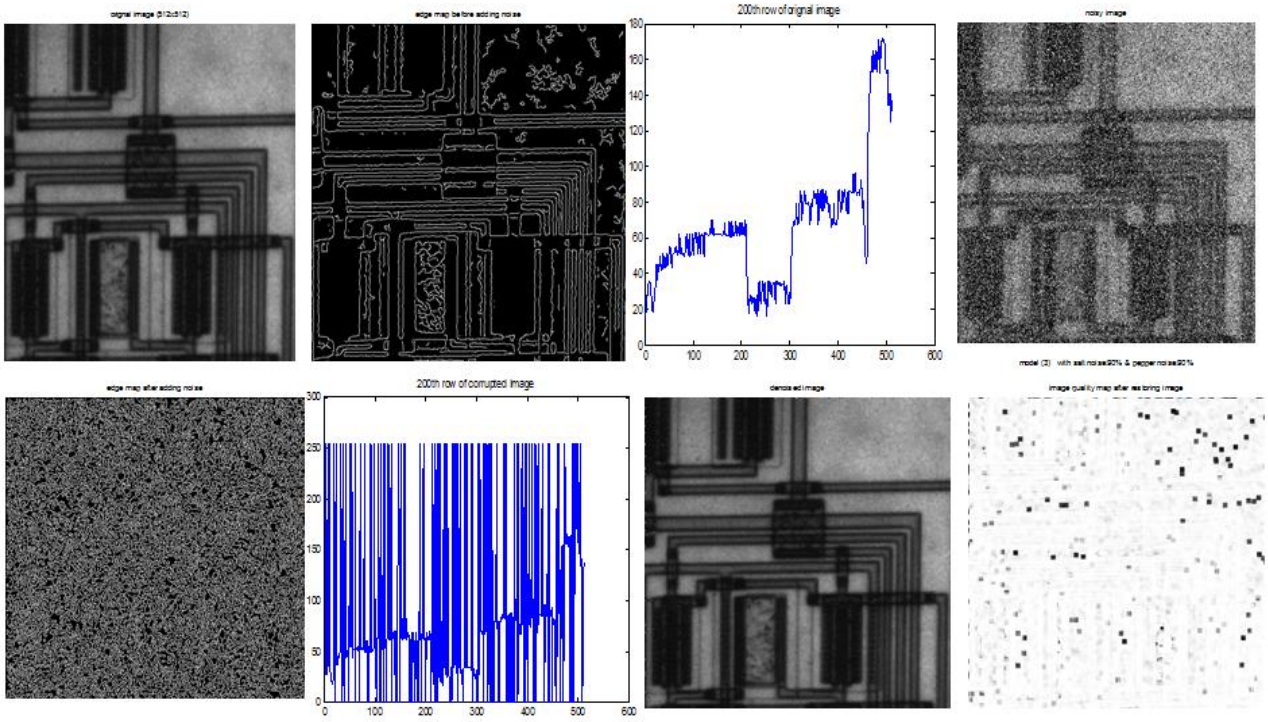


Fig. 14. Circuit image corrupted by 90% noise based on noise model 3

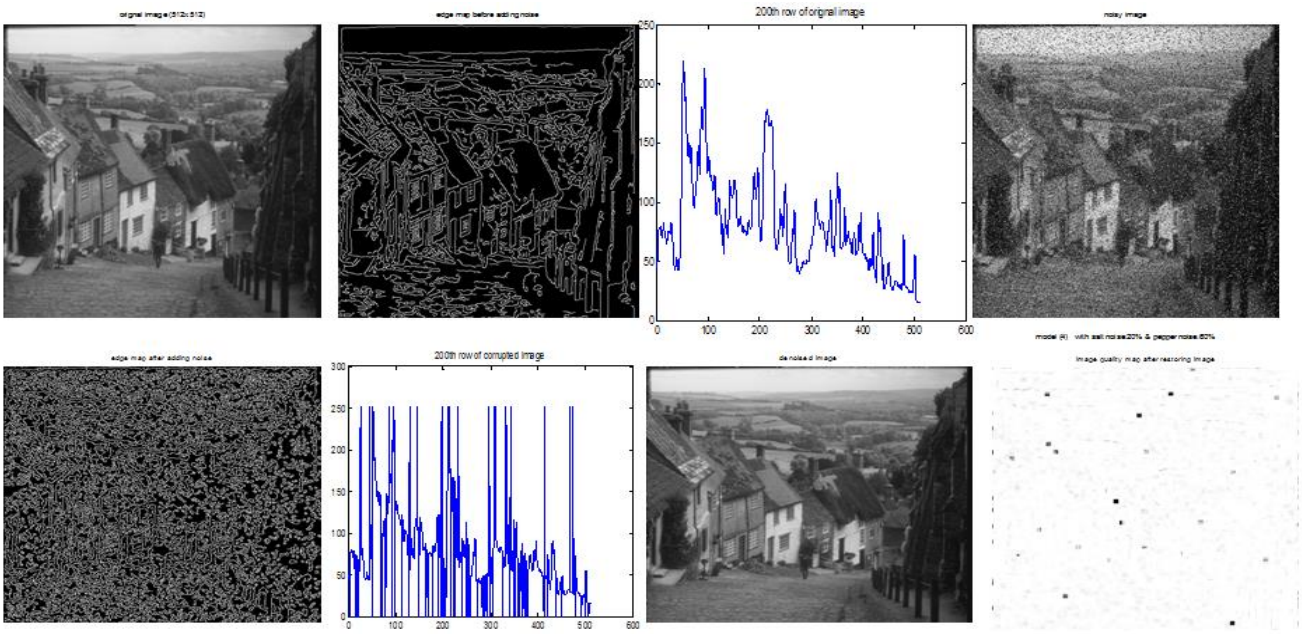


Fig. 15. Goldhill image corrupted by 20% Salt and 60% Pepper noise based on noise model 4

VI. Conclusion

Proposed filter works efficiently on highly corrupted images. It handles noise up to 90%. Experimental results show that this method produces good output as compare to fuzzy based filter. In model 1, 46% improvement in PSNR and 53% improvement in IQI has been found as compared with PASMF filter. PSNR value is improved upto 14% and IQI value is increased 28% when noise density is 20%S + 60%P. PSNR increased by 24% and IQI improved 44% when 90% noise is introduced according to model 3. PSNR value is improved by 33% and IQI value is improved 48% , as compared with PASMF when noise level is 20%S+ 60%P according to noise model 4. It is the ultimate filter for removing salt and pepper noise. Even at a very high noise levels Image Quality Index is very light which shows good quality restored image.

Some points can be discussed for further research.

1) It can also compare with more parameters as Image Enhancement Factor (IEF), Structure Similarity Index measure (SSIM).

2) Proposed filter can also test on colour images.

3) This technique may also be implemented on medical Images.

References

[1] R. C.Gonzalez, R. E. Woods " Digital Image Processing. 3rd edition. Prentice-Hall, " Englewood Cliffs, 2009.

[2] S. Akkoul, R. Ledee, R. Leconge, and R. Harba, "A new adaptive switching median filter," *IEEE Signal Processing Letters*, vol.17, pp.587-590, 2010.

[3] A.Eduardo, L. Michael,: " A new Efficient Approach for removal of the impulse noise from highly corrupted images," *IEEE Transaction on image processing*, vol.5, pp.1012-1025, June 1996.

[4] B. Deka, D.Baishnab:"A Linear Prediction Based Switching Median Filter for the Removal of Salt and Pepper Noise from Highly Corrupted Image," *IEEE conference on Computational Intelligence and signal processing*, pp. 99-102, 2012.

[5] K-K. Ma, H.L. Eng: "Noise Adaptive Soft-Switching Median Filter," *IEEE Transaction on image processing*, vol.10, pp.242-251, February 2001.

[6] S. Zhang, M.A. Karim, : "A new impulse detector for switching median filters," *IEEE Signal Processing Letters*, vol. 9, pp.360–363, 2002.

[7] R. Garnett, T. Huegerich, C. Chui : "A universal Noise Removal Algorithm with an impulse Detector," *IEEE Transaction on Image Processing*, vol.14, pp.1747-1754, 2005.

[8] P. E. Ng, K-K Ma : "A switching Median Filter With Boundary Discriminative Noise Detection for Extremely Corrupted Images," *IEEE Transaction on Image processing*, vol.15, pp.1506-1516, June 2006.

[9] S. Stefan, N. Mike, D.W. Valerie, V. Dietrich, E. K. Etienne : " A Fuzzy Impulse Noise Detection and Reduction Method," *IEEE Transaction on Image processing*, vol. 15 , pp.1153-1162, 2006.

[10] A. Bigand, O. Colot : "Fuzzy filter based on interval-valued fuzzysets for image filtering," Fuzzy Sets Systems. 161, pp.96–117, 2010.

[11] M.S. Nair, V. Shankar : " Predictive-based adaptive switching median filter for impulse noise removal using neural network-based noise detector", *Springer-london Signal, Image and Video processing*, vol. 7, pp.1041-1070, April 2012.

[12] R. Eberhat , Y.Shi : " Particle Swarm Optimization : Development, Applications and Resources," *IEEE conference on Evolutionary computation*, vol.1, pp.81-86,2001.

[13] J. Kennedy , R. Eberhart ,: "Particle Swarm Optimization," *Proceedings of IEEE Conference on Neural Networks*, pp. 1942-1948 ,1995.

[14] J.L. Marquez, J.L. Arcos: "Adapting Particle Swarm Optimization in Dynamic and Noisy Environments," *IEEE conference on evolutionary computation*, pp.1-8, 2010.

[15] Kennedy J. : "Small world and mega-minds: Effects of Neighborhood Topology onParticle Swarm Performance," IEEE conference on Evolutionary computation, pp.1931-1938, 1999.

[16] Z. Wang, A.C. Bovik,: "A universal image quality index". *IEEE Signal Processing. Letters*, vol. 9, pp.81–84 , March 2002.

[17] G. Kaliraj, S. Baskar,: "An Efficient approach for the removal of impulse noise from the corrupted image using neural network based impulse detector". *Image and Vision Computing*, vol. 28, pp.458–466, July 2009.

Switched Control of a Time Delayed Compass Gait Robot

Elyes Maherzi
National School of Engineers of Carthage,
University of Carthage,
Research Unit: Signals and Mechatronic Systems
Tunis , Tunisia

Walid Arouri
Higher Institute of Industrial Systems
University of Gabes

Mongi Besbes
Higher Institute of Information and Communication Technologies
University of Carthage,
Research Unit: Signals and Mechatronic Systems
Tunis, Tunisia

***Abstract*—the analysis and control of delayed systems are becoming more and more research topics in progress. This is mainly due to the fact that the delay is frequently encountered in technological systems. Most control command laws are based on current digital computers and delays are intrinsic to the process or in the control loop caused by the transmission time control sequences, or computing time. In other hand, the controls of humanoid walking robot present a common problem in robotics because it involves physical interaction between an articulated system and its environment. This close relationship is actually a common set of fundamental problems such as the implementation of robust stable dynamic control. This paper presents acomplete approach, based on switched system theory, for the stabilization of a compass gait robot subject to time delays transmission. The multiple feedback gains designed are based on multiple linear systems governed by a switching control law. The establishment of control law in real time is affected by the unknown pounded random delay. The results obtained from this method show that the control law stabilize the compass robot walk despite a varying delay reaching six times sampling period.**

Keywords*—Biped robot; delayed system; Switched system; Stability; Lagrange formulation; Lyapunov method; Relaxation; Linear matrix inequalities (LMI); bilinear matrix inequalities (BMI)

I. INTRODUCTION

Research on mobile robots during the last three decades has a huge progress. The biped robots are a relevant class of mobile robots due to their adaptability to various floors grounds and movement in rough environments. The non-linearity of biped walking makes the conventional control methods obsolete.

The stable walking of a biped robot can be defined as a stable oscillation around dynamic equilibrium points [1]. Other researchers are based mainly on the decomposition of a gait cycle indifferent main phases; flight, single and double support, with instant impact phase [2]. In this case the objective is to find a stabilizing control law which run between multiple operating modes, where each mode is governed by its own dynamics. The overall feedback control must stabilize each mode separately and the transitions between them. Therefore, we can formulate a switched model that includes the description of different mode and switching between them [3].

The study of the stability of biped robots under the effect of communication delays is currently the subject of intense research in the branch of the automatic delayed systems. By the way, research in the field of systems controlled via computer networks is growing. The analysis and control synthesis of delay systems are becoming more and more research topics in progress [4] [5].This is mainly due to the fact that the delay is frequently encountered in technological systems and can affect their behaviors significantly. Most are base don current digital computers and delays may occur intrinsically to the process or in the control loop caused by the transmission time control sequences, or computing time. The delay may affect one or more states of the considered system. It may also affect the establishment of the output. Several studies have modeled the linear systems with delays by differential equations covering both the present and the past states of the system, assuming that the derivative of the vector of states can be explained at every time t. Other studies consider delay systems as nonlinear and no stationary [6], [7] with parameters varying depending on time or the state of the system. The representation of such variation may be continuous or piecewise continuous [8].Modeling a delayed discrete time system as switched system is a new approach emerging from researches on lines supports and telecommunications systems. The idea is to build a set of several systems where each set constraints a value of delay[9]. Applied to the case of biped walking the overall model must be represented by a switched system submitted to two switching law. The first one is depending on gait cycle phases and known in real time, which allow us to choose between the appropriate feedback gains. The second one is unknown and depending on the delay value, which is bounded and integer (multiple of the sampling time).The feedback control synthesis approach is considered as a problem of robust control and leads us to aset of non-linear matrix inequalities conditions. To overcome this difficulty, we propose original relaxations stabilizing the robot running despite the delays and the non-linearity's.

In the first part of this article, we present a feedback control synthesis method for the command of delayed discrete time systems, based on the second method of Lyapunov.

The second part is dedicated to the modeling of compass gait biped robot. In the last part we show the results of the method applied to the obtained model.

II. STABILIZABILITY OF A DELAYED SYSTEM

A. Stability and stabilization of switched system

The stability of switched system analysis is assumed using a sufficient (but relatively nonrestrictive compared to the quadratic approach) stability condition using the poly-quadratic approach [10][11]. This approach is drawn primarily from a parameter dependent Lyapunov function [12].

Let's consider the following switched system defined as hybrid systems represented by a set whose elements are dynamic discrete time models with commutation law which define, in time, the switch between the elements:

$$x(k+1)=\sum_{i=1}^{N}\mu_i(k)A_i(k)x(k)+\sum_{i=1}^{N}\mu_i(k)B_i(k)u(k) \quad (1)$$

Where the parameters $\mu_i(k)$ replace the commutative law such as $\sum_{i=1}^{N}\mu_i=1$ the feedback control is written in the following form:

$$u(k)=\sum_{i=1}^{N}\mu_i(k)K_i(k)x(k) \quad (2)$$

The closed loop system is described by the following equation:

$$x(k+1)=\sum_{i=1}^{N}\mu_i(k)(A_i+B_iK_i)x(k) \quad (3)$$

The poly-quadratic stability analysis of the switched systems was proposed by [10]. It is possible to write the system (3) under the same following expression

$$\xi_i=\begin{cases}1:\text{If the model is described by the matrix } A_i\\ 0:\text{otherwise}\end{cases}$$

$$A(\xi_k)=\sum_{i=1}^{N}\xi_k{}^{i}A_i;\ \xi_k{}^{i}\geq 0;\ \sum_{i=1}^{N}\xi_k{}^{i}=1 \quad (4)$$

We can thus write the system according to the following form:

$$x(k+1)=\sum_{i=1}^{N}\xi_k{}^{i}A_i \quad (5)$$

The system (5) is poly-quadratically stable only if there are N symmetric matrices defined positively $S_1 \ldots S_N$ and N matrices $G_1 \ldots G_N$ of appropriate dimensions confirming:

$$\begin{bmatrix}G_i+G_i{}^{T}-S_i & G_i{}^{T}A_i{}^{T}\\ A_iG_i & S_j\end{bmatrix}>0,\ \forall i,j\in\{1\ldots N\} \quad (6)$$

The parameter dependent Lyapunov function used is written as:

$$V(x(k)\xi(k))=x^{t}(k)\sum_{i=1}^{N}P_i(\xi_k{}^{i})x(k) \quad (7)$$

With: $P_i=S_i^{-1}$

Replacing A_i by $(A_i+B_iK_i)$ and linearizing the matrix disparity by the change of variable $R_i=G_iK_i$. We reach the following condition expressed in LMI terms:

$$\begin{pmatrix}G_i+G_i{}^{t}-S_i & G_i{}^{t}A_i{}^{t}+R_i{}^{t}B_i^{t}\\ A_iG_i+B_iR_i & S_j\end{pmatrix}>0 \quad (8)$$
$$\forall i,j\in\{1\ldots N\}$$

The closed loop system is asymptotically stabilizable by state feedback if there are symmetric matrices $S_{ij}\succ 0$, Matrices R_i, G_i of appropriate dimensions such as the gain of return of state is given by:

$$K_i=R_iG_i^{-1} \quad (9)$$

B. Stability analysis of delayed switched system

When the delay affects commands, $u(k)=K_ix(k-\tau(k))$ where $\tau(k)\in[\tau_{\min},\tau_{\max}]=[i_{\min}T_e,i_{\max}T_e]$ is a variable delay.

Then we consider the augmented state vector:

$$\chi(k)=\left[x(k)^{T}\ldots x(k-\tau_{\min})^{T}\ldots x(k-\tau_{\max})^{T}\right]^{T} \quad (10)$$

Condition (6) can be used for the stability analysis of discrete delay system. The equivalence between the Lyapunov-Krasovskii functional approach for discrete delay systems and the stability conditions (6) was proved in [13].

The dynamic of the system can be represented by set of a state matrix:

$$\bar{A}_i=\begin{bmatrix}A_i & 0 & \cdots & 0 & B_iK_i & 0 & \cdots & 0\\ I & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0\\ 0 & I & 0 & \cdots & \cdots & \cdots & \cdots & 0\\ \vdots & & & & & & & \vdots\\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & I & 0\end{bmatrix} \quad (11)$$

The position of B_iK_i is depending in the value of *i*, it is found on the $(i+1)^{th}$ column of the first line of $\bar{A}_i$.

The augmented switched discrete delay system can be seen as an uncertainty switched system. Where any element of the set is a polytope whose apex are depending on the value of delay as shown in figure 1. Where SS_n represent the subsystems of the delayed switched system.$A_{ni}$the different apex $\bar{A}_i$ of the subsystem n.

Then the condition of stability analysis of the switched discrete delay system is the same than the one used for a polytopic uncertainty switched system [14].

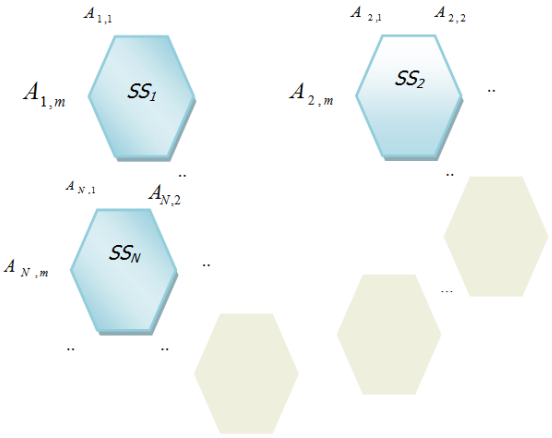


Fig. 1. Uncertainty switched system

Considering in condition (6) a commune matrix G of appropriate dimensions:

$$G = \begin{bmatrix} g_{1,1} & g_{2,1} & \cdot & \cdot\cdot & g_{1,i_{\max}+1} \\ \cdot\cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ g_{i_{\max}+1,1} & \cdot & \cdot\cdot & \cdot & g_{i_{\max}+1,i_{\max}+1} \end{bmatrix} \quad (12)$$

The condition (6) becomes:

$$\begin{bmatrix} G + G^t - S_i & \Xi_i^T \\ \Xi_i & S_j \end{bmatrix} > 0 \quad (13)$$
$$\forall i, j \in \left[\tau_{\min} / T_e, \tau_{\max} / T_e\right]$$

With:

$$\Xi_i = \begin{bmatrix} A_i g_{1,1} + B_i K_i g_{i+1,1} & \cdot & \cdot & \cdot & A_i g_{i_{\max}+1} + B_i K_i g_{i+1,i_{\max}+1} \\ g_{1,1} & g_{1,2} & \cdot & \cdot & g_{1,i_{\max}+1} \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot \\ \cdot & \cdot\cdot & \cdot & \cdot & \cdot \\ g_{i_{\max},1} & \cdot & \cdot & \cdot & g_{i_{\max},i_{\max}+1} \end{bmatrix} \quad (14)$$

The condition (13) is a *BMI*, which is impossible to solve by the current solvers. That is why it is essential to introduce relaxations in order to obtain linear conditions:

First we start with the change of variables $R_i = K_i g_{i+1,i+1}$, then we consider $g_{i,j} = 0 \forall i \neq j$ for all $i, j \in \left[\tau_{\min} / T_e, \tau_{\max} / T_e\right]$.

Then we replace in the conditions (13) Ξ_i (14) by :

$$\Xi_i = \begin{bmatrix} A_i g_{1,1} & 0 & A_i g_{i+1,i+1} + B_i R_i & \cdot & 0 \\ g_{1,1} & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot \\ \cdot & \cdot\cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & g_{i_{\max},i_{\max}} & 0 \end{bmatrix} \quad (15)$$

The feedback gains are given by: $K_i = R_i g_{i+1,i+1}^{-1}$

In case we are looking for a common feedback gain:

$$K = R_i g_{i+1,i+1}^{-1} \ \forall i \in \left[\tau_{\min} / T_e, \tau_{\max} / T_e\right] \quad (16)$$

Then it is necessary to satisfy the equality (16), to overcome this non-linearity one can consider:

$$R_i = R_j, g_{i,i} = g_{j,j} \forall i, j \in \left[\tau_{\min} / T_e, \tau_{\max} / T_e\right]$$

The matrix G becomes:

$$G = \begin{bmatrix} \begin{bmatrix} g_{1,1} & \cdots & g_{1,i_{\min}} \\ \vdots & \ddots & \vdots \\ g_{i_{\min},1} & \cdots & g_{i_{\min},i_{\min}} \end{bmatrix} & & 0 & \cdots & 0 \\ & 0 & g & 0 & \vdots \\ & \vdots & 0 & \ddots & 0 \\ & 0 & \cdots & 0 & g \end{bmatrix} \quad (17)$$

In particular case where $\tau_{\min} = 0$: $G = g \times I_{i_{\max} \times i_{\max}}$

with I is the identity matrix of appropriate dimensions.

III. Switched System Model Forwalking Compass Gait Robot

In this part, we are interested, in the modeling of the walking compass gait biped robot. The control sequence is transferred via a network data transmission. This transfer prevents the establishment of command sequence in the real time. We study more particularly the linear system stability when the pair applied by actuators, switches between the hip, the ankle, or the pair ankle - hip at the same time.

First, we have considered as a walker simplified model a bi-pendulum formed of point masses, containing only a single articulation in the hip, which is capable of reproducing the running. It is called compass gait robot. The figure1 presents this geometrical conception of the compass robot (fig.2). [15]

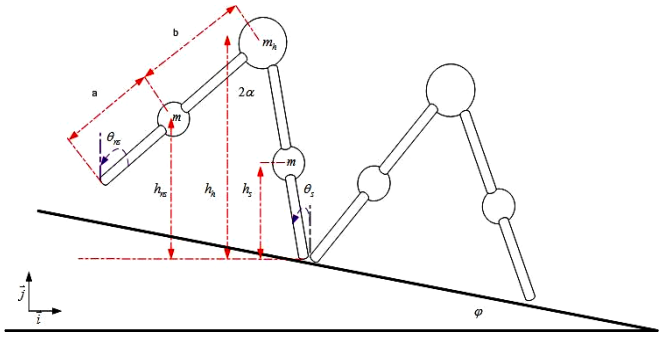


Fig. 2. Compass robot

θ_s : Absolute angle of the leg in touch with the ground (indication 's' is for support leg)

θ_{ns} : Absolute angle of the leg during flight (indication 'ns' is for no swing leg)

α: The half inter leg angle

φ: The slope angle

h_{ns}, h_s: Height separating both legs with regard to the point of biped contacting the ground

h_h: Height between hip and the point contacting the sole of compass

m,m_h : Masse of the pendulums which represents the leg and the hip

The equations of the compass during the phase of simple support are obtained by using the following Equations of Euler-Lagrange:

$$\frac{d}{dt}\left(\frac{\partial L(\dot{\theta},\theta)}{\partial \dot{\theta}}\right)-\left(\frac{\partial L(\dot{\theta},\theta)}{\partial \theta}\right)=F \tag{18}$$

With: $L(\dot{\theta},\theta)$: the Lagrangian of the system :

$$L(\dot{\theta},\theta)=E_c(\dot{\theta},\theta)-E_p(\dot{\theta},\theta) \tag{19}$$

F: External forces applied to the system.

The correspondent relations between the actuator pairs and the robot degrees of freedom are represented as follows:

$$\frac{\partial P_u}{\partial \dot{\theta}}=Torc_i \tag{20}$$

The Power P_u is given by the following relation:

$$P_u=T_c\dot{\theta}_s+T_h(\dot{\theta}_s-\dot{\theta}_{ns}) \tag{21}$$

$$\frac{\partial P_u}{\partial \dot{\theta}_s}=T_c+T_h \tag{22}$$

$$\frac{\partial P_u}{\partial \dot{\theta}_{ns}}=T_c-T_h \tag{23}$$

The equations of Euler-Lagrange are written by:

$$M(\theta)\ddot{\theta}+N(\theta,\dot{\theta})+G(\theta)=J*Torc_i \tag{24}$$

With:

$$M(\theta)=\begin{pmatrix} mb^2 & -mbl\cos(\theta_s-\theta_{ns}) \\ -mbl\cos(\theta_s-\theta_{ns}) & m_hl^2+m(a^2+l^2)\end{pmatrix} \tag{25}$$

$$N(\theta)=\begin{pmatrix} -mbl\sin(\theta_s-\theta_{ns})\dot{\theta}_{ns}{}^2 \\ mbl\sin(\theta_s-\theta_{ns})\dot{\theta}_s{}^2\end{pmatrix} \tag{26}$$

$$G(\theta)=\begin{pmatrix} mgb\sin(\theta_{ns}) \\ -(m_hl+mg(a+l)\sin(\theta_{ns}))\end{pmatrix};J=\begin{pmatrix}-1 & 0\\ 1 & 1\end{pmatrix} \tag{27}$$

$$Torc_i=\begin{pmatrix}T_h\\ T_c\end{pmatrix} \tag{28}$$

The torc is applied to the hip and the ankle.

The Lagrange equation (1) can, thus, be written in the following form:

$$\begin{pmatrix} mb^2 & -mbl\cos(\theta_s-\theta_{ns}) \\ -mbl\cos(\theta_s-\theta_{ns}) & m_hl^2+m(a^2+l^2)\end{pmatrix}\begin{pmatrix}\ddot{\theta}_{ns}\\ \ddot{\theta}_s\end{pmatrix}+ \begin{pmatrix} -mbl\sin(\theta_s-\theta_{ns})\dot{\theta}_{ns}{}^2 \\ mbl\sin(\theta_s-\theta_{ns})\dot{\theta}_s{}^2\end{pmatrix}+\begin{pmatrix} mgb\sin(\theta_{ns}) \\ -(m_hl+mg(a+l)\sin(\theta_{ns}))\end{pmatrix}=J\begin{pmatrix}T_h\\ T_c\end{pmatrix} \tag{29}$$

The state vector $q=\begin{pmatrix}\theta_{ns}\\ \theta_s\\ \dot{\theta}_{ns}\\ \dot{\theta}_s\end{pmatrix}$ the linear representation of the compass model by the jacobian method is thus written as:

$$\dot{q}=\begin{pmatrix}0&0&1&0\\ 0&0&0&1\\ \cdots A_{jacob}&\cdots\end{pmatrix}\begin{pmatrix}\theta_{ns}\\ \theta_s\\ \dot{\theta}_{ns}\\ \dot{\theta}_s\end{pmatrix} \tag{30}$$

With:

$$A_{jacob}=\frac{\partial}{\partial q}\left[(-M(\theta)^{-1}N(\theta,\dot{\theta})-M(\theta)^{-1}G(\theta))\right]_{q=q_e=0} \tag{31}$$

The linear representation of the compass model is written as follows:

$$\dot{q}=\begin{pmatrix}\dot{\theta}_{ns}\\ \dot{\theta}_s\\ \ddot{\theta}_{ns}\\ \ddot{\theta}_s\end{pmatrix}=\begin{pmatrix}0&0&1&0\\ 0&0&0&1\\ (m_hl^2+m(a^2+l^2))mg & -mbl(m_hl+m(a+l)g) & 0&0\\ m^2b^2l\,\mathrm{g} & -mbl(m_hl+m(a+l)g) & 0&0\end{pmatrix}\begin{pmatrix}\theta_{ns}\\ \theta_s\\ \dot{\theta}_{ns}\\ \dot{\theta}_s\end{pmatrix}+\begin{pmatrix}0&0\\ 0&0\\ ..J..\end{pmatrix}\begin{pmatrix}T_h\\ T_c\end{pmatrix} \tag{32}$$

The pair applied by actuators, switches between the hip, ankle, or the pair ankle - hip at the same time. This switching is described by the selection matrix *J*.

- Hip is commanded: In this case, where only the hip is commanded, the selection matrix J is then written:

$$J=\begin{pmatrix}-1&0\\ 1&0\end{pmatrix} \tag{33}$$

The model of compass is:

$$\dot{q}=\begin{pmatrix}\dot{\theta}_{ns}\\ \dot{\theta}_s\\ \ddot{\theta}_{ns}\\ \ddot{\theta}_s\end{pmatrix}=A\begin{pmatrix}\theta_{ns}\\ \theta_s\\ \dot{\theta}_{ns}\\ \dot{\theta}_s\end{pmatrix}+B_h\begin{pmatrix}T_h\\ T_c\end{pmatrix}\text{with } B_h=\begin{pmatrix}0&0\\ 0&0\\ -1&0\\ 1&0\end{pmatrix} \tag{34}$$

- The ankle is commanded: the selection matrix is:

$$J = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad (35)$$

The model command matrix B becomes:

$$B_c = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \quad (36)$$

- The ankle and the hip are commanded.

The selection matrix in this case becomes:

$$J = \begin{pmatrix} -1 & 0 \\ 1 & 1 \end{pmatrix} \text{ and } \mathrm{B}_{ch} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \quad (37)$$

In the case of the compass gait robot model, the application of the switching system approach has to find the set of feedback gainK_i, such as the closed loop system for a well determined delay is stable, has to take into account the command swaying between the hip, the ankle, or both at the same time.

IV. Simulation Results

The compass robot described the owing discrete time system, with:

$$a = b = 0.5m, l = 1m, m = 5Kg, m_h = 10Kg, g = 9.8m/s^2,$$

$$\varphi = \alpha = 3^\circ$$

$$\tau_{\min} = T_e, \tau_{\max} = 6T_e$$

$$\begin{cases} x(+1) = A_d x(k) + B_d u(k - \tau(k)) \\ y(k) = C_d x(k) \end{cases}$$

Where:

$$A_d = \begin{pmatrix} 1 & -1.4e^{-07} & 0.0001 & -4.667e^{-012} \\ -2e^{-08} & 1 & -6.667e^{-013} & 0.0001 \\ -0.0026 & -0.0028 & 1 & -1.4e^{-07} \\ -0.0004 & -0.0014 & -2e^{-08} & 1 \end{pmatrix}$$

Thematrix of control $B_d$is depending in which part is controlled hip, ankle or hip and ankle.

$$B_{dh} = \begin{pmatrix} -5e^{-009} & 0 \\ 5e^{-009} & 0 \\ -0.0001 & 0 \\ 0.0001 & 0 \end{pmatrix}, B_{dc} = \begin{pmatrix} 0 & -1.167e^{-016} \\ 0 & 5e^{-009} \\ 0 & -4.667e^{-012} \\ 0 & 0.0001 \end{pmatrix},$$

$$B_{dch} = \begin{pmatrix} -5e^{-009} & -1.167e^{-016} \\ 5e^{-009} & 5e^{-009} \\ -0.0001 & -4.667e^{-012} \\ 0.0001 & 0.0001 \end{pmatrix}$$

$$C_d = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, D_d = 0$$

The commutation due to the control choice is known at real time. In other part the commutation law due to the time delay is considered as unknown in real time but bounded. So the difficulty is to determinate three different state feedback gains stabilizing the walk despite the delay and the switch between control positions.

The studies of the stability of the delay compass gait biped robot controlled with feedback state rely on the use of parameter dependent Lyapunov function. The formulation of the gain calculation problem has led to nonlinear matrix inequality. Allaying the relaxation method presented in partII lead to a LMI's conditions. The use of Matlab (c) for the resolution of LMI allows calculating three feedback gains K_d (Table 1) ensuring the stability of the walking with an arbitrary delay affecting the input from one to six times the sampling period.

TABLE I. Gains K_d

Hip Command: B_d= B_{dh}
$K_{dH} = 1.0e+08 * \begin{bmatrix} 0.9516 & -1.2486 & 0.5235 & -0.2266 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
Ankle Command: B_d= B_{dc}
$K_{dC} = 1.0e+09 * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.4100 & -1.5177 & -0.4661 & 1.5738 \end{bmatrix}$
(Hip and Ankle) Command: B_d= B_{dch}
$K_{dCH} = 1.0e+07 * \begin{bmatrix} 1.2906 & -0.9221 & -0.9084 & 0.5399 \\ -1.3848 & 1.1161 & 0.6100 & -0.3413 \end{bmatrix}$

The simulation of the system around the balance point for various command matrices K_d , applied at the level of the ankle and of the hip, gives the following signals:(fig2, fig3, fig4)

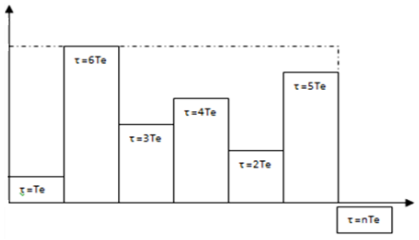


Fig. 3. Switching of the command according to the delays

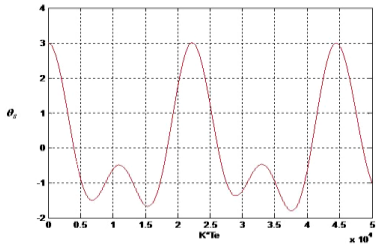


Fig. 4. θ_s : Angle of the leg when it touch the ground

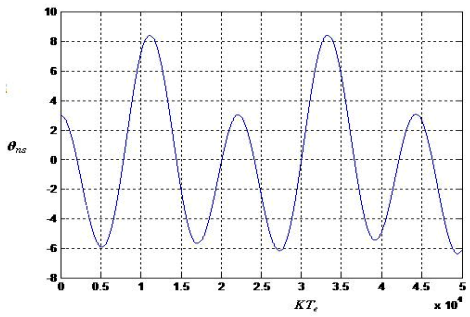


Fig. 5. θ_{ns} : Angle of the leg during flight

According to the curves of signals, θ_s , θ_{ns} ,the system remains stable, the gains by state feedback K_i calculated by the switching system method stabilizing the closed loop system for a delay reaching six times the sampling period in the three cases of command swaying. This calculus method offers thus an important stability domain because it tolerates the uncertainties on the system model aspect in real time. The simulation under Matlab simulink of the compass model verifies the compass stability conditions before the flight: $\theta_{ns} + \theta_s = -2\varphi$ and $\theta_{ns} + \theta_s = 2\alpha$, and the condition after one step movement: $\theta_{ns}^{-} - \theta_s^{-} = \theta_s^{+} - \theta_{ns}^{+} = 2\alpha$. for the various gains of K_d command.

V. CONCLUSION

We are interested in this product to the application of switched system approach to the compass robot control, controlled via a data transmission network seat of delays which prevent the establishment of control sequences in real time. Specifically, we studied the stability of the switched system of this model, when the transmitted control switches between the hip, ankle, hip or ankle at the same time. The simulation results justified the stability of the robot model for a delay varying from one to six times sampling period.

REFERENCES

[1] Yildirim Hurmuzlua, Frank Genotb, Bernard Brogliatoc, Modeling, stability and control of biped robots—a general frame work.Automatica 40 (2004) 1647 – 1664.

[2] A. Chemori and A. Loria, commande d'un robot bipède sur un cycle complet de marche. CIFA (Conférence international francophone d'automatique). France, 2002.

[3] Arouri, W., Maherzi, E., Besbes, M., Safya Belgith. Switched Control for the walking of a Compass Gait Biped Robot, Research Journal of Applied Sciences, Engineering and Technology. 2014.

[4] L. Hetel. Stabilité et commande robuste des systèmes à commutation, Institut National Polytechnique de Lorraine.2007.

[5] Seuret, Alexandre. Commande et observation des systèmes à retards variables : théorie et application. Thèse de doctorat. Universités des Sciences et technologie de Lille, école centrale de Lille, 2006.

[6] Fridman, E. and U. Shaked, 2002. An improvedstabilization method for linear time-delay systems,IEEE Trans. Automatic Control, 47: 1931-1937.DOI: 10.1109/TAC.2002.804462

[7] Cloosterman, M.B.G., N.V.D. Wouw, W.P.M. Heemelsand H. Nijmeijer, 2007. Stability of networkedcontrol systems with large delays. Proceedings ofthe 46th IEEE Conference on Decision andControl, Dec. 12-14, IEEE Xplore Press, NewOrleans, LA., pp: 5017-5022. DOI:10.1109/CDC.2007.4434669

[8] Ariba, Y. and F. Gouaisbaut, 2007. Delay-dependentstability analysis of linear systems with timevaryingdelay. Proceedings of the 45th IEEEConference on Decision and Control, Dec. 12-14,IEEE Xplore Press, New Orleans, LA., pp: 2053-2058. DOI: 10.1109/CDC.2007.4434619

[9] L.Hetel, J. Daafouz, and C. Iung. Stability analysis for discrete time switched systems with temporary uncertain switching signal. In Proceedings of 46th IEEE Conference on Decision and Control, 2007.

[10] Daafouz J., Riedinger P.et lung C. Stability analysis and control synthesis for switched systems: A switched lyapunov function approach.IEEE transactions on automatic control vol 47 n° 11, 2002, p 1883-1887.

[11] Daafouz, J., G. Millerioux and C. Lung, 2002b. A polyquadraticstability based approach for linearswitched systems. Int. J. Control, 75: 1302-1310.DOI: 10.1080/00207170210000237 35

[12] Daafouz, J. and J. Bernussou, 2001. Parameterdependent Lyapunov functions for discrete timesystems with time varying parametricuncertainties. Syst. Control Lett., 43: 355-359. DOI: 10.1016/S0167-6911(01)00118-9.

[13] L. Hetel, J. Daafouz, and C.Iung. Equivalence between the lyapunov-Krasovskii functional approach for discrete delay system and the stability conditions for switched system. Nonlinear Analysis Hybrid Systems 2(3):697-705.

[14] Maherzi, E., J. Bernussou and R. Mhiri, 2007. Stabilityand stabilization of uncertain switched systems, apolyquadratic Lyapunov approach. Int. J. Sci.Techniques Automatic Control, 1 : 61-74.

[15] Ahmed Keramane. Cycles de marche des robots de type compas, Analyse et commande. Thèse à l'Institut National Polytechnique de Greno.

An Evolutionary Stochastic Approach for Efficient Image Retrieval using Modified Particle Swarm Optimization

Hadis Heidari
Department of Computer Engineering
Razi University
Iran, Kermanshah

Abdolah Chalechale
Department of Computer Engineering
Razi University
Iran, Kermanshah

***Abstract*—Image retrieval system as a reliable tool can help people in reaching efficient use of digital image accumulation; also finding efficient methods for the retrieval of images is important. Color and texture descriptors are two basic features in image retrieval. In this paper, an approach is employed which represents a composition of color moments and texture features to extract low-level feature of an image. By assigning equal weights for different types of features, we can't obtain good results, but by applying different weights to each feature, this problem is solved. In this work, the weights are improved using a modified Particle Swarm Optimization (PSO) method for increasing average Precision of system. In fact, a novel method based on an evolutionary approach is presented and the motivation of this work is to enhance Precision of the retrieval system with an improved PSO algorithm. The average Precision of presented method using equally weighted features and optimal weighted features is 49.85% and 54.16%, respectively. 4.31% increase in the average Precision achieved by proposed technique can achieve higher recognition accuracy, and the search result is better after using PSO.**

Keywords—color moments; content based image retrieval; particle swarm optimization (PSO); texture feature

I. INTRODUCTION

The development of different images obligates the use of efficient techniques of managing the visual information by its content [1]. An image retrieval system is used the color, shape, and texture features to exact retrieve images from datasets [2]. Content based image retrieval (CBIR) is an open area research for retrieval of information using its contents [3]. From past decade, studies on CBIR have been an active research because in many large image databases, traditional techniques of image retrieval have proven to be insufficient. CBIR system extracts visual information of each image in the dataset and stores in features form and the system extract the related images that are similar to the query image. Color, shape and texture features are used in CBIR systems and practical applications [4].

One of the famous image retrieval systems is QBIC [5]. Shape feature represents the geometrical information [6] and is divided into boundary based shape and region based shape descriptors. The shape of the outer boundary is considered by boundary based shape descriptors. Zernike moments descriptors is a region based shape descriptors that describe the entire region of a shape [7]. Texture information can be used for recognizing an object [8] and structural methods are used [9] to describe it. The fine feature descriptor is applied to reach the truly matched images [10].

Color histogram is a color feature [11] that captures the number of pixels having proper properties [12]. A combined use of color and texture would provide better performance than that of color or texture alone [13] and the feature vector consists of the color and texture features [14]. Most of the image retrieval methods are not stochastic; consequently, searching in different solution space is not possible [15].

By using equal weights for the features we can't have appropriate average Precision, and Recall but applying different weights to each feature is a proper solution. For example, Particle Swarm Optimization (PSO) is an appropriate approach. Applying different weights to each feature and optimizing PSO algorithm is a method to increase the average precision in image retrieval system.

Discrete wavelet transform and particle swarm optimization was proposed by Quraishi et al. for optimizing image retrieval system [16]. The multilevel thresholds image segmentation approach and improved particle swarm optimization was proposed by Hongmei et al. [17]. A multilevel threshold-based image segmentation method and new particle swarm optimization was proposed by Jiang et al. [18]. A color image enhancement method was presented by Gorai et al. [19] in image retrieval system. Also, a histogram equalization approach and PSO algorithm was presented by Masra et al. [20]. Luo et al. introduced a wavelet-histogram image retrieval technique and PSO in CBIR systems [21]. A novel method based on PSO algorithm was proposed by Ye et al. for image retrieval [22]. Also, a new approach based on PSO and wavelet was proposed by Wei et al. that PSO was employed to optimize the weights [23].

Most of the CBIR systems may not perform robustly on image retrieval using the different features. Consequently, the key motivation in this work has been to develop a more robust and accurate image retrieval method which can be effective. In this paper modified PSO is used to effective retrieval in CBIR

Dr. Abdolah Chalechale, Department of Computer Engineering, Faculty of Engineering, University of Razi, Kermanshah, Iran.

systems. In fact, in this paper, a novel method based on color and texture image retrieval technique and modified PSO is presented to retrieval of images in huge databases to do color textured image retrieval.

It is important to mention that one of the appropriate methods for extraction of the color features is color moments. In this work, color moments including mean, standard deviation, and skewness are used and entropy, standard deviation, local range and contrast is applied to extraction of the texture features

In fact, the key contribution of this paper is given in the following:

- A proposed of an image retrieval system using the color and texture features.
- A proposed of an optimization algorithm for increasing average precision of CBIR system.

The reminder of this work is organized as follows. Section II discusses algorithms conventional PSO and the modified PSO algorithm. Section III explains a novel method for the image retrieval systems. Section IV illustrates the experimental results and finally, Section V provides conclusion and future work.

II. The Particle Swarm Optimization Algorithm

In this section, some information about algorithms conventional PSO and the modified PSO algorithm used in this study is provided.

A. The Standard Particle Swarm Optimization Algorithm

PSO is a heuristic technique and an evolutionary computation model developed by Kennedy and Eberhart in 1995 [24] that is related to genetic algorithms and evolutionary programming. PSO is considered robust in solving problems featuring non-differentiability, non-linearity, and high dimensionality [25] and is used in neural networks [26].

If X be the decision vector in a cost function f (X) then it must be minimize in the optimization problem. In the PSO algorithm, all particles have random coordinates in n-dimensional space. For each particle, pbest and gbest are the best coordinates each particle and the best coordinates among overall particles, respectively that the particles move based on pbest and gbest. X and V are current position vector and velocity vector for each particle, respectively. At the k^{th} time step (iteration), the velocity vector is updated as follows:

$$V_{id}^{k+1} = W \times V_{id}^{k} + c_1 \times rand_1^{k} (pbest_{id}^{k} - X_{id}^{k}) + c_2 \times rand_2^{k} (gbest^{k} - X_{id}^{k}) \quad (1)$$

Also, the position vector of the i^{th} particle is changed as follows:

$$X_{id}^{k+1} = X_{id}^{k} + V_{id}^{k+1} \quad (2)$$

$$k = 1,2,..., \max iteration$$

$$id = 1,2,..., p$$

That p is the number of particles. c_1 and c_2 are the relative attraction toward pbest and gbest, respectively and $rand_1$ and $rand_2$ are random numbers uniformly distributed between [0,1]. Also, W is inertia weight parameter. The PSO algorithm can be expressed as Fig. 1.

```
k=1
%initialize random particles
for i=1 to p
   particle(i).X=random value between X_LowerBond and X_UpperBound
   particle(i).V=random value between V_LowerBond and V_UpperBound
   particle(i).cost=cost_function(particle(i).X)
   particle(i).pbest=particle(i).cost
   particle(i).best_position=particle(i).X
end
[global_best_position gbest]=minimum_cost(particle)
%find best position
while(k<=maximum iteration)
   for i=1 to p
      update particle(i).X
      update particle(i).V
      particle(i).cost=cost_function(particle(i).X)
      if (paticle(i).cost<particle(i).pbest)
         particle(i).pbest=particle(i).cost
         particle(i).best_position=particle(i).X
      end
   end
   [global_best_position gbest]=minimum_cost(particle)
   k=k+1
end
return [global_best_position gbest]
```

Fig. 1. Standard PSO algorithm

Each dimension of X and V must be limited between lower bound and upper bound that are determined based on the parameter of the problem. These parameters must be optimized in optimization algorithms.

B. The Improved Particle Swarm Optimization Algorithm

An important problem in PSO method is to determine limits in search space [27]. In standard PSO algorithm, execution time increases with larger search space. Consequently, the domain of each dimension of vector X is limited and the standard PSO routine is called. Improved PSO algorithm is given in Fig. 2 that d is applied to limit search space.

```
k=1 %k is a global counter
do
{
  [global_best_position gbest]=standard_pso(k,X_bond)
  X_lower bound=global_best_position - ((X_UpperBound - X_LowerBound)/d)
  X_Upper bound=global_best_position + ((X_UpperBound - X_LowerBound)/d)
  } while(k<=maximum iteration)
  return [global_best_position gbest]
```

Fig. 2. Improved PSO algorithm

III. The Proposed Image Retrieval System

In this section, proposed image retrieval procedure is provided. In addition, because of color and texture are two of the most widely used features.

A. The Basic Concepts of CBIR Systems

One of the appropriate ways of accessing visual data is image retrieval that use to color, shape, and texture [28]. Feature extraction is one of important steps in CBIR systems. Extraction of features of the images is stored in feature vectors form. The input image is called the query image. The query image feature vector is compared with all feature vectors in the dataset. Consequently, the appropriate images retrieve using distance measurement technique. Fig. 3 illustrates the architecture of CBIR systems. The user interface is consists of a query formulation part and a visualization part, is the front page of most systems dealing with input and output. The matching process does similarity measuring and the necessary comparisons. The indexes of those images which are selected to retrieve are passed into the image pointers process. It obtains image pointers (image id's), and the fetching process physically retrieves the images from the dataset.

B. CBIR Systemsn using the Fusion of Texture Features and Color Moments

In this section, texture features and color moments are investigated.

Entropy, local range, standard deviation and contrast measures are used to extract the texture features.

Texture = (Entropy + Local Range + Standard deviation + Contrast)

Entropy can be used to describe the texture of the input image that can be calculated as:

$$ENT = \sum_{k=1}^{M} P_k \log \frac{1}{P_k} \tag{3}$$

Where, ENT, M, and P are entropy, total number of samples, and probability of occurrences, respectively. Maximum value of chosen pixel-minimum value of chosen pixel is called local range. Standard deviation can be calculated as follows.

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}} \tag{4}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

That, n is number of pixels in the image. Contrast represents the quality of picture in an image and is calculated by (5)

$$F_{con} = \frac{S^2}{\sqrt[4]{\mu_4}} \tag{5}$$

$$\mu_4 = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (X(i,j) - \overline{X})^4$$

Where, μ_4 is the 4th moment of the mean $\overline{X}$, S^2 is the variance of the gray values in image.

In this work, mean, standard deviation, and skewness to extract color features is used. Mean, standard deviation, and the skewness are effective in representing color distributions of images. Color moments are describe as follows.

Moment 1: Mean

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} P_{ij} \tag{6}$$

That P_{ij} is the value of the i-th color channel at the j-th image pixel

Moment 2: Standard deviation

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N} (P_{ij} - \mu_i)^2}{N}} \tag{7}$$

Moment 3: Skewness

$$S_i = \sqrt[3]{\frac{\sum_{j=1}^{N} (P_{ij} - \mu_i)^3}{N}} \tag{8}$$

In this work, the combination of texture features and color moments is used. Entropy, local range, standard deviation and contrast measures are used to extract the texture features and 13 features are applied.

Features = (Texture Features + Color Features)

C. Stochastic CBIR using Improved Particle Swarm Optimization

Cost function must be minimized in optimization algorithm. A minimization tool is the stochastic PSO method. The solutions space is made of the features f=1,…,F, that are calculated on every dataset image. In the algorithm, the value of F is adjusted to 13. These 13 features are entropy, standard deviation, local range, contrast, mean of red component, standard deviation of red component, skewness of red component, mean of green component, standard deviation of green component, skewness of green component, mean of blue component, standard deviation of blue component, and skewness of blue component, respectively.

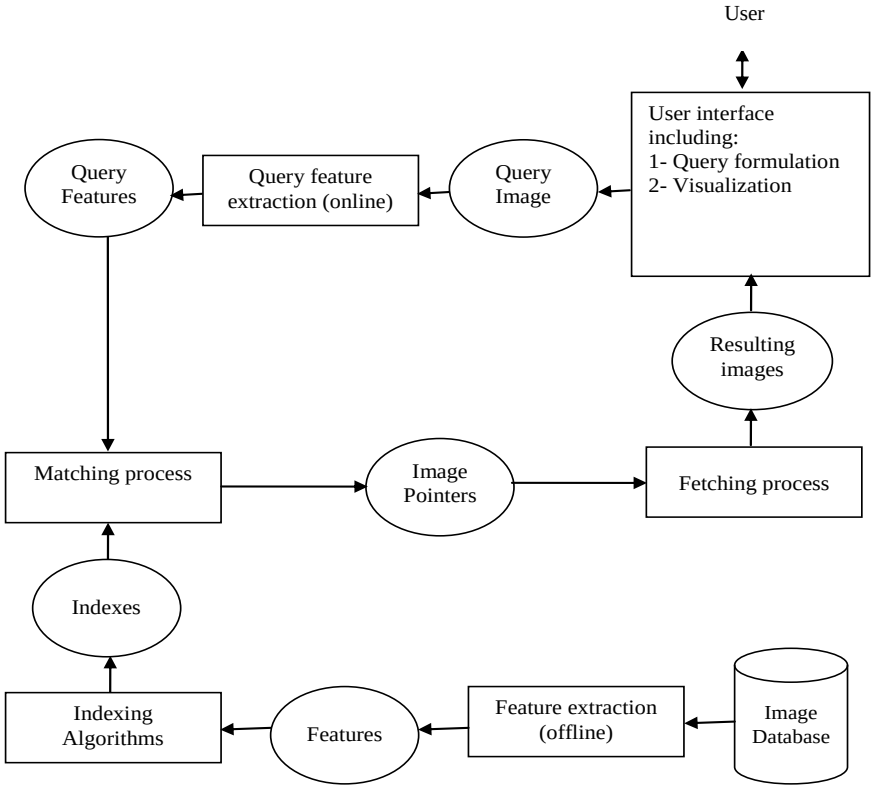


Fig. 3. Architecture of CBIR systems

The images of the database x_j, j=1,…, N_{DB} represent a discrete set of points and the particles can move within the features space. To associate every particle with the nearest image, a weighed city block distance (WCBD) or the Manhattan distance is used which is expressed as follows.

$$WCBD(x_q, x_j) = \sum_{f=1}^{F} | x_q^f - x_j^f | \times w^{k,f} \quad (9)$$

By using equal weights for each feature we can't have good average Precision, and Recall. Different weights to each feature are a good solution that is optimized using PSO algorithm. Weighing vector w^k calculate again at each iteration. The proposed algorithm shows in Fig. 4.

In the first iteration (*k=1*), a query image with a feature vector $x_q=[x_q^1,\ldots, x_q^f,\ldots, x_q^F]$ is selected. Then, the distances from all the dataset images x_j; j=1,…,N_{DB} are computed as WCBD(x_q,x_j). The speed vector each particle is set by randomly selecting a value over the features space, and then the stochastic optimization is done. The related and the irrelevant images are updated in each iteration and the new features weights are computed.

After classify the swarm based on the fitness of each particle, the k^{th} iteration is completed. Finally, each particle to the nearest image in the dataset is associated and the best N_{FB} is shown to the user. While a predefined numbers of iterations are reached, the optimization process ends. Then, the relevant solutions are shown.

IV. Experimental Results

The performance of an image retrieval system is computed using the Recall and Precision values. The Recall is defined as the ratio of the number of relevant images retrieved and the number of relevant images in class. The Precision is defined as the ratio between the number of relevant images retrieved and the total number of images retrieved [29]. Precision and Recall is computed as:

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}} \quad (10)$$

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \quad (11)$$

Also, the average Precision is computed by:

$$Average\_Precision = \sum_{k \in A_q} \frac{p(i_k)}{|A_q|} \quad (12)$$

That item belongs to the qth category (A_q). The fusion of color and texture features with optimal weights to a subset of MPEG-7 dataset is used.

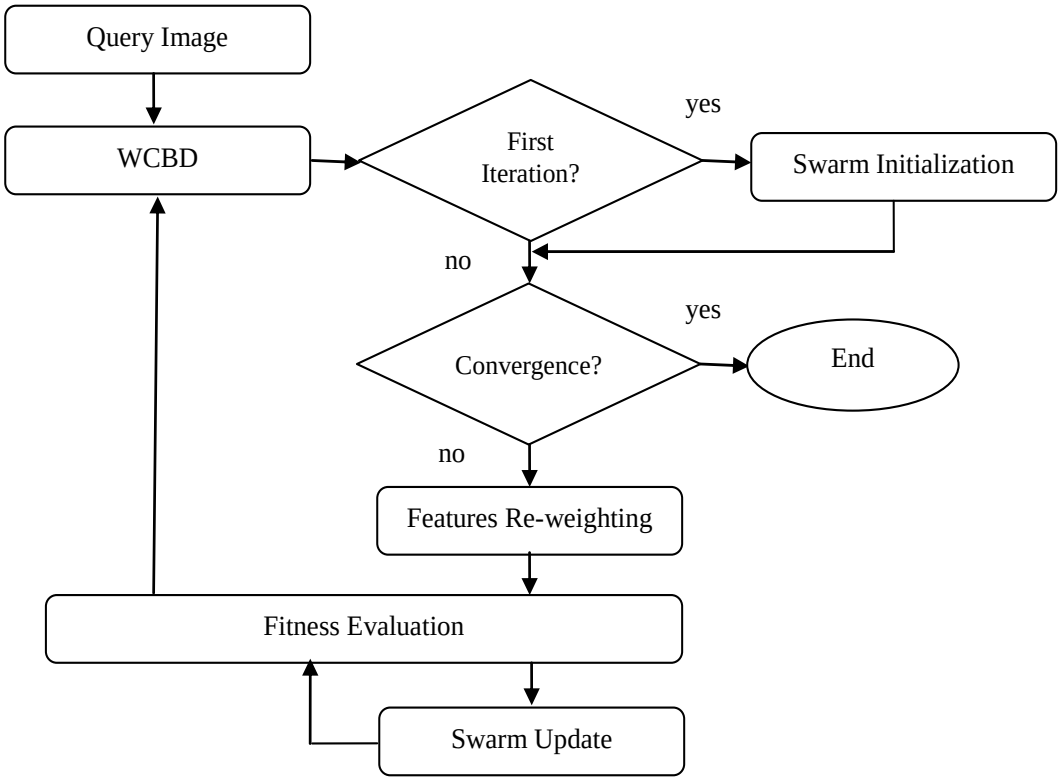


Fig. 4. Flowchart of the proposed method

The images in the dataset are categorized in 23 classes and each class contains 10 pictures in JPEG format that samples of MPEG-7 image dataset are shown in Fig. 5.

Fig. 5. Samples of MPEG-7 image database

Four texture features include entropy, standard deviation, local range, and contrast and nine colour features (mean, standard deviation, and skewness, for R, G, and B components in RGB space).

Precision and Recall are evaluation parameters in our experiments and implementations is done using a PC with Intel Pentium 2.5 GHz and 4 GB RAM. Fig. 6 illustrates the results generated from proposed system using optimal weighted features that show the efficiency of proposed method.

These results show that the performance of the proposed method is better than the other methods.

In experiments, Precision versus Recall curves to evaluate retrieval efficiency is adopted.

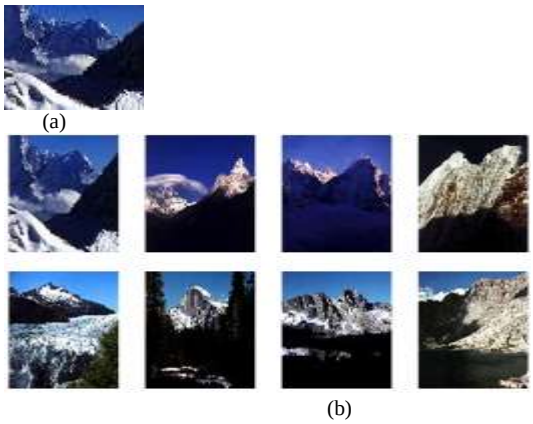


Fig. 6. Content based image retrieval results. (a) input image for retrieval. (b) using the improved PSO

Fig. 7 shows average Precision for when texture features (TF), the color moments (CM), the combination of the texture features and color moment using equally weighted features (TCEW) and optimal weighted features (TCOW), respectively extracted from images.

The average Precision in these four methods is 41.87, 45.64, 49.85, and 54.16 percent, respectively.

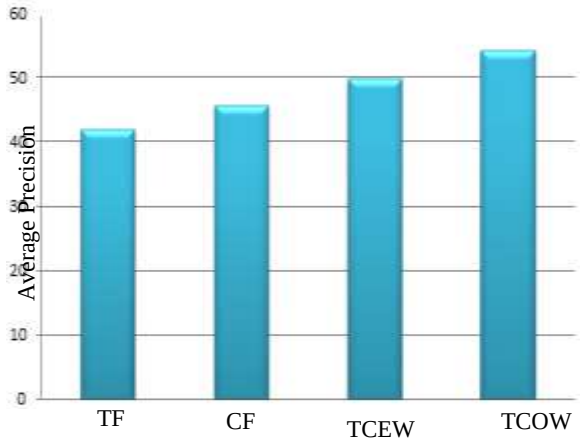
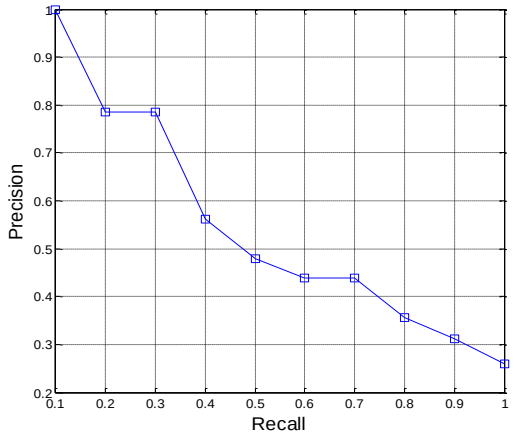
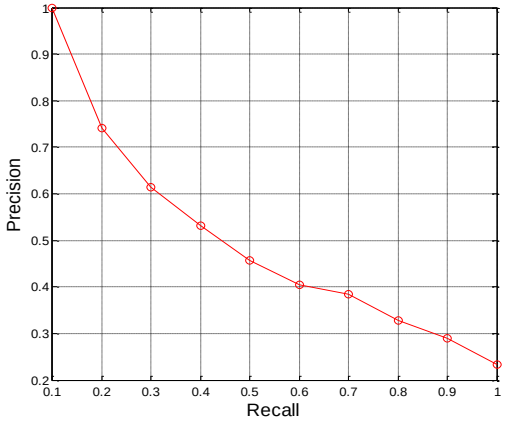


Fig. 7. The Average precision chart for CBIR system using the texture features, the color moments, equally weighted features and system with using improved PSO method.



(a)



(b)

Fig. 8. The Average precision/recall chart for (a) CBIR system using equally weighted (b) CBIR system with using improved PSO method

Fig. 8 (a) and (b) show the Precision-Recall graph for the proposed image retrieval system using equally weighted features and using the improve PSO, respectively that the results of optimal weighted features show better average Precision and Recall.

The average Precision of proposed approach using improved PSO are 54.16%. Also, for the proposed method, the maximum average Precision of 100% at Recall value is 10%, and the Precision value decreases to 25.92% at 100% of Recall. Table I shows the quantitative results obtained by the optimal weighted features to the dataset that a total average of 54.16% retrieved images is achievable using improved PSO algorithm. Fig. 9 illustrates the comparison of average precision the proposed method with the other methods in [8], [9], and [10].

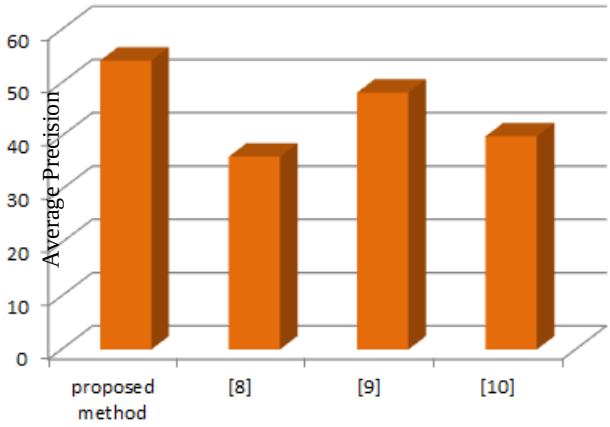


Fig. 9. The Average precision for the proposed method and the other methods

Fig. 10 shows the Precision versus Recall result of 30 query images that the proposed method is better than the fixed weighting.

TABLE I. PRECISION AND RECALL OF THE PROPOSED METHOD

Recall (%)	Precision (%) for the equally weighted features	Precision (%) for the presented method using improved PSO
10	100	100
20	74.05	78.51
30	61.48	78.50
40	53.22	56.15
50	45.76	47.97
60	40.54	43.90
70	38.49	43.89
80	32.78	35.58
90	28.90	31.18
100	23.26	25.92
AR = 55%	**AP = 49.85%**	**AP = 54.16%**

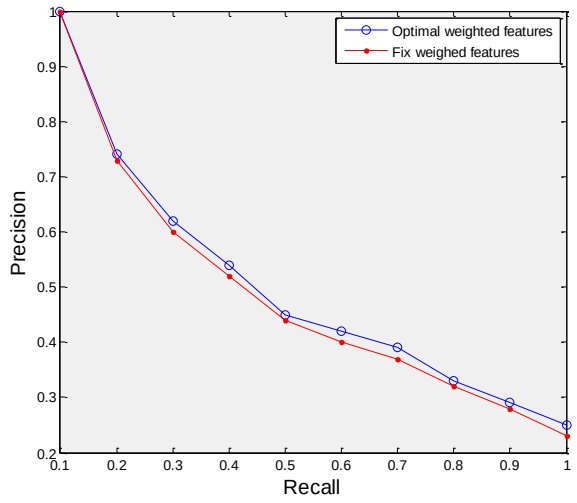


Fig. 10. Average precision vs. recall

In our experiments, size of swarm is 260 particles and sum of c_1 and c_2 variables is smaller 3. Also, the average precision in different iterations is not same. Fig. 11 shows the Precision chart in different iterations.

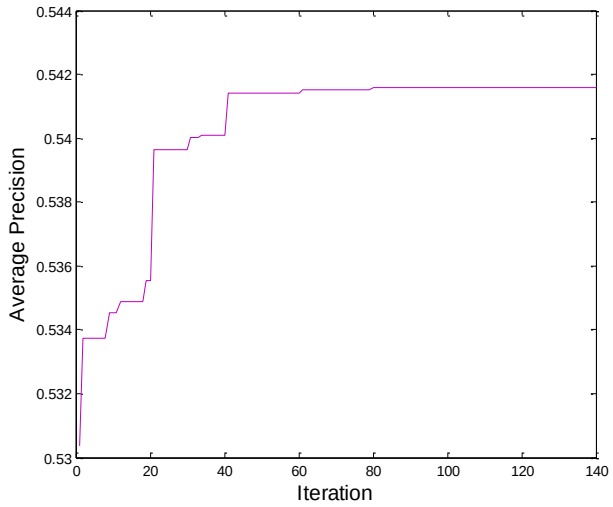


Fig. 11. Precision chart in different iterations

V. Conclusion and Future Work

In this paper, an method PSO to improve the accuracy and ability for image retrieval is presented and the improved PSO algorithm for image retrieval to get higher accuracy is employed. In fact, the use of a stochastic optimization algorithm to achieve a proper CBIR system was investigated. The experimental results showed that the proposed method was effective to the similarity search in images dataset after using PSO. To enhance the retrieval performance, cost function was minimizing. Proposed method was evaluated using Precision, Recall, and average Precision that the average Precision and the average Recall of proposed method are 54.16% and 55.00%, respectively.

In this work, to show the effectivity of proposed PSO algorithm, only color and texture features are selected, selecting more features will achieve better retrieval effect, which is our further work. Furthermore, the approach can be extended for translation and rotation properties, so that the retrieval efficiency can be increased.

References

[1] S. Agarwal, A. Verma, and P. Singh, "Content Based Image Retrieval using Descrete Wavelet Transform and Edge Histogram Descriptor", International Conference on Information Systems and Computer Networks, 2013, pp. 19-23.

[2] D. Pedronette and R. Torres, "Unsupervised measures for estimating the effectiveness of image retrieval systems", International Conference on Graphics, Patterns and Images, 2013, pp. 341-348.

[3] A. Ahmadi, A. Chalechale, and H. Heidari, "Parallelized Computation for Edge Histogram Descriptor Using CUDA on the Graphics Processing Units (GPU)", The 17th CSI Internattional Symposium on Computer Architecture & Digital Systems, 2013, pp. 1-6.

[4] S. Youssef, "ICTEDCT-CBIR: Integrating curvelet transform with enhanced dominant colors extraction and texture analysis for efficient content-based image retrieval", Computers and Electrical Engineering 38, 2012, pp. 1358-1376.

[5] A. Bhagat and M. Atique, "Design and Development of Systems for Image Segmentation and Content Based Image Retrieval", IEEE, 2012, pp. 1-5.

[6] C. Singh and Pooja, "An effective image retrieval using the fusion of global and local transforms based features", Optics & Laser Technology 44, 2012, pp. 2249-2259.

[7] P. Manipoonchelvi and K. Muneeswaran, "Significant Region Based Image Retrieval Using Curvelet Transform", International Conference on Recent Advancements in Electrical, Electronics and Control Engineering, 2011, pp. 291-294.

[8] K. Kim and S. Kwon, "Image Retrieval Scheme Based on Adaptive Feature Weighting", IEEE, 2012, pp. 747-751.

[9] W. Yuan, C. Feng, and Y. Jiao, "An effective method for color image retrieval based on texture", Computer Standard & Interfaces 34, 2012, pp. 31-35.

[10] C. Lin, D. Huang, Y. Chan, K. Chen, and Y. Chang, "Fast color-spatial feature based image retrieval methods", Expert Systems with Applications 38, 2011, pp. 11412-11420.

[11] W. Chen, W. Liu, and M. Chen, "Adaptive Color Feature Extraction Based on Image Color Distributions", IEEE Transaction on Image Processing, 2010, pp. 2005-2016.

[12] B. Syam, S. Victor, and Y. Rao, "Efficient Similarity Measure via Genetic Algorithm for Content Based Medical Image Retrieval with Extensive Features", IEEE, 2013, pp. 704-711.

[13] H. Heidari, A. Chalechale, and A. Ahmadi, "Accelerating of Color Moments and Texture Features Extraction Using GPU Based Parallel Computing", 8th International Conference on Machine Vision and Image Processing, 2013, pp. 1160-1165.

[14] A. Salahuddin, A. Naqvi, K. Mujtaba, and J. Akhtar, "Content based Video Retrieval using Particle Swarm Optimization", 10th International Conference on Frontiers of Information Technology, 2012, pp. 79-83.

[15] M. Broilo, P. Rocca, and F. Natale, "Content-Based Image Retrieval by a Semi-Suppervised Particle Swarm Optimization", IEEE, 2008, pp. 666-671.

[16] M. Quraishi, K. Dhal, J. Paul, and M. De, "A Novel Hybrid Approach to Enhance Low Resolution Images Using Particle Swarm Optimization", 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, 2012, pp. 888-893.

[17] T. Hongmei, W. Cuixia, H. Liying, and W. Xia, "Image Segmentation Based on Improved PSO", IEEE International Conference on Computer and Communication Technologies in Agriculture Engineering, 2010, pp. 191-194.

[18] F. Jiang, M. Frater, and M. Pickering, "Threshold-based Image Segmentation Through an Improved Particle Swarm Optimization", IEEE, 2012, pp. 1-5.

[19] A. Gorai and A. Ghosh, "Hue-Preserving Color Image Enhancement Using Particle Swarm Optimization", IEEE, 2011, pp. 563-568.

[20] S. Masra, P. Pang, M. Muhammad, and K. Kipli, "Application of Particle Swarm Optimization in Histogram Equalization for Image Enhancement", IEEE Colloquium on Humanities, Science & Engineering Research, 2012, pp. 294-299.

[21] T. Luo, B. Yuan, and L. Tan, "Blocking Wavelet-histogram Image Retrieval by Adaptive Particle Swarm Optimization", The 1st International Conference on Information Science and Engineering, IEEE, 2009, pp. 3985-3988.

[22] Z. Ye, B. Xia, D. Wang, and X. Zhou, "Weight Optimization of Image Retrieval Based on Particle Swarm Obtimization Algorithm", IEEE, 2009, pp. 1-3.

[23] K. Wei, T. Lu, W. Bi, and H. Sheng, "A Kind of Feedback Image Retrieval Algorithm Based on PSO, Wavelet and Sub-block sorting thought", 2nd International Conference on Future Computer and Communication, 2010, pp. 1-6.

[24] J. Kennedy and R. Eberhart, "Particle swarm optimization", IEEE International Conference Neural Networks, 1995, pp. 1942-1948.

[25] A. Taher, A. Karimian, and M. Hasani, "A new method for optimal location and sizing of capacitors in distorted distribution networks using PSO algorithm", Simulation Modeling and Theory 19, 2011, pp. 662-672.

[26] M. Broilo and F. Natale, "A Stochastic Approach to Image Retrieval Using Relevance Feedback and Particle Swarm Optimization", IEEE Transaction on Multimedia, 2010, pp. 267-277.

[27] R.A. Vural, O. Der, and T. Yildirim, "Investigation of Particle Swarm Optimization for Switching Characterization of Inverter Design", Expert Systems with Applications, Vol. 38, No. 5, 2011, pp. 5696-5703

[28] F. Malik and B. Baharudin, "Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain", Computer and Information Science 38, 2012, pp. 1-12.

[29] C. Rubert, L. Cinque, "Decomposition of two-dimensional shapes for efficient retrieval", Image and Vision Computing 27, 2009, pp. 1097-1107.

Evaluating Web Accessibility Metrics for Jordanian Universities

Israa Wahbi Kamal
Software Engineering Department
IT Faculty, Zarqa University
Zarqa, Jordan

Izzat M. Alsmadi
Computer Science Department
University of New Haven
West Haven, CT 06516, USA

Heider A. Wahsheh
Computer Science Department
College of Computer Science,
King Khaled University
Abha, Saudi Arabia

Mohammed N. Al-Kabi
Computer Science Department
Zarqa University
P.O. Box 132222
Zarqa 13132, Jordan

***Abstract*—University web portals are considered one of the main access gateways for universities. Typically, they have a large candidate audience among the current students, employees, and faculty members aside from previous and future students, employees, and faculty members. Web accessibility is the concept of providing web content universal access to different machines and people with different ages, skills, education levels, and abilities. Several web accessibility metrics have been proposed in previous years to measure web accessibility. We integrated and extracted common web accessibility metrics from the different accessibility tools used in this study. This study evaluates web accessibility metrics for 36 Jordanian universities and educational institute websites. We analyze the level of web accessibility using a number of available evaluation tools against the standard guidelines for web accessibility. Receiver operating characteristic quality measurements is used to evaluate the effectiveness of the integrated accessibility metrics.**

Keywords—web accessibility, web ranking, web evaluation, web testing

I. INTRODUCTION

University webmasters enhance their methods and techniques to present their content in the best possible way and to enable users to access and satisfy their needs easily. Search engines use many factors that affect web accessibility, thereby allowing the enhancement and improvement of these factors to further improve the websites' level of appearance in the search engine results page (SERP). Furthermore, university websites require obtaining the highest possible position in the SERP to increase their number of users (depending on web search engines such as Google, Bing, and Yahoo) who search for information. Web accessibility and search engine optimization share many features, such as using keywords in the following HTML tags: headings (e.g. <h1>, <h2> etc.), page titles <title>, anchor texts <a>, and alt attributes on images, which play an important role in improving the ranking of websites. Ivory et al.'s [1] study indicates that results of web search engines for blind users should be different from the results dedicated to normal users because the relevance of blind users is different from that of normal users relative to a certain query. Therefore, web search engines must consider web accessibility as a factor affecting their ranks for disabled web search engine users. Google Inc. launched in 2006 an experimental search engine called Google Accessible Search Engine (http://www.google.com/accessibility/labs/search/), but this search engine is no longer supported by Google Inc. An observational study on search behavior is conducted by Sahib et al. [2] to design and implement more accessible and usable search interfaces for both visually impaired and sighted web searchers. They further propose in their study the necessary design guidelines for interfaces to make them usable and accessible to screen readers. Using web accessibility within commercial websites by webmasters may enable more people to visit these websites and, in most cases, more customers (more sales). Furthermore, adopting web accessibility by webmasters means that websites that will be easier to maintain, update, and redesign and different devices will be improved. Web accessibility has a positive effect on search engine optimization (SEO). Moreno and Martinez indicate in their study that accessible web pages regularly appear within the top of SERP without applying SEO techniques because of similarities and overlapping characteristics between many SEO factors and web accessibility guidelines [3]. Commercially, decision makers consider struggling for accessibility does not bring return on investment, that is, the costs of ensuring accessibility are larger than the revenue provided by new customers. Independently, adopting web accessibility does not mean easy maintenance upgrade and design.

Web accessibility means making the web more accessible to people with different abilities. Furthermore, different machines, including assistive technologies, can be used to access the web. Web accessibility enables disabled people/people with special needs to use the web, thereby making web accessibility an important issue in web development. The disabilities affecting web access includes visual (blindness, Kalnienk vision, and low vision), auditory (hard of hearing, deafness, and deaf-blindness), speech (low speech, high speech, stuttering problems, influent, articulation problems), physical (arthritis, Parkinson's disease, essential

This research was funded by the Deanship of Research at Zarqa University/ Jordan.

tremor, multiple sclerosis, broken arm), learning (emotional disturbance, intellectual disability, dyslexia), cognitive, and neurological disabilities. Web accessibility is defined as the ability of users to universally access websites and obtain their information needs. According to the World Health Organization (WHO), over a billion people (approximately 15% of the world's population) have some form of disability [4]. Thus, this issue is an important one on web accessibility for this portion of the world population.

Web accessibility overlaps with but is not the same as usability, and it is different from device independence. A number of individuals and organizations have proposed various standards and guidelines for content accessibility throughout the web's history. Only the Web Content Accessibility Guidelines (WCAG) have been reviewed by the World Wide Web Consortium (W3C) members, software developers, and other W3C groups and interested parties, and are endorsed by the W3C director as web standards. The goals of web accessibility guidelines have to be achieved to assist us in developing and evaluating web accessibility.

WCAG 1.0 was released on May 5, 1999 and WCAG 2.0 was released on December 11, 2008 to assist web programmers to build web pages accessible to disabled people and web content accessible from different environments or platforms. They later became an ISO standard. WCAG 1.0 and WCAG 2.0 were proposed by W3C (www.w3.org). Each WCAG 1.0 guideline is divided into checkpoints, which indicate good practices for constructing accessible web content and assist web developers in avoiding barriers that may prevent users from accessing the web. Compared with that of WCAG 1.0, the scope of WCAG 2.0 is broadly applicable to modern and future web technologies. Furthermore, WCAG 2.0 can be evaluated more accurately, manually, and automatically. In other words, it is more testable than WCAG 1.0. International efforts contribute to the production of WCAG 2.0 to harmonize a single standard for web content. WCAG 1.0 and WCAG 2.0 are currently the most broadly accepted qualitative measures for web accessibility [5–6].

Universities, as an essential part of the world of academia, have to be universal and include all disabled people. Therefore, their websites have to be universal and accessible to all web users. Many studies have focused on evaluating the web accessibility features of many websites. However, we observe that web accessibility plays an important role in enhancing the ranking of websites. A study by Schmetzke finds that only a few American university websites are accessible [7], but later studies show that most of these websites are adopting accessibility policies. In this study, we aim to discover the accessibility of Jordanian university websites by analyzing and evaluating the metrics of web accessibility as a case study.

Vigo et al. [8] find that evaluations based solely on automated web accessibility tools are unreliable in terms of finding all errors or reporting errors that do not exist. Nevertheless, these tools are used in this study to analyze extracted web accessibility features using various tools and to find the common main shared features among the different tools. In the case study, we use selected features to evaluate web accessibility levels in a case study of university websites and higher educational institutes in Jordan. We further evaluate these features using four machine learning classifiers.

The rest of this paper is organized as follows. Section 2 presents the literature review and discusses the issue of web accessibility metrics for Jordanian universities. Section 3 discusses the methodology, and Section 4 presents the experiments and results. Section 5 outlines the main findings of this study and the planned future works.

II. Related Work

Recently, interest in measures of web accessibility has grown, and thus the literature has witnessed a substantial increase in the number of web accessibility metric studies. This section presents the previous studies on web accessibility metrics using different methods. The last part of this study presents some studies conducted on Arabic websites.

The first published study on accessibility measurement is that by Sullivan and Matson [9]. In their study, they mention that 95% or more of all websites are inaccessible and that it is a huge problem facing disabled users of the web. This finding indicates a clear ignorance to the issues of universal design and content accessibility. Therefore, they conduct a content accessibility compliance audit of the top 50 websites' most highly trafficked sites according to Alexa.com to determine whether these websites are accessible or not. They use a Lift Online tool in their study and determine if a substantive relationship exists between content accessibility and usability through the Spearman rank–order correlation coefficient. The statistical analysis shows that a weak relationship exists between content accessibility and overall usability [9]. They use the failure rate metric that considers the concept of "potential problems," and this metric for a given web page simply represents the ratio between the total number of real errors in a web page and the total number of potential errors. Furthermore, they use another metric that penalizes web pages with a large number of elements that cause accessibility errors. This metric is calculated by multiplying the number of accessibility opportunities (potential points of failure) by their failure rate. Penalizing accessibility opportunities is caused by the probable inclusion of accessibility barriers.

The Web Quality Evaluation Method (WebQEM) was proposed for the first time by Olsina [10] and then used by Olsina and Rossi [11] to evaluate and compare quality requirements for websites and applications. González et al. [12] adopt another important accessibility metric that considers the concept of "weight" for barriers. They aim to enhance the accessibility of web pages for visually impaired web users. This important metric is based on the WebQEM model, and a global ratio is calculated for a given web page. This global ratio is later multiplied by a weight, which is defined according to the effect of each barrier. The concept of barrier weight coefficients was first proposed by [12], and thus, later metrics used this concept.

Parmanto and Zeng propose another quantitative metric called the Web Accessibility Barrier (WAB) to measure the content accessibility of different web pages for disabled web users [13]. WAB considers the concept of potential problems and weights for barriers. The WAB metric considers the size

(total number of pages) contained in a given website. A high WAB score indicates that barriers exist, and a low WAB score indicates that the website under study complies better with the WCAG guidelines. A WAB score of zero means that the website under study has no barriers. The study conducted by Hackett and Parmanto [14] refers to the WAB score as a proxy of web accessibility and concludes that WAB is unable to differentiate between barriers posing minimal limitations and those posing absolute inaccessibility.

The Unified Web Evaluation Methodology (UWEM) was developed in Europe by the Web Accessibility Benchmarking (WAB) Cluster to be mainly used there. It was developed to become a standard for evaluating web accessibility. UWEM is a completely automatic accessibility metric; therefore, it provides its users with suitable methods and advice to carefully evaluate a set of websites or a single website. For a single web page, the UWEM final value depicts an approximation of the probability of discovering a barrier in a website that could prevent a user from finishing a specific task [15][16]. In their study, Buhler, Heck, Perlick, Nietzio, and Ulltveit-Moe [15] observe that WAB does not support different disability groups. Furthermore, web pages with a low number of various barriers are considered by WAB to be more accessible, and this aspect is undesirable. Accordingly, they propose a new aggregation metric (A3) to adapt the measurement to different disability groups, and this metric represents an improvement of UWEM 0.5 [15]. The A3 metric is similar to other metrics on the verification of checkpoint conformance. It uses some probability properties and aggregated some issues related to the complexity of the web page under consideration. The A3 metric considers the number of violations of a given checkpoint in relation to the total number of violations [15].

The study of Vigo, Arrue, Brajnik, Lomuscio, and Abascal [17] shows the importance of quantitative accessibility measurements and proposes the three different applications: information retrieval, quality assurance within web engineering, and accessibility monitoring. They propose an automatic quantitative metric to evaluate accessibility called Web Accessibility Quality Metric (WAQM) based on the reports of automatic evaluation tools. Fifteen websites (1363 web pages) and two automatic evaluation tools (EvalAccess and LIFT) are used to verify the reliability of their proposed metric. They conclude that their metric results are highly dependent on evaluation tools, and a high correlation exists among the results of different tools. Therefore, they deduce that their metric can be used by information retrieval systems to rank results, and that this metric is beneficial for accessibility monitoring scenarios and partially beneficial for web engineering scenario [17]. A total of 918 web pages belonging to 10 European, United States, and African university websites, as well as 445 web pages belonging to five newspaper websites, are utilized in their tests.

Freire, Fortes, Turine, and Paiva's study [18] reviews six web accessibility metrics used in previous years and compares them. They discuss the strengths and pitfalls of these six web accessibility metrics. Therefore, they present the first known web accessibility metric (Failure Rate) [9].

Buenadicha et al. [19] examine accessibility as a subcategory of web assessment index, which includes the following four categories: speed, navigability, and content, apart from accessibility. Through a detailed literature review, these authors identify the key factors considered as determinants of website quality and use them in their index to evaluate all websites of Spanish universities.

Kane et al. [20] present an evaluation of the previous state of the university website accessibility of the 100 top universities' home pages worldwide. They analyze the compliance of these 100 home pages with image accessibility, accessibility standards, text-only content, quality of web accessibility statements, and alternate language. Their study is limited to and based only on the analysis of only 100 web pages. The results of their study [20] show that many top universities have accessibility problems, and a significant variation in accessibility exists among these universities across different countries and geographic regions. This study [20] concludes that the accessibility of websites of universities in non-English-speaking countries is either low or does not exist.

Many accessibility studies include US higher education establishments, such as those conducted by Harper et al. [21], and aim to raise awareness on accessibility issues in higher education websites. These authors invited webmasters of higher education institutions to evaluate the overall accessibility of their websites using freeware. Study [21] indicates that most of the university homepages under study were non-compliant with the WCAG, and that only one establishment satisfied all W3C guidelines and gained a Triple A. Bradbard et al. [22] examine the accessibility of 58 well-known US universities. Data from Peterson's Four-year Colleges (2007) on these top 58 US universities were used in their study. Results show that only 50 of these universities adopt accessibility policies and that 78% and 88% of these universities neglect the timeframe to implement their policies and violate these standards, respectively. According to [22], only two US universities (Purdue University and University of California) have good accessibility policies that could serve as models for other universities worldwide.

Accessibility in higher education institutions is not restricted to websites as it also includes Learning Content Management Systems (LCMSs). Therefore, a number of web-based open-source LCMSs have been explored by researchers. One of these studies was conducted by Iglesias et al. [23] to evaluate three web-based open-source LCMSs (ATutor 1.6.2, Moodle 1.9.4, and Sakai 2.6.0). This type of study requires an assessment and monitoring of LCMS accessibility to guarantee the universal accessibility of this type of systems. Similar studies were conducted by [24], [25], and [26], among others.

III. METHODOLOGY

A convenience sample of Jordanian higher education websites was studied for the year 2015. In this study, we used various web accessibility tools to analyze the web accessibility metrics for the Jordanian university websites and to evaluate their level of web accessibility. The following steps describe the methodology of this study:

1) Select a number of popular web accessibility tools according to the W3C (http://www.w3.org/WAI) guidelines, which provides various recourses belonging to web accessibility guidelines, tools, and standards.

2) Apply the selected tools on Jordanian universities websites as a case study.

3) Extract a number of web accessibility features and metrics, and find common and shared features among all the tools.

4) Evaluate the selected shared metrics using the receiver operating characteristic (*ROC*) quality measurements. *ROC* is an essential evaluation of prediction metrics used to identify the possible best selected metrics.

A. This study is based on a dataset that consists of 36 Jordanian universities, including 9 public universities and 27 private universities and institutes. Each university website under study is represented by the two most visited web pages, namely, the homepage and the registration web page (for a typical university website). The total available web pages are 72. Table I lists the names of the university websites under study excluding the type of university, whether it is public or private.

TABLE I. LIST OF JORDANIAN UNIVERSITIES, COLLEGES, AND CENTERS UNDER STUDY

I	University Name	University Type
1	University of Jordan	Public
2	Yarmouk University	Public
3	Jordan University of Science & Technology	Public
4	Hashemite University	Public
5	Mutah University	Public
6	Al Balqa Applied University	Public
7	Al Al-Bayt University	Public
8	Al Hussein bin Talal University	Public
9	Tafila Technical University	Public
10	German Jordanian University	Public
11	Jordan Institute of Diplomacy	Public
12	Princess Sumaya University for Technology	Private
13	University of Petra	Private
14	American University of Madaba	Private
15	Philadelphia University at Jordan	Private
16	Zarqa University	Private
17	Al Isra Private University Amman	Private
18	Irbid National University	Private
19	Amman Arab University	Private
20	Al Ahliyya Amman University	Private
21	Al Zaytoonah University	Private
22	Applied Science University	Private
23	Middle East University Jordan	Private
24	Jerash Private University	Private
25	Ajloun National University	Private
26	Al Quds College	Private
27	American Language Centre	Private
28	Oval Office for Studies and Research	Private
29	Jordan Academy of Music Higher Institute of Music	Private
30	Jordan Applied University College of Hospitality and Tourism Education	Private
31	Institute of Banking Studies	Private
32	Queen Noor Civil Aviation Technical College	Private
33	World Islamic Sciences and Education University	Private
34	Jadara University	Private
35	Jordan Media Institute	Private
36	Arab Academy for Banking and Financial Sciences	Private

TABLE II. WCAG 2.0 (LEVEL AA) FOR A-CHECKER TOOL [27]

WCAG 2.0 (Level AA)	Guideline	Description
1.1	Text Alternatives	Provide text alternatives for any non-text content.
1.2	Time-based Media	Provide alternatives for time-based media.
1.3	Adaptable	Create content that can be presented in different ways (for example simpler layout) without losing information or structure.
1.4	Distinguishable	Make it easier for users to see and hear content including separating foreground from background.
2.1	Keyboard Accessible	Make all functionality available from a keyboard.
2.2	Enough Time	Provide users enough time to read and use content.
2.3	Seizures	Do not design content in a way that is known to cause seizures.
2.4	Navigable	Provide ways to help users navigate, find content, and determine where they are.
3.1	Readable	Make text content readable and understandable.
3.2	Predictable	Make Web pages appear and operate in predictable ways.
3.3	Input Assistance	Help users avoid and correct mistakes.
4.1	Compatible	Maximize compatibility with current and future user agents, including assistive technologies.

B. Accessibility guidelines present the methods to improve website visibility for users. We select the following seven free web accessibility tools to be used in this study:

1) A Checker tool: This tool is used to assess web accessibility problems for the HTML content, where the complete HTML source code for each web page under study is

pasted. AChecker tool divides problems into the following three main parts: Known, Likely, and Potential problems [27]. Table II presents the web accessibility guidelines (WCAG 2.0 [Level AA]) for this tool.

2) Cryptzone Cynthia Says is a joint education and outreach project of Cryptzone portal and the Internet Society Disability and Special Needs Chapter. It is used to evaluate and provide feedback on website accessibility according to the US Access Board's Section 508 or the W3C's WCAG 2.0 A-AAA Accessibility Guidelines [28].

3) Functional Accessibility Evaluator (FAE) is developed to assess web pages' accessibility according to the W3C Web Content Accessibility Guidelines 2.0 Level A and AA requirements using the accessibility features and techniques associated with W3C ARIA 1.0 and HTML5 specifications [29].

4) HERA is designed to evaluate web pages' accessibility according to the specification WCAG 1.0. HERA applies a set

5) of tests on the web page and finds any automatically detectable problems, which require further manual verification [30].

6) Validator tool from the W3 organization is used to evaluate the mark-up validity of web documents in HTML, XHTML, SMIL, MathML, and so on [31].

7) Wave WebAIM tool is used to assess and enhance web developers to improve the accessibility of their web content. It is easy to use by simply entering a URL and waiting for the results [32].

8) TAW is an online free tool used to evaluate website accessibility according to the W3C WCAG 1.0 [33].

IV. Experiments and Results

In our experiment, we applied these seven free online tools on Jordanian university websites. Thus, for every website, we tested the two most visited web pages, namely, the homepage and the registration web page, which are considered highly visited web pages by users. We combined the features for every tool and found the common features among all the tools. Then, we summarized the values for each university. The dataset was divided manually into three main categories based on the optimal and worst results (high, medium, and low). We evaluated the selected metrics using *ROC* quality measurements to assess the effectiveness of the common extracted features. Fig. I shows part of the homepage of Zarqa University using the AChecker tool.

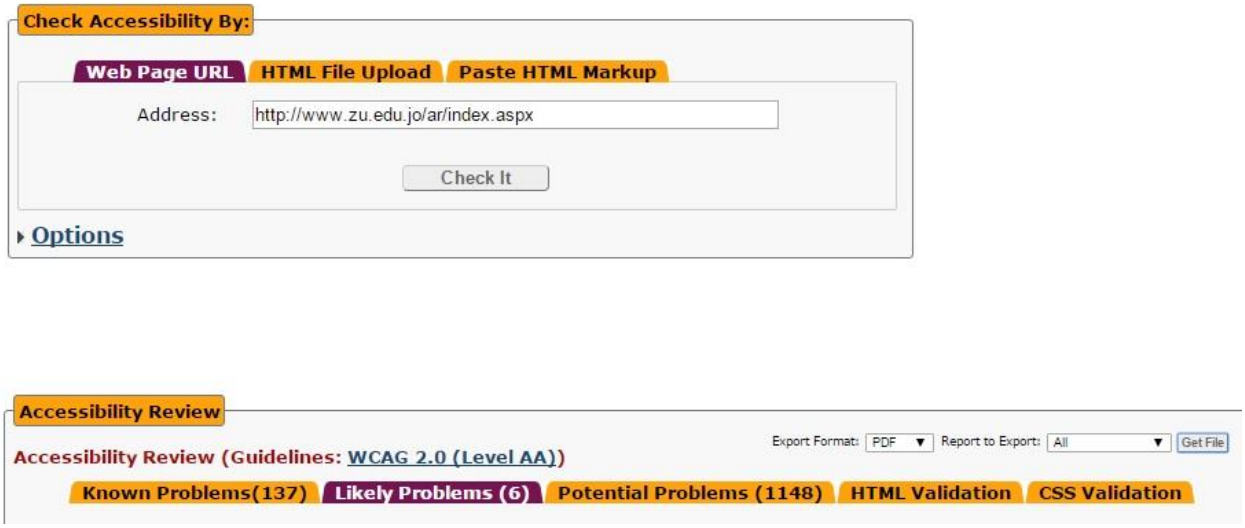


Fig. 1. First GUI of A-Checker tool

Fig. II shows part of the Yarmouk University homepage results using the Cryptzone Cynthia Says tool.

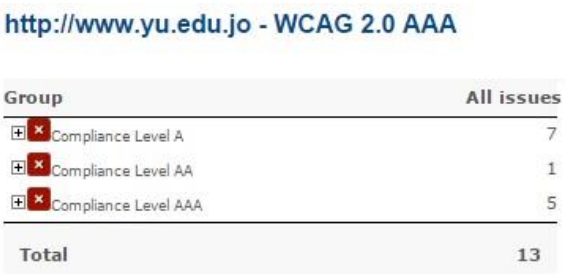


Fig. 2. Cryptzone Cynthia Says tool GUI

Fig. III shows the University of Jordan homepage results using the FAE tool.

Fig. III presents the FAE tool that defines the number of rules according to web accessibility guidelines, such as heading, tables, links, and styling content. The FAE tool provides detailed results for every rule. Fig. IV illustrates the HERA results of the German Jordanian University.

The University of Jordan :: Amman :: Jordan

Ruleset: HTML5 and ARIA Techniques

URL: http://ju.edu.jo/home.aspx

	Violations	Warnings	Manual Checks	Passed
Number of Rules	9	3	27	4

Default Sort

	Number of Rules			
Rule Category	**V**	**W**	**MC**	**P**
Landmarks	3	-	2	-
Headings	1	2	-	1
Styling/Content	-	-	8	1
Images	1	-	4	2
Links	1	1	1	-
Tables	1	-	-	-
Forms	-	-	-	-
Widgets/Scripting	2	-	-	-
Audio/Video	-	-	1	-
Keyboard Support	-	-	4	-
Timing	-	-	3	-
Site Navigation	-	-	4	-
Totals	9	3	27	4

Fig. 3. Summary of the results for the Jordanian University using the FAE tool

Summary

- *URL*: http://www.gju.edu.jo/
- *Date/time*: 11/10/2015 - 18:20 GMT
- *Total*: 859 elements
- *Automatic analysis*: 25 seconds
- *Errors*: **8 errors**
- **To check manually: 41 checkpoints**
- *Tester*: (unknown)
- *Navegador*: Sin identificar

Navigate by results

Use the links in the table to test each of the checkpoints manually or to check the results of automatic testing

Status of checkpoints

Priority	Needs checking	Pass	Fail	N/A
P1 HERA WCAG 1.0	9	--	1	7
P2 HERA WCAG 1.0	19	2	4	4
P3 HERA WCAG 1.0	13	1	3	2

Fig. 4. Summary of the results for the Jordanian University using the HERA tool

The HERA tool provides detailed results for every point in the results. The green tick symbol means that HERA has a positive reflection of applying web accessibility, the x-mark refers to missing web accessibility metrics, and the gray tick symbol refers to inapplicable metrics. Fig. V presents part of the results for Al Al-Bayt University using the Validator tool.

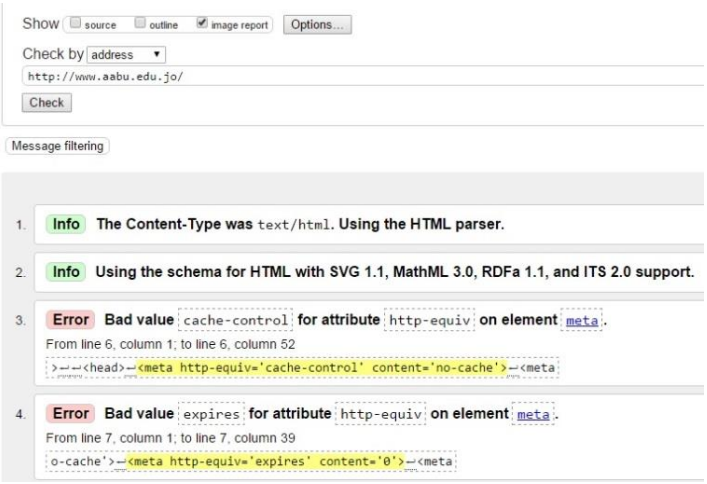


Fig. 5. Results for the Al Al-Bayt University using the Validator tool

Fig. VI illustrates part of the results for the World Islamic Sciences and Education University using the Wave WebAIM tool, which presents a summary of errors, alerts, features, and the structural element.

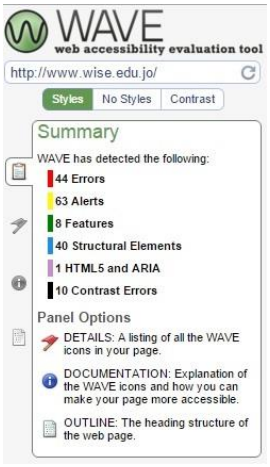


Fig. 6. Summary of the results for the WISE University using the Wave WebAIM tool

Fig. VII. Presents part of the results for Al Quds College using the TAW tool.

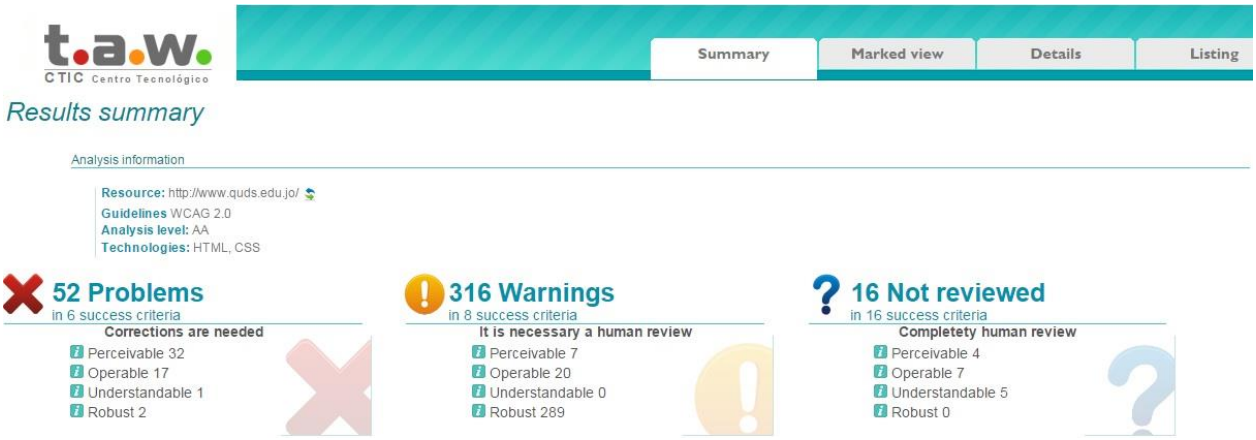


Fig. 7. Summary of the results for Al-Quds College using the TAW tool

The TAW free online tool presents a summary of the results as shown in Fig. VII. Through this tool, we can obtain the detailed results for every note. Applying the seven tools on the dataset was part of the process to find the common accessibility features among the different tools. Table III presents the combined features for web accessibility.

TABLE III. COMBINED FEATURES FOR WEB ACCESSIBILITY [27 – 34]

Feature	Description
1. Images using alt element.	Verify if the images using alt element or keep it empty.
2. Distribute the color.	The harmony between using font and the background.
3. Web page title element.	The title should provide description for the content of the Web page.
4. Navigate to relevant websites.	Check whether that the Website points to another universities.
5.Using of Heading elements.	There is a balance needed to using headings tags on a Web page, to provide useful structure and outlines to users.
6. Primary language.	Language tags use a primary code to indicate the language such as: language and XML: language attributes. Furthermore, optional sub codes to indicate variants of the language.
7. Using submit buttons.	Provide a technique that allow users to explicitly request changes of context, such as: generate an http to submit data that entered in a form.
8. Labels description.	Ensure that the label for any interactive component within Web content makes the components aims transparent.
9. Enough Time.	Provide users enough time to read and use content.
10. Readable.	Make text content readable and understandable.
11. Using anchor tags.	This allow the users to navigate in large Web pages easily.
12. Image links.	Ensure that the image link missing or available.
13. Web page size.	Size in Kilobytes.
14. Content Visibility.	The visible page fraction inside the <page> element, against hidden text inside a specific Web page.

15. Compress.	The total size of compressed files inside a specific Web page, and the total size of compression ratio inside a specific Web page.
16. Images.	Total number of Images inside a specific Web page.
17. Image size.	Total Image Size.
18. Structural elements.	Ensure structural elements such as: scroll, section, header, footer, article, and aside elements, whether missing or available.
19. Style sheets.	Style sheets code is correct or not.
20. Broken link.	The total number of broken links in Web page.
21. Empty link.	Total number of links without anchor text, and anchor text without links within the Web page.
22. Redirected link.	The total number of redirected links in Web page.

We used these features that we consider the main web accessibility metrics to rank our dataset. We divided the data into three main categories based on the optimal and worst results (i.e., high, medium, and low). Table IV shows the ranking of universities based on the combined features.

TABLE IV. RANKING OF UNIVERSITIES USING THE COMBINED FEATURES

University Rank	**Category Level**
1. Yarmouk University	High
2. Mutah University	High
3. Queen Noor Civil Aviation Technical College	High
4. Middle East University Jordan	High
5. Al Isra Private University Amman	High
6. Philadelphia University at Jordan	High
7. German Jordanian University	High
8. Jordan Academy of Music Higher Institute of Music	High
9. American Language Centre	High
10. World Islamic Sciences and Education University	High
11. Institute of Banking Studies	High
12. University of Jordan	High
13. Princess Sumaya University for Technology	High
14. Jadara University	High
15. Jordan University of Science & Technology	High
16. Tafila Technical University	Medium
17. Al Hussein bin Talal University	Medium
18. Ajloun National University	Medium
19. University of Petra	Medium
20. Amman Arab University	Medium
21. Irbid National University	Medium
22. Jerash Private University	Medium
23. American University of Madaba	Medium
24. Al Al-Bayt University	Medium
25. Jordan Applied University College of Hospitality and Tourism Education	Medium
26. Al Quds College	Low
27. Zarqa University	Low
28. Al Zaytoonah University	Low
29. Applied Science University	Low
30. Al Balqa Applied University	Low
31. Jordan Institute of Diplomacy	Low
32. Jordan Media Institute	Low
33. Oval Office for Studies and Research	Low
34. Arab Academy for Banking and Financial Sciences	Low
35. Hashemite University	Low
36. Al Ahliyya Amman University	Low

Then, we used the ROC quality measurements to evaluate the effectiveness of the following selected features after the 10-fold cross-validation process: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision, Recall, and F-Measure (F-M). The ROC represents a graphical plot that illustrates the performance of a binary classifier system [35]. Formula 1 represents the Accuracy formula [35]:

$$Accuracy_i = \frac{TP + TN}{TP + FP + TN + FN} \quad . (1)$$

Formula 2 represents the Recall formula [35]:

$$Recall_i = \frac{TP}{TP + FN} \quad . (2)$$

Formula 3 represents the Precision formula [35]:

$$Precision_i = \frac{TP}{TP + FP} \quad . (3)$$

Formula 4 represents the F-Measure formula [35]:

$$F-measure = \frac{2(TP)}{2(TP)+FP+FN} \quad . (4)$$

Table V presents the experiment results for the selected common features using the *K* nearest neighbor (*K-NN*) when the $k = 1$ classifier yields an accuracy of 88.9%.

TABLE V. DETAILED RESULTS FOR *K-NN*

Class	*TP*	*FP*	*Precision*	*Recall*	*F-Measure*	*ROC*
High	0.933	0.048	0.933	0.933	0.933	0.963
Medium	0.800	0.077	0.800	0.800	0.800	0.865
Low	0.909	0.040	0.909	0.909	0.909	0.942
Weighted AVG	0.889	0.053	0.889	0.889	0.889	0.930

Table VI presents the results for the selected common features using the support vector machine (*SVM*).

TABLE VI. DETAILED RESULTS FOR THE *SVM*

Class	*TP*	*FP*	*Precision*	*Recall*	*F-Measure*	*ROC*
High	0.867	0.429	0.591	0.867	0.703	0.798
Medium	0.300	0.192	0.375	0.300	0.333	0.513
Low	0.545	0	1	0.545	0.706	0.873
Weighted AVG	0.611	0.232	0.656	0.611	0.601	0.742

The detailed results for the *SVM* show high accuracy results for the high class, medium accuracy results for the medium class, and very low accuracy results for the low class. Thus, the overall weighted average results yield an accuracy of 61.11%.

Table VII presents the results for the selected common features using Decision Tree (*J*48). *J*48 yields an accuracy of 94.4% and an error rate of 5.6%.

TABLE VII. DETAILED RESULTS FOR *J48*

Class	*TP*	*FP*	*Precision*	*Recall*	*F-Measure*	*ROC*
High	1	0.048	0.938	1	0.968	0.976
Medium	0.9	0.038	0.9	0.9	0.9	0.931
Low	0.909	0	1	0.909	0952	0.955
Weighted AVG	0.944	0.031	0.946	0.944	0.944	0.957

TABLE VIII. DETAILED RESULTS FOR BAGGING CLASSIFIER

Class	*TP*	*FP*	*Precision*	*Recall*	*F-Measure*	*ROC*
High	1	0.048	0.938	1	0.968	0.957
Medium	0.9	0	1	0.9	0.947	0.908
Low	1	0	1	1	1	1
Weighted AVG	0.972	0.02	0.974	0.972	0.972	0.957

Table VII shows that *J*48 yields better results than the previous classifiers *K-NN* and *SVM*.

Table VIII presents the results using a bagging classifier, which splits the dataset into many sub-datasets. The bagging classifier computes the prediction in each sub-dataset, selects the most frequently predicted results, and finally considers them as the final dataset prediction [35]. Table VIII indicates that the bagging classifier yields an accuracy of 97.2% and an error rate of 2.8%, with the low class providing the optimal results for all the measurements. The bagging classifier shows better results than the three previous classifiers.

Finally, Table IX presents the effectiveness measurements of the four previous classifiers, namely, Kappa Statistic (*KS*), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE).

TABLE IX. DETAILED RESULTS FOR THE BAGGING CLASSIFIER

Classifier	***KS***	***MAE***	***RMSE***	***RAE***	***RRSE***
K-NN	0.8306	0.1045	0.2633	23.799%	56.1349%
SVM	0.3854	0.3272	0.4182	74.498%	89.1731%
J48	0.9149	0.037	0.1925	8.4337%	41.0356%
Bagging	0.9574	0.0802	0.1888	18.2531%	40.2507%

The results in Table IX indicate that the bagging classifier yields the best KS effectiveness measure results for the selected features, whereas SVM is the lowest classifier with *KS* value.

V. CONCLUSION AND FUTURE WORK

Websites are important especially for higher educational institutes. They are considered as the main gateway to the world. The academic ranking of a university in general can be dependent on the quality of its main website and its ability to provide relevant information and services to users. Web accessibility is considered one of the major important quality goals in designing websites in particular and software applications in general. It ensures that a developed website can be equally accessed by a large category of users regardless of

their physical abilities, skills, locations, languages, backgrounds, and so on.

In this paper, we evaluated most of the websites of Jordanian universities in terms of accessibility. The results showed a significant number of weaknesses in most of the universities. Furthermore, a variation of web accessibility standards was found when the websites were measured using different accessibility tools.

ACKNOWLEDGMENTS

This research was funded by the Deanship of Research and Graduate Studies in Zarqa University.

REFERENCES

[1] M. Y. Ivory, S. Yu and K. Gronemyer, "Search result exploration: a preliminary study of blind and sighted users' decision making and performance," in Proceedings of the CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04). ACM, New York, NY, USA, pp. 1453-1456, 2004.

[2] World Health Organization (WHO), Disability and health, Fact sheet No. 352, December 2014. [cited 12 Febraury 2016]. Available from: http://www.who.int/mediacentre/factsheets/fs352/en/

[3] L. Moreno and P. Martinez, "Overlapping factors in search engine optimization and Web accessibility," Online Information Review, 37 (4), pp. 564-580, 2013.

[4] Web Content Accessibility Guidelines 1.0, 1999. [cited 12 Febraury 2016]. Available from: http://www.w3.org/TR/WCAG10/

[5] Web Content Accessibility Guidelines (WCAG) 2.0, 2008. [cited 12 Febraury 2016]. Available from: http://www.w3.org/TR/WCAG20/

[6] How WCAG 2.0 Differs from WCAG 1.0, 2009. [cited 12 Febraury 2016]. Available from: http://www.w3.org/WAI/WCAG20/from10/diff.php

[7] A. Schmetzke, "Web accessibility at university libraries and library schools," Library Hi Tech, Vol. 19 No. 1, pp. 35-49, 2001.

[8] T. Sullivan and R. Matson, "Barriers to Use: Usability and Content Accessibility on the Web's Most Popular Sites," in Proceedings of the Conference of Universal Usability, ACM, 6 pages, 2000.

[9] J. González, M. Macías, R. Rodríguez and F. Sánchez, "Accessibility Metrics of Web pages for Blind End-Users," in Proceedings of the 2003 International Conference on Web Engineering, Oviedo, Spain, Lecture Notes in Computer Science, Vol. 2722, Springer Berlin / Heidelberg, pp. 374-383, 2003.

[10] B. Parmanto and X. Zeng, "Metric for Web Accessibility Evaluation," Journal of the American Society for Information Science and Technology. Vol. 56, Issue 13, pp. 1394–1404, 2005.

[11] C. Buhler, H. Heck, O. Perlick, A. Nietzio and N. Ulltveit-Moe, "Interpreting Results from Large Scale Automatic Evaluation of Web Accessibility," In K. Miesenberger, J. Klaus, W. Zagler, A. Karshmer (Eds.), Computers Helping People with Special Needs, Springer-Verlag Berlin Heidelberg, volume 4061/2006, pp. 184–191, 2006.

[12] M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio and J. Abascal, "Quantitative Metrics for Measuring Web Accessibility," in Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A), pp. 99-107, 2007.

[13] A. P. Freir, R. P. M. Fortes, M. A. S. Turine, and D. M. B. Paiva, "An Evaluation of Web Accessibility Metrics based on their Attributes," in Proceedings of the 26th annual ACM international conference on Design of communication (SIGDOC '08), ACM, New York, NY, USA, pp. 73-80, 2008.

[14] S. Hackett and B. Parmanto, "A longitudinal evaluation of accessibility: higher education Websites," Internet Research, 15(3), pp. 281–94, 2005.

[15] A. Nietzio, C. Strobbe and E. Velleman, "The Unified Web Evaluation Methodology (UWEM) 1.2 for WCAG 1.0," in K. Miesenberg, J. Klaus, W. Zadger, and A. Karshmer (Eds.), Lecture notes in computer science. International conference on computers helping people with special needs, Berlin, Germany: Springer-Verlag, Vol. 5105, pp. 394–401, 2008.

[16] A. P. Freir, T. J. Bittar and R. P. M. Fortes, "An approach based on metrics for monitoring Web accessibility in Brazilian municipalities Websites," in Proceedings of the 2008 ACM symposium on Applied computing (SAC '08). ACM, New York, NY, USA, pp. 2421-2425, 2008.

[17] L. Olsina, "Web Engineering: A Quantitative Methodology for Quality Evaluation and Comparison of Web Applications," Doctoral Thesis (in Spanish), Ciencias Exactas School, UNLP, La Plata, Argentina, 2000.

[18] L. Olsina and G. Rossi, "Measuring Web Application Quality with WebQEM," IEEE MultiMedia, 9(4), pp. 20–29, 2002.

[19] M. Buenadicha, A. Chamorro, F.J. Miranda and O.R. Gonzalez, "A new Web assessment index: Spanish universities analysis," Internet Research: Electronic Application and Policy, 11(3), pp. 226-234, 2001.

[20] S. K. Kane, J. A. Shulman, T. J. Shockley and R. E. Ladner, "A Web accessibility report card for top international university Websites," in Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A) (W4A '07). ACM, New York, NY, USA, pp. 148-156, 2007.

[21] K. Harper, J. DeWaters, "A Quest for Website accessibility in higher education institutions," The Internet and Higher Education, 11, pp. 160–164, 2008.

[22] Bradbard D. A., Peters C. and Y. Caneva, "Web accessibility policies at land-grant universities," The Internet and Higher Education, 13(4), pp. 258-266, 2010.

[23] A. Iglesias, L. Moreno, P. Martínez and R. Calvo-Martin, "Evaluating the Accessibility of Three Open-Source Learning Content Management Systems: A Comparative Study," Computer Applications in Engineering Education, 19(2), 320–328, 2011.

[24] R. Calvo, A. Iglesias, and L. Moreno, "Is Moodle Accessible for Visually Impaired People?" in Web Information Systems and Technologies, pp. 207-220, 2012.

[25] R. Calvo, A. Iglesias and L. Moreno, "Accessibility barriers for users of screen readers in the Moodle learning content management system," Universal Access in the Information Society, 13(3), pp. 315-327, 2014.

[26] T. Calle-Jimenez, S. Sanchez-Gordon and S. Luján-Mora, "Web Accessibility Evaluation of Massive Open Online Courses on Geographical Information Systems," in Proceedings of the 2014 IEEE Global Engineering Education Conference (EDUCON), 2014, pp. 680-686.

[27] AChecker. [cited 12 Febraury 2016]. Available from: http://achecker.ca/checker/index.php CynthiaSays. [cited 12 Febraury 2016]. Available from: http://www.cynthiasays.com/

[28] Functional Accessibility Evaluator 2.0: Testing (version 0.9.9). [cited 12 Febraury 2016]. Available from: http://fae20.cita.illinois.edu/

[29] Testing Accessibility with Style. [cited 12 Febraury 2016]. Available from: http://www.sidar.org/hera

[30] The W3C Markup Validation. [cited 12 Febraury 2016]. Available from: http://validator.w3.org/

[31] Wave WebAIM tool. [cited 12 Febraury 2016]. Available from: http://wave.Webaim.org/.

[32] TAW. [cited 12 Febraury 2016]. Available from: http://www.tawdis.net/ingles.html

[33] W3C. [cited 12 Febraury 2016]. Available from: http://www.w3.org/

[34] I. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann Series in Data Management Systems, 2005.

Ontology for Academic Program Accreditation

Ontology of Accreditation Board of Engineering and Technology(ABET) Process

Jehad Sabri Alomari
Taif University, KSA
Taif, KSA

Abstract—**Many educational institutions are adopting national and international accreditation programs to improve teaching, student learning, and curriculum. There is a growing demand across higher education for automation and helpful educational resources to continuously improve student outcomes. The student outcomes are the required knowledge and skill set that graduates of any accredited program have to gain in order entry into the workforce or for to continue with their future education. To evaluate student outcomes, each assessment activities must map to a course learning outcomes which maps students' outcomes. The problem is that all course learning outcomes and student outcome mapping are placed in documents or database which requires more work and time to access and understand. This paper proposes an ontology based solution to enable visual discover of all course learning outcomes that maps to a particular student outcome and related assessments to help faculty or curriculum committees avoid over mapping or under mapping students' outcomes.**

Keywords—Accreditation; Ontology; Semantic Web; classification; Education

I. INTRODUCTION

Ontologies have played a major role in knowledge representation in many domains and considered one of the pillars of semantic web (Vocabulary). Semantic web is a formal conceptualization that represents new technologies used to help in web search. It provides knowledge about a real world domain and enhances understanding by using entities, relationships, and attributes [1, 2]. Furthermore, ontology use is becoming more effective in information retrieval, robots, knowledge management, and electronic commerce [3, 4]. Ontologies contributed to these domains and more due to providing shared and common understanding among people and applications. However, creating ontologies is complicated due to ambiguity of concepts and semantics heterogeneity in communication [5]. Many Academic institutions are investing in national and international accreditations to ensure the quality of educational programs. Programs or institutions accreditations have a Board of Directors (BOD). For example, The Accreditation Board for Engineering and Technology (ABET) sets policy and approves all accreditation criteria that are used to evaluate programs. To evaluate a program, the evaluation process has to establish criteria for evaluations. The criterion applies to students, program, curriculum, facilities, assessment and evaluation to increase the quality of and inspire confidence in the program. Each individual program set its own criteria with continuous improvement and institutional support [6, 7]. The Accreditation process requires mapping and defining concepts for each course in the program. The problem is that all course learning outcomes and student outcomes mapping are placed in documents or databases which requires more work and time to access and understand. This paper proposes an ontology based solution to enable visual discover of all course learning outcomes that maps to a particular student outcome and related assessments to help faculty or curriculum committees avoid over mapping or under mapping students' outcomes. The following section will focus on the back of using ontology in diverse fields; Section III describes the semantic framework of the accreditation model and the hierarchy of the ontology.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

II. BACKGROUND

A. Ontology

Thomas Gruber [1] has defined ontology as "formal, explicit specification of a shared conceptualization". This is the most common definition of ontology which means a description of concepts and relationships in a domain such as education, medicine, knowledge management, etc. The description of concept is prepared by explicitly naming the concepts and the relationship. This description is more precise structure than just being taxonomy by providing relations and constraints between concepts.

B. Applications

Ontologies' applications have common usages in many different fields:

- Natural language processing: There are models that support semantics for natural language expressions such as Generalized Upper Model (GUM) that are semantics for natural language expressions to arbitrate between systems and natural language technology. GUM can provide mapping structure in multilingual generation systems [8]. To enhance reasoning for deeper understanding of texts used by machine translation, SENSUS project was developed. SENSUS [9] is a framework into which additional knowledge can be added to a system. It is an extension that uses WordNet at the top level containing nodes from the Penman Upper Model. The Penman Upper Model is a class structure of concepts organized originally in three sub-classes: Object, Process, and Quality [10]. WordNet is a lexical database of English developed by Princeton University.
- Educational Ontologies: Learning resources is widely available via the Web and the private network of educational institutions. Considering the constant increase of learning resources, Ontology is a key

enabler of supporting educational systems using conceptualizations. It is becoming strongly useful in the domain of Web-based Educational Systems (WBES). Many WBES concepts were developed by [12]-[14] such as subject domain, repository of learning resources, etc. It supports the representation of a domain ontology which provides formal definition of concepts for domain knowledge representation. Ontology designed for course learning has to identify formal rules for concept representation from a given course content. Ontology allows the visualization of a course content and concepts with syntactic and semantic meaning for learners [15]. Designing ontologies based on sharing learning environment was the focus of the Ontologies for the use of digital learning resources and semantic annotations on an online (OURAL) project. This project was based on real case studies to help teachers in describing learning domain problem solving and critical analysis [16].

- Tagging of Resources: Represents a link between objects for future use and collaboration with other users. This type of application allows users to add their cognitive information to resolve ambiguity and have consensus by using general classification. This type of classification leads to automatic discovery of new information, and improve precision in searching. Many users use tag to attract attention, show their interest, and make contributions to an object [17, 18]. On the web, users can tag objects based on their understanding using unstructured classification. Folksonomy unstructured classification system that pretence in this type of tagging is a real challenge of information retrieval by making many semantic tags and many abstractions levels [19, 20].

III. Ontological Semantic Accreditation Framework

This section describes the proposed semantic framework. The accreditation framework is a knowledge-based approach that requires a comprehensive analysis of the entire domain concepts which includes course domain, institution domain, and accreditation domain. Normally, the course domain is represented by a course syllabus. This document has many concepts such as description, course objectives, course learning outcomes, topic, book, policies, etc. The institution domain has individuals, programs, facilities, technology, policies, etc. Many accreditations, such as the Accreditation Board for Engineering and Technology (ABET), support and encourage institutions to adopt and use their own terminology. Also, the accreditation domain has its own concepts. For example, ABET has defined some concepts such as program educational objectives, student outcomes, assessments and evaluations [21]. The domain's concepts have to be mapped according to the ontology model to enable machine-tractable representation and adhere to the rule. The ontology can be accessed from a knowledge acquisition system. The knowledge acquisition system contains an ontology editor and a visualization plug-in. Figure 1 illustrates the Proposed Semantic Accreditation Model.

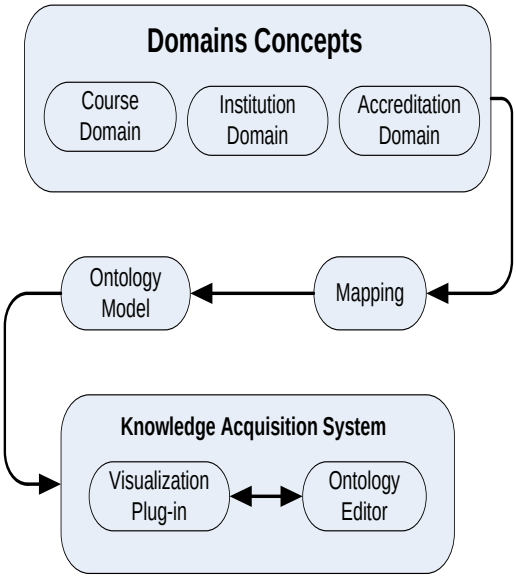


Fig. 1. Semantic Accreditation Model

Accreditation taxonomies or concept hierarchies are crucial for any knowledge-based system to structure information into categories and enable reasoning based on knowledge. These concepts hierarchies formulate relationships and rules to enable reasoning and reuse of knowledge based system.

The Accreditation model is implemented with a plug-in for Protégé 4.3. This Protégé tool has been widely used in many research projects in the area of semantic-web and modeling. Protégé 4.3.tool is extendible and fully supports the second version of the Ontology Web Language (OWL 2). The semantic web of an accreditation process helps student and faculty to understand the accredidation process. Furthermore, it promotes knowledge discovery and knowledge reuse about recruitment and evaluation. The definition of terminologies is debatable between domain experts. Therefore, the paper adopted the approach of defining accreditation terminologies based on defining relationships between terms used mosly by domaim experts. This approach allows realistic understanding of terms and avoids definitions conflicts. However, ambiguity terminologies are used. For example, in writing a course learning outcome, the statement could include "Students understand" or "Students Know". This ambiguity will be discussed in the future papers. In the mean time, we focused on the terms used in this ontology by clarifying and visualizing these terms making them easy to understand. The accreditation concepts have been adopted in this model are formally used by the Accreditation Board for Engineering and Technology (ABET). Figure 2 illustrates the Accreditation Ontology Model.

Fig. 2. ABET Ontology Hierarchy

The hierarchy of ABET in the ontology is debatable among domain specialist. Some specialist suggested that all this domain classifications should be listed under the super class Program. Their justification is that accreditation process applies to program. Others, argue that domain classifications should be separate because in real world education institutions are using these classifications for another purposes than ABET Accreditation. Therefore, this ontology adopted this hierarchy shown in Figure 1. It consists of seven classes: Evaluation, Assessments, Knowledge Domain, Person, Program, Report, and Facility. Figure 2 illustrates the ABET ontology hierarchy.

A. Evaluation: Consists of processes for interpreting a course data and evidence to determine the attainment level which a program educational objectives and student outcomes has improved. The data and evidence collection come from the assessment practices during the course period. Before initiating the evaluation process of a program, the program must have met the eligibility requirements [29] of ABET and apply for Request for Evaluation (RFE). The accreditation of a program may be granted to students who graduated before the on-site visit if their samples work and transcripts have been evaluated. There are two types of RFE:

- Requesting Initial Accreditations: The program must submit the RFE with one official graduate's transcript.
- Renewing Existing Accreditations: The program must submit the RFE to renew the existing accreditation.

B. Assessment: A valuable assessment uses relevant method (direct, indirect, qualitative, and quantitative) to the objective or outcome being measured. The result of the evaluation processes is used as a base for the decisions to improve the program [21]. Figure 3 illustrates the assessment class hierarchy.

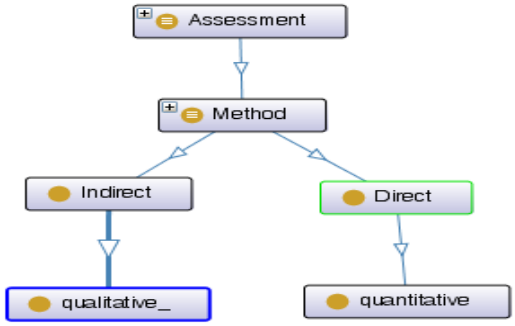


Fig. 3. Assessment Class Hierarchy

C. Knowledge Domain: It has teaching material (textbooks) and topics to be covered, assessed, and evaluated to determine the attainment level of the program educational objectives and student outcomes as shown in Figure 4.

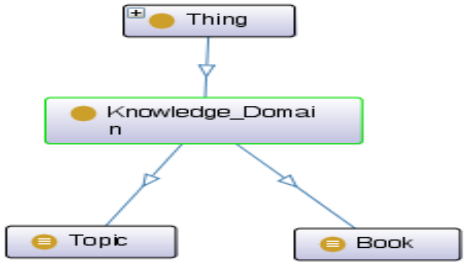


Fig. 4. Knowledge Domain Class Hierarchy

D. Person: This class hierarchy is evaluator, faculty (lecturer, teaching assistant), staff, and student. The main focus of the ABET accreditation process is the students in a program and their continuous improvement. Figure 5 illustrates the hierarchy of a person.

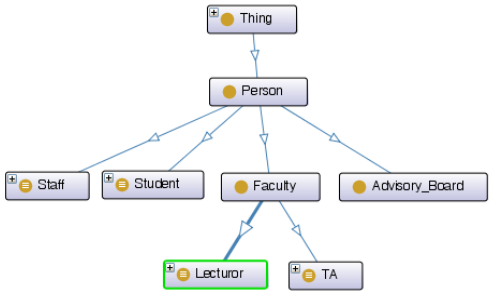


Fig. 5. Person Class Hierarchy

E. Program: ABET defines it as "An integrated, organized experience that culminates in the awarding of a degree" [25]. It has courses, objective, Student Outcome (SO) to be measured and evaluated, and a status. A course has key performance indicators (KPI) or course learning outcomes

(CLO), sections, syllabus, and samples. A course can have at least one section which is related to a knowledge domain. At most the course has exactly one syllabus which has many key performance indicators, teaching material, and covers some topics. A course also has samples to measure a particular KPI or CLO and it belongs to exactly one student. Program education objective statements are a description of what students are expected to attain within a few years of graduation. This objective should be carefully written based on the program constituencies [21] and it is related to an evaluation. The Student Outcomes statements are descriptions of what students are expected to know and able to do by the time of graduation such as skills, knowledge and behaviors and also related to an evaluation. Each program has status includes the followings:

- Accredited: The program is granted ABET accreditation since it satisfies accreditation criteria.
- Not to Accredit: The program is denied ABET accreditation since it has deficiencies that are not compliance with the accreditation criteria. This decision is taken only after a Show Cause Report or a Show Cause Visit to review the status of a new and unaccredited program. The accreditation is not extended as a result of this decision which is the only decision that can be appealed.
- Observations: The suggested statements offered by ABET to assist the institutions in the continuous improvement of the program. These statements are not related directly to the accreditation process.
- Concern: The program's current situation satisfies ABET's criterion, policy, or procedure, but the possibility exists for this situation to change negatively.
- Weakness: The program lacks strength of not being in compliance with the accreditation criterion, policy, or procedure. The institution is required to respond to this weakness with the corrective of actions to show the compliance before the next review.
- Deficiency: A Statement that indicates that the program is not in compliance with the ABET criterion, policy, or procedure.
- Satisfactory: A Statement that indicates that the program is in compliance with the ABET criterion, policy, or procedure.

Figure 6 illustrates the hierarchy of the Assessment Class.

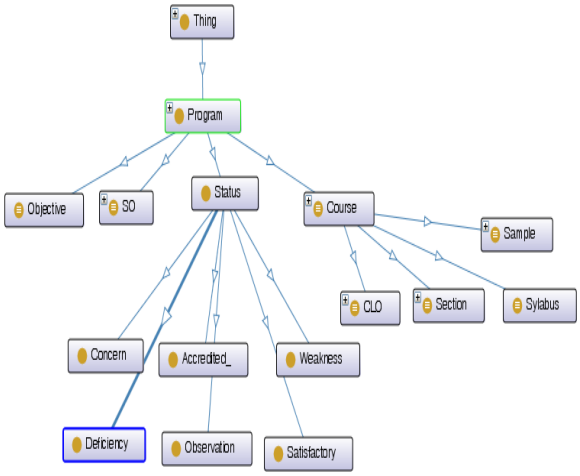


Fig. 6. Program Class Hierarchy

F. Report: According to the business dictionary a report is a document containing information about events, occurrences, or subjects.[26] In Figure 1 report class has four subclasses:

- Show Cause (SC): Is an action which indicates that the currently accredited program has one or more deficiencies.
- Show Cause Visit (SCV) action indicates that a currently accredited program has one or more deficiencies. Therefore, the deficiencies require an on-site visit to make sure a corrective of actions has been taken by the institution within typical duration of two years. This action cannot be for the same deficiency.
- Interim report (IR): Is an action which indicates that the program has one or more weaknesses. The institution is required to take a corrective of actions to these weaknesses typically within duration of two years.
- Report Extended: Is a satisfactory action taken bb the institution with respect to weaknesses identified IR. This report has typical duration of either two or four years which extends accreditation to the Next General Review.
- Self Study: Is the "Primary document that a program prepares to demonstrate compliance with ABET criteria" [27].
- Show Cause Report (SCR): Is an action which indicates that a currently accredited program has one or more deficiencies but this action cannot follow previous SC

action for the same deficiency. The institution is required to take a corrective of actions to these deficiencies typically within duration of two years [21].

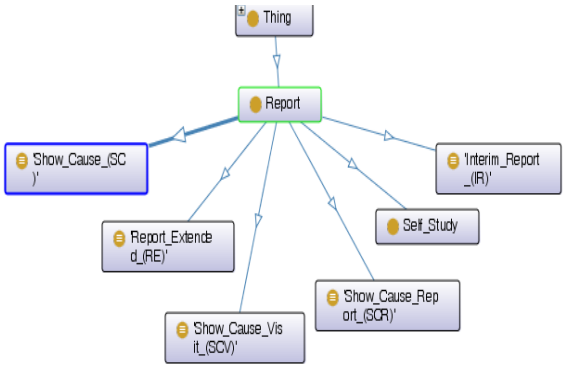


Fig. 7. Report Class Hierarchy

IV. Conclusions and Future Work

The development and implementation of an ontological model allows the integration of all available data into a specific and unique information system. The ontological approach allows improving the decision making process to improve the quality of education and information management. The future work will focus on fuzzy ontology to define terms used by the accreditation process to eliminate conceptual and terminological confusion and come to a shared understanding.

Acknowledgment

This work was supported by many professionals from academia and ABETS Program Evaluators who care about sustaining their respective through quality education.

References

[1] Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5, 199–220 (1993).

[2] Guarino, N., Giaretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: Mars, N. (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25–32. IOS Press, Amsterdam (1995) .

[3] D. Fensel, Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce,‖ Springer-Verlag, 2001.

[4] Lee, C.H.L. et Liu, A. Designing Robot Services with Ontology and Learning, Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, TX, USA, 2009.

[5] F. Hakimpour and A. Geppert,: Resolving semantic heterogeneity in schema integration: An ontology base approach,‖ in C. Welty and B. Smith, editors, Formal Ontology in Information Systems: Collected Papers from the Second Int'l Conf, FOIS'01, pp. 297–308. ACM Press, (2001).

[6] Cain , J.: Engineering and Computer Science Accreditation . 978-1-4244-2929-5, IEEE (2008).

[7] "www.abet.org/".

[8] The Generalized Upper Model. http://www.fb10.unibremen.de/anglistik/langpro/webspace/jb/gum/index.htm

[9] Sensus ontology.
http://www.isi.edu/natural-language/projects/ONTOLOGIES.html

[10] Penman Upper Model.
http://www.fb10.uni-bremen.de/anglistik/langpro/kpml/um89/um89-root.htm

[11] R. Mizoguchi and J. Bourdeau, "Using ontological engineering to overcome common AI-ED problems," International Journal of Artificial Intelligence in Education, vol. 11, no. 2, pp. 107-121, 2000.

[12] Aroyo, L., & Dicheva, D. (2001). AIMS: Learning and Teaching Support for WWW-based Education. International Journal for Continuing Engineering Education and Life-long Learning, 11 (1/2), 152-164.

[13] Brusilovsky, P. (2004). KnowledgeTree: A Distributed Architecture for Adaptive E-Learning. In Proceedings of the Thirteenth International World Wide Web Conference (WWW 2004), New York: ACM Press, 104-113.

[14] Brusilovsky, P., Eklund, J., & Schwarz, E. (1998). Web-based education for all: A tool for developing adaptive courseware. Computer Networks and ISDN Systems, 30 (1-7), 291-300.

[15] Alomari, J, Hussain, M, Turki, S; Masud, M.: Well-formed semantic model for co-learning Computers in Human Behavior, ISSN 0747-5632, 12/2014.

[16] M. Grandbastien, F. Azouaou, C. Desmoulins, R. Faerber, D. Leclet, and Q.-J. Céline, "Sharing an ontology in education: Lessons learnt from the OURAL project," presented at 7th IEEE International Conference on Advanced Learning Technologies, 2007.

[17] A. Marchetti, M. Tesconi, F. Ronzano, M. Rosella, and S. Minutoli, "Semkey: A semantic collaborative tagging system," May 2007.

[18] M. Gupta, R Li, Z. Yin, and J. Han, "Survey on social tagging techniques," SIGKDD Explor. Newsl., vol. 12, pp. 58-72, November 20 1 0

[19] K. Lee, H Kim, H Shin, and H-J. Kim, "Folksoviz: A semantic relation-based folksonomy vi sualization using the wikipedia corpus," Software Engineering, Artificial Intelligence, Networking, and ParallellDistributed Computing, ACIS International Conference on, vol. 0, pp. 24-29, 2009.

[20] A. Mathes, "Folksonomies - cooperative classification and communication through shared metadata," December 2004.

[21] www.abet.org

[22] https://www.ohio.edu/mechanical/chair/ABET_Terminology.pdf

[23] http://www.abet.org/accreditation/get-accredited-2/get-accredited-step-by-step/self-study-report/

[24] http://assessment.uconn.edu/docs/HowToWriteObjectivesOutcomes.pdf

[25] http://www.thefreedictionary.com/Domain+knowledge

[26] http://www.merriam-webster.com/dictionary/course

[27] http://www.dictionary.com/browse/objective

[28] http://www.abet.org/wp-content/uploads/2015/05/E001-15-16-EAC-Criteria-03-10-15.pdf

[29] http://www.abet.org/accreditation/new-to-accreditation/eligibility-requirements/

A Dual Cylindrical Tunable Laser based on MEMS

Ahmed Fawzy
Department of Electrical Engineering
Minia University
Nanotechnology central lab
Electronic Research Institute
Cairo, Egypt

Osama M. EL-Ghandour
Department of Elect., Commu., and Comp. Engineering
Helwan University
Cairo,
Egypt

Hesham F.A. Hamed
Department of Electrical Engineering
Minia University
Minia,
Egypt

***Abstract*—Free space optics is considered the topic of the day and have a large variety of applications which free space separates source from destination such as External cavity tunable laser (ECTL). In ECTL, laser source emits Gaussian beam that propagates in plane with substrate until reach external reflector. The efficiency of these applications depends on the amount of light that coupled back to the laser, called coupling efficiency. Increasing coupling efficiency depends on using assembled lens's or any optical part in the path between laser front facet and external reflector, which result in increasing the cost and integration effort. We innovate here anew configuration of external cavity tunable laser based on cylindrical (curved) Mirrors. The usage of cylindrical mirror affects on the amount of light that coupled back to laser and that decreases the alignment requirement in the laser assembly as compared to another configurations based on flat mirror. The fabrication of cylindrical mirror is simple with respect to spherical mirror so it can be used in batch fabrication. Tuning achieved by using micro electro mechanical system MEMS technology. The system consists of a laser cavity and a two filter cavities for wavelength selection. The formation of cylindrical microstructures were made into the substrate volume. So we report also the micromachining method that used for fabricating the cylindrical mirror. Anisotropic etching and the deep reactive ion etching (DRIE) are especially useful for the batch fabrication of large optical mechanical devices. The characteristics of the laser's spectral response versus laser facet reflectance variations are described via simulations. The diffraction of light in ECTL formed by the laser front facet and the external reflector are taken into account. Here we report all things about the model including the fabrication steps and simulation analysis.**

Keywords—Dual ECT; wavelength tuning; MEMS; DRIE

I. INTRODUCTION

Optical communications are the most important branch in communication engineering field as it introduces high speed communication, high bandwidth and low interference. Tunable laser will be discussed as one of the most important topic in optical communication[1-7]. We report here ECTL as a main stream technology in a tunable laser due to its advantages of simple configuration, small device dimensions, high tuning speed, wide tuning range, spectrum purity and high power. There are types of external cavities introduces wide tuning range but suffers from large device dimensions such as Littrow and Littman configurations which the tuning is achieved by rotating and translating an external grating simultaneously[8-9]. The rapid progress of the MEMS technology allows for miniaturization of a bulky tunable laser, improvement of the tuning speed and the mechanical reliability at lower fabrication cost. Designs based on the simple ECTL with a movable external mirror suffer from the tradeoff between tuning range and side mode suppression SMS [2,3,11-13]. So we present a model that compensates between the tuning range and SMS taking into account the diffraction effects for the light between laser facet and external mirror, and tuning range calculations. The paper is organized as follows: in section II materials and fabrication methods is explained in detail. In section III, theoretical analysis is explained and show the effective reflectivity concept and how it is affected by curved mirror. In Section IV, the results which shows the difference between single ECTL and dual ECTL the wavelength shift in the optical output of the laser is calculated based on that calculation.

II. MATERIALS AND FABRICATION METHODS

ECTL uses Fabry perot (FP) filters which exists between laser facet and external reflector. The tunability achieved by using FP filter, which comprises of two reflectors one fixed and another movable separated by free space or air gap. The external reflectors (mirrors) made from crystalline silicon(C-SI) for many reasons [14], First unique material properties of crystalline silicon in combination with wet anisotropic etching or advanced DRIE permit for the fabrication of improved optical components with mechanical parts. Second micromechanics needed some features in materials to provide its operation such as flat alignment, exact surface orientation and low oxygen content and low crystal defect density all of this exists in C-SI as well as unusual crystal cuts such as [110]-silicon, So we consider silicon material the most important one in MEMS systems that used in optics and Commercially C-SI uses as the standard substrate for microelectronics.

C-SI has also mechanical properties useful for all optical application that depend on moving parts. Stress-strain behavior of C-SI is unique which the absence of plasticity for temperature under 700 °C help in stable operation of mechanical component. Young's modulus E of C-SI affected extremely with the crystal orientation .

There are many micromachining techniques used in MEMS[1, 14-15]. we report here the most important of them. Surface Micromachining (SM) which used in mirrors fabrication. we can consider SM as a direct extension of semiconductor manufacturing technology. SM can manufacture devices in order of 50 - 100 micrometer. The key in this technique is the use of sacrificial layer which this layer are usually SIO2. This technology is based on depositing and

etching structure and sacrificial films. After deposition of thin film, sacrificial layer is etched away, leaving a completely assembled microstructure as shown in fig (1). By using of this technology free space components can be realized. The important applications of this technology are optical cross-connect switches [16-19] and micro-optical free-space benches. The main advantage for this technique that the components after fabrication become self-assembled. The functional layer in SM almost made from poly SI and this consider disadvantage in mirrors. Also maximum possible thickness of the microstructure is limited to that of the deposited film. The main problem in this technique is how the mirror lifted off from substrate to the desired position. Usually we use microprobes to do this action and due to this action the process needed more time and money.

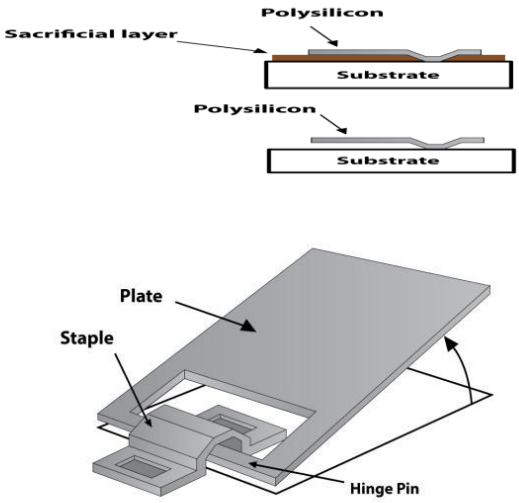


Fig. 1. General schematic diagram for the fabrication steps of micromirror by using SM

Second techniques Silicon-on-insulator (SOI) wafers consists of SiO2 sandwiched between two C-SI wafers. The sacrificial layer here is also SiO2 layer. Removing of sacrificial layers allows for mechanical movement of structures. Here the complete configuration can putted on chip that includes mirrors, actuator and fiber alignment structure as well as laser source assembled on chip and the fabrication of all components can made in a single DRIE etch process[20].

We introduce here a new monolithically dual silicon ECTL configuration based on SOI and DRIE with a standard FP semiconductor laser coupled with two ends with highly reflective movable curved mirror in one end and another end with fixed curved mirror as shown in fig (2)[15, 21-22]. The length of the gap between the laser and the movable mirror is varied by means of a actuator fabricated also on the same chip. The output of the laser is collected after transparent mirror. However spherical mirrors can be used, This paper concerned with curved mirrors only. This is because steps of fabrication of spherical mirrors cannot be used in batch fabrication. Also spherical mirror fabrication causes the wafer to be fragile due to long etching time. So our geometry depends on simplicity in fabrication as well as miniaturization [20].

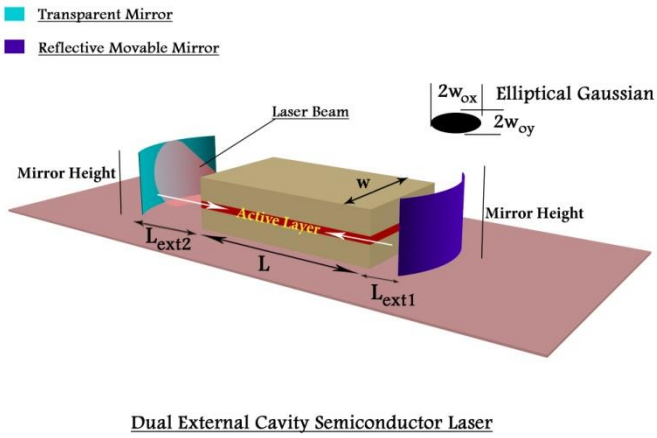


Fig. 2. General schematic diagram for the Dual ECTL source with external curved mirror

III. THEORITICAL ANALYSIS

The most important parameters in the ECTL designs are the effective reflectivity of the external cavities r_{eff} and the optical cavity length L_{ext} . These parameters affects on the performance of FP filters which determined by free spectral range FSR, cavity finesse F and cavity filter bandwidth BW. FSR can be defined simply as the spacing between adjacent maxima of a transmitted wavelength. $FSR = \lambda^2/2L_{ext}$ where λ is operating wavelength. From this relation we can show that a shorter cavity provides wider FSR. F is determined also by the following relation $F=\pi\sqrt{r_{eff}}/1 - r_{eff}$ that shows F depend only on r_{eff}. BW of the filter defined as the sharpness of each transmission peak BW=FSR/F.

Now we will focus on another parameter that called mirror loss term α_M that depend also on $r_{eff}$and L_{ext} .For FP laser α_M can expressed as follow:

$$\alpha_M = -(1/2L)\ln(|r_1||r_2|) \quad (1)$$

where L is the length of the semiconductor laser and r_1 ,r_2 are the amplitude reflectances of the two facets of the laser. For external cavity, the effective reflectance is used to describe the effect of the external reflector [2, 11-12].

$$r_{eff}(\nu) = r_2 + (1 - |r_2|^2)r_{ext}\exp(-j2\pi\nu\tau_{ext}) \quad (2)$$

where τ_{ext} refers to the round trip delay through the external cavity of length (Lext), ν is the frequency and the r_{ext} is the reflection of external reflector. Here we neglected multiple reflections in the external cavity for simplicity. The introduction of the external reflector can be thought to change the laser mirror reflectance r either r_1 or r_2 by an amount that is proportional to the ratio of light coupling back to the laser to the light going out. Both the amplitude and the phase of the reflected light is changed. This can be described by replacing r_1 and r_2 by r_{eff1} and r_{eff2}

respectively that is obtained via a coherent superposition of r of laser and the external reflectance. In other words The effective reflectivity is the coherent sum of the internal laser facet reflectance and the external ECTL reflectance, $r_{eff} = r_{int} + r_{ext}$. The interference between the internal and external reflections on the laser facets usually the important part of the wavelength dependency of the effective reflectivity, and therefore is the basic reason for the wavelength selectivity in many ECTL lasers. α_M now can expressed as follow:

$$\alpha_M = -(1/2L)\ln(|r_{eff1}||r_{eff2}|) \quad (3)$$

The external reflectance can be calculated by using the overlap integral at the facet of the reflected from external reflector after round trip ψ_{ext} and the field emitted from FP laser chip ψ_G by the following relation:

$$\eta = \iint_{-\infty}^{\infty} \psi_G(x,y)\psi_{ext}^*(x,y)dxdy / \iint_{-\infty}^{\infty}|\psi_{ext}(x,y)|^2 dxdy \quad (4)$$

In our study, the emitted field from FP laser was taken to be a normalized elliptic Gaussian field, with waist dimensions corresponding to the output beam of the solitary laser.

$$E(x,y) = E_m e^{\frac{-x^2}{w_{ox}^2}+\frac{-y^2}{w_{oy}^2}} \quad (5)$$

where E_m is the field amplitude at the center of the beam, w_{ox} is the spot size in the x direction and w_{oy} is the spot size in the y direction. The laser spot size is 3.5 micrometer in the horizontal direction and 1 micrometer in the vertical direction , in an elliptical shape , the laser wavelength is 1550 nm. The external reflector is considered to be curved mirror as shown in fig (3). The curved mirror not only reflect the beam but also add a phase due to the curvature of a mirror. The reflection coefficient of a curved mirror R_c can be expressed as:

$$R_c = R_m e^{-2jk_o(d_o-(R-\sqrt{(R^2-x^2)}))} \quad (6)$$

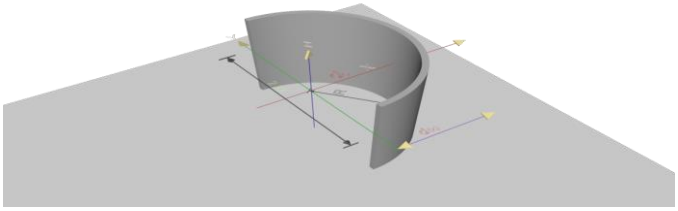


Fig. 3. 3D model of external curved mirror with radius of curvature R and mirror width 2w

Thus for a mirror of reflection coefficient R_m , and a radius of curvature R in the wafer plane where d_0 is the mirror sag at the plane of the wafer given by:

$$d_0 = R - \sqrt{R^2 - w^2}$$

with 2w the mirror width assumed to be 32μm in our calculations. By using the equations (2), (3), (4),(5) and(6) we can calculate the coupling power of external reflector when the distance between the laser and the micromirror surface was varied.

For the sake of comparison we use two kinds of mirror flat mirror and curved mirror as shown in fig(4).

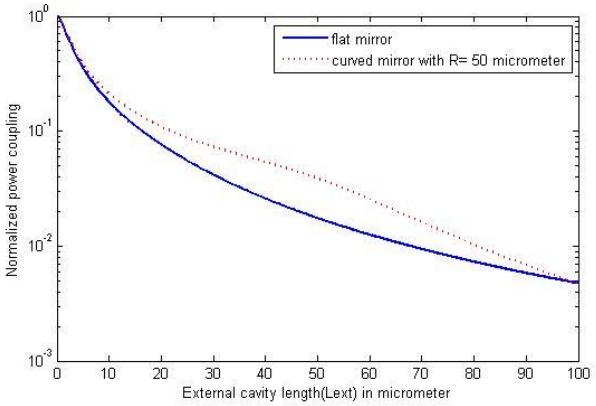


Fig. 4. Normalized power coupling efficiency of the curved micromirror with radius of curvature of 50 micrometer and flat micromirror versus External cavity tunable laser

In our model, a plane wave expansion method is used [8,19]. In this method the effect of transfer function of free space between laser facet and external reflector, called diffraction effect, is taken into account. As shown in figure (4), the curved mirror makes collimation for the light in one direction so the amount of light that coupled back has been increased with respect to the flat mirror. In other words, the amplitude and phase of effective reflectivity will affect on not only the power coupled to laser, but also the frequency shift. This is because the phase of round trip of laser cavity will be changed due to existence of external cavity. The change in the phase can be calculated from following equation:

$$\Delta\emptyset_l = 2\pi\tau_l(U - U_{th}) + \emptyset_r - \alpha[\ln\left(\frac{1}{|r_{eff1}||r_{eff2}|}\right) - \ln\left(\frac{1}{|r_1||r_2|}\right)] \quad (7)$$

where τ_l is the round trip time delay of the optical beam inside the primary laser cavity, $\emptyset_r$ represents the total phase of effective reflectivity due to two cavities, U is the new oscillation frequency, U_{th} is the oscillation frequency before using an external cavities and α is the linewidth enhancement factor in the semiconductor material. From this relation, we can find a new formula for the frequency shift of the laser emission that result due to using dual cavities as follows:

$$\text{freq shift} = U - U_{th} = [\alpha[\ln\left(\frac{1}{|r_{eff1}||r_{eff2}|}\right) - \ln\left(\frac{1}{|r_1||r_2|}\right)] - \emptyset_r] / 2\pi\tau_l \quad (8)$$

This expression for the frequency shift is valid for both weak and strong feedback reflections. It can thus be used to study the effect of the diffraction on the frequency shift and hence on the resonance frequency of the laser. To calculate this frequency shift, it is required to evaluate both the amplitude and phase of the reflected beam after travelling in the external cavity. The phase difference and the balance between the amplitudes of the internal and external reflections play an important roles in many of the sensing application as shown in figure(5) [23-25].

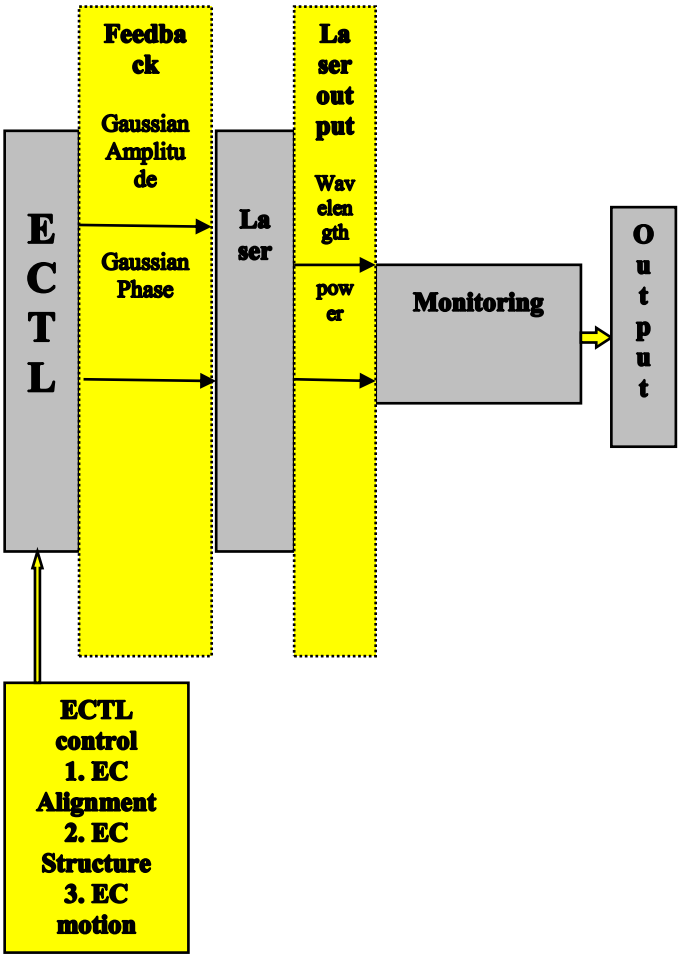


Fig. 5. General concept of sensing in ECTL

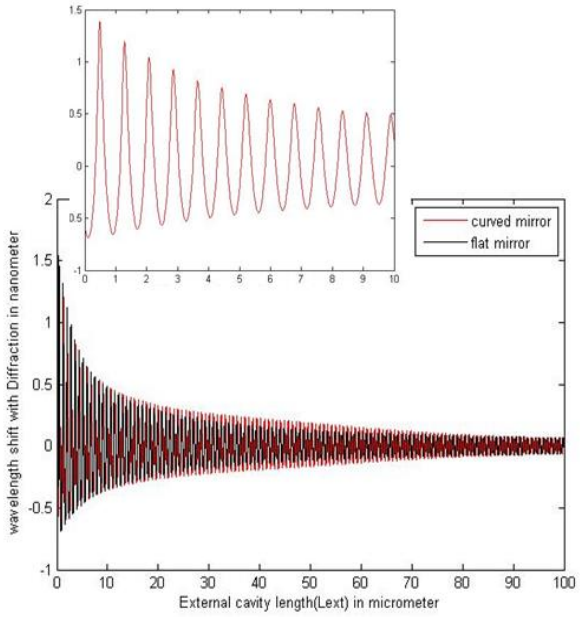


Fig. 6. Wavelength shift versus external cavity length of single ECTL which its external cavity is comprised by micromirror and the front facet of the semiconductor laser. When the micromirror flat (black line) and when the micromirror curved with radius of curvature 50 micrometer (red line)

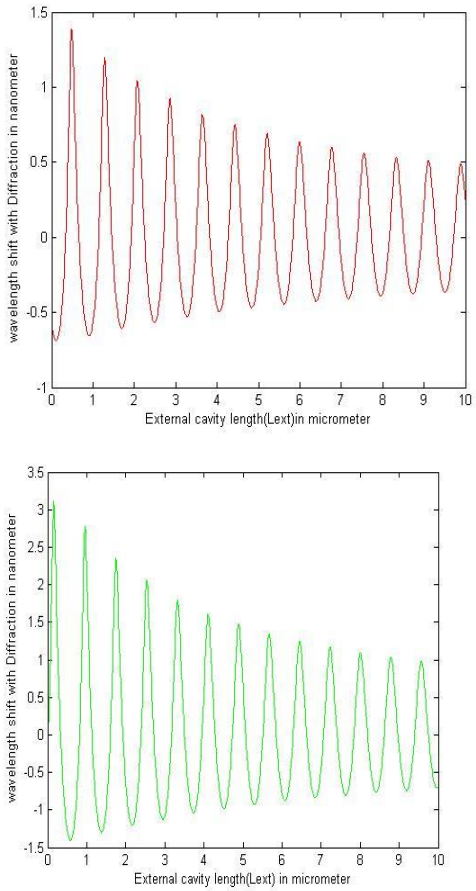


Fig. 7. Wavelength shift versus external cavity length is shown in (a) single ECTL formed by the external cavity comprised by curved micromirror and the front facet of the semiconductor (Red line) (b) Dual ECTL with two cavities formed by curved mirror in both sides. Radius of curvature of both cases a , b= 50 micrometer and reflectivity of mirror= .32 (green line)

IV. RESULTS

For a curved external mirror, the coupled power that come back to active cavity from external mirrors is affected by phase that is introduced by external mirror as shown in figure(6). For flat mirror there is no collimation as curved mirror; which makes collimation for light in one direction at the point of radius of curvature. The amount of power that coupled back from curved mirror is larger than the flat mirror. In a dual configuration, we assume only one round trip in the calculation after one round trip; the optical intensity may still be strong enough to perturb the operation of the laser. In figure 7 we see the difference in range of wavelength shift between single ECTL and dual ECTL. In case of a dual ECTL the range of wavelength shift is greater than the double of single ECTL.

V. CONCLUSIONS

In this paper, we propose a new simple unique model based on curved mirror. This model uses two cavities instead of single cavity. Our study reveals several important features that we utilized in the design of miniaturized MEMS tunable laser. These features include the importance of curved mirror which achieves efficient coupling of light than flat mirror. A micromachining method, that is used to fabricate curved

mirror, is very simple comparable with spherical mirror so it can be used in batch fabrication. A new analytical approach is developed to determine the performance of dual ECTL taking into account diffraction effect. Analytical results show that dual ECTL can achieve larger tuning range than that of single cavity. All these factors make our model the best miniaturized tunable laser between all models.

REFERENCE

[1] O. Solgaard, A.A. Godil, R.T Howe, L.P. Lee, Y.-A. Peter and H. Zappe, "Optical MEMS: From Micromirrors to Complex Systems," Journal of Microelectromechanical Systems, vol.23, no.3, pp.517-538, June 2014

[2] A. Q. Liu," Photonic MEMS Devices Design, Fabrication and Control", CRC Press, 2009.

[3] A. Q. Liu, and X. M. Zhang, "A review of MEMS external-cavity tunable lasers", Journal of Micromechanics and Microengineering, Vol. 17, No. 1, 2007.

[4] O. Solgaard, "Photonic Microsystems: Micro and Nanotechnology Applied to Optical Devices and Systems",Springer New York, 2009.

[5] M. Ren, H. Cai, Y. D. Gu, P. Kropelnicki, A. B. Randles and A. Q. Liu, "A tunable laser based on nano-opto-mechanical system", 27th IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2014), San Francisco.

[6] Komljenovic, T., Srinivasan, S., Davenport, M., Norberg, E., Fish, G., Bowers, J.E." Widely-tunable narrow-linewidth lasers with monolithically integrated external cavity", Conference on Lasers and Electro-Optics (CLEO), 2015.

[7] W. M. Zhu, W. Zhang, H. Cai, J. Tamil, B. Liu, T. Bourouina, and A. Q. Liu "A MEMS Digital Mirror for Tunable Laser Wavelength Selection", International Conference on Solid-State Sensors, Actuators and Microsystems, pp. 2206-2209, 2009.

[8] A. Q. Liu, X. M. Zhang, D. Y. Tang, and C. Lu, "Tunable laser using micro machined grating with continuous wavelength tuning", Applied Physics Letters, Vol. 85, No. 17, pp. 3684-3686, 2004.

[9] X. M. Zhang, A. Q. Liu, D. Y. Tang, and C. Lu, "Discrete wavelength tunable laser usingmicroelectromechanical systems technology", Appl. Phys. Lett., Vol. 84, pp 329, 2004.

[10] X. M. Zhang, A. Q. Liu, C. Lu and D. Y. Tang, "A real pivot structure for MEMS tunable lasers," IEEE Journal of Microelectromechanical Systems, Vol. 16, no. 2, pp. 269-278, 2007.

[11] A., Fawzy, S., El-sabban, , I., Ismail and D., Khalil," On the Modeling of an External Cavity Tunable Laser ECTL Source with Finite Mirror Dimensions", Progress in Electromagnetic Research Symposium Proceedings, Stockholm, 12 -15 August, pp. 691- 694, 2013.

[12] J. Aikio, and D. Howe, "Extremely short external cavity laser: profilometry via wavelength tuning ", Conference of laser and electro optics, pp 484-485, 2001.

[13] Aikio, J.K., Kataja, K.J., Alajoki, T., Korioja, P., Howe, D.G., "Extremely short external cavity lasers: The use of wavelength tuning effects in near field sensing", Proceedings of SPIE – The international Society for Optical Engineering, pp. 235-245, 2002.

[14] Hoffmann, M., Voges, E., "Bulk silicon micromachining for MEMS in optical communication systems", Journal of Micromechanics and Micro Engineering, Vol. 12 pp.349-360, 2002.

[15] A. M. Abu-El-Magd, " Double Tuning of A Dual External Cavity Semi Conductor Laser For Broad Wavelength Tuning With High Side Mode Suppression", Master Thesis, McMaster University, Hamilton, 2011

[16] H. Cai, J. X. Lin, J. H. Wu, B. Dong, Y. D. Gu, Z. C.Yang, Y. F. Jin, Y. L. Hao, D. L. Kwongand A. Q. Liu, "NEMS optical cross connect (OXC) driven by optical force", 28th IEEE International Conference on Micro Electro Mechanical Systems (IEEE MEMS), 2015

[17] B. Dong, H. Cai, Y. D. Gu, Z. C. Yang, Y. F. Jin, Y. L. Hao, D. L. Kwong and A. Q. Liu, "NEMS variable optical attenuation (VOA) driven by optical force", 28th IEEE International Conference on Micro Electro Mechanical Systems (IEEE MEMS), 2015.

[18] W. M. Zhu, X. M. Zhang, T. Zhong, A. Q. Liu and M. Yu, "Micromachined optical well structure for thermo-optic switching," Applied Physics Letters, Vol. 91, No.26,261106, 2007.

[19] A. Q. Liu, A. B. Yu, M. F. Karim and M. Tang, "RF MEMS switches and integrated switching circuit," Journal of Semiconductor Technology and Science, Special issue on NANO/Microsystems Technology, Vol. 7, No.3, September, 2007.

[20] Y.M. Sabry, D. Khalil, B. Saadany, and T. Bourouina, "Silicon Micromirrors with Three-Dimensional Curvature Enabling Lensless Efficient Coupling of Free-Space Light", Light: Science & Applications, 2, e94,2013.

[21] X. Zhu, and D. T. Cassidy, "Liquid Detection with InGaAsP Semiconductor Lasers Having Multiple Short External Cavities", Applied Optics, vol 35, pp 4689-4693, 1996.

[22] A., Fawzy, O., M., EL-ghandour, H., F., A., Hamed.,"Performance Analysis on a Dual External Cavity Tunable Laser ECTL source", Journal of Electromagnetic Analysis and Applications,Vol.7, PP. 134-139.,2015.

[23] B. Dong, J. G. Huang, H. Cai, P. Kropelnicki, A. B. Randles, Y. D. Gu and A. Q. Liu, "An all optical shock sensor based on buckled doubly-clamped silicon beam", 27th IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2014), San Francisco.

[24] X. Zhao, J. M. Tsai, H. Cai, X. M. Ji, J. Zhou, M. H. Bao, Y. P. Huang, D. L. Kwong and A. Q. Liu, "A nano-opto-mechanical pressure sensor via ring resonator," Optics Express, Vol 20, pp.8535-8542, 2012.

[25] J. F. Tao, H. Cai, J. Wu, J. M. Tsai, Q. X. Zhang, J. T. Lin and A. Q. Liu, "Optical wavelength signal detector via tunable micro-ring resonator for sensor applications", 26th IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2013), Taipei.

[26] J.F. Tao, A. B. Yu, H. Cai, W. M. Zhu, Q.X. Zhang, J. Wu, K. Xu, J. T. Lin, ,A. Q. Liu, "Ultra-high coupling efficiency of MEMS tunable laser via 3-dimensional micro-optical coupling system" IEEE 24th International Conference on MEMS, pp. 13-16, 2011.

Function-Behavior-Structure Model of Design: An Alternative Approach

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

***Abstract*—The Function-Behavior-Structure model (FBS) of design conceptualizes objects in terms of function, behavior, and structure. It has been widely utilized as a foundation for modelling the design process that transforms posited functions to a description of behaviors. Nevertheless, the FBS model is still regarded as a subjective and experience-based process and it provides no theory about how a function is transformed into behavior. Research has shown that the critical concepts of function and behavior have many different definitions. This paper suggests a viable alternative and contrasts it with the FBS framework of design using published study cases. The results point to several benefits gained by adopting the proposed method.**

Keywords—conceptual design; FBS framework; flow-based model; function; behaviour; structure

I. Introduction

To develop a science of designing, design research aims at a better understanding of design, the development of tools to aid designers, and the potential automation of some design tasks. "Design exists because the world around us does not suit us, and the goal of designers is to change the world through the creation of artifacts" [1].

In engineering design, the product development process starts with the problem definition and requirements. This is followed by the phase of *conceptual design*, which focuses on what a design must do to realize the requirements. Conceptual design involves the creation of a design *description*, which is represented graphically, numerically, or textually [2-3]. "The conceptual design phase is acknowledged as particularly critical. It offers the greatest scope for significant enhancements and decisions made in this phase impact all subsequent design phases" [4]. This phase can be based on the framework called the *Function-Behavior-Structure model* (FBS).

The FBS-based design conceptualizes objects in terms of function, behavior, and structure. It has been widely utilized as a foundation for modelling the design process [1][5-6]. This process refers to transforming posited functions to a description of behaviors [1]. Many studies on function, behavior, and structure concepts have been conducted, resulting in several variants and extensions of the model.

The model is still regarded as a *problematic* approach. It is looked at as "a subjective and experienced-based process" [7] and it provides no theory about how function is transformed into behavior [1]. Research has shown that the critical concepts of function and behavior have many different definitions [8]. "There are debates on the suitability of these notions to the design model, which have left much confusion" [9].

[Such notions as function, behavior, and structure] have created some confusion about and debates on which one should be the most appropriate one. Naturally, a question one may ask is whether they are actually the same thing but with different names or whether they have different scopes of applications for different design problems. [9]

This paper suggests a viable alternative to the FBS model of design in terms of a diagrammatic language that is akin to specifications in software engineering. It then applies this alternative to the concepts of function, behavior, and structure. The two approaches are contrasted using published study cases. The results point to several benefits gained by adopting the proposed diagram representation.

For the sake of a self-contained paper, the next section briefly reviews the diagrammatic language that forms the foundation of the theoretical development in this paper. The model has been adapted to several applications [10--15]; however, the example given here is a new contribution.

II. Flowthing Model

The Flowthing Model (FM) is a language for representing "things that flow," called *flowthings*. Flow in FM refers to the exclusive (i.e., being in one and only one) transformation among five *states* (also called stages): transfer, process, creation, release, and receive, as shown in Fig. 1. A flowthing may be called, simply, a *thing*.

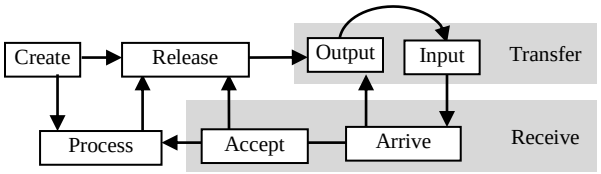


Fig. 1. Flow system

The fundamental elements of FM are as follows:

Flowthing: A thing that has the capability of being created, released, transferred, arrived, accepted, and processed while flowing within and between *flow systems*. In the FBS literature, an *object* is a thing which is observable by its properties. For example, "a power plant is an object which is observable by its

properties, e.g., generating power." [9]. In FM, power is a flowthing, as shown in Fig. 2. Suppose that we are interested also in representing the power plant as an existing physical thing that is being inspected. Fig. 3 shows the resulting diagram.

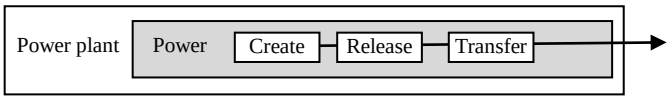


Fig. 2. Power is a flowthing in the plant sphere

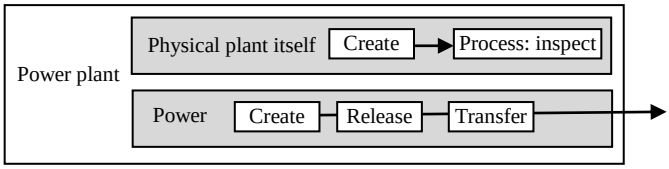


Fig. 3. The plant as an existing thing that is being inspected

A flow system is a system with five stages and connections between them. In FM, flows can be controlled by the progress (sequence) of a stream of events (create, release, transfer, transfer to another sphere, receive, …) or by triggering that initiates a new flow.

Spheres and subspheres: These are the environments of the flowthing that reflect *structure*, e.g., in Fig. 3, the *power plant* is a sphere with the two sub-spheres (flow systems) *power* and *physical plant itself*.

Triggering: *Triggering* in FM is the activation of a flow, denoted by a *dashed arrow*. It is a dependency among flows and parts of flows. A flow is said to be triggered if it is activated by another flow (e.g., a flow of electricity triggers a flow of heat) or activated by another point in the flow. Triggering can also be used to initiate events, such as starting up a machine (e.g., remote signal to turn on).

Example: According to Zhang Lin and Sinha [9], a system is a set of entities connected in a meaningful way. The entities are perceived in the form of their *states*, which change with respect to time. Fig. 4 shows a crank–slider linkage system where a powered motion is given to the crank and this motion is transferred to the coupler, which, in turn, is transferred to the slider. The angles are *state* variables. The movement of the slider is called *behavior.*

The behavior of a system is about the response of the system when it receives stimuli. Since the system (structure) is represented by its state, the stimuli and the response are further represented by the state variable. Therefore, the behavior is the relationship between the independent state variable and the dependent state variable… The above definition does imply that the behavior is *about the relation between inputs and outputs*. [9](Italic added.)

Fig. 5 shows the corresponding FM representation. The generation (creation) of a new Θ_2 (circle 1) trigger (2) and the creation of a new Θ_1 (3), which triggers (4) the creation of a new distance (5). *Process* in the diagram indicates a possible change in a current value, e.g., rotated angle.

Such a diagram provides new meaning for the concept of behavior and structure. We can define the *behavior* of a system not as a mere relationship between inputs and outputs but rather, in general: it is *stream(s) of flows and triggering from source(s) to destination(s) through spheres and sub-spheres*. A stream here is analogous to, say, the Nile river as a system that includes countries, districts, cities, dams, delta, etc., through which its water flows as spheres and sub-spheres (counting flow systems). The basin of flows and triggering, including interrelated sub-spheres, is the *structure* of the system.

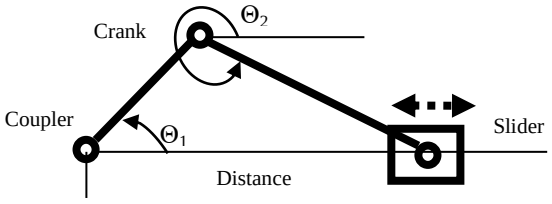


Fig. 4. A crank–slider linkage system (Modified – Re-drawn from [9]

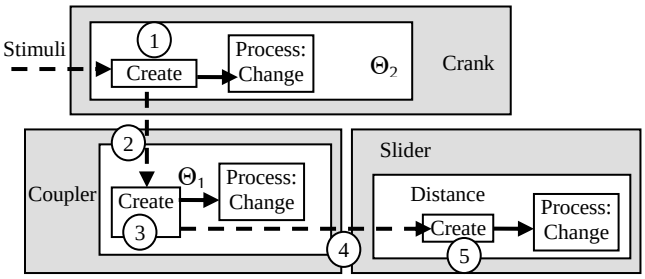


Fig. 5. FM representation of the crank–slider linkage system

III. Applying FM to the Design Process

This section applies the FM representation to some concepts that are utilized in the context of FBS-based design.

A. Functional structuring

According to Keuneke [16], functional structuring is useful because, to understand the functioning of a complex system, we often must decompose the system's function into its components' functions and decompose each component's functioning into the functions of subcomponents, and so on. Eventually, this decomposition terminates in behaviors by which these functions are achieved, which point to the functional components used. For example, Fig. 6 shows the functional structuring of a telephone [16].

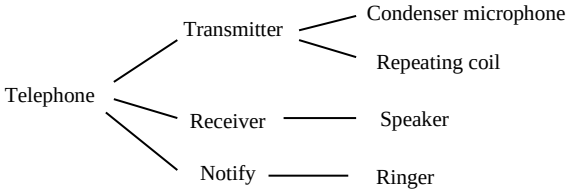


Fig. 6. Functional structural representation of a telephone (re-drawn from [16])

Fig. 7 shows a corresponding FM representation. As seen in the figure, the description is built around *flows* instead of functions. Due to the importance of identifying these flows in the process of design, we will start by describing these flows shown in the simplified diagram Fig. 8 before explaining Fig. 7.

Fig. 8 reflects the fact that receiving, sending (transmitting in Keuneke's terminology), and notifying (ringing) are processes (functions) that interweave, and it is difficult to separate them. Both sides of the communication process involved in receiving, sending, and ringing form three steams of flows. Accordingly, the sources and destination of the communication process involves:

- The outside caller, named *sender* (*outsider*), calls the *user* of the telephone, named *receiver (user)*,

- The telephone user takes the role of dialer, named *sender (user)*, who dials the outsider, named *receiver (outsider)*

In Fig. 7, the communication process starts at circle (1) when the caller, named *sender (outsider)*, send signals by dialing his/her telephone. The telephone receives (2) these signals and triggers the ringer of the telephone to create (3) the sounds that notify the person being called (4), named *receiver (user)*.

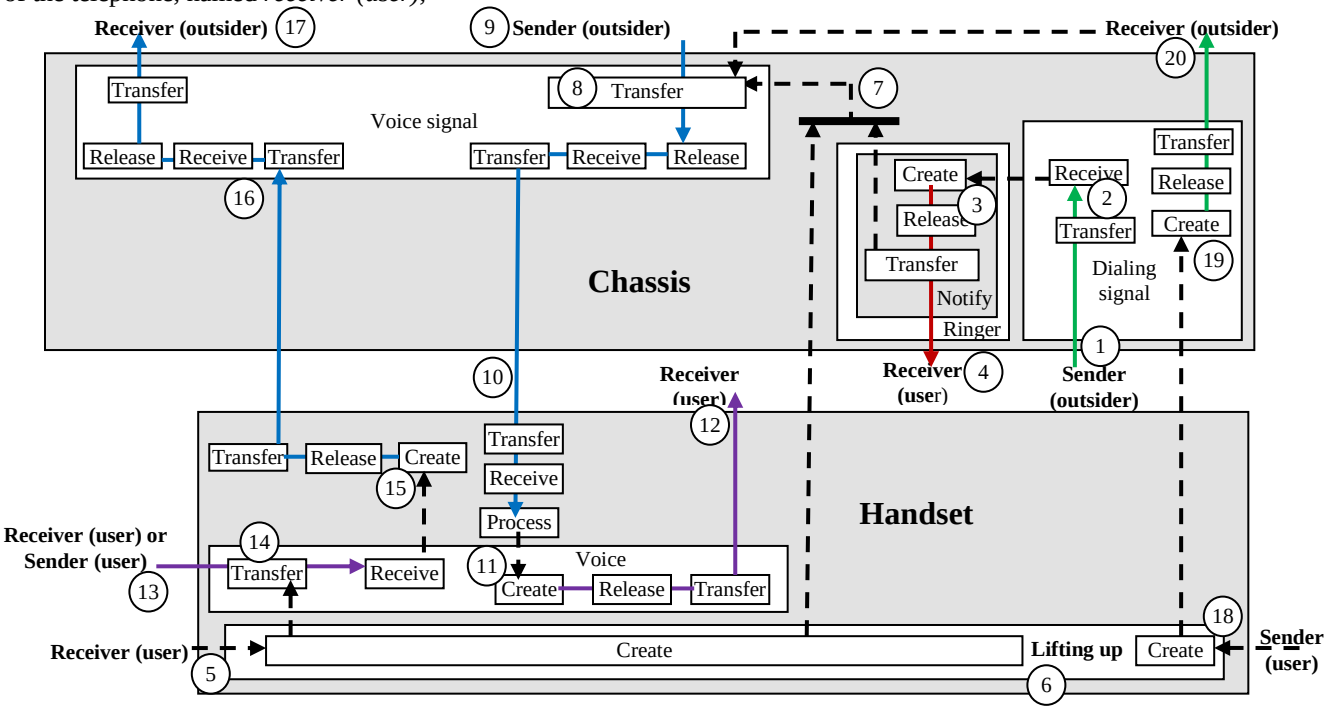


Fig. 7. FM representation of the telephone

Accordingly, the *receiver (user)* (5 – repeated mentioned in several places for simplification of the diagram) performs the lifting up (6) of the handset. These two states together, (a) the ringing (3), and (b) the lifting up (6) of the handset, trigger (7) the transferring (8) of the voice signals coming from the sender (outsider) (9), which flow to the handset (10) where they are converted into a sound (11) that flows to the receiver (user) (12). Now the telephone used plays the role of sender (user) (13) that creates the sound (14) that is converted into signals (15) that flows (16) to the caller in the role of receiver (outsider) (17).

As we can see, the conversation now has two sides:

- The sender (user) to the receiver (outsider) (circles 13, 14, 15, 16, and 17)
- The sender (outsider) to the receiver (user) (circles 8, 9, 10, 11, and 12). It is assumed that the transfer (8) will continue in the permission state after lifting the handset (7).

The case of ending the call is not included in this scenario because it is not mentioned in Keuneke's [16] specification.

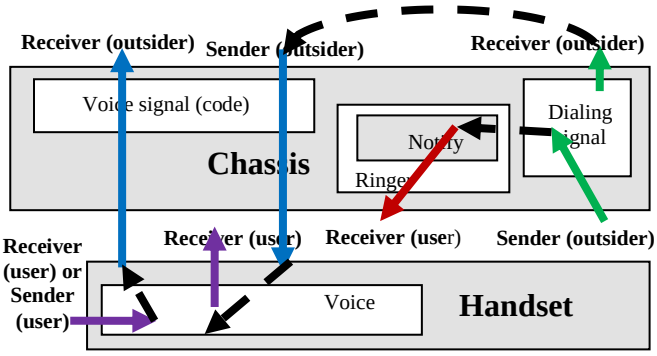


Fig. 8. Initial identified flows in the design of the telephone

When the telephone user initiates the call, the process starts at circle 18 where he/she plays the role of sender (user) by lifting up the handset (6) that triggers the dialing signals (19) and a number, which flows to the other telephone of the receiver (outsider) (20). Additionally, the transfer module (8) is set ON in anticipation of signals from the other end, i.e., the sender (outsider) (9). In this case, the same two ways of communicating (circles 8, 9, 10, 11, and 12) and (circles 13, 14, 15, 16, and 17) are open to exchanging the data.

It is clear that if there is any misunderstanding of the telephone functioning process because of the author's lack of knowledge in telephony, the diagram can easily be corrected. The point is that the diagramming language can express such functionality.

Now, we can ask about the *functions* of Keuneke's [16] Fig. 6 that are used as the base for building the design of the telephone. The FM representation shows that the functions Transmitter, Receiver, and Notify are interweaved with each other and that it is almost impossible to remove such entanglement. For example, Fig. 9 shows four diverse positions in the FM representation that involve the function of *sending*. Accordingly, this sharp division of functions seems to be needed at a later stage when our objective is to *optimize* the four portions of receiving in the FM diagram as one physical component motivated by cost considerations. In the initial *conceptual design*, cost is not taken into consideration, thus, it is possible to *implement* the design of the telephone as four transmission pieces, as in the case of cheap hardware or for the sake of simplicity in production.

We note that *Flows* is an alternative phenomenon that provides a base for a pure conceptual design (e.g., optimization of physical implementation is not involved). Over the steams of flows (i.e., flows trigger each other), it is possible to identify the basic structure of the design as shown in Fig. 10. Still, further details can be specified, as given in Fig. 11, where the Transfer module (stage) (circle 8 in Fig. 7) receives signals from Sender (outsider) (9) according to triggering 7 and 20. Fig. 11 also shows a possible pseudo-language level of details for this transfer stage.

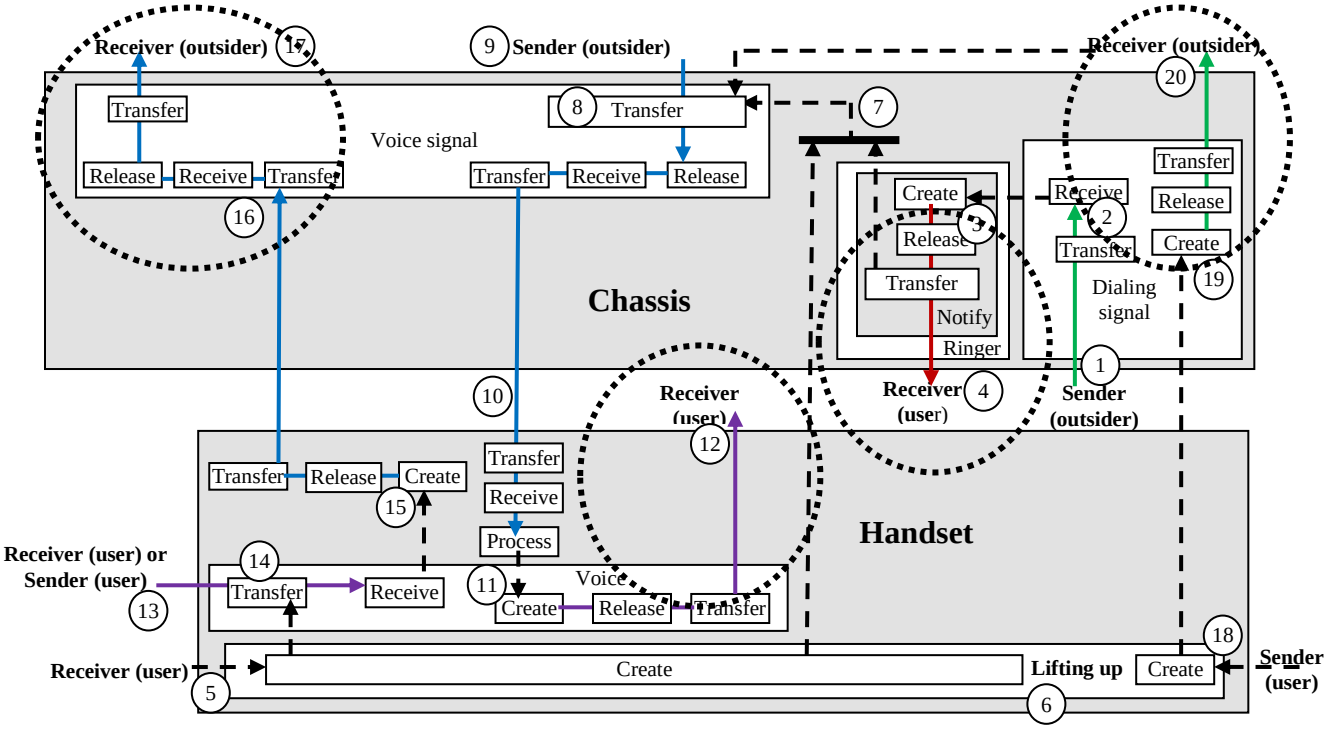


Fig. 9. Portions of the FM representation of the telephone that are related to *Receive*

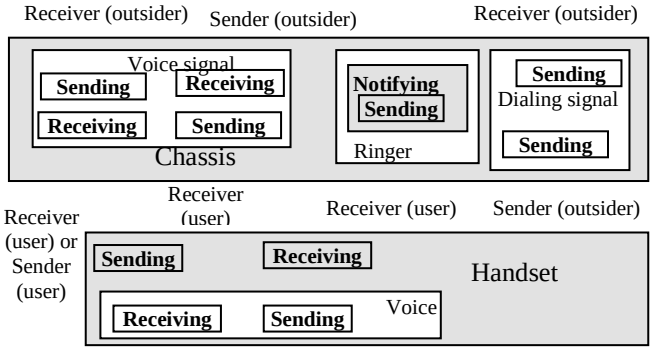


Fig. 10. Further decomposition of structure of the telephone

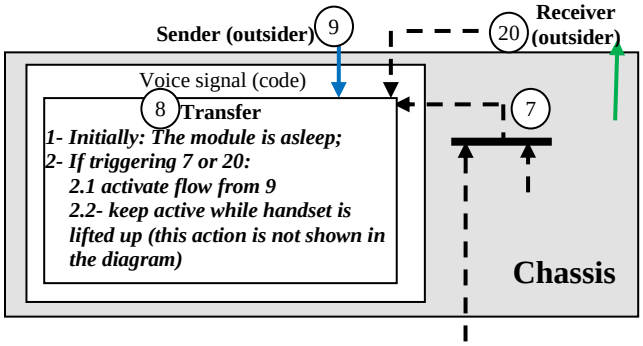


Fig. 11. Sample of the possible details of Transfer

B. Behavior and structure

According to Khanal [17], referring to his sources, a system can be described using two kinds of abstractions: structural abstraction and behavioral abstraction.

In physical models, the *structure* denotes the arrangement and relationship of components of the physical organization (the visible topology). Alternatively, the structure refers to a structural organization based on functional components [16].

In the structural view, the system consists of sub-systems that interact with each other to achieve functions that can be captured from the behavior view. To avoid complexity, systems are partitioned into minimally interacting sub-systems. Fig. 12 (taken from Eggert [18]) shows the synthesis of a cart transmission system and its functions from a behavioral view (left diagram) and a structural view (right diagram).

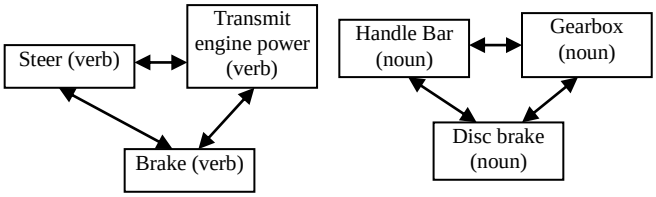


Fig. 12. Description by functions (left) and by structure (right) (Re-drawn from [18])

The designer is interested only in dealing with abstractions of both functions and structures. The abstract embodiment of a structure contains only the relevant information about the structure that is of interest to the designer. The abstract embodiment for the purpose can be expressed in terms of linguistic variables. [17]

Such sharp distension between structure and behaviors disturbs the Gestaltic depiction where behavior and structure express a holistic representation of the system. In FM, the structure is reflected in terms of spheres while the behavior is represented by the streams of flows. Fig. 13 shows the FM diagram of this cart transmission system. The cart (circle 1) has three spheres: Handle Bar (2), Gearbox (3), and Disc brake (4). However, these sub-systems receive instruction signals from the controller (e.g., driver - 5) who generates three types of signals (6, 7, and 8). Each of the sub-systems; Handle Bar (2), Gearbox (3), and Disc brake (4) embody two flow spheres: instruction signals and the execution of the instruction. For example, the controller generates a braking signal (8) that flows to the disc brake (9) where the signal is received and processed (10) in its flow system (for simplicity's sake, this has not been drawn in a box). The processing of the braking signal triggers (11) the generation of the brake action (12).

It is clear that flows are important factors in determining the structure of the system. Fig. 14 is an incomplete representation of the structure of the cart transmission system, which can be extracted from the FM representation, as shown in Fig. 13.

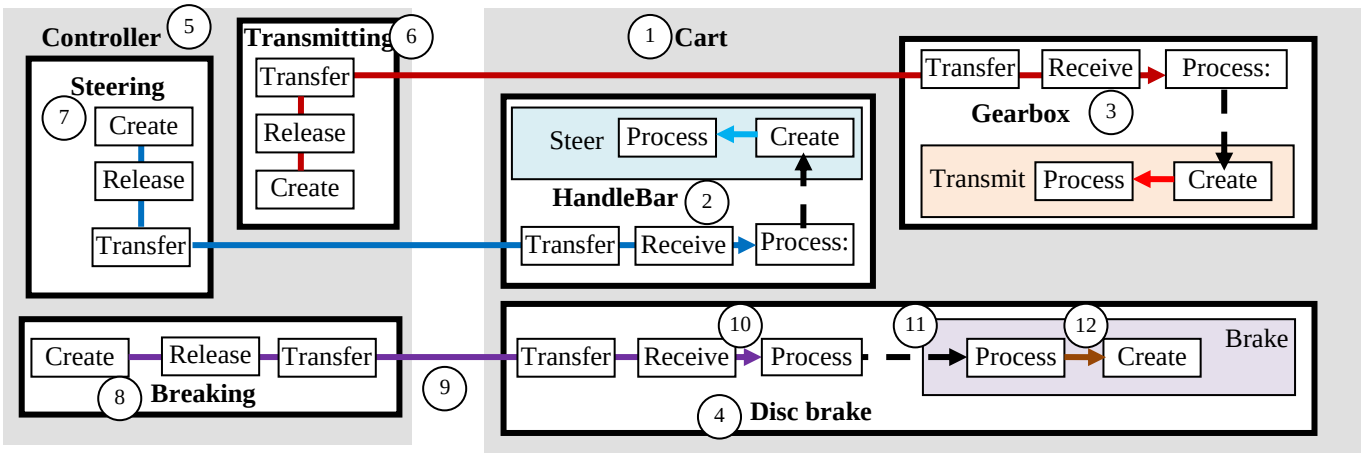


Fig. 13. The FM representation of the cart transmission system

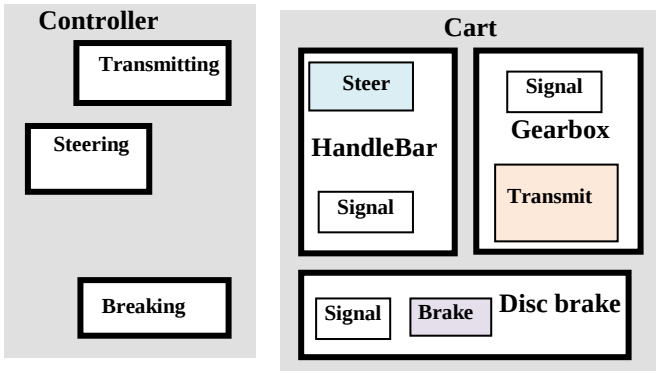


Fig. 14. The structure of the cart transmission system

C. Complex systems and purpose

According to Hmelo, Holton, and Kolodner [19], understanding complex systems is often difficult because of their multiple perspectives and the fact that their analysis may create conflict beyond the range of everyday experience. Design activities can be an excellent way to help students achieve a more systemic understanding of systems. Hmelo Holton, and Kolodner [19] give a simplified Structure-Behavior-Function Model of respiratory systems, which is shown in Fig. 15.

Structure	Behavior	Function
Lungs	Gas passes from high concentration to low across semipermeable membrane.	Bring in oxygen *for cells*, remove waste *from cells.*
Diaphragm	Lower pressure in chest by increasing volume.	Move *lungs* so they can take in fresh air.
Brain	Send signals *to respiratory system.* Receive and process *signals regarding body status.*	Control or regulate *movement of lungs in response to changing metabolic needs*

Fig. 15. A simplified Structure-Behavior-Function Model of the respiratory system (From [19])

Fig. 16 shows the corresponding FM diagram. First, the body sends signals regarding its status to the brain (circle 1), which are processed (2) to trigger the creation of an instructing signal (3) to the respiratory system (4). The diaphragm receives and processes (5) the brain instruction to either contract or expand (6 and 7, respectively). This causes the physical movements of the inhalation (8) or exhalation (9) of the lungs. The inhalation triggers (10) the pulling in of fresh air, which is processed (11) to generate (12) oxygen that can then flow to the cells (13). On the other hand, an exhalation causes the release (14) of carbon dioxide from the body's cells.

Over the years, there has been a great deal of functional representation research. A *function* of a system refers to its intended behavior [20] or *purpose* [16]. *What* a function intends to accomplish is achieved by *how* the behavior is implemented. In general, Kitamura and Mizoguchi [21] define the function of an entity as "a kind of abstraction of changes in objects associated with the entity."

In the FM diagram, the *purpose* can be specified according to corresponding sub-diagrams. Fig. 17 shows the *purpose* of the lungs in terms of two sub-diagrams:

- Inhalation of air to extract oxygen that is sent to cells
- Exhalation to remove carbon dioxide from cells

It should be noted that *purpose* is a flowthing that can be created, processed, released, transferred, and received. Fig. 18 illustrates *purpose* as one of the flow systems of the lung.

The point of these examples is to show that the notions of function, behavior, and structure can be discussed as features of the FM diagrammatic representation. All are obtained *uniformly* as global characteristics of the FM diagram in terms of spheres and sub-spheres, flow systems, flows, and triggering. This is in contrast to the FBS model in which doubts are raised about the meaning and suitability of these notions to the design process.

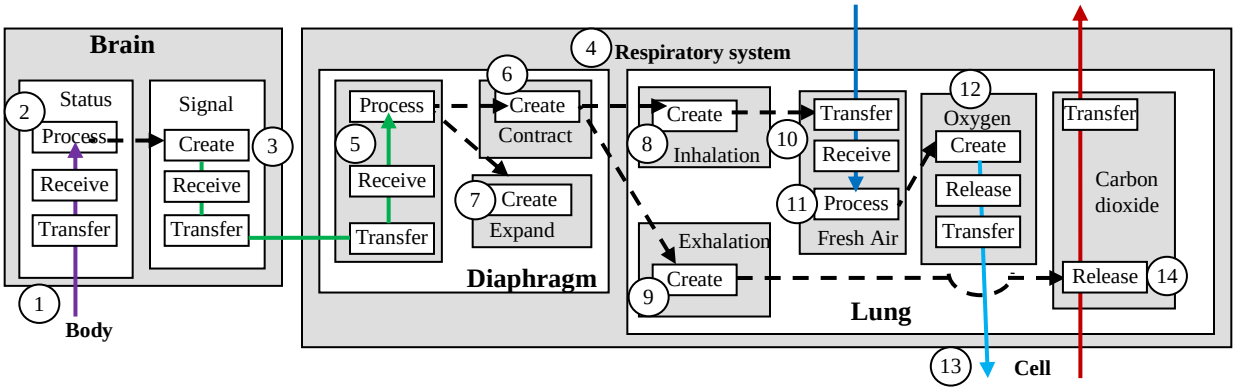


Fig. 16. FM representation of the simplified respiratory system

Function: Bring in oxygen *for cells*, remove waste *from cells.*

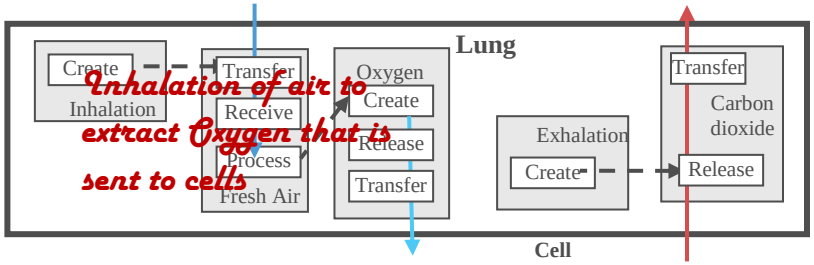


Fig. 17. The function as a sub-diagram

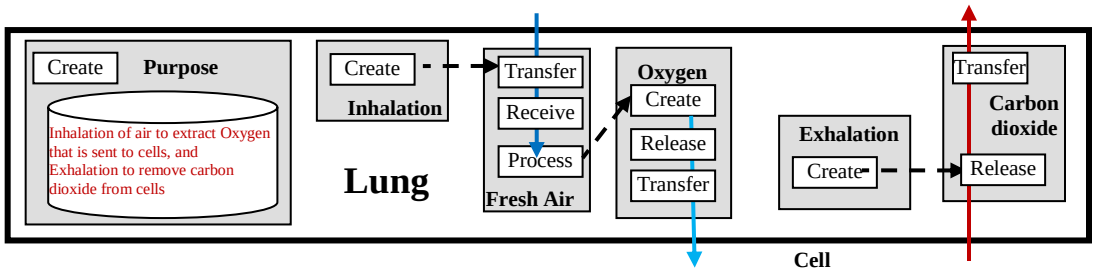


Fig. 18. *Purpose* is a flowthing that can be created, processed, released, transferred, and received

IV. Conclusion

The Function-Behavior-Structure model (FBS) of design provides no theory about how a function is transformed into behavior, and research has shown that its critical concepts have many different definitions. This paper suggests an alternative approach in which behavior is associated with streams of the flow of things in the system. Structure emerges as the "territories" of these flows, and function is associated with sub-diagrams of the total diagrammatic description. Behavior, structure, and function are all developed around the representation. The proposed conceptualization of design in terms of FM is still exploratory and needs more precise analysis; nevertheless, the approach seems to be promising as a unifying framework for the science of designing.

References

[1] J.S. Gero, "Design Prototype: A knowledge representation schema for design," AI Magazine, vol. 11, no. 4, pp. 26-36, 1990.

[2] M.J. French, Conceptual Design for Engineers, 3rd ed., London, United Kingdom: Springer, 1999.

[3] W.E. Eder and S. Hosnedl, Design Engineering: A Manual for Enhanced Creativity. Boca Raton: CRC Press, 2008.

[4] B. Helms, Object-Oriented Graph Grammars for Computational Design Synthesis, Ph.D. thesis, Technische Universität Müchen, March 2013.

[5] J.S. Gero and U. Kannengiesser, "The situated function behaviour structure framework," Design Studies, vol. 25, no. 4, 373e391, 2004.

[6] G. Cascini, G. Fantoni, and F. Montagna, "Situating needs and requirements in the FBS framework," Design Studies, vol. 34, no. 5, pp. 636-662, 2013.

[7] S. Borgo, M. Carrara, P.E. Vermaas, and P. Garbacz, "Behavior of a Technical Artifact: An Ontological Perspective in Engineering, Proceedings of the 2006 conference on Formal Ontology in Information Systems," Proceedings of the Fourth International Conference (FOIS 2006), pp. 214-225.

[8] Y. Chen, Z.Q. Lin, P.E. Feng, and Y.B. Xie, "Understanding and representing functions for conceptual design," Proc. International Conference on Engineering Design, ICED'07/223, Paris, August 2007.

[9] W.J. Zhang, Y. Lin, and N. Sinha. "On the function-behavior-structure model for design,' in Canadian Design Engineering Network Conference 2005, Kananaskis, July 2005.

[10] S. Al-Fedaghi, "Awareness of context of privacy," 7th International Conference on Knowledge Management, Pittsburgh, Pennsylvania, 22-23 October 2010.

[11] S. Al-Fedaghi and A. Alrashed, "Threat risk modeling," 2010 International Conference on Communication Software and Networks, Singapore, 26-28 February, 2010.

[12] S. Al-Fedaghi, "Privacy as a base for confidentiality," Fourth Workshop on the Economics of Information Security, Harvard University, Cambridge, MA, 2005.

[13] S. Al-Fedaghi, "Scrutinizing the rule: Privacy realization in HIPAA," International Journal of Healthcare Information Systems and Informatics, vol. 3, No. 2, pp. 32-47, 2008.

[14] S. Al-Fedaghi, "Scrutinizing UML activity diagrams," 17th International Conference on Information Systems Development, Paphos, Cyprus, 25-27 August, 2008.

[15] S. Al-Fedaghi, "Toward flow-based semantics of activities," International Journal of Software Engineering and Its Applications, vol. 7, no. 2, pp. 171-182, 2013.

[16] A.M. Keuneke, "A device representation: The significance of functional knowledge," IEEE Expert, vol. 24, pp. 22-25, 1991.

[17] Y.P. Khanal, "Object-oriented design methods for humancentered engineering design," Ph.D. thesis, Mechanical and Materials Engineering, University of Western Ontario, Ontario, Canada, 2010.

[18] R.J. Eggert, Engineering Design, Pearson: Prentice Hall, 2005.

[19] C.E. Hmelo, D.L. Holton, and J.L. Kolodner. "Designing to learn about complex systems," Journal of the Learning Sciences, vol. 9, no. 3, pp.247-298, 2000, DOI: 10.1207/S15327809JLS0903_2.

[20] M. Lind, "Modeling goals and functions of complex industrial plants," Applied Artificial Intelligence, vol. 8, pp. 259-283, 1994.

[21] Y. Kitamura and R. Mizoguchi, "Meta-functions of artifacts," Thirteenth International Workshop on Qualitative Reasoning, pp. 136-145, Loch Awe, Scotland, 1999.

Evolutionary Strategy of Chromosomal RSOM Model on Chip for Phonemes Recognition

Mohamed Salah Salhi
ISSAT-Mateur
LR SITI – ENIT

Nejib Khalfaoui
ISET Jendouba
LR SITI - ENIT

Hamid Amiri
ENIT- Tunis
Director of LR SITI Lab

***Abstract*—This paper aims to contribute in modeling and implementation, over a system on chip SoC, of a powerful technique for phonemes recognition in continuous speech. A neural model known by its efficiency in static data recognition, named SOM for self organization map, is developed into a recurrent model to incorporate the temporal aspect in these applications. The obtained model RSOM will subsequently introduced to ensure the diversification of the genetic algorithm GA populations to expand even more the search space and optimize the obtained results. We assigned a chromosomal vision to this model in an effort to improve the information recognition rate.**

Keywords—Information recognition; Recurrent SOM; Chromosomal RSOM model; Evolutionary RSOM; Implementation over SoC

I. INTRODUCTION

Voice recognition is complicated by the dynamic state highly variable of the speech signal. A technique often used is to decompose the signal into smaller atoms under stationary states representing phonemic entities able to be treated easily. This strategy motivates the ability to further improve the recognition score in stationary areas. A lot of paradigms were identified and competed in this area over this decade to bring more improvement on the phonemic recognition rate. This paper aims to contribute to the phonemes recognition of continuous speech. A neural model known by its power in static data recognition, named SOM for self organization map, is developed into a recurrent model to incorporate the temporal aspect in these applications. This idea overcomes a large adequacy between the ability robustness of the recognition tool and the speech signal to be processed. The obtained model RSOM will subsequently introduced to ensure the diversification of the genetic algorithm GA populations to expand even more the search space. This makes it possible to optimize results and avoid being trapped in a local search space. In each iteration, of the training or the test stage of the RSOM model by a certain input data, appears one BMU which is illustrated by one type of RSOM map. This type of map represents also one chromosome of the adopted genetic algorithm GA. It is designated by an individual. These individuals obtained through many iterations until a stop criterion involve populations which represent different phonemes of a continuous speech signal to be processed.

This work revolves around two points:

- The first one, about the modeling strategy of the GA-RSOM approach for phonemes recognition.
- The second point describe how to integrate this model into a system on chip Soc for a really practice.

II. THE SOM CHOSEN MODEL FOR PHONEMES RECOGNITION

1) Adopted Approach

The Experiments of phonemic recognition by SOM are operated on a multi-speakers speech basis. The process is as follows:

- Reading and segmenting the speaker's speech into 10ms samples to benefit from these stationary supports.
- Transform these sound supports into codes vectors consisting of MFCCs coefficients. Some approaches use the MFCCs of central windows of phonemes where energy is considered maximum [1].

In our approach, we considered the MFCCs of all windows of phonemes where all the energy will be presented. The choice of MFCC coefficients in determining the acoustic vectors, depends on comparative studies applied on phoneme recognition rate, can be generated under different characteristic parameters of a signal. The following example shows the scores on a test basis of 100 phonemes of TIMIT database realized by Bruno Gas in his habilitation at the University of Paris 6 in 2005.

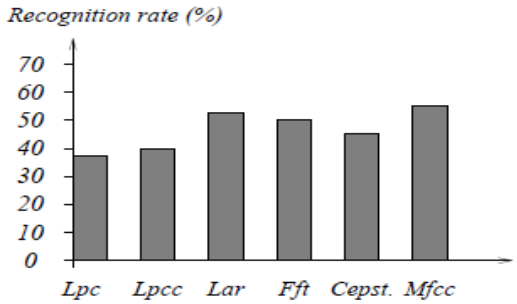


Fig. 1. Recognition Rates over different types of coding [2].

- Classification of the MFCC coefficients into lists of macro classes.
- Determination of phonemes list from each class.
- Conversion of phonemic data related to each class into a data structure 'sD' coherent in dimension to the SOM map structure. This step is very important because it adapts the input data to be accepted by the SOM map.

- Eventually, we train the SOM map, after study and choice of parameters such as size, network topology ... etc..
- Finally, we repeat the trainings of the SOM map through a TIMIT test basis which is specific and contain fewer samples. In this step, we train the SOM map on TIMIT training basis and we do the test of phonemic recognition program on the test data structure, to obtain the generalization rate. The notion of the SOM map generalization is based on the idea that to learn well data does not the fact of learning by heart, but it must be able to perform well before any points and in variables situations [3].

2) Dynamic extension of SOM

The temporal data processing is a very important task, for which there is no unified approaches. In RSOM algorithm, each prototype vector has a weight representing its position in the distribution, and also has another context vector representing the activation of the whole map in the previous iteration. The selection of the closest prototype in this case is based on a distance making account, on the one part, the difference between the data and the prototype weight and in the other part, the difference between the previous context and the actual context of the prototype. The prototype update requires changing the weight and context behind the winner prototype and its neighbors [3]. This idea was developed by Thomas Voegtlin in 2000, and it is represented as below in our phonemic recognition strategy.

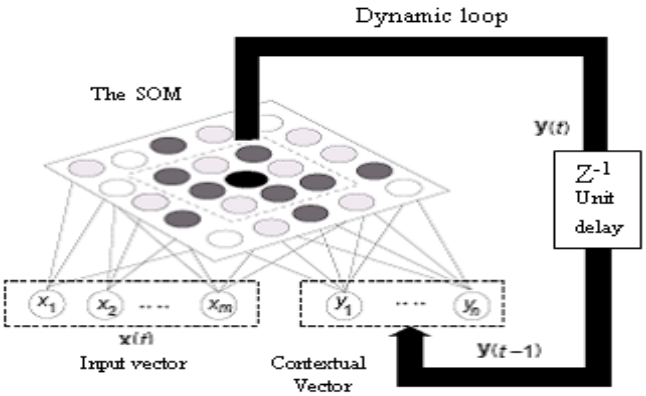


Fig. 2. The feedback loop over SOM introducing the temporal context (RSOM)

Also, this approach was deducted from the Elman SRN (Single Recurrent Network) structure invented in 1990. The SRN is a modified perceptron with one hidden layer, using a delayed copy of the activities of its hidden layer as an additional input. His task is to learn associations of input / output sequences. He trained with the error propagation algorithm (see fig.3):

The representation in the layer context of an SRN is the same as in its hidden layer. Therefore, we say that the hidden layer learns to represent his past activities, as it will receive each time its past activities. In this sense, the representation in the hidden layer is called self referent. This self reference will influence learning by acting on the error function between desired output and actual output located at a definite level of iteration [4].

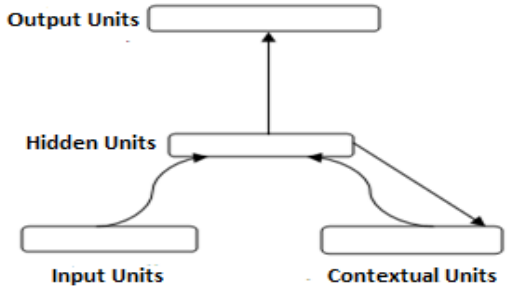


Fig. 3. Principle of Elman network structure SRN with feedback loop introducing the idea of self reference

Thus, the output of each neuron in the output layer can be modeled by a leaky integrator based on an active low pass filter which integrates the temporal aspect by operation of charging and discharging of the storage capacity.

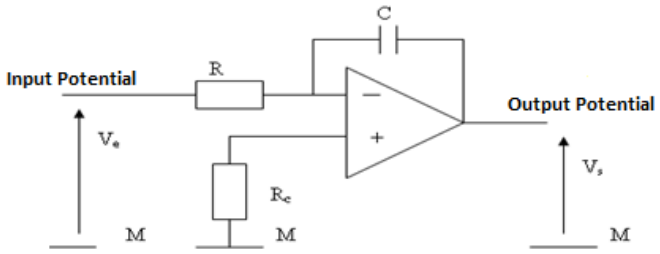


Fig. 4. Modeling of neuron outputs in the RSOM map by an electrical low-pass filter

The time constant of the RC circuit is expressed in seconds. This filter circuit added to an operational amplifier is referred to as integrator, which introduces the time factor, by effect of charge and discharge time constant T = RC.

The capacitor is initially discharged: Uc (0) = 0 volts. The discharge equation of initially charged capacitor to the potential value E is given by:

$$U_c(t) = E.e^{-t/RC} \quad (1)$$

3) Experimentation and Results

We are interested in a speech recognition tool with multi speakers regardless of the context. This requires having at our disposal a large amount of vocabulary, for learning and recognition in continuous.

There are currently many databases that words were recorded especially in English. Our research focuses on the application of SOM and RSOM for phonemes recognition in TIMIT databases. The wide dissemination of this base in the international community allows an objective assessment of performance. This corpus of speech data called: DARPA TIMIT was prepared at the National Institute of Standards and Technology (NIST) funded by Defense Advanced Research Projects Agency Information Science and Technology Office, (DARPA-ISTO) to study the acoustic variability of American English on different dialects and different regions for multiple

speakers. These dialects are referred by 8 directories DR1, DR2, ..., DR8, which contain the records of 630 American speakers (from the U.S.) saying 10 sentences each.

The total vocabulary base is 6300 sentences, shared between 630 speakers, including 438 men and 192 women, as follows:

- 462 speakers were including 326 men and 136 women, for the learning set.
- 168 speakers were including 112 men and 56 women for the test set.

This database contains a phonemic segmentation and accurate labeling that affect learning models.

The proposed method gave phonemic recognition results, envisaged at the table below:

TABLE I. RPECOGNITION RATE OVER SOM FOR TIMIT DATABASE PHONEMES

Phonemes Rec. by SOM	Self-coherence Rates in %	Generalization Rates in %	Margin Rates in %
Vowels	75.07	54.16	20.91
Semi-vowels	87.10	73.00	14.10
Nasals	95.65	66.14	29.51
Fricatives	83.58	62.32	21.26
Stops	76.05	51.12	24.93
Others	76.34	55.23	21.11
Affricates	91.17	84.38	6.79
Global Rates	**83.56**	**63.76**	**19.80**

These experiments show that the ability of Kohonen algorithm depends on many parameters such as the vector input dimension, the SOM map dimension, the iteration numbers, the speech's sample numbers taken in training or in test stages, and the phoneme classes. It means that each neural unit in the map specializes in a particular kind of data. This model gives a result of generalization rate 61.66% compared to a training rate of 82.96%. This result seems trapped in a local maximum. It is then necessary to extend our research space.

We propose then to introduce dynamic recurrent loops over the map to integrate the time aspect. Eventually, we will offer a possible hybridization schematic of the SOM map with the genetic algorithm GA.

The results are given by the both comparative figures:

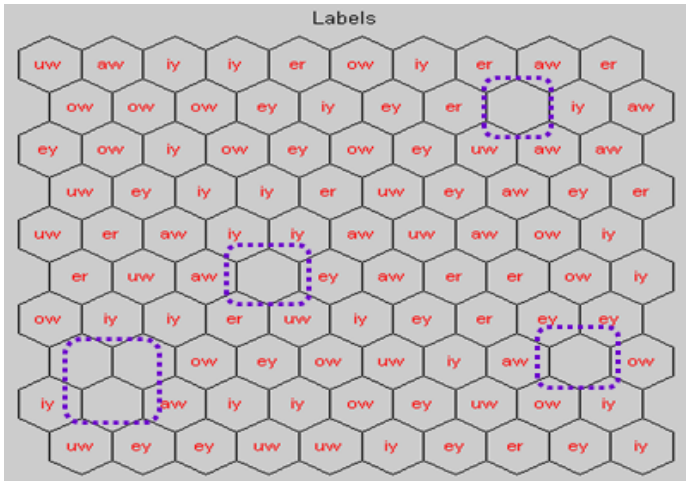


Fig. 5. Phonemes recognition results in static SOM.

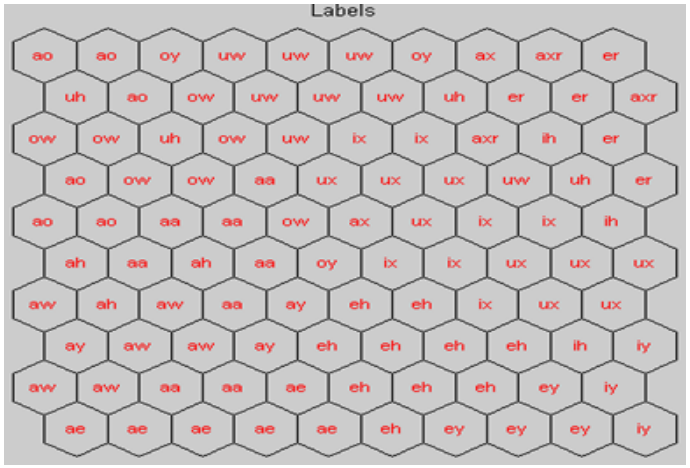


Fig. 6. Phonemes recognition results in dynamic SOM.

III. ADOPTED STRATEGY FOR AN EVOLUTIONARY RSOM

The basic strategy of the GA RSOM concept that we proposed for raising the phonemic recognition rate revolves around the following main ideas:

- Establish recurrence loops on the neuronal map SOM designed to collect and recognize static data; each neural unit of the map is considered as static combinatory circuit. The recurrence loop can integrate the memory effect to every cell, so the integration of the temporal aspect during processing of received data. This latter provides a dynamic quality for each neuron that will be compatible with the pattern shape of each phonemic support variability to be recognized.
- The resulting model RSOM will have a certain time diversity and a certain winner neural diversity taking into account certain parameters such as the number of phonemic data inputs, the number of neurons in the RSOM map that determines the frequency of the phonemic input positions, the extent of the neighborhood function that represents the bandwidth

enlargement of the selective filter BP for different feature vectors corresponding to phonemes, and the number of iterations during the learning phase or network test.

- The iteration will end with a BMU: a winner neuron that specializes in feature vectors sequentially provided to the RSOM network input.
- Each BMU representing a phoneme characteristic vector is considered as a chromosome vector. This vector carries the singularity of individual traits giving a diverse population of the RSOM maps.
- The diversity of BMU samples promotes the field to apply the genetic algorithm GA over the different speech phonemes. This allows extending the search space and avoiding to be trapped in a local optimum solution by confirming the survival and the recognition for the best phonemic individual. [5], [6], [7], [8].

This strategy is described following the below diagram:

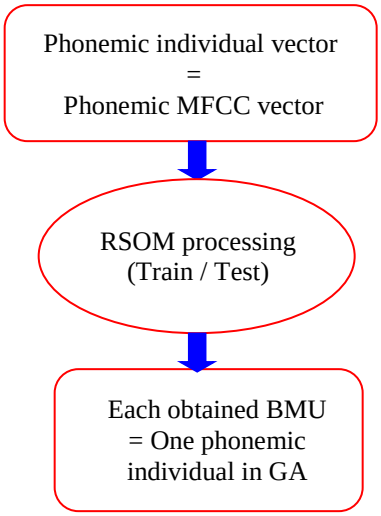


Fig. 7. Technique of GA-RSOM Hybridization

This idea is abstracted on the following algorithm.

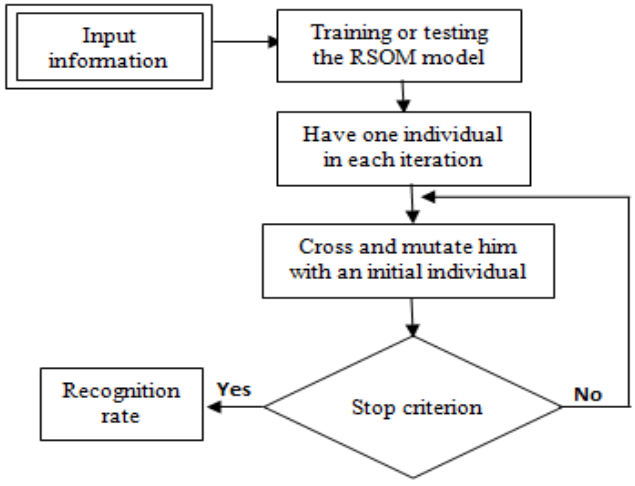


Fig. 8. The UML Unified Modeling Language for adopted Strategy

Each individual is assigned to one obtained BMU over iteration. It will be represented by a concatenated MFCC vector corresponding to one phonemic chromosome related to an RSOM map [9].

IV. The GA-RSOM Paradigm

If you choose effective parents, it is very probably that the offspring have an efficacy at least as important as their parents; it is the selection and crossing principle of species to assure the survival at best. In this way, the application of genetic algorithm on maps RSOM for phonemic recognition, performs a global search for solutions by avoiding local minima and can estimate many parameters varying in ranges of important values.

In our approach, the data are necessarily temporal and the recognition tool is recurrent Kohonen map RSOM which the time factor is introduced by its differential equation.

The genetic evolution of the map optimizes recognition rate in a search space as large as possible. The research will be guided by a cost function associated to the developed model and reflects the individuals' effectiveness in a given generation. The chosen function is described by the following expression [10], [11], [12]:

$$fitness = \frac{1}{1+\sum_{i=0}^{N}\sum_{t=0}^{T}\left(x_i(t)-w_i(t)\right)^2} \quad (2)$$

$X_i(t)$ represents the characteristic coefficients of the input data vector.

$W_i(t)$ represents the synaptic weight vector of each neuron.

This objective function provides a means of evaluating scores of individuals in a generation. It is between 0 and 1, and is even greater than the map weight is close to the data input. That is to say the difference between the input observations and the output solutions will be reduced.

1) The selection technique

The individuals' selection for reproduction is made by random bringing following the given distributions by the Fitness function, more the fitness of a chromosome solution is good, more it is closer to 1, and more the chance of bringing it at random is higher. The random selection is made according to the empirical distribution of the relative fitness of individuals. The selection algorithm is presented as follows:

a) Calculate the fitness of each candidate at the selection.

b) For each candidate i, we associate the value:

$$P_s(i) = \frac{\sum_{j=1}^{i} fitness(j)}{\sum_{j=1}^{N} fitness(j)} \quad (3)$$

Where, fitness (j) is the score of candidate j; N is the candidates number. The quantity Ps(i) is situated between 0 and 1. We select randomly the 0 and 1, then we chose the candidate i number (n Ps) between such as:

$$P_s(i-1) \le nP_s < P_s(i) \quad (4)$$

The selection must be served to the cross and mutate operation. [13], [14].

2) Algorithm of the GA-RSOM proposed Model

The speech is constituted by a phoneme set. Every phoneme represents a sound atom characterized by certain stationarity. This specificity limits the ability of the RSOM tool which incorporates the temporal aspect in phonemes recognition. To overcome this constraint, we tried to consider, in our experiments, all the phonemic support windows instead of taking the central window where concentrates the maximum energy of the signal. This solution secures all information of a phoneme even in adverse environmental conditions. Similarly, the consideration and implementation of a recurrent dynamic model such as GA-RSOM for phonemes recognition allows better identifying the variability of speech. This idea opens another way of research such as the recognition of an isolated word or some keywords, where recursive models are very interesting.

Else, our evolutionary model around the RSOM map has the principle of an adaptative tool; it is scalable in the objective of optimizing the obtained recognition results. The tracking algorithm begins by creating an initial population which consists of initializing the neural weights of a developed RSOM map. Sequences of serial data will be provided in the entrances for their identification. Each phonemic coded vector will transmitted towards each neuron of the considered map. After a learning or test phase, each neural unit specializes in one type of input vectors having the closest form.

Our evolutionary model takes advantage of the diversity of neuronal units winners BMU obtained during the iterations of the RSOM maps.

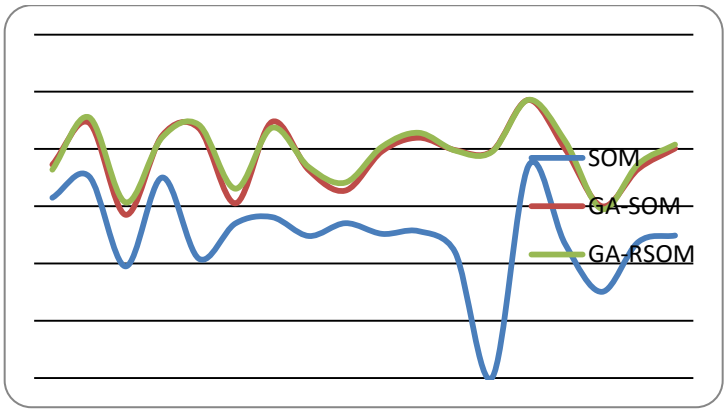


Fig. 9. Comparative Rates between Models for Vowels Test

The diversity of the BMU constitutes chromosome diversity for different individuals of an RSOM population.

This diversity offers more chance to expand the search space and to have the best descendants may participate in ensuring the survival to the best along the generations until a stopping criterion which result in an optimization of an improve recognition rate.

The pseudo-code of our evolutionary algorithm revolves around the following points:

a) Linear initialization of the RSOM network to create an initial population.

b) Admission sequential of phonemic feature vectors to the RSOM map.

c) Prototyping of BMU representing individuals in the population.

d) Computing the individual scores seeing a fitness function.

e) Applying a geometric selection to cross parents.

f) Computing the genes quality of each parent.

g) Applying the crossover between parents to obtain a child.

h) Mutate every child with certain probability.

i) Looping from selection step until the production of new population step.

j) Define an individual's 'BMU' through iteration.

k) Computing the winning recognition rate.

l) Check the algorithm stopping criterion.

3) Experiment results

The experimentation of our proposed model is performed on phonemic class richest in energy of vocal cords vibration; containing twenty vowels extracted from TIMIT database. A comparative study was established under different phonemic recognition tools such as SOM, GA-SOM and GA-RSOM to target the most appropriate model to optimize the phonemic recognition rate.

These values are listed in the two following tables then represented by figure 9 and figure 10.

TABLE II. VOWELS RECOGNITION RATE FOR TIMIT TRAINING BASIS

Phonemes	SOM	GA-SOM	GA-RSOM
'aa'	72.60	96.17	96.83
'ae'	76.37	95.51	96.02
'ah'	70.50	62.00	59.17
'ao'	77.10	80.50	83.11
'aw'	83.75	71.82	76.99
'ax'	72.60	44.30	45.16
'axr'	74.15	82.58	83.00
'ay'	73.83	52.89	54.10
'eh'	68.41	84.85	86.88
'er'	82.71	84.00	81.45
'ey'	78.10	77.47	80.34
'ih'	63.91	80.43	80.50
'ix'	54.23	87.29	89.66
'iy'	75.48	93.17	95.23
'ow'	72.86	80.00	81.19
'uh'	85.28	63.16	65.72
'uw'	90.40	78.36	77.94
'ux'	83.25	80.64	82.38
Means rate	74.95	78.85	79.11
Established time	**t = 14 h for GA-SOM t = 14 h 30 min for GA-RSOM**		

TABLE III. VOWELS RECOGNITION RATE FOR TIMIT TEST BASIS

Phonemes	SOM	GA-SOM	GA-RSOM
'aa'	62.97	74.54	72.77
'ae'	70.46	89.18	91.09
'ah'	38.98	57.00	61.34
'ao'	70.04	84.74	84.12
'aw'	41.66	87.25	88.45
'ax'	54.01	61.00	66.13
'axr'	56.19	89.52	87.44
'ay'	49.59	72.64	73.69
'eh'	54.08	65.48	68.28
'er'	50.35	79.20	80.80
'ey'	51.28	83.96	85.67
'ih'	43.80	79.61	79.44
'ix'	60.41	79.20	79.00
'iy'	74.20	97.16	97.21
'ow'	46.78	80.00	82.73
'uh'	30.08	60.00	59.01
'uw'	47.54	73.07	74.78
'ux'	49.76	80.19	81.55
Means rate	52.89	76.35	77.64
Established time	**t = 13 h for GA-SOM t = 13 h 25 min for GA-RSOM**		

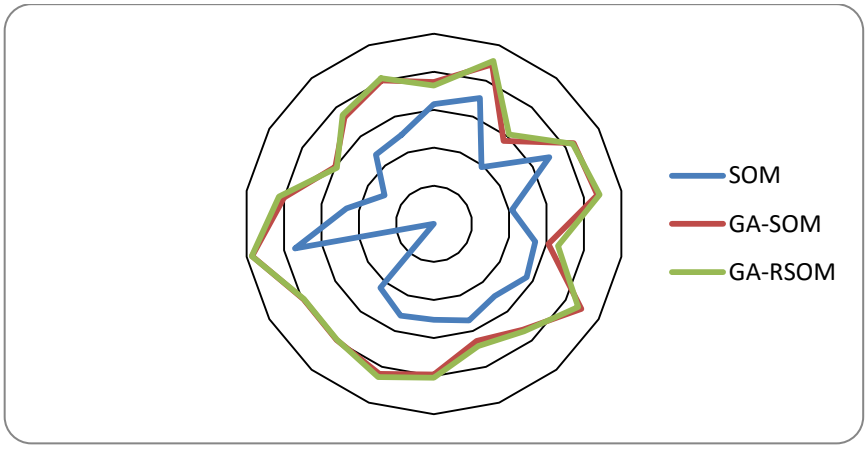


Fig. 10. Explanatory Diagram of Comparative Rates between Models

These results show that the model SOM is limited to static data. While hybridization of SOM by the GA gives a slight improvement in recognition rate because it is always around a model of SOM core which can handle only static data.

However, the application of recursive evolutionary GA-RSOM model promises more in the results as it integrates dynamic aspect comparably to the variability of phonemic support.

V. MODELING SOM ON FPGA

1) Similarity factors between the SOM map and the FPGA

An FPGA is an integrated circuit, which based on configurable logic bloc (CLB), Programmable connection matrix and RAM blocks to implement complex digital computations [15], [16], [17]. CLB is a configurable element, permit to FPGA to be a configurable tool and to be used for hardware verification. Each CLB, which characterized by its logic architecture containing a combinatorial static part and a dynamic part used for memorizing can be simulated for a neuron, which featured by its effects memory, erase and re memorization skills; This specificity allows us to think of shaping CLBs in a way to mimic the behavior of the SOM map.

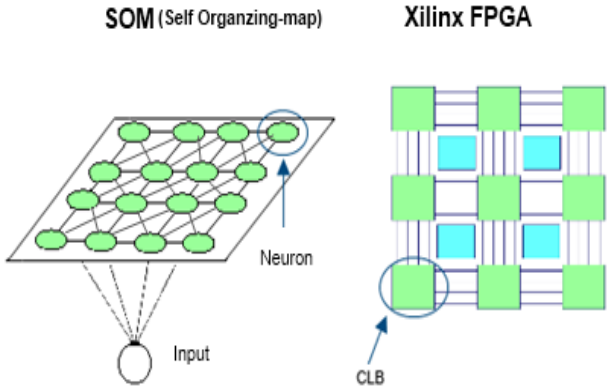


Fig. 11. Comparative block diagrams between the SOM and FPGA

The main feature of Kohonen neural network manifests in that each neuron of it is connected to all other neurons. In parallel on FPGA, each Configurable Logic Bloc can be linked to all other CLBs owing to the programmable connection matrix.

Even the algorithm of Kohonen is suitable to the FPGA due to the fact that both of it work on cycled mode. Therefore, iteration of Kohonen algorithm will be a clock cycle on the FPGA achievement.

2) Modeling strategy of SOM on FPGA

This work will start by modeling and implementing an artificial neuron in hardware, which will be a Xilinx FPGA. Secondly, it will proceed to modeling the whole SOM and implement it on FPGA.

The SOM will be basically composed of one unit artificial neuron either called processor. It has some inputs, which will represent entries that mimic the dendrites of a biological neuron and an output that mimics the axon and which serves to spread information to other neurons.

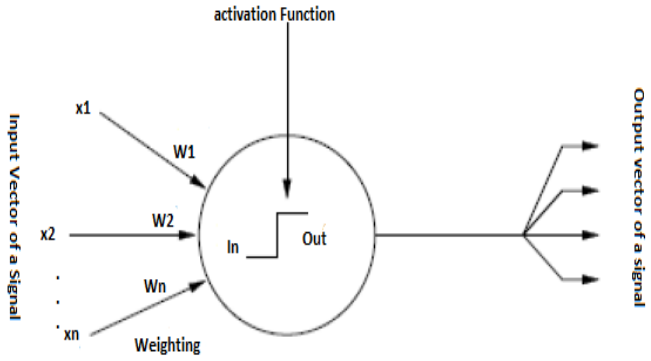


Fig. 12. Representation of an artificial neuron

The cells of the competitive neural layer of SOM are grouped according to their learning similarities. Therefore, the cells are sorted so that their neighbor has almost the same characteristics.

The operating principle of each artificial neuron is described as follows by this algorithm.

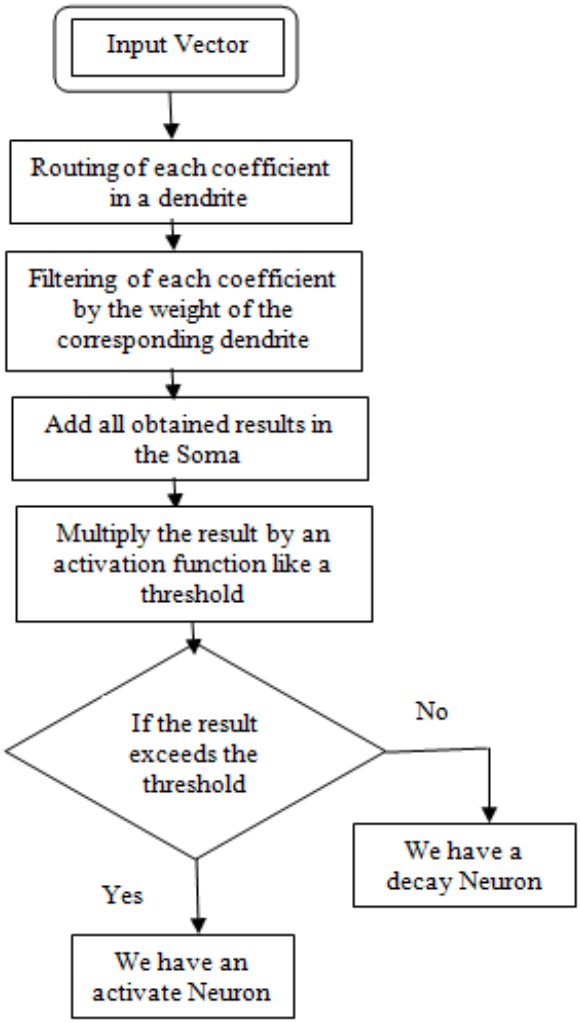


Fig. 13. The UML Diagram of an artificial neuron behavior

After this algorithm accomplished by each neuron within a SOM map, all activate neurons were compared sequentially one by one to the input vector.

Therefore, the neuron which best represents the input vector is the winner called the BMU for Best Matching Unit in a considered iteration. It is the neuron having the minimum of Euclidian distance between the input vector coefficients and the vector of weight codebook at the beginning of Soma. This idea will be clarified by the following algorithm.

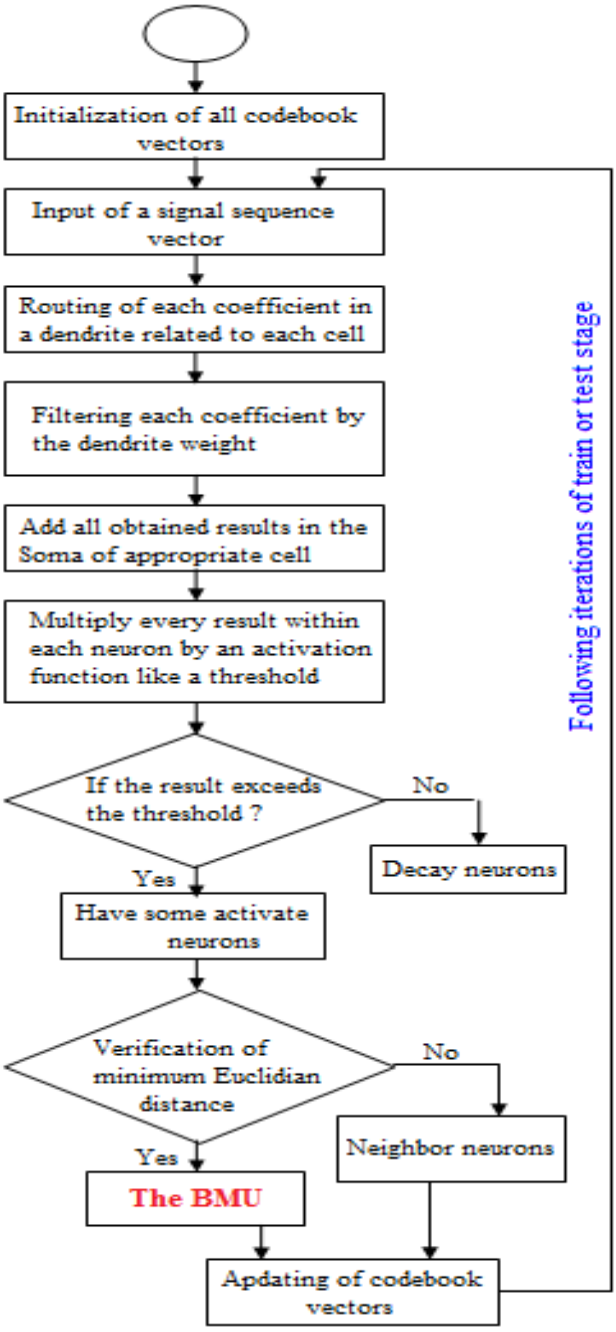


Fig. 14. The UML Diagram of a SOM map behavior

The behavior of our modeled system of SOM will be coded on FPGA. Thereby, because the waveform which happens to stimulate a biological neuron is represented by spikes, so we will use square pulse as it is shown in Figure 15 and as it is a digital signal which is suitable to be processed by every CLB within the FPGA.

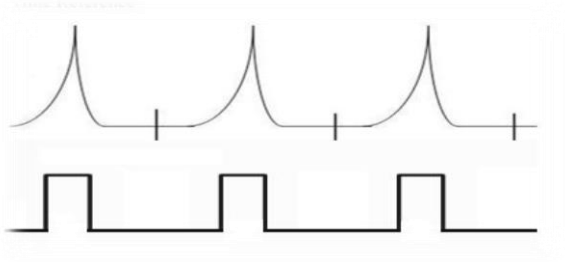


Fig. 15. Representation of spikes by square pulses to be processed by neurons

When a spike arrives, the soma has two functions; to generate the potential action according to the input data and to compare, if the addition of all potential actions in this instance is over a threshold, in that case, it will have to generate a pulse through the axon.

In order to facilitate the design task for neuron modeling, the method "top / down" is applied.

The way to cope with the problem is using a top-down method that consists to divide a complex design in easier designs or modules; each module is redefined with more details or divided in subsystems. For that, a general idea over the system is made in the first view, and if you go down through the subsystems, you can see with more detail how each block works.

In the developed system, each artificial neuron will have only three dendrites as inputs to facilitate the tasks.

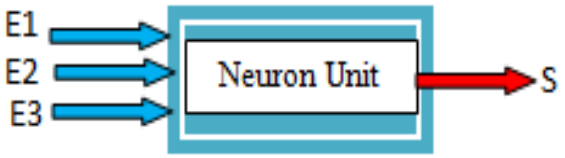


Fig. 16. A basic modeling of neuron

Each neuron unit is compared to a cell memory; which likes a D (Data) flip-flop. It is therefore considered as a processor unit that can handle only binary information.

For processing information and analog signals, these intake dendrites must be preceded by an analog / digital converter ADC. The received pulses are therefore a square wave as defined previously.

The received binary signal is weighted by the weight of each dendrite to be summing and thresholding using an activation function that is defined in the operation of the neuron unit algorithm.

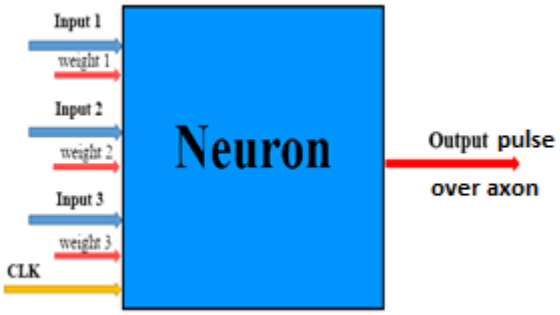


Fig. 17. Modeling different ports I / O of a formal neuron

Each neuron is modeled by a block containing multipliers floors of sample coefficients related to the input signal by the appropriate weights to each dendrite; followed by a block as the sum of the obtained products and a threshold ensuring the neuron activation decision.

This will be illustrated by the following figure.

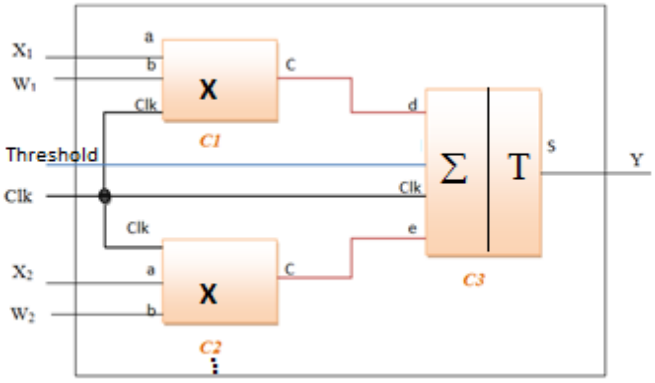


Fig. 18. Representation of the neuron's internal behavior

The output Y from previous diagram is considered as the input of a D (Data) flip-flop.

In our system, the SOM map is represented by four neuronal units on four D flip-flop. The four outputs are compared using a comparator. The first activated D flip-flop will be considered the Best Matching Unit BMU. She is the one with the closest information of the input vector.

In a deeper analysis, we will include other important blocks. We have to introduce more inputs to get the system can learn quickly. Also, it is necessary to introduce a clock signal in the system. All the signals will be digitals, for that, it is recommended a signal that synchronizes the global system.

The implementation of the FPGA device under Xilinx follows this diagram.

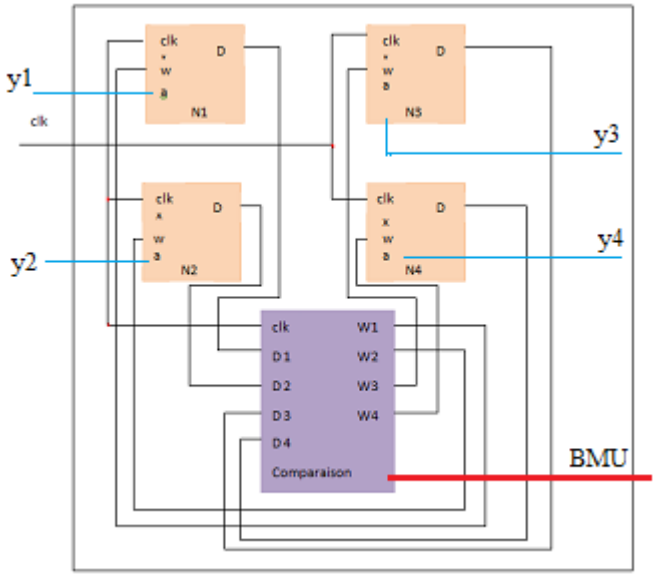


Fig. 19. Modeling of the SOM map by the D flip-flops and a comparator

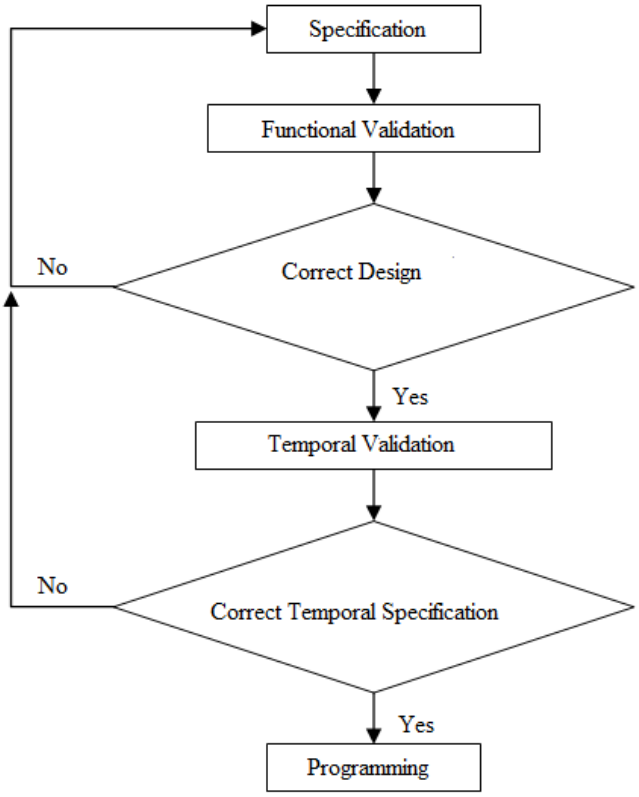


Fig. 20. The design flow diagram of a circuit on FPGA

The specification step comprises the choice of the logic device using a syntactic specification in the language VHDL, or Verilog, or in graphics mode.

The functional validation is based on a functional simulation of the concept. This is to see problems of inputs/outputs, loops, etc. This audit does not take into account the temporal aspects of the Device.

The temporal validation includes the temporal and functional simulation of the created device on FPGA, such as the propagation time, the signal overlap, etc.

At the implementation stage, the program will be brought physically on the created circuit, as a project on FPGA, according to the specifications specified by the programmer in the light of pins allocation and internal behavior.

VI. Simulation Results

Our experiment is based on the development of an FPGA type cyclone II over the Xilinx software. Then we created the SOM map adopted model on this FPGA by applying the VHDL language and following the design protocol developed in previous Section.

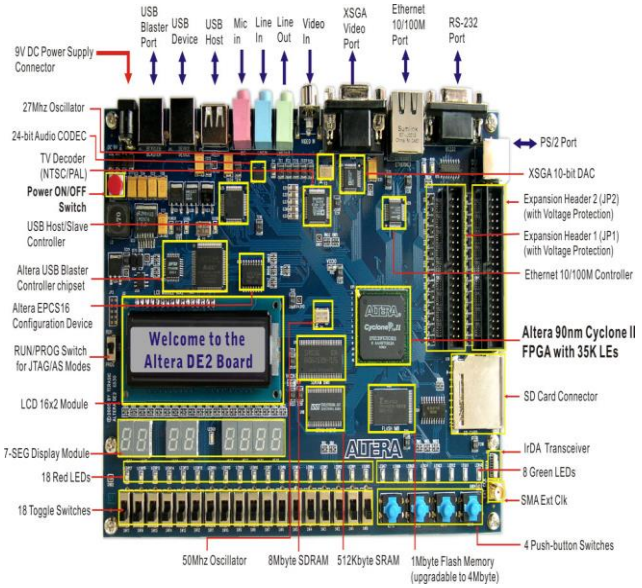


Fig. 21. The schematic of FPGA type Cyclone II

The simulation result of a SOM map on FPGA, made of four neural processor units, is established at the following two figures.

During this simulation, we took all the weight w1, w2 and w3 to a value of 1. Thus; a period of 10 ms is allocated for the clock signal (clk).

Similarly, we chose different periods for i1, i2 and i3 which are the three dendrites of a neuron.

S1, S2, S3 and S4 are the neurons outputs.

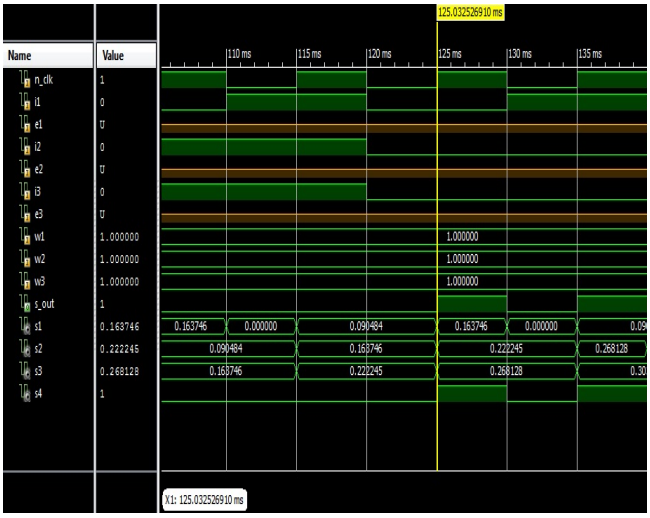


Fig. 22. The simulation result of SOM on FPGA without activated neuron

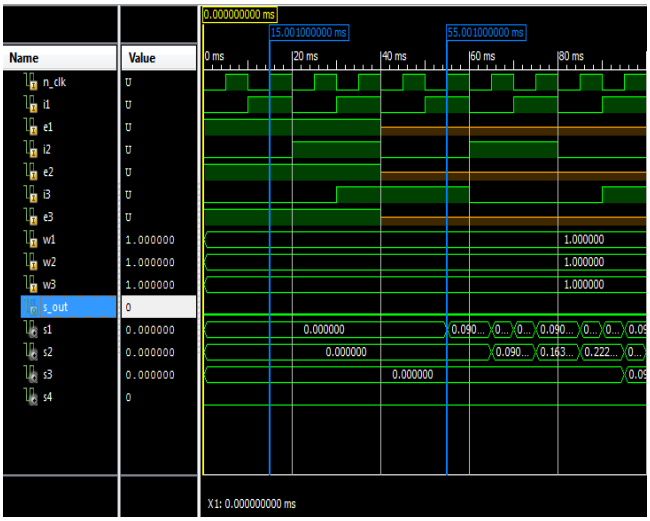


Fig. 23. The simulation result of SOM on FPGA with an activated neuron (The BMU)

We find in the first marker at 15 ms from Figure 22 that although i1 equal to 1 and the clock is at the leading edge, the output S1 remains at 0 because e1 is not enabled.

In the second marker at 55 ms, we note that the first dendrite is excited (i1 = 1), the clock is also at the leading edge and e1 is enabled, this implied a generated result on the signal S1.

This simulation shows that at 115 ms the sum of 3 signals S1, S2 and S3 doesn't pass the neuron threshold (threshold = 0.5), which means the exit "Sout" remains 0,

Against, by the next rising edge at 125 ms, we find that the sum of the latter outputs exceeds the neuron threshold, which shows at Figure 23 the activation of a neuron that called the BMU and "Sout" became equal to 1.

VII. CONCLUSION

In this paper, we have developed an evolutionary model GA-RSOM. His experimentation gives promising results (table 2 and table 3) by appearing to mean recognition rate for other models such as the SOM, the GHSOM, and GA-SOM for static data.

By applying a recursive loop on SOM we could introduce the dynamic temporal aspect of this model RSOM.

Similarly, we considered the best matching unit BMU, obtained in each RSOM iteration, as a chromosome bearing the characteristics of an individual selected from a diverse population by application of genetic algorithm GA.

This idea increases the research space and avoids the rapid convergence of the algorithm related to our hybridized paradigm.

Subsequently, we proposed a technical implementation of this model on an FPGA to materialize the adopted approach.

REFERENCES

[1] S. Mallat: 'Une exploration des signaux en ondelettes', Editions de l'Ecole Polytechnique, 2000.

[2] Dan Weaver and Nick Mount: 'The use of Kohonen mapping for the elucidation of space-time trajectories of multiple parameters: potential applications in fluvial geomorphology', School of Geography, University of Nottingham, Nottingham NG7 2RD. 2007.

[3] Teuvo Kohonen, Helsinki University of technology: 'Self-organizing maps', november 16, 2000.

[4] Thomas voegtlin: 'Recursive Principal Components Analysis', Inria-campus scientifique, Nancy-France, 2002.

[5] Mohamed Salah Salhi, Noureddine Ellouze: 'A suitable model of evolutionary SOM for phonemes recognition', journal IRECOS Napoly-Italy, (Indexé COMPENDEX – Elsevier–Copernicus). Laboratoire LSTS- ENIT_Tunis, septembre 2009.

[6] J.H. Holland: 'Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to biology', control and artificial intelligence. MIT Press, ISBN 0-262-58111-6. 1998.

[7] Christopher Houk: 'A genetic algorithm for function optimization: a matlab implementation', north Carolina state university-1994.

[8] Mohamed Salah Salhi, Noureddine Ellouze: 'Principal temporal extension of SOM', journal IJSIP Koréa, laboratoire LSTS-ENIT_Tunis, septembre 2009.

[9] Mohamed Salah Salhi, Noureddine Ellouze: 'The SOM Robustness Capacity for Phonemes Recognition in Adverse Environment', International Journal of Computer Applications (0975 – 8887) Volume 60– No.1, December 2012.

[10] Mohamed Salah Salhi, Noureddine Ellouze: 'Ability of Evolutionary and Recurrent SOM model GA-RSOM in Phonemic Recognition Optimization', International Journal of Mathematics Trends and Technology- Volume4 Issue 6 - July 2013.

[11] N. Khalfaoui, M. S Salhi, H Amiri ''Anomaly Detection in Induction Machines'' 2nd International Conference on Automation, Control, Engineering and Computer Science; ACECS-2015 Sousse, Tunisia.

[12] B. Trajin, J. Regnier, J. Faucher, Indicator for bearing fault detection in asynchronous motors using stator current spectral analysis, International Symposium on Industrial Electronics, Juin-Juillet 2008, pp. 570-575.

[13] Julien BRICHE: 'Adaptation d'un algorithme génétique pour la reconstruction de réseaux de régulation génétique : COGARE', thèse de l'Université du Sud Toulon, 2009.

[14] Heni Ben Amor and Achim Rettinger: 'Intelligent Exploration for Genetic Algorithms Using Self-Organizing Maps in Evolutionary Computation', February 4, 2005.

[15] E. Ayeh, K. Agbedanu, Y. Morita, O. Adamo, and P. Guturu: 'FPGA Implementation of an 8-bit Simple Processor, Department of Electrical Engineering University of North Texas, Denton, TX. 76207 USA, IEEE 2008.

[16] Amrinder Kaur, Mandeep Singh, Balwinder Singh: 'VHDL Implementation of Universal Encoder for communication', ISP Journal of Electronics Engineering,Vol.1,Issue 2,ISSN 2250-0537, .December 2011.

[17] Tanmay Biswas, Sudhindu Bikash Mandal, Debasree Saha, Amlan Chakrabarti: 'A Novel Reconfigurable Hardware Design for Speech Enhancement Based on Multi-Band Spectral Subtraction Involving Magnitude and Phase Components', School Of Information Technology, University Of Calcutta, 2015.

An Intelligent Agent based Architecture for Visual Data Mining

Hamdi Ellouzi[1], Hela Ltifi[1,2] ,Mounir Ben Ayed[1,3]
[1]Research Groups on Intelligent Machines
National School of Engineers (ENIS), University of Sfax, BP 1173, Sfax, 3038, Tunisia
[2] University of Kairouan, Faculty of sciences and techniques of SidiBouzid, Tunisia
[3] University of Sfax, Faculty of sciences of Sfax, Tunisia

***Abstract*—the aim of this paper is to present an intelligent architecture of Decision Support System (DSS) based onvisual data mining. This architecture applies the multi-agent technology to facilitate the design and development of DSS in complex and dynamic environment.Multi-Agent Systems add a high level of abstraction. To validate the proposed architecture, it is implemented to develop a distributed visual data mining based DSSto predictnosocomial infectionsoccurrence in intensive care units. The developed prototype was evaluated to verify the architecture practicability.**

Keywords—Multi Agent System; Decision Support System; Visualization; Knowledge Discovery from Data; Nosocomial Infection

I. INTRODUCTION

Decision-making in dynamic and complex environment faces several problems as a result of the increase in temporal data size and the diversity of heterogonous data sources. Integrating Data mining technology in Decision Support System (DSS) assists decision-makers in problem solving. Data mining algorithmsprovideuseful patterns to discover data associations [1] [2].Data mining as two main goals: (1) extracting relevant information from a set of raw data according to user's request to have coherent knowledge about the system's variations (2) transforming data from textual representation into meaningful formsto specify data associations. The complexity of the previoustasks realization increases in case of temporal data. The integrationof visualization techniques[5]in data mining based DSS provides an acceptable outcome in clear graphical formto better understands temporal data and extracted patterns variation[5] [6] [7]. Thus, we are interested to visual data mining based DSS. Such systems are characterized by their complexity and dynamic character. To support this complexity, intelligent agent technology is a promising solution [3] [4].

In this context, our work aims to propose a multi-agent architecture for visual data mining based DSS,which ensureshigher level of adaptability, mobility, and intelligence. It involves a set of active goal-oriented agents to play one or more roles in the decision environment.

This paper is organized as follow: in section 2, we present our researchcontext including decision support system, visual data mining and intelligent agent technology. Then in section 3, we introduce the suggested architecture based on cognitive intelligent agents. Next, we present the architecture applicationin the medical field. In section 5, we will apply a set of evaluation utility and usability tests to validate the developed prototype.

II. RESEARCH CONTEXT

A. Visual Intelligepnt Decision Support System

Decision doesn't refer to a specific step clearly identified[8]. It is based on several phases defined in previous works such as[9],[10] and [11]which relyon Simon'sdecision process called ICDRbased on four phases: (1) *Intelligence* to extractrelevant information, (2) *Design* to generate a set of related models presenting different scenarios that may occur, (3)*Choice* to opt for one solution among theproposed scenarios, and (4) *Evaluation*aimingat reviewing the results found in all previous phases.

Decision processis integrated in decisional tools called decision support system (DSS)to assist domain experts tofind solutions for problems andmake decisions to improve or to adjust a current situation. The DSS developed tools suggested in previous works are applied in several domains. Among these works, we state [12] that proposed a DSS which assist user to find the best route in case of travel challenge. A decisional tool having as goalsplan and manage support in energy companies was proposed in [13]. Furthermore,a DSS relying on objective and subjective criteria to improve quality of service in digital library by generating a set of recommendations was suggested in [14] works.

The data analysis for decision-makingallows defining a set of parameters restriction to limit the search space. It provides more efficiency and clarity. Data mining algorithms gives patterns that will be analyzed by decision-makers. We are interested thus in data mining based DSS.

Several works were interested in integrating visualization methods and techniques in data mining for decision-making.In fact, it is recommended to integrate the Human in the data exploitation process, which is known as visual data mining. In this case, we ensure the integration of Human knowledge with the biggest computer capacity storage. Raw data, data mining process and generated patterns can be interactively and graphically presented to the user.

As consequence,numerous visual tools were initiatedin different context of use; e.g. CAST (Clustering And visualizing Spatio-Temporal data)[15] is a visualization tools that ensures

moving entity control in order to study and evaluate them.A SIMID called tool developed by [16] provides decision-maker spatio-temporal visualization of infection. Visual temporal tool for distant monitoring was introducedMittelstädt and his colleagues [17].Based on this brief literature investigation, weset a first objective. It consists of seeking to make data-miner and decision-maker able to visualize data mining patterns, draw conclusions in real-time and interact with data in the different data miningsteps.

Visual data mining and DSSbecome more and more complex due to temporal data continuous progress.To face this complexity, the main problem can be divided into sub-problems; each one is assigned to a sub-system to reduce the complexity.So that, intelligence can be distributed into different parts of the system and their sub-systems.

B. *Intelligent agents technology*

The Multi-Agent technology consists in implementing a distributed intelligence in complex environment. Each task should undergo a localprocessing. An agentis a computer entity situated in an environment and capable of acting in an autonomous wayand can reach designed goals for which it was conceived [18]. We believe that each part of the whole complex and visual system can be assigned to a particular agent. Each agent belongs to one of the following categories:

1) Reactive agentdevoid from memory and environment representation, it relies on communication with agent's environment to solve problems [19]. It reacts according to its reflex without maintaining any internal state.

2) Cognitive agent: if the agent have a memory and able to realize an environmentsymbolic representation, and can take into account its past in order to reach an explicit purpose.

3) Hybrid agent: combines the twoprevious categories. The hybrid agent follows its plans. It can sometimes directly reacts to external events.

The agent technology, as we are mentioned previously, is based on autonomous andcognitive intelligent agents that havebeen applied in various domains to perform divers tasks. Agent technology is used on several DSS works to accomplish a define objective according the case. For exampleNeLH project [20] involves intelligent agents looking to find the available medical center of a given geographic area. A MAS based DSS was suggested by [21] to predict patient state according to context.

Contrary to the previous agent technology relatedworks; we will focus on proposing a set of cognitiveagents that aims to improve the effectiveness of the visual data mining based DSS. We call such system as ***v**isual **i**ntelligent DSS* (viDSS).So, our context dealswith viDSS based agent concepts that will be detailed in the next section.

III. VIDSS PROPOSED APPROACH

In section2, we are studying works related to viDSS conceptsand agent technology. We are claimedthat, as far as we know,there is no work thatbrings them together. So we propose to consider then inproposing a new architecture of viDSS relying on agent technology.

A. *viDSS modules*

According to [22], theviDSS architecture is based on four principal modules (cf.Fig.1)

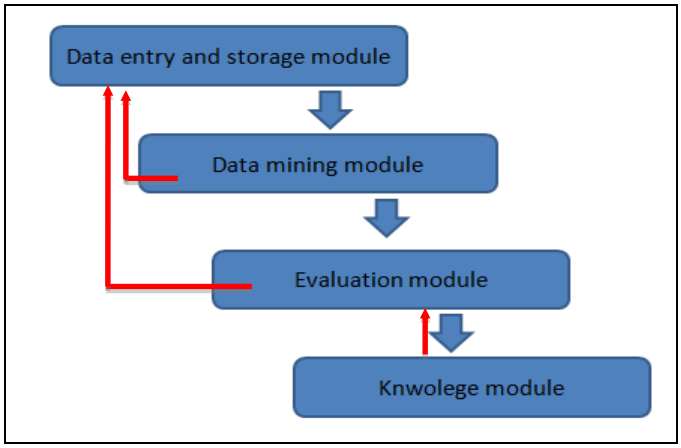


Fig. 1. viDSS principal modules

- Dataentry and storage module: this module consists ofdata selecting, pre-processing and transformation steps for data mining technique.
- Data mining module: in this module a data mining algorithm (e.g. association rules, Bayesian dynamic network, etc.) is applied totransformed data in order toextract useful patterns as output.
- Evaluation module: patterns provided by the previous module are evaluated. According to the evaluation result, it consists of moving to the next step or to make a feedback.
- Knowledge management module: after evaluation, in casepatterns are relevant and give a pertinent knowledge. This knowledge will be integrated for decision-making.

The viDSSprovides userwithvisualization of providedresults in different modules. For this reason, we added a new module to process the visualization tasks (c.f. Fig.2) called visualization module integrated in the four viDSS cited modules.

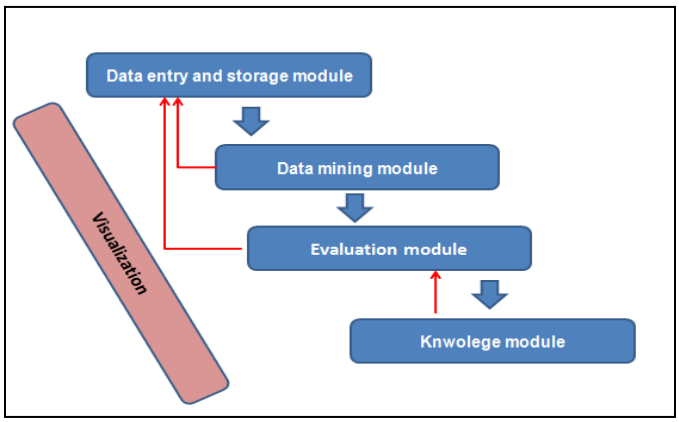


Fig. 2. Visualization module integration

B. *Intelligent agent integrated on viDSS archirecture*

After defining the different modules of viDSS architecture (cf. section III.A), we have to define the various intelligent agentsto be integrated,as well as the tasksassigned by each onein different modules (c.f. Fig.3).

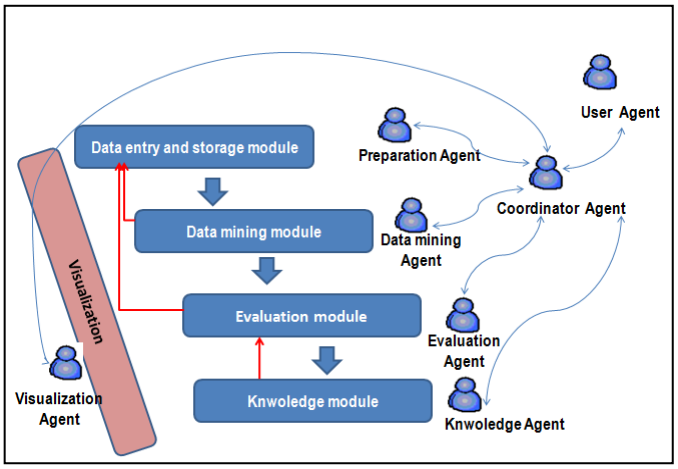


Fig. 3. viDSS integrated agents

The coordination between all the defined agents is controlled by in intelligent agent called "coordinator agent" (c.f. Fig.4).

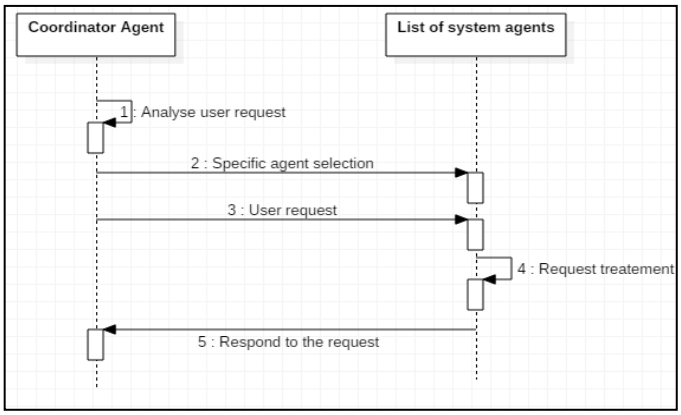


Fig. 4. Coordinator agent task

The coordinator agent is able to identify the user (Data-miner or Decision-maker), as well as the agent's works progress and specify for each one the task to do. Following, we present the agents by module.

1) Data entry and storage module involved agent

In this first module, an agent called "data preparation agent" ensures the cleaning and the pretreatment tasks to resolve the missing valuesin the raw data (c.f. Fig.5).

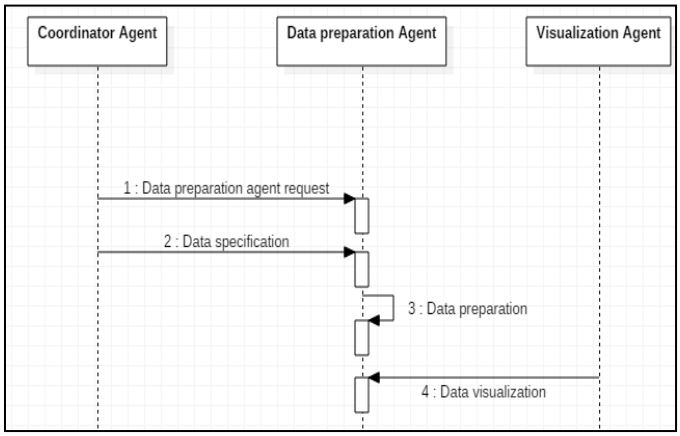


Fig. 5. Coordinator and data preparation agents interaction

The "data preparation agent" receives a request from "coordinator agent" to resolve the missing or invalid values in temporal data. The data preparation agent applies a specific algorithm to get back the missing values. The result of preparation process can be visualized by calling the "visualization agent".Prepared data provided by this agent will be used in themodule of data mining.

2) Data mining module involved agent

After receiving prepared data from "data preparation agent", the "data mining agent" applies a data-mining algorithm togeneratepatterns that allow the futureprediction(c.f. Fig 6).

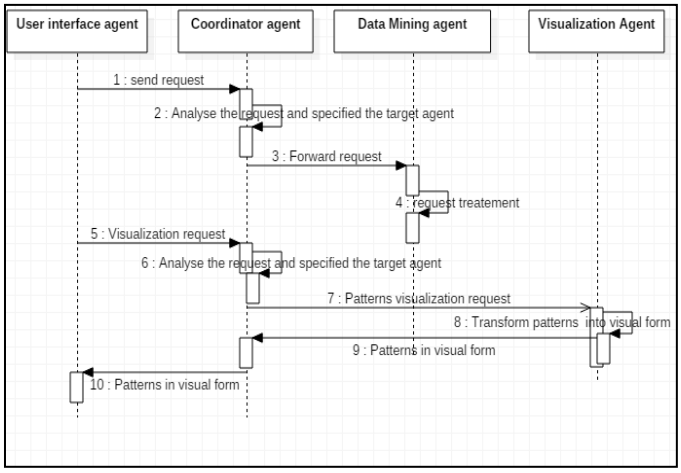


Fig. 6. Data mining agent interactions

The user (i.e. data miner / decision-maker) is able to visualize results of data treatment in different modules thanks to the "visualization agent" providing an easier interaction between user and system.

3) Evaluation module involved agent

An agent called "evaluation agent" achieves the patterns evaluation task.It ensures also thefeedback to the previous task if evaluation results are not valid by collaborating with the "coordinator agent"(c.f. Fig.7).The evaluation is realized through a set of criteria learned by the agent.

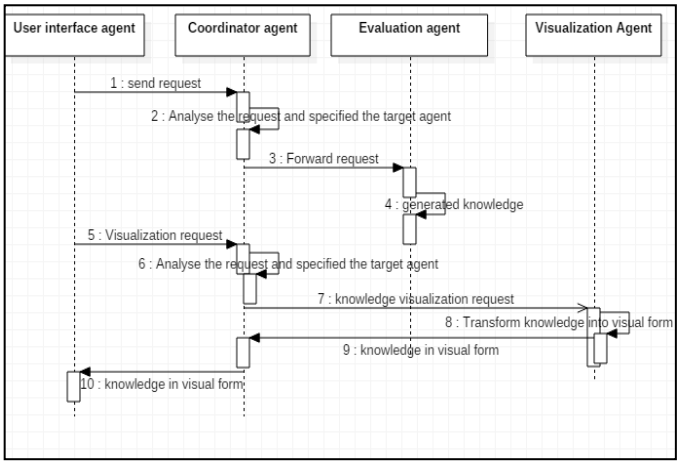


Fig. 7. Evaluation of data mining patterns

4) Knowledge management module

Knowledge management task is assigned to "knowledge agent".It has two roles; first it generates knowledge from patterns and second integrates it in knowledge base in case of it was validated by the "evaluation agent", else knowledge will be rejected and a new knowledge generation is started until we got a validate knowledge (c.f. Fig.8).

User is able to visualize the extracted knowledge, so he/she indicates the needs to visualize the extracted knowledge from models. As consequence, a communication between "Visualization agent" and "knowledge agent" occurred in order to present knowledge in a visual form.

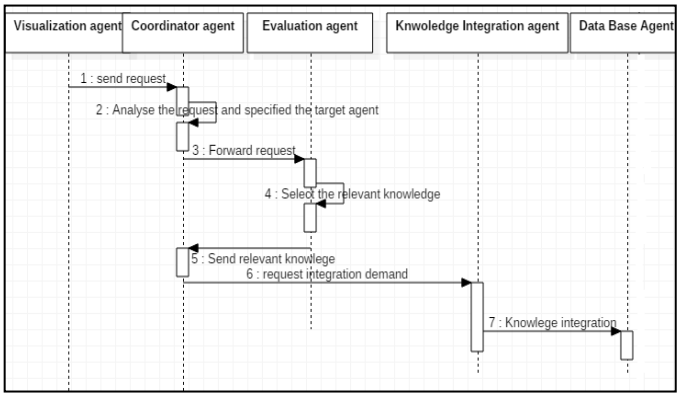


Fig. 8. Knowledge and evaluation agents interactions

C. Agents interaction modeling

The viDSS involved agents are cognitive, they can learn from their previous acts. Furthermore they have the capacityto discover environment in which they belong. Theycommunicate with each other to carry out a specific task(c.f. Fig.9).

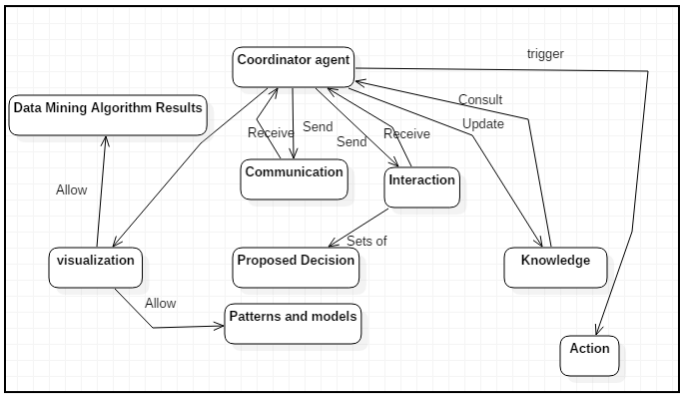


Fig. 9. Intelligent agent interaction

Each agent can send and receive message for another, it is a collaborative environment that looks for improving results quality and make the viDSS architecture center-user.

The « coordinator agent » is the principal agent of the system. It ensures the communication and the interaction between all the system agents. It receives requests and sends responses from/to agents according to the request nature. It aims to schedule the agent tasks to reach the objectives. The "coordinator agent" facilitates the interaction and collaboration between the user and the viDSS. It makes possible a quick and an easier visualization of data, patterns and generated knowledge.

IV. viDSS Developped Tool

The NI detection is acomplex process and presents a challenge because itis based on analysis and interpretation of temporal medicaldata such as performed actsand antibiotics doses. This complexity is proved by a study realized onhospitalized population that contains 280 patientsbetweenApril 17th 2002 (midnight) and April 18th2002 (midnight). This studyshown that 18 % of total number has beenaffected bythe NI [23].The NI detection is a complex task due to the big number of temporal data analyses continuously (e.g. antibiotic doses, clinical tests values, etc…). In order to supervise the NI, physicians are based on prevalence survey given by the means of several tools and devices. The divertools give different data values with different types, so that the temporal data visualization is not an easier task. Several works havepublished such as that of [24]and [25] whichlook for real solutions to decrease the number of affected persons and to detect infection in early stages. In spite of those efforts there is a lack ofdeveloped toolsbased on MAS for thepredictionofNIconcurrence.

In order to validate the intelligent agent based viDSS architecture introduced in the previous section, we have developed a prototype called "viDSS_NosDetc" for the physicians of the Intensive Care Unit (ICU) of theteaching hospitalHbibBourguibaSfax, Tunisia. This tool enables physicians toidentify infected patients inhospital intensive careunitthanks to: (1) visual data mining ofthe temporal data, providing knowledge and (2) the obtained results of temporal data interpretation.In addition, we intended toguarantee the

tool capacity toimproveinteraction between user and system by meansof visualization of handled data and patterns. The applied data mining technique is the Dynamic Bayesian Networks (DNB) [26], which relates variables to each other over adjacent time phases on any day, with t is between patient entry date and leaving date(in days)(c.f. Fig.10). We present the development of our MAS based viDSS architecture by module.

A. Data entry and storage module development

We have first developed a “database agent” to ensure the collection and the storage of temporal ICU data in the database. The collected data can contain missing values, which need a specific treatment. Based on “data preparation agent” intelligence, the missing values are recovered.

B. Data mining module development

The “data mining agent” applies the DBN algorithm in order to generate model that represent the pertinent information. As mentioned above, we have developed the DBN algorithm. The causal graph of the developed algorithm is shown in the following figure (c.f. Fig 10).

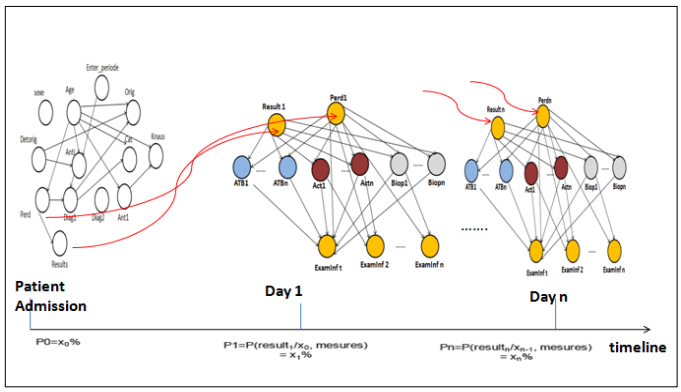


Fig. 10. Causal dependencies in a dynamic BN

The variables used as input of DBN is presented in the following table (cf. Table I).

TABLE I. TEMPORAL VARIABLES

Measure	*Description*
Result	The NI occurrence probability
Hospital stay	The period between the patient admission and exit from the ICU
Antibiotic (ATB)	The daily antibiotic catch during the patient stay in the ICU.
Infectious examination (ExamInf)	The daily performed infectious examination
Acts	The daily carried out acts in the ICU
Biological parameters (BioP)	The daily measured biological parameters

C. Evaluation module development

The generated patterns resulting from the data-mining agent are evaluated by the evaluation agent. We have developed an automatic data preparation algorithm that allowsto analyze the data and identifying corrections, eliminates problematic or unnecessary fields, deriving new attributes when necessary and improves performance with intelligent scanning techniques.

D. Knowledge managment module development

The developed knowledge agentalgorithm integrates the evaluated pattern as new element (NI occurrence probability) in the knowledge basefor further decision-making. Based on this probability, viDSS suggests solutions. The system output helps physicians to make the best decision in real-time.

The following figure (c.f. Fig.11),is the main interface ofour viDSS prototype.

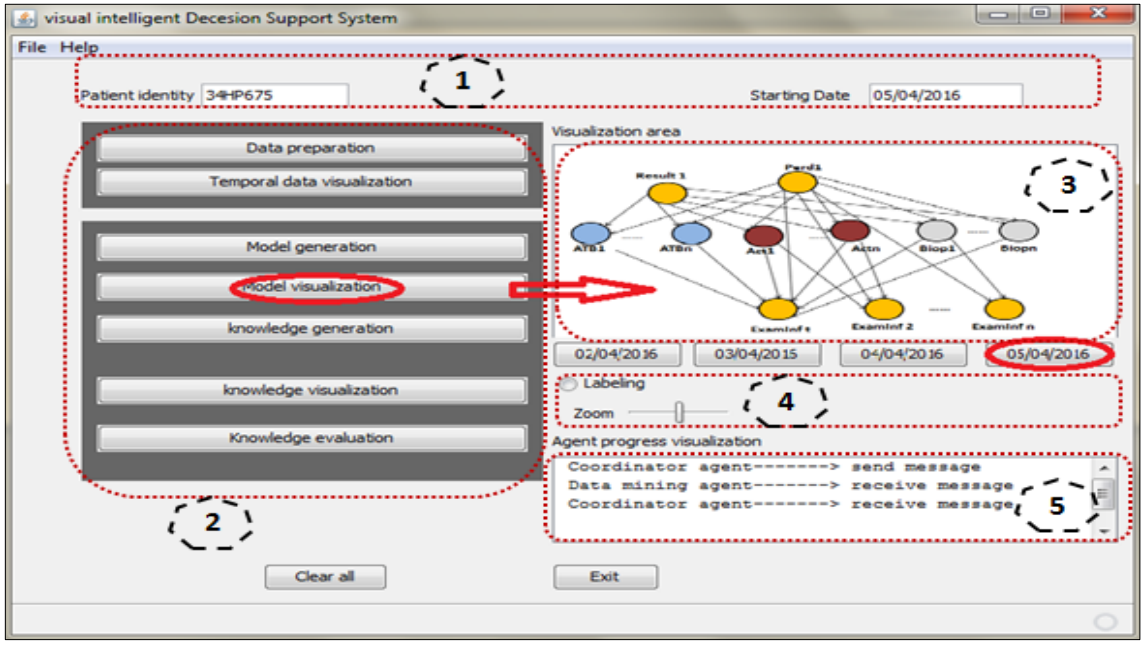


Fig. 11. viDSS tool

The user (decision-maker) specifies at the beginning the set of parameters related to the supervised patient (c.f. Fig.11(1)), then he/she chooses the task to do by clicking on the specific button (C.f. Fig.11(2)). Each task is performed by an assigned cognitive agent. The visualization task provides both temporal data (c.f. Fig.11(3)) and mining results visualization (c.f. Fig.11(4)).The user gets moredetailed abstract of temporal data variation at different time granularities by choosing one among interaction options(c.f. Fig.11(6)). After having achieved the mining tasks, interestingpatterns areextracted (c.f. Fig.11(5)) and integrated for possible decision solutions suggestion. If the physician acceptedthe suggested decision and the evaluation resultshave been acceptable, knowledge will be integrated in knowledge base to be used for future decision-making.

V. EVALUATION

After developing the "viDSS_NosDetc" tool, we move to its validation that deals with utility and usability evaluation[27] [28] [29] [30] tomake sure that the user is satisfiedwith the provided services.We begin with the first evaluation dimensionaiming at measuring the system performance through confusion matrix.

A. Utility evaluation

The utility evaluation intended to verify if results provided by different agents, such as data preparation, data mining and data knowledge agents are significant by considering the observed NI probability. This comparison is realized by means ofconfusion matrix application.

TABLE II. CONFUSION MATRIX

MAS-VIDSS			
		Observed results	
		Yes	No
Predicted results	Yes	10	4
	No	4	46
The accuracy rate: 79%			
The negative capacity of prediction: 87%			
The positive capacity of prediction: 74%			
Previous version of the system [31]			
		Observed results	
		Yes	No
Predicted results	Yes	9	8
	No	7	34
The accuracy rate: 77%			
The negative capacity of prediction: 90%			
The positive capacity of prediction: 60%			

As visible in the table 2, ourdeveloped viDSStoolprovided significant and interesting results.The tool's usability evaluation consists of assessing the visualization agent results in terms of visual representations.

B. Usability evaluation

The evaluation test checks the quality tool usability. To be sure from the user satisfaction, so we asked him/her about the degree of satisfaction (using a questionnaire). The evaluation test gives the following results visible in the figure 12:

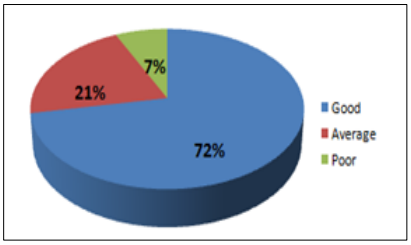


Fig. 12. Evaluation rate

This usability evaluation shows that 72% of the interviewees judged that the system was ''Good'' for decision-maker or data-miners (cf. Fig 12). We can conclude that a great number of users are satisfied with tool interfaces and the ergonomics of system.

VI. CONCLUSION

In the following paper, we have focused on the design and development of visual Intelligent Decision Support based on Multi-Agent architecture. The used agents are intelligent and cognitive.For each agent we have affecteda specific KDD task. The viDSS developed agents communicate using the coordinator agent. The proposed architecture make user able to easily interact with the viDSSoutputs (e.g. temporal data, patterns and knowledge) using the visualization agent.

The suggested approach is implemented for the development of amedical tool called "viDSS_NosDetc". It is used to detect the NI for the patients residing in the Intensive CareUnit of the teaching hospital HbibBourguibaSfax Tunisia. The viDSSutility and usability evaluation proves that the developed tool satisfies its users.

In our future works, we intend toextend the current architecture to support the storage and the treatment of the big data concepts. As a second contribution, we intend to develop the viDSS architecture to be applied on other distributed environments.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of this research by grants from the ARUB program under the jurisdiction of the General Direction of Scientific Research (DGRST) (Tunisia).

REFERENCES

[1] H. Ltifi, G. Trabelsi, M. Ben Ayed, and A.M. Alimi, "Dynamic Decision Support System Based on Bayesian Networks, application to fight against the Nosocomial Infections", International Journal of Advanced Research in Artificial Intelligence (IJARAI), Vol. 1, no. 1, pp. 22_29, 2012.

[2] K. Rajesh, A.K. Pujari and D.S. Reddy, "Clustering techniques in data mining- A survey", *IETE Journal of Research*, Vol. 47, no. 1, pp. 19_28, 2001.

[3] R. Bose, and V. Sugumaran, V. "Application Of Intelligent Agent Technology For Managerial Data Analysis And Mining", *Data Base*, Vol. 30, pp. 77_94, 1999.

[4] T. Rahwan, T. Rahwan, I. Rahwan, and R Ashri, (Eds.), Agent-Based Support For Mobile Users Using Agentspeak (L), Germany, Lncs, Springer, 2004.

[5] D. Keim, "Information Visualization and Visual Data Mining", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, no. 1, pp. 1_8, 2002.

[6] H. Ltifi, E. Ben Mohamed and M. Ben Ayed, "Interactive visual KDD based temporal Decision Support System", Information Visualization, Vol. 14, no. 1, pp. 1_20, 2015.

[7] E. Ben Mohamed, H. Ltifi and M. Ben Ayed, "Using visualization techniques in knowledge discovery process for decision-making", The 13th International Conference on Hybrid Intelligent Systems (HIS 2013), Tunisia, pp. 94_99, 2013.

[8] A.Akharraz, Acceptabilité de la décision et risque décisionnel : Un système explicatif de fusion d'informations par l'intégrale de Choquet. Thèse de doctorat, Universitéde Savoie, 2004.

[9] P.Lévine, et J.C. Pomerol. Systèmes interactifs d'aide à la décision et systèmes experts, Hermès, 1989.

[10] E. Turban, " Decision Support and Expert Systems ". Macmillan, New York, 1993.

[11] D.J.Power,"Decision support systems: concepts and resources for managers".Westport, Conn., Quorum Books, 2002.

[12] Y.Xie,H.Wang,J.Efstathiou,A research frame work for web-based open decision Support systems, Knowlege.-Based Syst.18, 2005, pp.309–319.

[13] H.K.Bhargava,D.J.Power,D.Sun,Progressinweb-based decision support technologies, Decis.SupportSyst.43(2007)1083–1095.

[14] F.J.Cabrerizo,J.A.Morente,Molinerab,I.JPérezc,J.LópezGijónd,E.Herrera-Viedma,
"Adecisionsupportsystemtodevelopaqualitymanagementinacademicdigitallibraries", Journal of Information Sciences, 2015, pp. 48–58

[15] H. Munaga, L. Ieronutti, and L. Chittaro, "CAST: A Novel Trajectory Clustering and Visualization Tool for Spatio-Temporal Data", Proceedings of the First International Conference on Intelligent Human Computer Interaction, 2009 , pp 169-175.

[16] L. Ramírez, Y. R. Gel, M. Thompson, E. de Villa and M. McPherson, "A new surveillance and spatio-temporal visualization tool SIMID: SIMulation of Infectious Diseases using random networks and GIS", Computer Methods and Programs in Biomedicine, vol. 110, no 3, 2013, pp. 455-470

[17] S. Mittelstädt, X. Wang, T. Eaglin, D. Thom, D. Keim, W. Tolone and W. Ribarsky, "An Integrated In-Situ Approach to Impacts from Natural Disasters on Critical Infrastructures", IEEE 48th Annual Hawaii International Conference on System Sciences, HICSS-48, 2015, Los Alamitos, CA: IEEE Computer Society Press.

[18] M.Wooldridge, "An Introduction to MultiAgent Systems". London: John Wiley& Sons,2002.

[19] J. Ferber. "Multi-Agent Systems : An Introduction to Distributed ArtificialIntelligence". Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[20] P. Kostkova, J. ManiSaada, G. Madle, J.Weinberg. "Applications of Software Agent Technology in the Health Care Domain", (Ch. AgentBased Uptodate Data Management in National Electronic Library for Communicable Disease, (2003) pp. 105-124.

[21] M. Tentori, J. Favela, M. Rodriguez, "Privacyaware autonomous agents for pervasive healthcare", IEEE Intelligent Systems 21 (6), pp. 55-62.,2006.

[22] H. Ltifi, G. Trabelsi, M. Ben Ayed, and A.M. Alimi, "Dynamic Decision Support System Based on Bayesian Networks, application to fight against the Nosocomial Infections", International Journal of Advanced Research in Artificial Intelligence (IJARAI), Vol. 1, no. 1, , 2012, pp. 22_29.

[23] H.Kallel, M.Bouaziz, H.Ksibi, H.Chelly, C.B.Hmida, A.Chaari, N.Rekik, et M.Bouaziz, "Prevalence of hospital-acquired infection in a Tunisian Hospital". Journal ofHospital Infection, vol. 59, 2005, pp. 343-347.

[24] S.Gafsi-Moalla,« Les infections acquises en réanimation, étude prospective réalisée dans le service de réanimation de Sfax sur une période de 3 mois ». Thèse de doctorat, Faculté de Médecine de Sfax, Tunisia, 2005.

[25] L. Hergafi, « Présentation et validation d'un nouveau système pour la surveillance de l'infection acquise en réanimation ». Thèse de doctorat, Faculté de Médecine de Sfax, Tunisia, 2006.

[26] A. Darwich, "Constant-space reasoning in dynamic Bayesian networks", International journal of approximate reasoning, Vol. 26, 2001, pp. 161-178.

[27] J. Nielsen, "Usability Engineering", Academic Press, Boston, 1993.

[28] M.Y. Ivory and M.A. Hearst. "The state of the art in automating usability evaluation of user interfaces", ACM Computing Surveys, vol. 33(4), 2001,pp. 470-516.

[29] A. Sears and J.A. Jacko, editors. "The Human Computer Interaction Handbook: Funda- mentals, Evolving Technologies and Emerging Applications". Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2008.

[30] R.Shibl, M.Lawley, J.Debuse, "Factors influencing decision support system acceptance", Decision Support Systems, Vol. 54 (2),2013, pp. 953–961.

[31] D. Godoy, S. Schiaffino, and A. Amanda, "Interface Agents Personalizing Web-Based Tasks", *Cognitive Systems Research*, Vol. 5, pp. 207_222, 2004.

A Zone Classification Approach for Arabic Documents using Hybrid Features

Amany M.Hesham, Sherif Abdou, Amr Badr
Faculty of Computers and Information
Cairo University
Cairo, Egypt

Mohsen Rashwan
Faculty of Engineering
Cairo University
Cairo, Egypt

Hassanin M.Al-Barhamtoshy
Computing and Information Technology
King Abdulaziz University (KAU)
Saudi Arabia

***Abstract*—Zone segmentation and classification is an important step in document layout analysis. It decomposes a given scanned document into zones. Zones need to be classified into text and non-text, so that only text zones are provided to a recognition engine. This eliminates garbage output resulting from sending non-text zones to the engine. This paper proposes a framework for zone segmentation and classification. Zones are segmented using morphological operation and connected component analysis. Features are then extracted from each zone for the purpose of classification into text and non-text. Features are hybrid between texture-based and connected component based features. Effective features are selected using genetic algorithm. Selected features are fed into a linear SVM classifier for zone classification. System evaluation shows that the proposed zone classification works well on multi-font and multi-size documents with a variety of layouts even on historical documents.**

***Keywords*—segmentation; layout analysis; texture features; connected component analysis; Arabic script; genetic algorithms**

I. INTRODUCTION

Layout analysis is the process of extracting text lines from a document image and identifying their reading order. It is a major step in converting documents into electronic text [1]. The first major step in document layout analysis is segmentation. In this step a documents are divided into distinct geometrical regions where each is classified as text or non-text region. Text regions contain text only while non-text regions may contain images, graphs, drawings, etc. as shown in Fig. 1. Text regions are sent to character recognition engine [2] which converts text images into digital text while there is no need to send images to recognition engine as it will give garbage output.

(a) (b) (c)

Fig. 1. Examples for regions in Arabic documents (a) text region (b) non-text (image) region (c) non-text (drawing) region

Our focus is on Arabic script documents. Arabic script languages –such as Arabic, Kurdish, Urdu, Persian etc.- is the second most used script after Latin script. Arabic script segmentation is a challenging problem as it is written in many different styles. In addition, Arabic script is cursive where letters are connected together forming ligatures.

In literature, many algorithms were presented for document images segmentation. Kumar et al. [3] evaluated the performance of six page segmentation algorithms on different scripts. The evaluated algorithms are X-Y cut, whitespace analysis, constrained text-line finding, Docstrum, Voronoi diagram and smearing algorithms. X-Y cut algorithm [4] is a recursive top-down algorithm. It is a tree based algorithm where the whole document image represents the root of the tree. A document is then decomposed recursively into rectangular blocks, until the document can't further be split, which forms tree nodes. Whitespace analysis algorithm [5] analyzes the structure of background in a document image. It extracts the maximal whitespace rectangles, to be merged subsequently. Merged rectangles cover the background and thus isolate text blocks. Constrained text-line detection algorithm [6] is also a top down algorithm. It performs a two-step text lines extraction algorithm. First, it extracts maximal empty rectangles covering the whitespace background as in whitespace analysis algorithm. Then, the extracted rectangles are considered obstacles that are used in text line detection step. Document spectrum, or docstrum, algorithm [7] is a bottom up algorithm. It works on connecting document components using the nearest neighbor clustering algorithm forming the regions. Voronoi-diagram based algorithm [8] is also a bottom-up algorithm. It extracts sample points from the boundaries of connected components. Voronoi cells are extracted surrounding the sample points. A decision is taken on the edges whether to be removed or not based on two features, distance and ratio of cell areas. The predefined algorithms are evaluated on different scripts including Arabic scripts particularly on Urdu documents. They give poor segmentation results on complex script documents such as Arabic scripts. As based on the algorithms evaluation on Urdu documents, maximum accuracy achieved was 71.5%. Another evaluation was done by Shafait et al. [9] on the same six algorithms but only on Latin English scripts. Their algorithms evaluation shows that constrained text-line finding, Docstrum and Voronoi give better results than X-Y cut, whitespace analysis and smearing algorithm.

Moreover, other approaches are used like multiresolution morphological operations as in Bloomberg algorithm [10]. Bloomberg algorithm performs well on text and halftone document segmentation. But except for halftones, it gives poor results on any other non-text like drawings, graphs, etc. Bukharia et al. [11] proposed an improvement on Bloomberg algorithm to be generalized. A hybrid between Voronoi and Docstrum is also proposed in [12].

On the other hand there are other text segmentation approaches based on classification of segmented text zones. Pixel based approaches [13] which work on classifying each pixel in a document into text or non-text. Connected component analysis based approaches [14] where each extracted component is classified into a text or non-text class. Edge based [15] approaches where edged images are portioned into blocks and each block is classified to a class. Texture analysis based [16] approaches which make use of the regular periodic texture property in text regions and irregular textures in non-text zones. In this work we propose a new approach for zone classification in Arabic documents. The proposed algorithm applies text and non-text segmentation using dilation morphological operation and connected component analysis. Features extracted from each segmented zone are then used in classifying text and non-text regions. Those features combine texture and connected component based features.

The rest of this paper is organized as follows. Page segmentation of Arabic script document images is described in Section II. Starting by preprocessing then followed by zone segmentation. Extracted zones classification is discussed in Section III. Performance evaluation and experimental results are discussed in Section IV. Results analysis and final conclusions are included in Section V.

II. Page Segmentation

A. Preprocessing

Collected document images are required to be in the same format for any further processing. Since, collected document images may be colored or grey scaled; images need to be unified in bi-level representation –black, white-. This binarization step is the main pre-processing step. In [17] an adaptive local thresholding method is used for image binarization. Image is then cleaned using a 5×5 median filter for noise removal. Marginal noise and document borders are also removed. Skewed images are corrected using Radon transform [18] to be aligned correctly for further processing. The result from preprocessing phase is a unified, cleaned and aligned binary image with text and non-text components only.

B. Zone Segmentation

This step works on segmenting binary image into a set of zones. Segmentation algorithm used in this step is based on multiresolution morphological operations, dilation operation. Connected component analysis is applied to extract mean height (μ_h) and mean width (μ_w) of the extracted connected components. Image is dilated using a rectangular structuring element with size relative to μ_h and μ_w. The resulting image contains each zone components connected together. Each zone is surrounded by a rectangle as shown in Fig. 2.

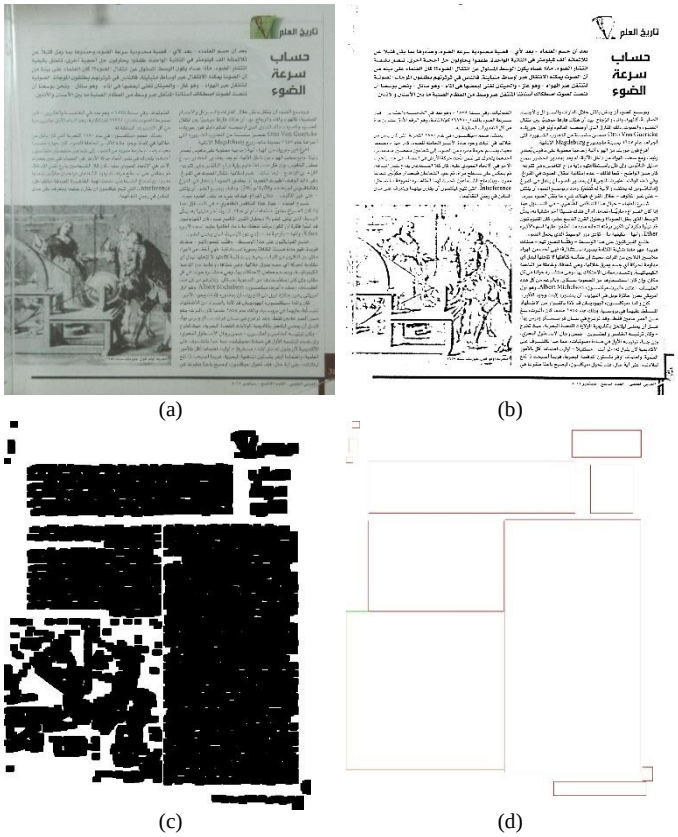


Fig. 2. Given (a) Original book page image (b) binarized and deskwed image using Sauvola algorithm (c) dilated image (d) image segmented into zones

III. Zone Classification

Classifying each zone into text and non-text requires extraction of a set of features. Those features are a mixture of connected component and textural based features. Textural features are features which contains spatial distribution of an image pixel intensities. Given the regular periodicity of the intensity distribution in text regions unlike non-text regions [16], textural features are good choice in zone classification. Textural features are extracted from sub-blocks where each sub-block represents part of a component e.g. letter or word as shown in Fig 3. Connected components based features are added to represent the whole zone characteristics. It was found that letters and words have common features e.g. height of connected components are near, while non-text regions have random structure e.g. big scattered objects and small noise.

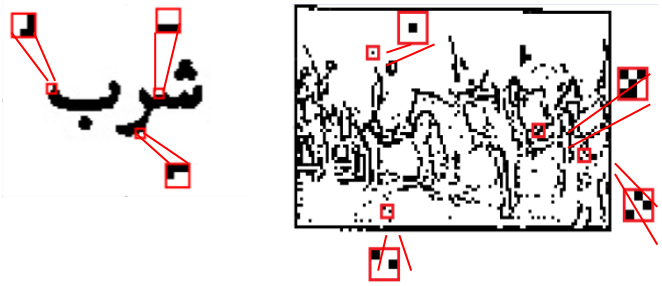


Fig. 3. Example for textural feature extraction for (a) text (b) non-text region

b_8	b_7	b_6
b_5	b_4	b_3
b_2	b_1	b_0

Fig. 4. Order of pixels in DSE block

C. Textural based features

Feature used in the proposed system is the document structure element DSE [19]. It is based on a 3x3 blocks. Pixels distribution in a given image block is shown in Fig. 4. There exist 2^9 = 512 different DSEs range from 0 to 511. For each pixel a 3x3 block, with this pixel as center b_4, is converted into integer which is called document structure element characteristic number DSECN [19]. DSECN is calculated using the following formula:

$$DSECN_{(x,y)} = \sum_{i=0}^{8} b_i 2^i \quad (1)$$

where b_i = *{0,1}*; 0 for black pixels and 1 for white pixels.

Features are calculated by counting the frequency of each of the 512 blocks. Some of the features are irrelevant which causes conflict in the classification between classes -text and non-text-. It is infeasible to extract all 2^{512} different possible combination between features of the DSEs to find the relevant and most effective features. As a result, a feature selection algorithm is used for dimensionality reduction without reducing accuracy. Only subset of features, which hold sufficient information used in discriminating classes, are selected. There exist many feature selection algorithms in literature, some apply linear transformations on the extracted feature vectors like principle component analysis (PCA) [20]. Others use evolution-based optimization techniques like Genetic algorithms (GA) [20] which focus on feedback result from the classifier to the feature selector.

In the proposed system GA is applied. Binary chromosome is used to represent the 512 features. Fitness used for the selection process is the accuracy of the model generated from a linear kernel support vector machine [21] (SVM) classifier. Input to the SVM is a subset of selected features and the output is the accuracy of the input features on a training set. The subset of features giving the highest accuracy is considered to hold the most effective subset of features. GA results in 223 features which is almost half the original set. Accuracy of the classification process using the selected features is 97.83% on cross validation data. Experimenting the whole 512 features and the 223 selected features gives equal results.

D. Connected Component based features

As previously explained, features are mixture of textural and connected component based features. Connected component analysis is applied and some statistical features are extracted to enhance the accuracy of the classification process. Features extracted from each zone are:

1) *Aspect ratio between width and height*
2) *Percentage of foreground to total zone pixels*
3) *Normalized maximum components height*
4) *Normalized maximum components width*
5) *Normalized mean width of the components*
6) *Normalized Standard deviation for width of the components*
7) *Normalized mean height of the components*
8) *Normalized standard deviation for height of the components*
9) *Mean for aspect ratio of the components*
10) *Standard deviation for aspect ratio of the components*

Features are normalized by dividing its value by the document image value, e.g. width, height. Adding the 10 connected component based features to the selected textural features improves the accuracy of zones classification to reach 98.54% on cross validation data. This means that connected component features increased the classification by 0.71%.

IV. System Evaluation

Evaluation of the current system requires collecting document images because of the absence of standard Arabic document dataset. Our Data set was 85 scanned documents collected from different sources. Documents are scanned using 12 Mega pixel camera. Images are collected from 3 main sources – books, magazines and historical documents- with different layouts e.g. single and multi-column documents. Images are as follow, 60 images from books, 10 images from historical documents and 15 images from magazines. Evaluation of the proposed zone classification algorithm is shown in table I. Results are shown for each set separately. In Books set, 1400 segmented and tested zones, 650 text zones and 750 non-text zones. In magazine set, 350 segmented and tested zones, 210 text zones and 140 non-text zones. Finally, historical documents set, 150 segmented and tested zones, 40 text zones and 110 non-text zones. Accuracy is calculated from zone classified as text or non-text using precision and recall. The precision is the ratio of the number of text zones to the total number of zones classified as text. The recall is the ratio of the number of text zones to the total text zones.

$$recall = \frac{\#(text\ zones\ correctly\ recognized)}{\#(actual\ text\ zones)} \quad (2)$$

$$precision = \frac{\#(text\ zones\ correctly\ recognized)}{\#(zones\ recognized\ as\ text)} \quad (3)$$

Table I shows that books and magazines precision and recall are near. However, in historical documents the values are far. The high difference occurs as a result of high degradation in old historical documents. Low precision value results from big amount of noisy non-text zones detected as text. However, recall value is high which shows that only few text-zones are detected as non-text. Figure 5 shows some examples for documents sent to the system.

Our approach is also evaluated against the preprocessing components of commercial state of art Arabic OCR systems, RDI Clever Page [22] and Sakhr Arabic OCR [23]. RDI Clever Page is based on connected component analysis and whitespace analysis. Evaluation is applied on a set of 109 books and magazines images. Results in table II show that our proposed classification approach has high recall value; which means that few text zones were missed. RDI system results in higher precision value but with smaller recall value compared

to our system. This means that RDI system misclassifies many text zones as non-text than our system which may result in losing important information from original document. On the other hand, our system overpasses the values of Sakhr system in both precision and recall.

V. CONCLUSION

In this paper we proposed a framework for documents layout analysis. It consists of two main phases 1) segmentation and 2) classification. Collected documents are binarized using an adaptive global binarization method, Sauvola. Images are then cleaned and skewed for further processing. Binarized images are segmented using morphological dilation operation into zones. Features are extracted for classification of segmented zones. Features used are combination between connected components and textural based features. Textural features are reduced using genetic algorithm and added to connected component features. Each zone is classified into text or non-text by sending features to a linear SVM classifier. Only text zones are sent for character recognition engine OCR to avoid garbage resulting from sending non-text regions. Evaluation results show that the proposed approach managed to achieve high accuracy for text zones classification in Arabic documents. Comparing those results with state of art commercial Arabic OCR systems it achieved the best recall while saving high accuracy.

TABLE I. EVALUATION OF ZONE CLASSIFICATION

Document type	Precision	Recall
Books	97.5%	96.7%
Magazines	94.5%	94.5%
Historical Documents	77%	95.2%
Total	**95.7%**	**96.1%**

TABLE II. COMPARISON BETWEEN PROPOSED SYSTEM AND OTHER SYSTEMS

Methods	Precision	Recall
Proposed system	93.5%	99.1%
RDI	97.9%	94.1%
Sakhr	91.6%	95.7%

ACKNOWLEDGEMENT

The teamwork of the "Arabic Printed OCR System" project was funded and supported; by the NSTIP strategic technologies program in the Kingdom of Saudi Arabia- project no. (11-INF-1997-03). In addition, the authors acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support.

REFERENCES

[1] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Layout analysis of Arabic script documents," in Guide to OCR for Arabic Scripts, V. Märgner and H. El Abed, Eds. London: Springer London, 2012, pp. 35 – 53.

[2] M. Attia, M. a a Rashwan, and M. S. M. El-Mahallawy, "Autonomously normalized horizontal differentials as features for HMM-based omni font-written OCR systems for cursively scripted languages," in International Conference on Signal and Image Processing Applications, 2009, pp. 185–190.

[3] K. S. S. Kumar, S. Kumar, and C. V. Jawahar, "On segmentation of documents in complex scripts," in International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 1243–1247.

[4] R. M. Haralick and I. T. Phillips, "Recursive X-Y cut using bounding boxes of connected components," in International Conference on Document Analysis and Recognition, 1995, vol. 2, pp. 952–955.

[5] H. Baird, "Background structure in document images," Int. J. Pattern Recognit. Artif. Intell., vol. 8, pp. 1013 – 1030, 1994.

[6] T. Breuel, "Two geometric algorithms for layout analysis," Doc. Anal. Syst. v, pp. 188 – 199, 2002.

[7] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, 1993.

[8] K. Kise, A. Sato, and M. Iwata, "Segmentation of Page Images Using the Area Voronoi Diagram," Comput. Vis. Image Underst., vol. 70, no. 3, pp. 370–382, Jun. 1998.

[9] F. Shafait, D. Keysers, and T. Breuel, "Performance comparison of six algorithms for page segmentation," Doc. Anal. Syst. VII, pp. 368 – 379, 2006.

[10] D. S. Bloomberg, "Multiresolution morphological approach to document image analysis," in International Conference on Document Analysis and Recognition, 1991, pp. 1–12.

[11] S. Bukhari, "Improved document image segmentation algorithm using multiresolution morphology," IS&T/SPIE Electron. Imaging, 2011.

[12] M. Agrawal and D. Doermann, "Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features," in International Conference on Document Analysis and Recognition, 2009, pp. 1011 – 1015.

[13] M. a. Moll, H. S. Baird, and C. An, "Truthing for Pixel-Accurate Segmentation," in International Workshop on Document Analysis Systems, 2008, pp. 379–385.

[14] S. S. Bukhari, M. Ibrahim, F. Shafait, and T. M. Breuel, "Document Image Segmentation using Discriminative Learning over Connected Components," in International Workshop on Document Analysis Systems, 2010, pp. 183 – 190.

[15] M. Pietikäinen and O. Okun, "Edge-Based Method for Text Detection from Complex Document Images," in International Conference on Document Analysis and Recognition, 2001, pp. 286–291.

[16] O. Okun and M. Pietikäinen, "A survey of texture-based methods for document layout analysis," in Texture analysis in machine vision, World Scientific, 1999, pp. 165–177.

[17] F. Shafait, D. Keysers, and T. M. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images," Doc. Recognit. Retr. XV, Proc. SPIE, vol. 6815, p. 681510, 2008.

[18] J. Dong, D. Ponson, A. Krzyżak, and C. Y. Suen, "Cursive word skew/slant corrections based on Radon transform," in 8th International Conference on Document Analysis and Recognition, 2005, pp. 478 – 483.

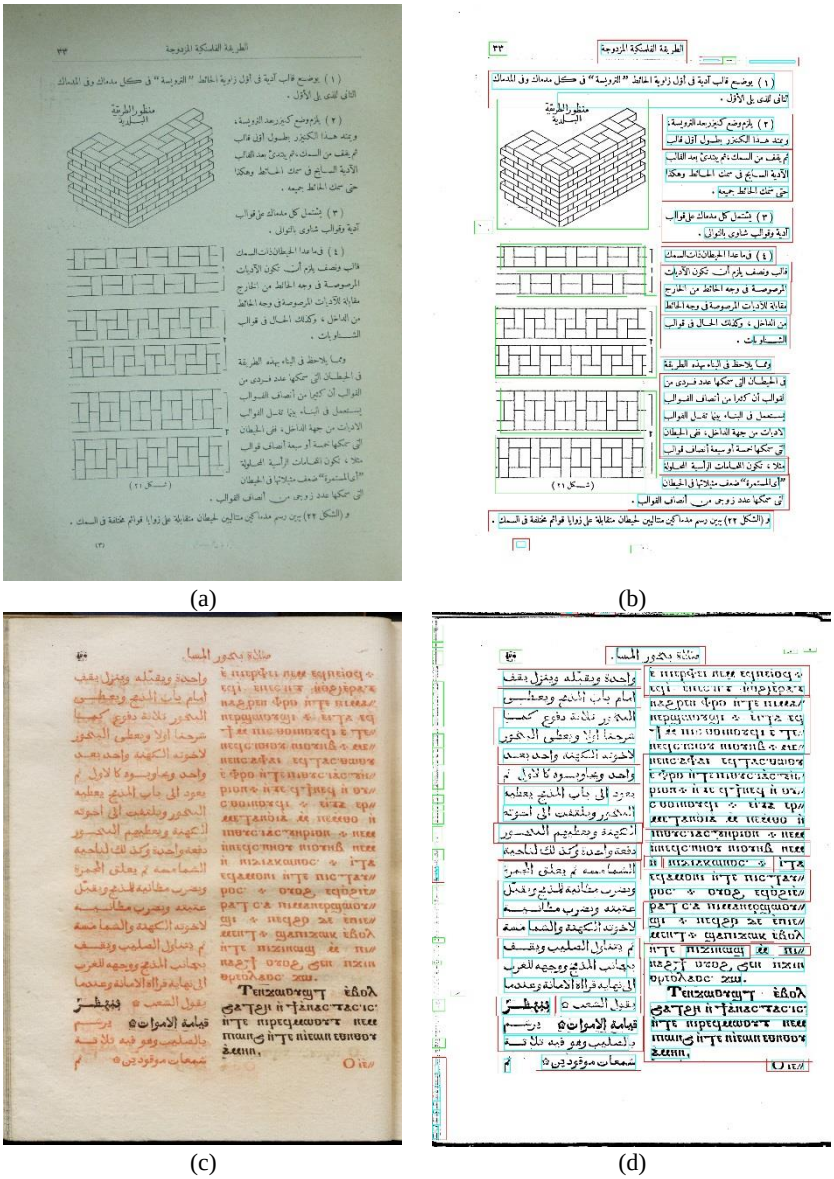
[19] C. Strouthopoulos and N. Papamarkos, "Text identification for document image analysis using a neural network," Image Vis. Comput., vol. 16, pp. 879–896, Aug. 1998.

[20] M. Raymer and W. Punch, "Dimensionality reduction using genetic algorithms," IEEE Trans. Evol. Comput., vol. 4, no. 2, pp. 164–171, 2000.

[21] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, pp. 27 – 66, 2011.

[22] RDI, "Clever Page - Arabic omni font-written OCR," http://www.rdi-eg.com/projects/OCR.htm.

[23] Sakhr, "OCR (Optical Character Recognition)," http://www.sakhr.com/index.php/en/solutions/ocr.



(a) (b) (c) (d)

Fig. 5. Examples for applying algorithm on different documents

Air Pollution Analysis using Ontologies and Regression Models

Parul Choudhary
Post Graduate of Computer Science & Engineering
JSSATE
Noida, India

Dr. Jyoti Gautam
HOD & Associate Professor (CSE)
JSSATE
Noida, India

***Abstract*—Rapidly throughout the world economy, "the expansive Web" in the "world" explosive growth, rapidly growing market characterized by short product cycles exists and the demand for increased flexibility as well as the extensive use of a new data vision managed data society. A new socio-economic system that relies more and more on movement and allocation results in data whose daily existence, refinement, economy and adjust the exchange industry. Cooperative Engineering Co - operation and multi -disciplinary installed on people's cooperation is a good example. Semantic Web is a new form of Web content that is meaningful to computers and additional approved another example. Communication, vision sharing and exchanging data Society's are new commercial bet. Urban air pollution modeling and data processing techniques need elevated Association. Artificial intelligence in countless ways and breakthrough technologies can solve environmental problems from uneven offers. A method for data to formal ontology means a true meaning and lack of ambiguity to allow us to portray memo. In this work we survey regression model for ontologies and air pollution.**

Keywords—Ontologies; Air pollution Analysis; Regression Models; Linear Regression

I. INTRODUCTION

The term "ontology" Data Science and research in 1990 across countless man-made intelligence (AI) research community has to. AI researchers have adopted the term "ontology" is usually depicted in a scheme to code they even (from the view point of the computational society) should be a fair representative of the world.

Ontology is a fair; a public explanation of the concept is clear. A "concept" of an event, the event is prepared by identifying the relevant considerations of a hypothetical idea."Formal" way of ontology languages are always involved in the use of machine-readable form. For eg., in the areas of health, disease and symptom of ideas, they are the relationship between cause and not just a nuisance that can cause a disease.

An ontology is a "share" concept that the use of ontologies for the target group aimed to embody the vision of consensus. Ideally ontology terms for its use independently and in a way that the public may be universal, but useful functions arrest uneven and unequal representation of vision calls for an ontology uses.

Meaning how ontologies are helpful in data recovery?

Budding accelerated digitization processes and globally relevant databases that are transpiring in the current year, the focus of the problems of data is modified.

This matter is now the subject of a precise vision to find the data and commands, but only the most relevant agents is difficult to choose from the huge piles of data. Such a syntactic conventional engine, hunting for keywords and its abundance by data elements in order to explore the processes in place is used. A situation that relevant data, including but not a precise data object is different because it uses words to portray - These methods suffer from setbacks such as terminology inconsistency and "contrast" that irrelevant data because of the similarities in wording has been taken.

Recently, however, is moving in a new way - meaning approach. Data about data - - order an additional satisfactory method for data retrieval and exploration of those "need to answer this is to use meta data.

Ontologies are useful in two ways in improving the process:

1) It allows to abstract the information and represent it explicitly- highlighting the concepts and relations and not the words used to describe them. [1]

2) Ontologies can possess inference functions, allowing more intelligent retrieval. For example, a "basketball player" is also a "professional athlete", and an Ontology that defines the relations between these concepts can retrieve one when the other is queried.

II. WHAT IS ONTOLOGY?

"Although the ontology is currently a fashionable word, there is no agreement on the precise meaning of the term: it is reported that the main purpose of the ontology is to regulate the meaning of words, the term" ontology "is not clearly defined humorous "and" concerning AI (Artificial Intelligence) to produce a lot of controversy in the discussion "feel. F karma, vision and multimedia ontology engineering request finds applicability in many areas, and each one of them has its own meaning.

1) A bit of etymology and philosophy.

The term "ontology" last a very long time in philosophy, starting with the works of Aristotle. As being "science", from the Greek "ontos" This way both speech and reason for being

and "Logo" is described. From the point of thinking about the event, an additional current philosophy is to start with German theorists of the 19th century; the existence of ontology is a systematic report. This idea does not mean different things back (phenomenological approach) are going to be and despite the appearance, these are two ways that cannot be included (phenomenological reduction concerning contemplate). To keep in mind, the larger epistemological sense can help us to understand what must be ontology. Vision, language and reasoning: As a matter of fact, one can maintain three dimensions. In other words: to clear speech about the world, to manipulate our understanding and representation throughout the world to understand the concept.

2) The knowledge engineering point of view.

Multimedia and vision engineering sectors nineties instead of a systematic nature, a spiritual way of being able to focus more on early reports of the presence of 1 used the word ontology. As a matter of fact, man-made system of intelligence, what exists is a declarative language that can be embedded in. Ontology next how objects, ideas and connections in a short period of interest are agreeing to continue to embody a clear specification is appropriate. Gruber shouted ideas and connections that it is an agent or agents can continue for a description of the area (a program like a reasonable specification) that "the concept of a specification is". The definition of terms used in engineering include semantic web, we can say that to begin again: "An ontology is a computer-readable language to express ideas in a common description of the connection." Fig.1 is an example of general higher ontology.

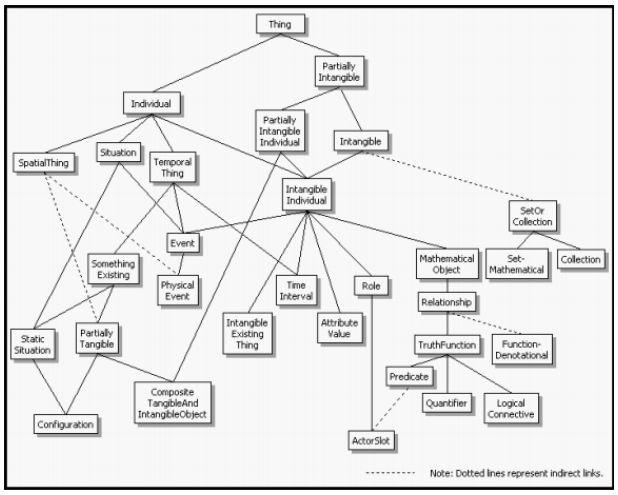


Fig. 1. The Upper Ontology

III. What is Ontology for?

"What is vital is what ontology is for" (T.R Gruber). As a matter of fact, the ontology's necessities depend on the intention of the final application. The main goal of an ontology is to enable contact and vision allocating amid computer arrangements by seizing a public understanding of words that can be utilized by humans and programs. We can next recognize disparate groups of uses for ontologies:

- communication amid people and organizations;
- inter -operability (communication) amid arrangements, including enterprise modeling and multi agent arrangements;
- system engineering, for specification, reliability and reusability;
- knowledge association like data retrieval, document association, vision center systems;
- natural speech treatment for semantic research, lexical construction;

In connection with all requests and Internet electronic transactions and the semantic web, we are well past the one that shows the trend of the real ontology allow inadequate word to say confronting the increasing use of the web, additional human and computer programs to interact with the demand for data on websites. Described as an extension of the current one, a new form of web content was necessary. "Semantic web is a representation of the globe vast web of data".

(W3C's definition): Semantic Web on Resource Description Framework (RDF) for words, XML for syntax and URIs for shouting has been established.

IV. Air Pollution

Air pollution and prepare for every metropolis outstanding importance for nature is an environmental shock. And in the industry spans the development of motor-car traffic over the past year has been the severity of the situation. Numerous authors and institutions solve the problem of air pollution in the disturbances. Emissions of pollutants that affect air quality, air quality, very much. Lung infections and a high degree of air pollution, with spans of respiratory diseases in the population decline of the sensitivity of the air quality reasons. Car density and air pollution caused by the industry, cars and industry emissions, depending on the type of uneven geographical location, such as temperature, wind conditions, according to the weather conditions, and complementary factors. For the calculation of air pollution emissions, the dense monitoring systems are working in many areas of city. Intended area of monitoring indoor air quality is to tell about the powers accountable. Surveillance data from shock and such small words to solve the contamination level forecasting, emergency response measures, deciding to eliminate air pollution in the form of work is used by experts.

Decision-making has been installed on quantitative data and measurements and observations, presented data concerning the frequency and size of the change and weather standards are maintained from the beginning is consolidated.

I for cooperation in air quality decision support to offer an unfailing, a mathematical model has to be designed, environment protection and air pollution (NO, NO2, mist, SO2, etc.) between connection established. Obviously, there are unforeseen factors that could impact the degree of air pollution are a lot of, and it is certain that the institution due to a cut or an increase of air pollutants is difficult pointer. Install heavy numerical modeling methods and computational resources when running complex data that often are not

attainable need to rush. AI methods, expert systems and knowledge-based methods employ the way in this area are not enthusing. The successes of these methods are used in various fields of science. Environmental Law Association and the Association of air quality models and data processing components and modules encompass blow to solve. A knowledge-based numerical model is another route proposal; a vision center is an inference engine that can deal with uncertainty in terms of use by many original features. An ontology way, employing a consistent, coherent and non-redundant central vision can be estimated.

V. Urban Air Quality Monitoring

Wise accurate environmental data, uneven metropolitan locations, the resort affords elevated alarm system possible scenarios and to advocate appropriate countermeasures should notice. Mobility is possible to exploit the inherent web nodes, consequently, a low cost environmental monitoring system with high spatial coverage to apply.

With a vehicular sensor web in potential environmental data collectable, air quality plays a second role. It's really a concern in modern cities because of air pollution effect on population and environment is an important encounter.

Local factors, the city's terms, and conditions hot spot: the air quality in the city spread is the result of three factors. In rural spans, pollution levels with medium to long-distance voyaging complementary areas of air pollutants from the crowd mainly depend on transportation. Emerging compression levels are significantly less in these areas. The city spans, the air contamination such warming settings and lighting off, region and private vehicular transport, or assembly activities as human hobby, is related to the set. City pollution accordingly human passion, the topography, and the innate micrometeorology, spatial varies as it is reasonable to anticipate.

The importance of this topic is so high that the Directive 96/62 / EC, establishes the principles that Frank is a public strategy and delineate or to circumvent the order, cut close to the ambient air quality targets in the European is manipulated by the commission on the human condition and nature of harmful consequences, Associate States assess ambient air standard in the region to inform and enhance the quality of air, while it is undesirable. Air pollution is usually very good web is monitored by permanent stations. A tracking station just a huge scope of pollution standard research tool can calculate work. However, continuous monitoring stations repeatedly so as to measure the surrounding background concentrations or possible hotspot areas are assigned which normally are in addition to the countless kilometers. In addition, the purchase and maintenance of such a heavy price to limit the number of capacities manipulating the dimensional resolution of the pollution map is emerging. In fact, a stable sensor capable of spatial coverage scope is manipulated, so it's a huge interest in a vast number of detectors for monitoring period should seize. These failures, we advocate a profitable and feasible channel sensor networks win. If requested efficiently, a VSN can offer a wide dimensional coverage, and saw a better quality of the characteristics. Instead of employing fixed sensors, detectors in cars or transport services sector can be installed on vehicles. Sensors secured on advancing vehicles periodically track the air standard and send the taken data to a central storage system. Though, there is a trade-off for this gain in dimensional coverage. Time-related coverage of detected information in a particular locale will be less contrasted to fixed sensors readings as a feature will be measured in the alike locale only when the vehicle will cross once more that point. This lack of time-related coverage can be grasped by rising web nodes density and so climbing sensors on extra vehicles, or allocating sensors on area buses, so that environmental features can be tracked unceasingly alongside their routes. One more setback that ought to not be underrated concerns degraded measurements inside sensor networks. Sensor nodes could sporadically produce wrong measurements due to battery exhaustion, dust on sensors, tampering and supplementary causes. Amid the countless mathematical methods and algorithms in works, suitable instruments to pre-process such gathered data are Bayesian Networks.

In the subsequent serving we debate a little continuing air contamination monitoring period bestowing a brief overview of the state of the fine art.

VI. Ontological Approach

Urban air corruption affiliation needs raised demonstrating and information preparing procedures. Synthetic mind gives incalculable techniques and advances that can resolve usefully dissimilar ecological issues. AI techniques present increases above additional set up numeric demonstrating ways that need substantial processing assets and interest as info convoluted data, much of the time not easily accessible. It is critical to settle on choices in natural security affiliation, so a multi-specialist game plan (MAS) way may be asked for better arrangement. A multi-specialist game plan is a web of interactive media operators that conveys to determine mishaps that are surpassing the individual limits or vision of each and every misfortune solver. A specialist can be a physical or neighboring presence that can deed autonomously and has mastery to finish its points and slant. The learning of smart specialists and multi-operator game plans gives the interactive media preparation to the execution of disseminated environment courses of action that can track and control the ecological standard.

In our setting, the vehicular sensor web can be demonstrated as a multi-operator framework made by a set out of astute substances. Each and every vehicle has a bit of running media specialists that screen and analyze the advancement of natural information like air quality, and report to a manager operator if somewhat basic mishap can happen. As per Chomsky speculations, information, assembled by sensors introduced on vehicles, must be epitomized in an appropriate strategy so that keen choice operators can understandable and reason on it. A technique to formalize information is the significance of a right cosmology that grants us to depict reminders close by importance and lacking uncertainty. One of the additional open points in developing ontologisms is apportioning open comprehension of the development of information in the midst of individuals or interactive media specialists. The contact in the midst of specialists depends by and large on the reception of a

conceptualization, that is, a legitimate representation of the truth of a particular circumstance, so oncologic pace is straight to the point in our course of action, keeping in mind the end goal to portray an open vocabulary for scientists who interest to allot information around there, yet also for learning based arrangements outlining that inquiries and attestations are traded in the midst of operators. The utilization of distinguishing components and remote sensor networks is rising, so a rising volume of heterogeneous information, information organizations, and estimation methods is produced. The ontological perfect gives a technique to deal with the sensors and the convoying volume of created information. It can moreover be used to accept sensor readings and to deal with imperfect sensor information as portrayed in, in that the W3C Semantic Sensor Web Incubator group (SSNXG) depicted OWL 2 cosmology to outline the aptitudes and properties of sensors, the deed of identifying and the developing perceptions.

So as to formalize air quality contemplations, philosophy could be used to development air sullying zone ideas. Since a standout amongst the most crucial philosophy elements is its reusability, we accepted to development our vision range as per AIR POLLUTION Ontology. This cosmology is committed to air defilement examination and mastery, and has been genuinely used in Air Pollution.

Urban air pollution organized information will be justifiable and process capable by specialists whose point is to screen and moderate defilement results crosswise over different sorts of uses. Case in point, joining 3D city models, meteorological elements and air tainting information, it is plausible to guesstimate the nature of air in a little city zone. One additionally fascinating solicitation may be street activity administration. Utilizing the information enriched by natural and street metaphysics, it ought to be likely to build up somewhat smart mixed media specialist that can set a proposed most extreme velocity for a vehicle, or effect activity lights keeping in mind the end goal to get a handle on the quantity of vehicles on the streets of a specific city zone. These activities may provide for scatter meteorological poisons in this way cutting their pressure in building appended ranges.

VII. Building of Ontology for Air Pollution Control

Making metaphysics is not a paltry issue. It needs not simply aptitudes in information advances but rather also profound vision in the demonstrated space. The strategy of vision course of action advancement is organized in several models that must be made. On the "setting" level of reflection three models is proposed: Organizational perfect, Task perfect and Agent's model. The authoritative perfect portrays the relationship close by the objective to see the mishaps and chances of vision administration. The Task perfect epitomizes errand that are gave inside the association. An undertaking is whatever that must be given by a specialist. The operator perfect portrays all specialists – agents of assignments - their demonstrations, capabilities, abilities, and confinements. Metaphysics progress systems help making ontologies in grouped region arranged applications. Endless procedures have been industrialized so as to formalize making Ontologies for assembling or supplementary applications. Despite the fact that cosmology progresses systems are not develop bounty, they can be useful in developing philosophy set up vision courses of action.

VIII. Related Works

J., Fenger, et. al. (1999) [1] In this paper, following 1950 the world populace has dramatically increased, and the worldwide number of autos has expanded by a component of 10. In the same time frame the part of individuals living in urban ranges has expanded by a component of 4. In year 2000 this will add up to about portion of the world populace. Around 20 urban areas will each have populaces above 10 million individuals. Seen over longer periods, contamination in real urban communities tends to increment amid the developed stage, they go through a most extreme and are of course diminished, as reduction methodologies are created. In the industrialized western world urban air contamination is in a few regards in the last stage with successfully diminished levels of sulfur dioxide and residue. In late decades in any case, the expanding movement has changed the consideration regarding nitrogen oxides, natural mixes and little particles. In a few urban areas photochemical air contamination is a critical urban issue, however in the northern piece of Europe it is a huge scale wonder, with ozone levels in urban avenues being ordinarily lower than in rustic territories. Urban communities in Eastern Europe have been (and by and large still are) intensely dirtied. After the late political change, trailed by an impermanent subsidence and a consequent presentation of new innovations, the circumstance seems to make strides. Be that as it may, the rising number of private autos is a rising issue. In most creating nations the fast urbanization has so far brought about uncontrolled development and falling apart environment. Air contamination levels are here as yet ascending on numerous fronts.

Noboru, Yamazoe, Go Sakai et. al. (2003) [2] In this paper, semiconductor gas sensors use permeable polycrystalline resistors made of semiconducting oxides. The working guideline includes the receptor capacity played by the surface of every oxide grain and the transducer capacity played by every grain limit. Furthermore, the utility variable of the detecting body additionally participates in deciding the gas reaction. In this manner, the ideas of sensor outline are controlled by considering each of these three key variables. The necessities are choice of a base oxide with high versatility of conduction electrons and attractive steadiness (transducer capacity), choice of an outside receptor which improves surface responses or adsorption of target gas (receptor capacity), and creation of a very permeable, meager detecting body (utility component). Late advance in sensor plan taking into account these elements is depicted.

Mihaela. Oprea et. al. (2005) [3] In this paper, air contamination control in urban districts is one of the fundamental bearings of exploration in the ecological sciences. For every area the contamination causes and contamination scattering are distinctive, contingent upon the

modern movement, on vehicles activity, on residential sources et cetera, and also on the topographical area, temperature of the air, velocity and heading of the wind, and other climate elements. A few scientific models are utilized for the portrayal of the connections between natural insurance and meteorological elements. An option way to deal with the numerical models is an information based methodology that incorporates numerous wellsprings of learning in an information base. The paper portrays a contextual investigation of learning displaying in an air contamination control choice emotionally supportive network that utilizations DIAGNOZA_MEDIU, a model master framework committed to air contamination examination and control in urban areas. We have created metaphysics, AIR_POLLUTION, for the application area. A few AI procedures were utilized as a part of the learning demonstrating process. A counterfeit neural system gives prescient information to the actualities base, and a part of the principles from the tenet base are removed by utilizing inductive learning.

Michael, Blaschke, Thomas Tilleet. al. (2006) [4] In this paper, microelectromechanical-framework (MEMS) metal-oxide gas sensors have achieved a full grown stage, which makes mass business sector applications in the car region conceivable. As opposed to the officially settled fold control framework, which controls the entrance of (ignition) gasses from outside the vehicle to the auto lodge, the framework contemplated here distinguishes scent occasions made inside the auto lodge. The occasions under study have been tobacco smoke, fast-food smell, excrement, and bio effluents (tooting). As the reference can't be a "straightforward logical estimation," a human test board for evaluating the hedonic impact on a scale from 0 to 5 is utilized as reference. The specialized framework is a MEMS metal-oxide-sensor cluster comprising of three unique sensors. The information assessment approach utilized here is consolidating the human-tactile information and the MEMS sensor information. The assignment is performed by the blend of two free calculations, where one is identified with the standardized conductance and the other to flag change. Utilizing a consolidated methodology has the favorable position that "false" occasions are smothered. After the calculation was effectively exchanged onto a microcontroller, genuine information were recorded and grouped. A few functional cases are given in this paper. The general gas-sensor framework achieves great agreement with the human-tactile impression, which is spoken to via air quality levels. This empowers the outline of an interest controlled ventilation framework.

H. K. Eminir, Hala Abdel-Galil et. al. (2006) [5] In this paper, the absence of natural information is a typical component of numerous creating nations. This is truth in Egypt, where air quality is starting to be efficiently observed in a few spots of the nation. To defeat these issues, the requirement for exact assessments of air quality levels turns out to be always vital. To accomplish such forecast errands, the utilization of fake neural system (ANN) is viewed as a practical procedure better than conventional factual techniques. In this paper, ANN prepared with a back spread calculation is utilized to appraise the surely understood poisons, from promptly discernible neighborhood meteorological information. The outcomes demonstrate that the ANN model anticipated air contamination focuses with great exactness of roughly 96 %.

Uichin, Lee, Biao Zhou, et. al. (2006) [6] In this paper, vehicular sensor systems are developing as another system worldview of essential importance, particularly for proactively assembling observing data in urban situations. Vehicles normally have no strict imperatives on preparing force and capacity abilities. They can sense occasions (e.g., imaging from roads), process detected information (e.g., perceiving tags), and course messages to different vehicles (e.g., diffusing significant warning to drivers or police operators). In this novel and testing versatile environment, sensors can produce a sheer measure of information, and conventional sensor system approaches for information reporting get to be unfeasible. This article proposes MobEyes, a proficient lightweight backing for proactive urban checking taking into account the essential thought of abusing vehicle versatility to artfully diffuse rundowns about detected information. The reported exploratory/scientific results demonstrate that MobEyes can collect outlines and fabricate an ease appropriated file with sensible fulfillment, great versatility, and restricted overhead

Jakob, Eriksson, Lewis Girod et. al. (2008) [7] In this paper researches an utilization of versatile detecting: recognizing and reporting the surface states of streets. We portray a framework and related calculations to screen this vital common base utilizing a gathering of sensor-prepared vehicles. This framework, which we call the Pothole Patrol (P2), utilizes the characteristic versatility of the taking an interest vehicles, deftly assembling information from vibration and GPS sensors, and preparing the information to survey street surface conditions. We have sent P2 on 7 taxis running in the Boston range. Utilizing a straightforward machine-learning approach, we demonstrate that we can recognize potholes and other extreme street surface abnormalities from accelerometer information. By means of cautious determination of preparing information and sign components, we have possessed the capacity to manufacture a finder that misidentifies great street fragments as having potholes under 0.2% of the time. We assess our framework on information from a great many kilometers of taxi drives, and demonstrate that it can effectively recognize various genuine potholes in and around the Boston region. In the wake of bunching to encourage diminish spurious identifications, manual assessment of reported potholes demonstrates that more than 90% contain street abnormalities needing repair.

Yajie, Ma, Mark Richards et. al. (2008) [8]In this paper, we exhibit a conveyed base taking into account remote sensors system and Grid processing innovation for air contamination checking and mining, which means to grow minimal effort and pervasive sensor systems to gather ongoing, expansive scale and exhaustive ecological information from street movement emanations for air contamination observing in urban environment. The primary informatics challenges in appreciation to building the high-throughput sensor Grid are talked about in this paper. We display a two-layer system structure, a P2P e-Science Grid engineering, and the disseminated information mining calculation as the answers for location the difficulties. We mimicked the framework in

TinyOS to inspect the operation of every sensor and the systems administration execution. We likewise display the appropriated information mining result to analyze the viability of the calculation.

Giuseppe, Anastasi, Giuseppe Lo Re et. al. (2009) [9] In this paper, observing basic soundness of verifiable legacy structures might be an overwhelming undertaking for structural specialists because of the absence of a prior model for the building solidness, and to the nearness of strict limitations on checking gadget arrangement. This paper provides details regarding the experience maturated amid an undertaking with respect to the outline and usage of an imaginative innovative system for observing basic structures in Sicily, Italy.

Giuseppe, Anastasi, OrazioFarruggiaet. al. (2009) [10] In this paper, this work reports the experience on the outline and organization of a WSN-based framework for checking the profitable cycle of fantastic wine in a Sicilian winery. Other than giving the way to pervasive observing of the developed range, the undertaking depicted here is intended to bolster the maker in guaranteeing the general nature of their generation, as far as exact arranging of mediations in the field, and protection of the put away item. Remote Sensor Networks are utilized as the detecting base of a dispersed framework for the control of a prototypal profitable chain; hubs have been sent both in the field and in the basement, where wine maturing is performed, and information is gathered at a focal unit with a specific end goal to perform surmisings that propose auspicious mediations that protect the grapes' quality.

Artis, Mednis, Girts Strazdinset. al. (2010) [11] In this paper, Road surface investigation including pothole reports is an essential issue for street maintainers and drivers. In this paper we propose a technique for pothole location utilizing portable vehicles outfitted with off the rack receiver and worldwide situating gadgets connected to an on-board PC. The methodology is sufficiently nonexclusive to be stretched out for other sort of occasion identification utilizing diverse sensors also. The vehicles are driving on open roads and measuring pothole prompted sound signs. Our methodology was tried and assessed by genuine tests in a street portion for which we had built up the ground truth previously. The outcomes show pothole discovery with high precision regardless of the foundation commotion and other sound occasions.

A. R., Al-Ali, Imran Zualkernan et. al. (2010) [12] In this paper, an online GPRS-Sensors Array for air contamination observing has been composed, actualized, and tried. The proposed framework comprises of a Mobile Data-Acquisition Unit (Mobile-DAQ) and an altered Internet-Enabled Pollution Monitoring Server (Pollution-Server). The Mobile-DAQ unit incorporates a solitary chip microcontroller, air contamination sensors exhibit, a General Packet Radio Service Modem (GPRS-Modem), and a Global Positioning System Module (GPS-Module). The Pollution-Server is a top of the line PC application server with Internet availability. The Mobile-DAQ unit accumulates air toxins levels (CO, NO2, and SO2), and packs them in an edge with the GPS physical area, time, and date. The casing is in this manner transferred to the GPRS-Modem and transmitted to the Pollution-Server by means of the general population versatile system. A database server is connected to the Pollution-Server for putting away the poisons level for further utilization by different customers, for example, environment security offices, vehicles enrollment powers, and vacationer and insurance agencies. The Pollution-Server is interfaced to Google Maps to show continuous toxins levels and areas in vast metropolitan territories. The framework was effectively tried in the city of Sharjah, UAE. The framework reports constant contaminations level and their area on a 24-h/7-day premise.

Alessandra, De Paola et. al. (2011) [13] In this paper, normal tactile gadgets for measuring natural information are commonly heterogeneous, and present strict vitality limitations; also, they are likely influenced by commotion, and their conduct may differ crosswise over time. Bayesian Networks constitute an appropriate device for pre-handling such information before performing more refined simulated thinking; the methodology proposed here goes for getting the best exchange off amongst execution and expense, by adjusting the working method of the basic tangible gadgets. Besides, self-design of the hubs giving the proof to the Bayesian system is done by method for an on-line multi-target enhancement.

S.C., Hu, Y.C. Wang, Huang et. al. (2011) [14] In this paper, considers a small scale atmosphere checking situation, which more often than not requires conveying a substantial number of sensor hubs to catch ecological data. By misusing vehicular sensor systems (VSNs), it is conceivable to prepare less hubs on autos to accomplish fine-grained observing. In particular, when an auto is moving, it could lead estimations at various areas, in this way gathering bunches of detecting information. To accomplish this objective, this paper proposes VSN design to gather and measure air quality for miniaturized scale atmosphere observing in city zones, where hubs' portability might be wild, (for example, taxis). In the proposed VSN engineering, we address two system related issues:

1) how to adaptively change the reporting rates of versatile hubs to fulfill an objective observing quality with less correspondence overhead and

2) how to misuse deft interchanges to decrease message transmissions. We propose calculations to unravel these two issues and check their exhibitions by reenactments.

What's more, we additionally build up a ZigBee-based model to screen the convergence of carbon dioxide (CO2) gas in city zones.

Alessandra, De Paola, Salvatore Gaglio et. al. (2012) [15] In this paper, surrounding Intelligence frameworks are normally portrayed by the utilization of pervasive gear for observing and adjusting the earth as indicated by clients' needs, and to internationally characterized imperatives. Our work depicts the usage of a testbed giving the equipment and programming instruments for the advancement and administration of AmI applications taking into account remote sensor and actuator arranges, whose fundamental objective is vitality putting something aside for worldwide maintainability. An example application is introduced that locations temperature control in a workplace, through a multi-objective

fluffy controller checking clients' inclinations and vitality utilization.

Srinivas, Devarakonda, ParveenSevusu et. al. (2013) [16] In this paper, customarily, contamination estimations are performed utilizing costly gear at settled areas or committed versatile hardware labs. This is a coarse-grained and costly approach where the contamination estimations are few and far in the middle. In this paper, we show a vehicular-based portable methodology for measuring fine-grained air quality continuously. We propose two practical information cultivating models - one that can be conveyed on open transportation and the second an individual detecting gadget. We introduce preparatory models and talk about execution challenges and early examinations.

IX. Conclusion and Future Scope

Air pollution control in city ranges is one of the fundamental requests of examination in the natural sciences. For each and every traverse the defilement reasons and pollution scattering are divergent, dependent on the assembling consideration, on vehicles activity, on inner sources et cetera, and additionally on the topographical area, temperature of the air, rate and relationship of the wind, and supplementary meteorological conditions components. Innumerable scientific models are used for the portrayal of the associations in the midst of natural security and meteorological elements. An option route to the scientific models is a learning based way that consolidates a few starting points of vision in a dream base. The paper portrays a case find of vision demonstrating in an air contamination utilizing ontologies. In upcoming we will endeavor to use regression models and ontology for city air contamination estimation.

References

[1] J., Fenger, 1999. Urban air quality. Atmospheric environment, 33(29), pp.4877-4900.

[2] Noboru, Yamazoe, Go Sakai, and KengoShimanoe. "Oxide semiconductor gas sensors." Catalysis Surveys from Asia 7, no. 1 (2003): 63-75.

[3] Mihaela. Oprea, "A case study of knowledge modelling in an air pollution control decision support system." Ai Communications 18, no. 4 (2005): 293-303.

[4] Michael, Blaschke, Thomas Tille, Phil Robertson, Stefan Mair, Udo Weimar, and Heiko Ulmer. "MEMS gas-sensor array for monitoring the perceived car-cabin air quality." Sensors Journal, IEEE 6, no. 5 (2006): 1298-1308.

[5] H. K. Eminir, Hala Abdel-Galil: Estimation of Fair Pollutant Concentrations fromMeterological Parameters Using Artificial Neural Network, Journal of Electrical engineering, Vol. 57, No. 2, 2006, pp. 105-110

[6] Uichin, Lee, Biao Zhou, Mario Gerla, Eugenio Magistretti, Paolo Bellavista, and Antonio Corradi. "Mobeyes: smart mobs for urban monitoring with a vehicular sensor network." Wireless Communications, IEEE 13, no. 5 (2006): 52-57.

[7] Jakob, Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, and Hari Balakrishnan. "The pothole patrol: using a mobile sensor network for road surface monitoring." In Proceedings of the 6th international conference on Mobile systems, applications, and services, pp. 29-39. ACM, 2008.

[8] Yajie, Ma, Mark Richards, MoustafaGhanem, YikeGuo, and John Hassard. "Air pollution monitoring and mining based on sensor grid in London." Sensors 8, no. 6 (2008): 3601-3623.

[9] Giuseppe, Anastasi, Giuseppe Lo Re, and Marco Ortolani. "WSNs for structural health monitoring of historical buildings." In Human System Interactions, 2009. HSI'09. 2nd Conference on, pp. 574-579. IEEE, 2009.

[10] Giuseppe, Anastasi, OrazioFarruggia, G. Lo Re, and Michele Ortolani. "Monitoring high-quality wine production using wireless sensor networks." In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on, pp. 1-7. IEEE, 2009.

[11] Artis, Mednis, Girts Strazdins, Martins Liepins, AndrisGordjusins, and Leo Selavo. "RoadMic: Road surface monitoring using vehicular sensor networks with microphones." In Networked Digital Technologies, pp. 417-429. Springer Berlin Heidelberg, 2010.

[12] A. R., Al-Ali, Imran Zualkernan, and FadiAloul. "A mobile GPRS-sensors array for air pollution monitoring." Sensors Journal, IEEE 10, no. 10 (2010): 1666-1671.

[13] Alessandra,De Paola, Salvatore Gaglio, Giuseppe Lo Re, and Marco Ortolani. "Multi-sensor fusion through adaptive Bayesian networks." In AI* IA 2011: Artificial Intelligence Around Man and Beyond, pp. 360-371. Springer Berlin Heidelberg, 2011.

[14] S.C., Hu, Y.C. Wang, Huang, C.Y., Tseng, Y.C.: Measuring air quality in city areas by vehicular wireless sensor networks. J. Syst. Softw. 84(11), 2005–2012 (2011)

[15] Alessandra, De Paola, Salvatore Gaglio, Giuseppe Lo Re, and Marco Ortolani. "Sensor 9 k: A testbed for designing and experimenting with WSN-based ambient intelligence applications." Pervasive and Mobile Computing 8, no. 3 (2012): 448-466.

[16] Srinivas, Devarakonda, ParveenSevusu, Hongzhang Liu, Ruilin Liu, LiviuIftode, and BadriNath. "Real-time air quality monitoring through mobile sensing in metropolitan areas." In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, p. 15. ACM, 2013.

Decision Support System for Diabetes Mellitus through Machine Learning Techniques

Tarik A. Rashid
Software & Informatics Engineering
College of Engineering
Salahadin university-Erbil
Hawler, Kurdistan

Saman . M. Abdulla
Software Engineering
College of Engineering
Koya university
Hawler, Kurdistan

Rezhna . M. Abdulla
Software & Informatics Engineering
College of Engineering
Salahadin university-Erbil
Hawler, Kurdistan

***Abstract*—recently, the diseases of diabetes mellitus have grown into extremely feared problems that can have damaging effects on the health condition of their sufferers globally. In this regard, several machine learning models have been used to predict and classify diabetes types. Nevertheless, most of these models attempted to solve two problems; categorizing patients in terms of diabetic types and forecasting blood surge rate of patients. This paper presents an automatic decision support system for diabetes mellitus through machine learning techniques by taking into account the above problems, plus, reflecting the skills of medical specialists who believe that there is a great relationship between patient's symptoms with some chronic diseases and the blood sugar rate. Data sets are collected from Layla Qasim Clinical Center in Kurdistan Region, then, the data is cleaned and proposed using feature selection techniques such as Sequential Forward Selection and the Correlation Coefficient, finally, the refined data is fed into machine learning models for prediction, classification, and description purposes. This system enables physicians and doctors to provide diabetes mellitus (DM) patients good health treatments and recommendations.**

Keywords—Diabetes disease; Blood sugar rate and symptoms; ANN; Prediction and Classification models

I. INTRODUCTION

The International Diabetes Federation stated that within the next 20 years, the figure of diabetic persons will stretch to 285 million in the world [1, 2]. Consequently, numerous research works have been conducted to analyze and categorize the DM patient types [3, 4]. Most of researchers have depended further on artificial intelligence (AI) and data mining (DM) techniques for constructing their classifier or forecaster models. They aimed at targeting two important objectives to AI classifier models; first is to point out the most related features and predictors or statically so called independent variables that should have no correlation among each other and have strong correlation with the desired target. Second is to select a suitable AI technique as a classifier or predictor tool which would possibly produce highest accuracy rate [5, 6]. Thus, at this stage, most of AI models would not provide or improve something to the knowledge of the physicians and medical staffs who are observing DM cases. The only support that they can provide is to categorize the type of DM cases or predict glucose rate in the blood. On the other hand, the most important missing benefit to the physician staffs and even to the indigent patients themselves is to describe the future of DM patients. Thus, it is so crucial to study the symptoms of DM patients not only to categorize their types, but also to envisage what side-effects or more chronic diseases a patient should anticipate.

For the above reasons, the influences of this work can go further than just classifying DM cases. Thus, the main contributions are as follows: 1) It utilizes some independent variables (which are consisted of; a- independent variables that a consultant for DM has considered, b- independent variables that considered by researchers in their previous works and c- independent variables that considered by this work) to diagnose or predict the rate of blood surge for patients through ANN model. 2) After diagnosing or prediction, the work utilizes more variables (symptoms of the patients) and the predicted blood sugar rate for the same patients to find out the relation between the symptoms and five major chronic diseases that diabetic patients have high probability to get them.

The rest of this paper is structured as follows: The next section describes the background of DM and AI techniques. Section 3 describes the proposed method, and in Section 4 the discussion and conclusion of the paper are outlined, and finally, the future work is suggested.

II. DM AND AI TECHNIQUES

The most important AI techniques that have been used by researchers are Artificial Neural Network (ANN), Support Vector Machine, Fuzzy Logic systems, K-mean classifier, and many others [3, 7-11]. ANN is considered to be the most popular one among all. A review work on using ANN in medical diagnoses is accomplished by [7], and it has been displayed in [16] that ANN can have several practices and can have different algorithms for training. Most of research works utilized the multilayer ANN with feed-forwarded back propagation (FFBP) algorithms to achieve DM classification. A research work has been done by [5] to categorize diabetic patients into insulin and non-insulin. The work depended on datasets collected in India, called Pima Indian Diabetes Dataset [8]. Another research work used the same FFBP algorithm to diagnose the DM cases [6]. They collected the database from Sikkim Manipal Institute of Medical Sciences Hospital, Gangtok, Sikkim for the diabetic patients.

Fuzzy logic classifier model is another type of AI tools that has been utilized by researchers [11] to categorize cases into type-1 and type-2. Their work relied on a secondary type of

database called Pima dataset. The accuracy of their work was evaluated against the rate of misclassified cases. A particular work utilized Decision Tree (DT) which is also considered as an AI tool for diagnosing diabetes to achieve classification and compared to ANN. They concluded from their results that DT demonstrated better accuracy [13].

Several approaches and algorithms were used to extract hidden information from biomedical datasets. A research work conducted to classify diabetes cases using Principal Component Analysis and Neuro-Fuzzy Inference. The diabetes disease dataset that used in this study was taken from Machine Learning Database (Department of Information and Computer Science, University of California) and the obtained classification accuracy was 89.47% [14]. Another work conducted to classify diabetes cases and they obtained 78.4% classification accuracy with 10-fold cross-validation (FCV) using Evolving Self-Organizing Map [15]. A combination approach was followed to combine Quantum Particle Swarm Optimization (QPSO), Weighted Least Square (WLS) and Support Vector Machine to diagnose Type-II diabetes. More Research works recorded in this area as the one that applied c4.5 algorithm for classification and it obtained 71.1% accuracy rate [16].

As mentioned in all above works, researchers continuously were busy to classify or diagnose diabetes cases into some predefined categories. Nevertheless, lately, physicians and doctors are concerned about the likelihoods of more chronic diseases or problems that might attack the diabetic patients. Thus, this work is given more descriptive information about diabetic patients through predicting which chronic diseases (problems) more probably can attack a diabetic patient or case based on their detected symptoms.

III. The System Struture

The new proposed approach in this work has a combination form in which the flow operation is shown in Fig. 1. Details on the pro-posed system are elaborated in the next four parts:-

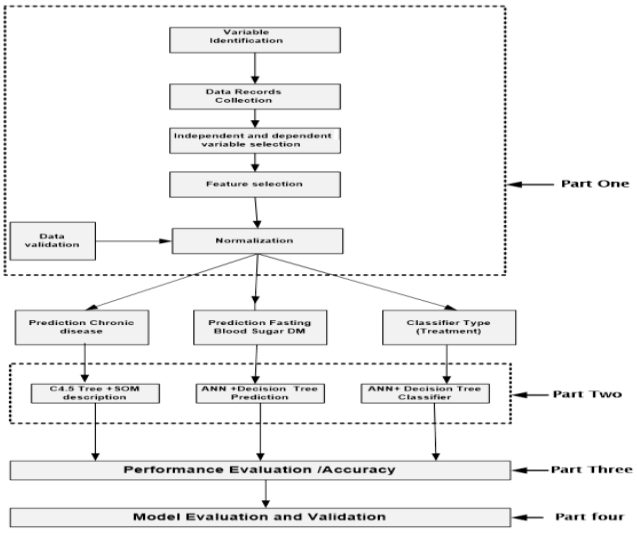


Fig. 1. The Flowchart of the Proposed Emotion Recognition System Structure

A. Part one: Data Collection and Pre-Processing

The part one can be divided into the following stages:-

1) Model's Variables Identification

From the relevant works that mentioned in section 2, many independent variables that are related to DM cases could be obtained. However, it is important to mention that each work has utilized the variables in a specific way. The collected independent variables can be grouped into physical, biological, and symptoms. Most physical records were about the height, weight, mass index, age and sex of the DM patients [17]. There are also some physical recodes that usually considered for specific cases, for instance, the number of pregnancy for a DM patient was only considered for female cases or if the research study is about Gestational Diabetes Mellitus GDM [13]. The biological related records are mostly Blood Glucose rate, Systolic blood pressure, and Diastolic blood pressure [6]. Researchers utilized these independent variables to build models that can classify DM cases or predict rates of glucose in blood.

The final group of independent variables is known as patient's symptoms. These records are more related to troubles and problems that a DM patient may suffer form, and together some of them are related to signs of DM disease. Some of these problems or symptoms are related to some chronic diseases. The second source of this work is to check the record list of DM patients in a specialist clinic. To achieve that, an official communication conducted between Salahaddin University-Erbil, College of Engineering and Layla Qasim Consultant Center Clinic for DM through the Department of Health in Erbil. Table 1, shows the most important records (independent variables) that a specialist clinic or consultant for DM has considered in his documentations. The table is also showing the most important independent variables that considered by researchers in their works. The third column in the table is about the independent variables that considered by this work to build an enhanced model for diagnosing and describing DM patients. The column shows how this work included the most important variables that are considered by DM consultant clinics and previous researches in DM diagnoses filed.

2) Data and Records' Collection

The data collection in this research work involves 26 variables through the process of building and simulating an enhanced medical model for diagnosing and describing DM patients. At this stage, it is necessary to collect records and data about all those variables. To achieve that, a second visit to Layla Qasim Center Clinic for DM patients has been made.

Through different processes, tests and interviewing, the record fields for each (501 patients) patient have been collected. The process of collecting data and records has been achieved carefully by the authors of this work under the supervision of specialist, medical and physician personnel.

After 60 days of getting data and records about DM patients. Two types of records obtained; the first type is those variables that have units such as Weight (kg), Height (cm), and S.BP (mg/dl). The second type is those variables that are logically recorded as 0 or 1. These logically variables are

related to some symptoms that DM patients might have felt them, such as Polyuria, Weakness, Numbness, and Coma.

3) Dependent and Independent Variable Selection

All records (there are 26 variables) that have been collected for DM patients are tabulated for each patient. It is necessary now to define the independent variables and dependent variables among these records. It is also necessary to find out the correlation between the dependent and independent variables in order to define the targets for both prediction and description sub-models.

4) Feature Selection

This work utilized two types of feature selection methods. Details of feature selection techniques and their implementation are outlined in the three points below:-

- Sequential Forward Selection (SFS): It attempts to find the best feature subset that decreases the feature space dimensionality with the smallest loss in classification accuracy. In other words, for a set of *D* features, the algorithm chooses a subset of size *d* <*D*, which has the greatest ability to discriminate between classes. The goodness of a particular feature subset is evaluated using an objective function, $J(Y_m)$, where Y_m is a feature subset of size *m* [18]. SFS is considered as a greedy search algorithm that chooses a top set of features for extraction through beginning from a void set and successively adding a distinct feature in the superset to the subset when increasing the value of the chosen objective function [19]. This type of algorithm has $O(n^2)$ worst-case complexity. Suppose we have a set of d_1 features X_{dI}, for each of the feature yet not selected ξ_j (i.e. in X- X_{dI}) the criterion function is evaluated according to the below equation [19]:-

$$J_j = J(X_{dI} + \xi_j) \quad (1)$$

The feature that yields the maximum value of J_j is chosen as the one that is added to the set X_{dI}. Thus, at each stage the variable is chosen, when added to the current set, and it maximizes the selection criterion. The feature set is initialized to the null set. Whenever the best improvement makes the feature set worst, or when the maximum allowable number of feature is reached, the algorithm terminates. Here, *J* can be given by the below equation [20]:-

$$j = X_k^T . S_k^{-1} \quad (2)$$

Where X_k is a *k* dimensional vector and S_k is a *KxK* positive definite matrix, where *K* features are used. At each stage of the search; sets of subsets are generated for evaluation. Variable ξ_j is chosen for which $J(X- \xi_j)$ is the largest. The new set is $(X- \xi_j)$. This process is repeated until the set of required cardinality remains. The following algorithm explains the whole procedure:-

a) initial X_{dI} *= null, m=1,*

X: is defined as a full set of feature

b) Choose new feature (ξ_j)

$$\xi_j = X - X_{dI} \quad (3)$$

c) K-fold checking learning performance

$$Avarage\ J_m = X^T . X_{dI} . \xi_j \quad (4)$$

d) Checking feature

$Avarage\ J_{m+1}$ *is better than* $Avarage\ J_m$

e) Go to step 2

Worst set of X_{dI} *is achieved*
Number m features are reached

f) Create training and testing sets

- The Correlation Coefficient *(r): r*, is a rapid portion that can define the scope of the statistical correlation between two variables. *r,* is scaled in a way that is constantly between -1 and +1. As soon as *r* is close to 0, it means that there is little correlation between the variables and the farther away from 0, *r* is, in either the positive or negative direction, the greater the correlation between the two variables. To compute the correlation between two variables, below steps can be followed [23]:-

a) Start with a set of data, x and y points. Each data point is kept in a separate row.

b) Find $\bar{x}$, $\bar{y}$, *the mean of x and y respectively. To do this, add the values of x and divide by the number of points; then, do the same process for y. Use equation below: - .*

c) $\bar{x} = \sum x/n$, $\bar{y} = \sum y/n$ (5)

d) Subtract $\bar{x}$ *from each value of x and subtract* $\bar{y}$ *from each value of y to get a new table of rows.*

e) Where, $y - \bar{y}$, $x - \bar{x}$

f) Compute the products of each row in step 3 and calculate the sum.

g) Take each x value in step 3, square it and calculate the sum of all points; do the same thing for y.

h) Calculate the square root of the product of the sums of the squares in step 5.

i) Calculate r by dividing the sum in step 4 by the value in step 6.

$$r = \frac{\sum (X - \bar{X}).(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 . (Y - \bar{Y})^2}} \quad (6)$$

The most significant variable of these is added to the model, so long as its *P-value* is below some pre-set level. It is customary to set this value above the conventional 0.05 level at say 0.10 or 0.15, because of the exploratory nature of this method. The number of variables that fed into feature selection technique are 24 variables, assume it as *X*. The first step is to check whether the validation of data is acceptable or not. This work utilized holdout validation method as a function, which depends on checking the re-substitution performance rate. The validation has been checked using Quadratic Discrimination Analysis *(QDA)* to check the covariance matrix of training set whether they are positive or negative. The acceptable value of the *(QDA)* should be positive. Assume V_{QAD} is the function of the validation check of features *X*.

$$P\ value = V_{QAD}(X) = Positive\ number \quad (7)$$

The first check that is done by this work is the value of the function of V_{QAD}, which represented as *P* value for variables.

The second check is doing t-test, which defines the relation between each *P* value that found for each variable and Cumulative Distribution, which denoted as *CDF* (cumulative distribution function), for each *P*. The relation between these two values should be as close as to one. For any variable, if the relation value is close to zero, the variable will be considered as insignificant. Or else, the variable will be significant if the *CDF* value of *P* is close to one.

Features are selected based on the correlation coefficient value between independent variables and target variable. Although, the correlation coefficient usually used to find the relation between features themselves, in this work it has been utilized to find the correlation between each features and the target variables. The feature selection implementation can be explained as follows:-

- Feature Selection Implementation for DM classification Sub model: The input data set includes 24 variables, excluding the Diabetes Type and the FSR columns as they are considered as targets or dependent variables. Based on the SFS algorithm, the significant features will be those variables that their P-Values are less than 0.09. Table 2, shows the output of the SFS and the features that can be selected for DM classification, which used to distinguish Type-I DM from Type-II. According to the *P* values of the all 24 variables, the won variables are the 16 variables that are highlighted in yellow color in the table. For the second test, this work has found the relation between *P* values and *CDF (P)* values. As explained, the value of this function should be as close as possible to one. According to Fig. 2, which shows the *P-CDF(P)* relation of features that collected by this work, is 33 % of all features, which is equal to eight features that are insignificant. However, the remaining 66 % which is equal to 16 features are significant.
- Feature Selection for Blood Surge Prediction Sub model: The P-Value in this case has the same range (only less than 0.09 is considered). Table 3 illustrates that P-value of all variables and the highlighted values are the considered features. According to the condition, only 14 features considered as significant variables.

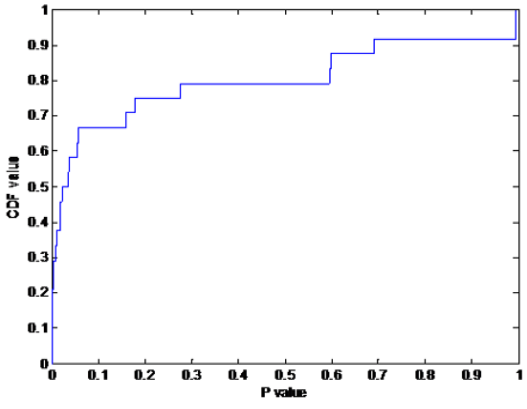


Fig. 2. P-CDF (P) relation of features with P value

5) Data Normalization

Techniques such as Min-Max and Z-Score are used to normalize data. The Min-Max involves the linear transformation on raw data. *Min A* and *Max A* are minimum and maximum value for the attribute *A* respectively. This technique maps the value of attribute *A* into range of [0,1], as in the following equation [21, 22]:-

$$v' = \frac{v - MinA}{MaxA - MinA} \qquad (8)$$

Z-Score is considered as a useful normalization technique when the actual minimum and maximum values of attribute are unknown. This is expressed in the below equation [22]:-

$$v' = \frac{v - \bar{A}}{\sigma A} \qquad (9)$$

Where, $\bar{A}$, σA and v are the mean, standard deviation and value of attribute of *A* respectively. The main advantage of this technique is to put the row data in specific range, so that models can map input-output relationship easily. The main problem that should be avoided in such process, is occurring zero redundant inside a specific attribute.

6) Data Validation

After normalizing input data, it is necessary to validate the data, to observe whether they are generalized or not. This work executed the 10-fold method to discover the performance of each fold of data. Both obtained normalized tables have been fed to the proposed ANN with different division of training and testing data sets. For both tables, same steps are followed. The process started to get 10 % of all data, which is 50 observations, as a testing part of data and the remaining 451 observations are used for training. Each time, the selected data for testing have been changed to another 50 observations. Through ten times, all data are used for testing and training. Tables 4 and 5 are illustrating the results of this 10-fold validation process. All performance records showed that collected data are validated data as there is no outlier performance among the records.

B. Part two: DM Medical Model

The main model has three sub models. The first part receives five independent variables for each patient, and it does the prediction and / or classification. The second sub-part is more important than the first one as it provides more important information to physician staffs. Details of each sub models are as follows:-

1) The classification sub model

An ANN is proposed to build a classifier model that can distinguish the type of treatment between insulin and noninsulin. Matlab program is used to build the DM treatment based ANN classifier sub model. The network has been trained based on back propagation algorithm, the network receives the 11 features [3, 24].

2) FBS prediction for DM patients

ANN is not only used for classification, it can be adopted for prediction too. In this work, the same algorithm is used for training the ANN to forecast the FBS rate for DM patients. The

only thing that changed with the presented ANN is the number of the selected significant features [3], the network receives 10 features.

3) DM description sub model

The last sub part of the main medical model is giving more details and information on a DM case with reference to symptoms with common chronic diseases relationships. The input part of this sub model is involved 10 features (Symptoms) and the output is 6 chronic diseases and problems, which are more common among DM cases. Through this prediction part, it will be easy for any physician to provide more details and information about the problems and chronic diseases that a DM patient might get them based on his/her symptoms. Two AI models were used for evaluations, these are namely; c45 and Self-Organizing Map (SOM) [7].

C. Part three: Performance Evaluations

There are five measuring parameters in this work, namely; True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), and the accuracy rate. The classification results are shown in Table 6. It can be seen from the table, trial 4 has the best TP and FN records, while trail 6 recorded the best TN and FN. And the best accuracy recorded at the trial 10. This is because, accuracy is directly changed with the ratio of all corrected classified objects to all object's number. The accuracy of the tree prediction sub model is shown in the Fig. 3. The figure shows the relation between a specific cost that required visiting a node and the number of nodes that have the same cost. Because the proposed decision tree is working as a regression mode, thus, this cost is the average squared error over the observations in that node. The dashed line in the cost is representing the minimum cost among the set of costs.

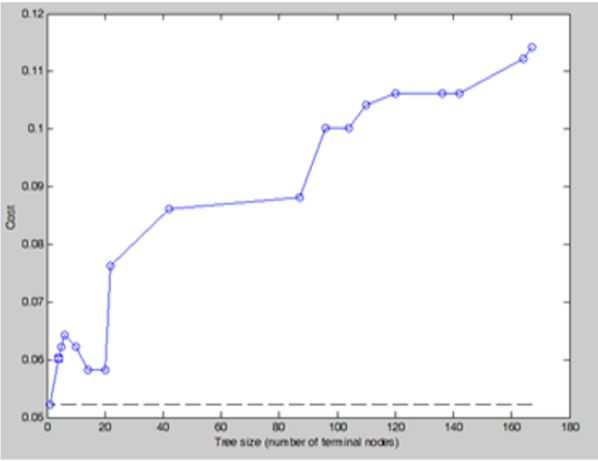


Fig. 3. The cost accuracy of the proposed decision tree sub model

D. Part four: Models Evaluations and Validations

In this work, the proposed model evaluated against supervised and unsupervised AI models. Fig. 4 shows the performance evaluation for classification AI models with / without feature selection process.

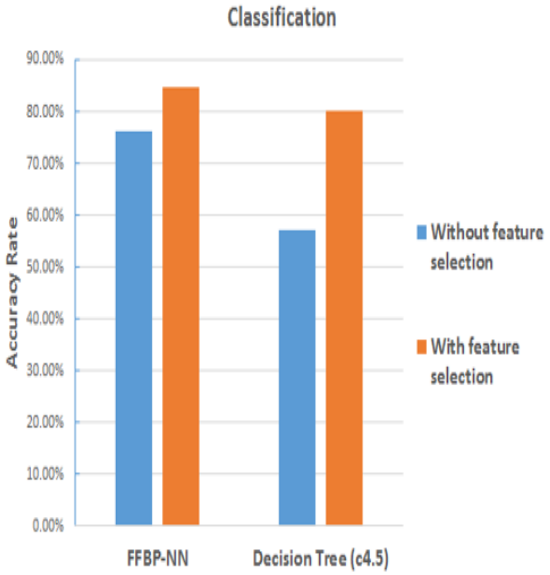


Fig. 4. Evaluation For classification

Fig. 5 show the performance evaluation for prediction FBS with /without feature selection process.

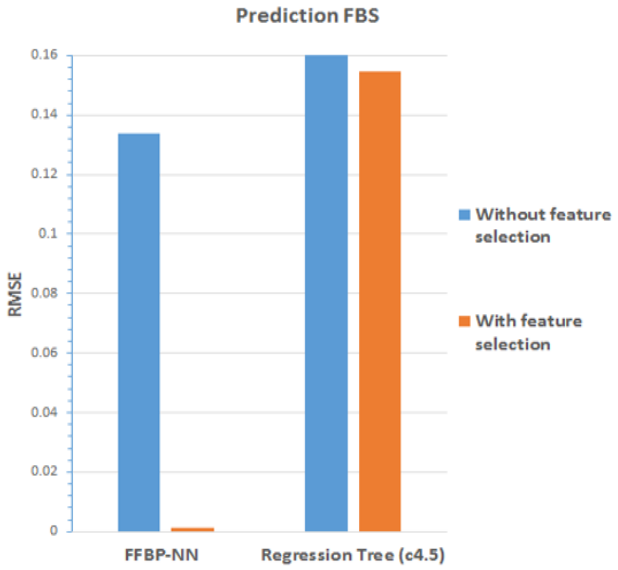


Fig. 5. Evaluation for Prediction

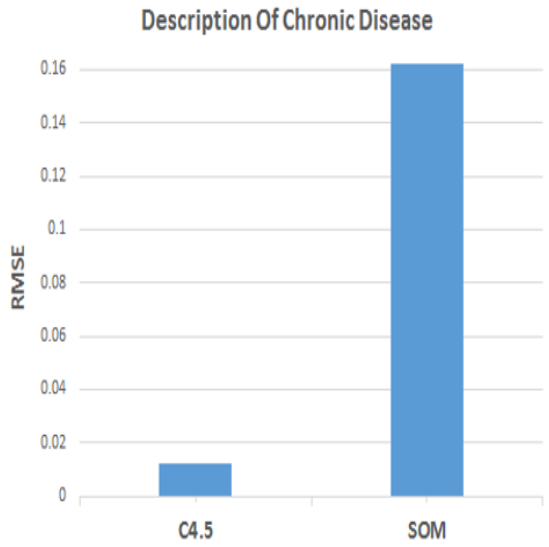
Fig. 6 shows the performance evaluation two AI models for Description of Chronic Disease.

Fig. 6. Evaluation for description

It is evident that ANNs trained with backpropagation learning algorithm using feature selection for the classification type of DM and for prediction FBS are better than c4.5 tree, however, for description of chronic disease the c4.5 is better than SOM.

IV. Discussion and Conclusion

In this paper, three intelligent models are designed for classification, prediction, and description purposes to offer complete knowledge about Diabetes Mellitus patients. Classification and prediction (classifying the types of MD patients and predicting FBS) models are very imperative for DM patients as doctors will be influenced by the outputs of these two models so that to espouse the type and dosage of treatments. Based on some symptoms extra health care recommendations are also specified by physicians to put DM patients away for some potential side effects.



It is worth to say that standard neural networks are good techniques for classification and easier to train when compared with deep neural networks that are regularly more difficult to train. This is a bad news, nevertheless, it is proven that if a deep neural network was, it would be much more controlling and prevailing than a standard neural network [25].

V. Recommendations for Future Work

While the suggested approach using artificial neural networks in this paper delivered encouraging outcomes for classification, prediction, and description purposes. As a consequence, the following future work can be suggested:-

1) It would be a good practice to use Deep Learning algorithms for classification problems instead of standard neural networks.

2) On the other hand, this research work can be upgraded in terms of learning algorithms such as using Grey Wolf Optimizer (GWO) and Bat algorithm (BA) which are freshly suggested swarm-based meta-heuristic.

TABLE I. The type of Variables used in Kurdistan's Clinics, Previous Research Works and in this Research Work

Variables of DM's Documented by consultants (Layla Qasim Clinic)	**Variables in DM Diagnosing and Classification in Previous Research Works**	**Variables in the Proposed Enhanced Medical Model**
Personal	***Personal***	***Personal***
1. Privacy records	1. Age 2. Weight 3. Height 4. Body index 5. Sex 6. Privacy records.	1. Age 2. Weight 3. Height 4. Sex 5. Privacy records.
Diabetic Related Records	***Diabetic Related Records***	***Diabetic Related Records***
1. FBS test 2. RBS test 3. HBA1C test 4. Insulin or tablet base (Treatment)	1. RBS test. (Glucose level) 2. Insulin or tablet (Treatment) 3. Diabetes pedigree function (pedi) (*Inheritance issue*) 4. HBA1C test	1. Fasting BS test. (Glucose level) 2. Insulin or tablet (Treatment) (class A or class B) 3. Diabetes pedigree function (pedi) (*Inheritance issue*) 4. Since when (year base)
Symptoms and problems Related Records	***Symptoms and problems Related Records***	***Symptoms and problems Related Records***
Not found	1. Polyuria 2. Nocturia 3. Polydpsia 4. Weakness 5. Paraesthesia 6. Frequency 7. Weight loss 8. Numbness 9. Polyhagia 10. Coma 11. Thirst 12. VD 13. Imp	1- S. Blood Pressure. 2- D. Blood Pressure. 3- Polyuria 4- Nocturia 5- Polydpsia 6- Weakness 7- Paraesthesia 8- Urinal Frequency 9- Weight loss 10- Numbness 11- Polyhagia 12- Coma 13- Heart Problem 14- Teeth Problem 15- Kidney Problem 16- Eyes Problem 17- Diabetic Foot (injury or damages)

TABLE II. THE P-VLAUE OF FEATURES (VARIABLES) FOR CLASSIFICATION SUB-MODEL

Feature Name	P-Value	Included or Excluded
Sex	0.623	No
Age	1.66 x E-33	Yes
Weight	1.38 x E-08	Yes
Height	0.021	Yes
S.BP	1.36 x E-09	Yes
Inheritance	0.020	Yes
D.B.P	0.021	Yes
Polyuria	0.110	No
Nocturia	0.032	Yes
Polydpsia	0.293	No
thirsty	0.127	No
Weakness	0.082	Yes
Par aesthesia	0.1812	No
Urinal Frequency	0.004	Yes
Losing Weight	0.025	Yes
Numbness	0.002	Yes
Polyhagia	0.915	No
Coma	0.060	Yes
Since When	0.0009	Yes
Eyes Problem	0.0004	Yes
Heart Problem	0.0885	Yes
Teeth Problem	0.0062	Yes
Kidneys Problem	0.9357	No
Injury Probe	0.3991	No

TABLE III. THE CORRELATION-VLAUE OF FEATURES (VARIABLES) FOR PREDICTION SUB-MODLE

Feature Name	P-Value	Included or Excluded
Sex	0.1956	No
Age	0.8169	No
Weight	0.1185	No
Height	0.0346	Yes
S.BP	0.8745	No
Inheritance	0.4134	No
D.B.P	0.0230	Yes
Polyuria	1.18 x E-05	Yes
Nocturia	0.00251	Yes
Polydpsia	1.19 x E-05	Yes
thirsty	5.36 x E-06	Yes
Weakness	0.00860	Yes
Paraesthesia	0.0012	Yes
Urinal Frequency	0.0014	Yes
Losing Weight	0.0615	Yes
Numbness	0.0005	Yes
Polyhagia	0.8836	No
Coma	0.2300	No
Since When	0.3106	No
Eyes Problem	0.0001	Yes
Heart Problem	0.0557	No
Teeth Problem	0.0422	Yes
Kidneys Problem	0.0923	No
Injury Problem	0.0059	Yes

TABLE IV. PERFORMANCE OF PREDICTION FASTING BLOOD SUGAR

Testing Data	Performance
1:50	0.0178
51:100	0.0171
101:150	0.0151
151:200	0.0210
201:250	0.0239
251:300	0.0149
301:350	0.0165
351:400	0.0274
401:450	0.0207
451:501	0.0158

TABLE V. PERFORMANCE FOR CLASSIFICATION TYPE OF DM

Testing Data	Performance
1:50	0.0298
51:100	0.0441
101:150	0.0279
151:200	0.0236
201:250	0.0400
251:300	0.0210
301:350	0.0195
351:400	0.0371
401:450	0.0278
451:501	0.0201

TABLE VI. THE CONFUSION MATRIX FOR THE MD CLASSIFIER SUB-MODEL

#	TP%	TN%	FP%	FN%	Accuracy %
1	148	234	44	75	76.2
	29.	46.7	8.8	15.0	
2	144	242	36	79	77.0
	28.8	48.3	7.2	15.8	
3	124	226	52	99	69.9
	24.8	45.1	10.4	19.8	
4	179	238	40	44	83.2
	35.7	47.5	8.0	8.8	
5	172	245	33	51	83.2
	34.3	48.9	6.6	10.2	
6	133	251	27	90	76.6
	26.5	50.1	5.4	18.0	
7	176	250	28	47	85.0
	35.1	49.9	5.60	9.4	
8	166	245	33	57	82.0
	33.1	48.9	6.6	11.4	
9	172	239	39	51	82.0
	34.3	47.7	7.8	10.2	
10	186	250	28	37	87.0
	37.1	49.9	5.6	7.4	

REFERENCES

[1] P. Srimani and S. Koti, "Medical diagnosis using ensemble classifiers-a novel machine-learning approach," J Adv Comput, vol. 1, pp. 9-27, 2013

[2] S. Wild, G. Roglic, Green, A., et al., "Global prevalence of diabetes estimates for the year 2000 and projections for 2030," American Diabetes Association, Diabetes care, vol. 27, no. 5, pp. 1047-1053, 2004.

[3] A. Kumari and R. Chitra., "Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications, 2013. vol. 3, no. 2, pp. 1797-1801, 2013.

[4] B. Deekshatulu, and P. Chandra., "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection, " Global Journal of Computer Science and Technology, vol. 13, no. 3, 2013.

[5] Z. Zainuddin, O. Pauline and C Ardil, "A neural network approach in predicting the blood glucose level for diabetic patients," International Journal of Computational Intelligence,. vol 5, no. 1: pp. 72-79, 2009.

[6] B. Adeyemo, and E. Akinwonmi, "On the Diagnosis of Diabetes Mellitus Using Artificial Neural Network Models Artificial Neural Network Models," African Journal of Computing & ICT Reference Format, vol. 4, no. 1, pp. 1-8, 2011.

[7] F. Amato, F. Amato, A. López., P-M. María, et al., "Artificial neural networks in medical diagnosis," Journal of Applied Biomedicine, vol.11, no. 2, pp. 47-58, 2013.

[8] G. Karegowda, V. Punya, A. Jayaram et al., "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4. 5," International Journal of Computer Applications, vol. 45, 2012.

[9] L. Liberti, C. Lavor, N. Maculan, et al., "Euclidean distance geometry and applications," SIAM Review, vol. 56, no. 1, pp. 3-69, 2014.

[10] R. Dey, V. Bajpai, G. Gandhi, et al, "Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus," in IEEE Region 10 and the Third international Conference on Industrial and Information Systems," ICIIS 2008. Kharagpur: IEEE, 2008.

[11] F. Baldwin and W. Xie, "Simple fuzzy logic rules based on fuzzy decision tree for classification and prediction problem," in Intelligent information processing II, Springer. pp. 175-184, 2005.

[12] B. Yegnanarayana, B., "Artificial neural networks, " PHI Learning Pvt. Ltd, 2009.

[13] E. Caballero-Ruiz, E., G. García-Sáez, M. Rigla, et al, "Automatic Blood Glucose Classification for Gestational Diabetes with Feature Selection: Decision Trees vs. Neural Networks," in XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, Springer, 2014.

[14] B. Yegnanarayana1, "Artificial neural networks for pattern recognition Sadhana, vol. 19, no. 2, pp. 189-238, 1994.

[15] A. Feizollah, N. Badrul Anuar, R. Salleh, et al., "A review on feature selection in mobile malware detection. Digital Investigation, vol. 13,pp. 22-37, 2015.

[16] E. Berglund and J. Sitte, J., "The parameterless self-organizing map algorithm. Neural Networks," IEEE Transactions on,. vol. 17, no. 2, pp. 305-316, 2006.

[17] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks," in Data Storage and Data Engineering (DSDE), 2010 International Conference, 2010.

[18] S. García, J. Luengo and F. Herrera, "Feature Selection, in Data Preprocessing in Data Mining, Springer. pp. 163-193, 2015.

[19] M. Fauvel, C. Dechesne, A. Zullo, et al., "Fast forward feature selection for the nonlinear classification of hyperspectral images," arXiv preprint arXiv:1501.00857, 2015.

[20] L. Burrell, G. Georgoulas, E. Marsh, E., et al., "Evaluation of Feature Selection Techniques for Analysis of Functional MRI and EEG," in International Conference on Data Mining 2007, Las Vegas, 2007.

[21] N. Mohd, H. Atomia and Z. Rehman, Z., "The effect of data pre-processing on optimized training of artificial neural networks", Procedia Technology, 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013, vol. 11, pp. 32–39, 2013.

[22] A. Semary, A. Tharwat, E. Elhariri, E., et al., "Fruit-Based Tomato Grading System Using Features Fusion and Support Vector Machine," in Intelligent Systems' 2014, Springer. pp. 401-410, 2015.

[23] R. Momberum, "A comparative Study of the capital structures of liquid and liquidity - stressed Banks," in Financial Managmnet 2012, University of Johannesburg: http://hdl.handle.net/10210/8563 p. 115, 2012.

[24] T. Rashid, S. Abdullah and R. Abdullah, "An Intelligent Approach for Diabetes Classification," Prediction and Description, Editors: Vaclav Sansel, Ajith Abraham, Pavel Kromer, Millie Pant, Azah Kamilah Muda, In book: Series : Advanced in Intelligent Systems and Computing, Edition: 424, Chapter: IBIA 2105 Proceeding, Publisher: Springier Verlag, , pp.323-335, 2015.

[25] N. George, "Deep Neural Network Toolkit & Event Spotting in Video us-ing DNN features," master thesis, department of computer science and engineering, Indian institute of technology madras, 2015.

An Efficient Application Specific Memory Storage and ASIP Behavior Optimization in Embedded System

Ravi Khatwal
Research scholar
Department of computer science
MLS University
Udaipur, India

Manoj Kumar Jain
Professor
Department of computer science
MLS University
Udaipur, India

***Abstract*—Low power embedded system requires effective memory design system which improves the system performance with the help of memory implementation techniques. Application specific data allocation design pattern implements the memory storage area and internal cell design techniques implements data transition speeds. Embedded cache design is implemented with simulator and scheduling approaches which can reduce the cache miss behavior and improve the cache hit quantities. Cache hit optimization, delay reduction and latency prediction techniques are effective for ASIP design. The design functionality is simply specifying the tradeoff among various design metrics like performance, power, size, cost and flexibility. ASIP behavior and memory storage area optimized for low power embedded system and implements cycle time with effective scheduling techniques which implements the system performance with low power consumption.**

Keywords*—Memory design; Compiler; Processor design; Scheduling Techniques; Memory storage

I. INTRODUCTION

Embedded systems uses some specific constraints such as Real time design metrics are a measurement of application features such as Cost, Size, Power and High Performances. Reactive and real time required to implement our system environments and computed application results in real time without any delay [Fig. 1]. Currently embedded system designer are being designed on a silicon chip and also design for critical applications like killer application (smart phone), smart card, video game, mobile internet, handheld embedded system, GBPS device, gigabyte per second LAN system. Embedded design technologies used to improve the design technology to enhance productivity has been a focus on software and hardware design mechanism.

In HLS design mechanism, Xilinx simulator software is used to verify all the functionality and timing custom peripheral design architecture [18, 20]. ASIP design used to implement the functional unit may then either be integrated on a chip or implements peripheral devices. Profiler is effectively used in Pre-allocation memory design and implements pre-allocation based execution delay time. Recently a memory implementation technique is attracting strong research interest in ASIP. ASIP is a heterogeneous platform composed of programmable processor core and used customized hardware environments [1, 2, 3]. ASIC architecture is not flexible for specific application design architecture. DSP processor is also flexible and fully programmable; it can't achieve high performance with low power consumption and not suitable for various complex application development mechanisms.

VLIW processor unit require compiler support and VLIW architecture is characterized by instructions such that each specifies several independent operations. This is compared to RISC instructions that typically specify one operation and CISC instructions that typically specify the several operations with sufficient registers, A VLIW machine can place the results of speculative executed instructions in temporary registers. The level of sophistication in VLIW compiler is significantly higher.

The heterogeneous vector width method use to expose the heterogeneous vector widths for VLIW ASIP [10, 13]. Effective automation is analyzed for VLIW ASIPs. The lower bound latency is effective for VLIW ASIP. Latency bound mechanism implements the data transfer delays [9]. By the help of these approaches a window data flow graph and lower bound deign mechanism reduce the delay penalties due to operation serialization or data transfer mechanism.

An effective emulation tool chain designed for ASIP design architecture [5]. The FPGA based emulator is alternative to pure software cycle-accurate simulation and this tool chain to reduce the design exploration time [13]. Fast and accurate processor simulator used for high performance ASIP simulation [4] and an integrated tool chain design also evaluated for ASIP systems [5]. ASIP architecture also design for a Discrete Fourier transform (DFT)/Discrete cosine transform (DCT) /Finite impulse response filters (FIR) engine[14].

Memory data storage and operational optimal delay frequency analyzed according application computational conditions. Embedded process system analysis is presented in the next section. Section 3 and 4 represents the application specific data storage and data storage is effectively optimized in memory system. Last section represents application specific data storage in ASIP system and implements system performances with various techniques such as delay reduction, latency prediction and operational scheduling mechanism.

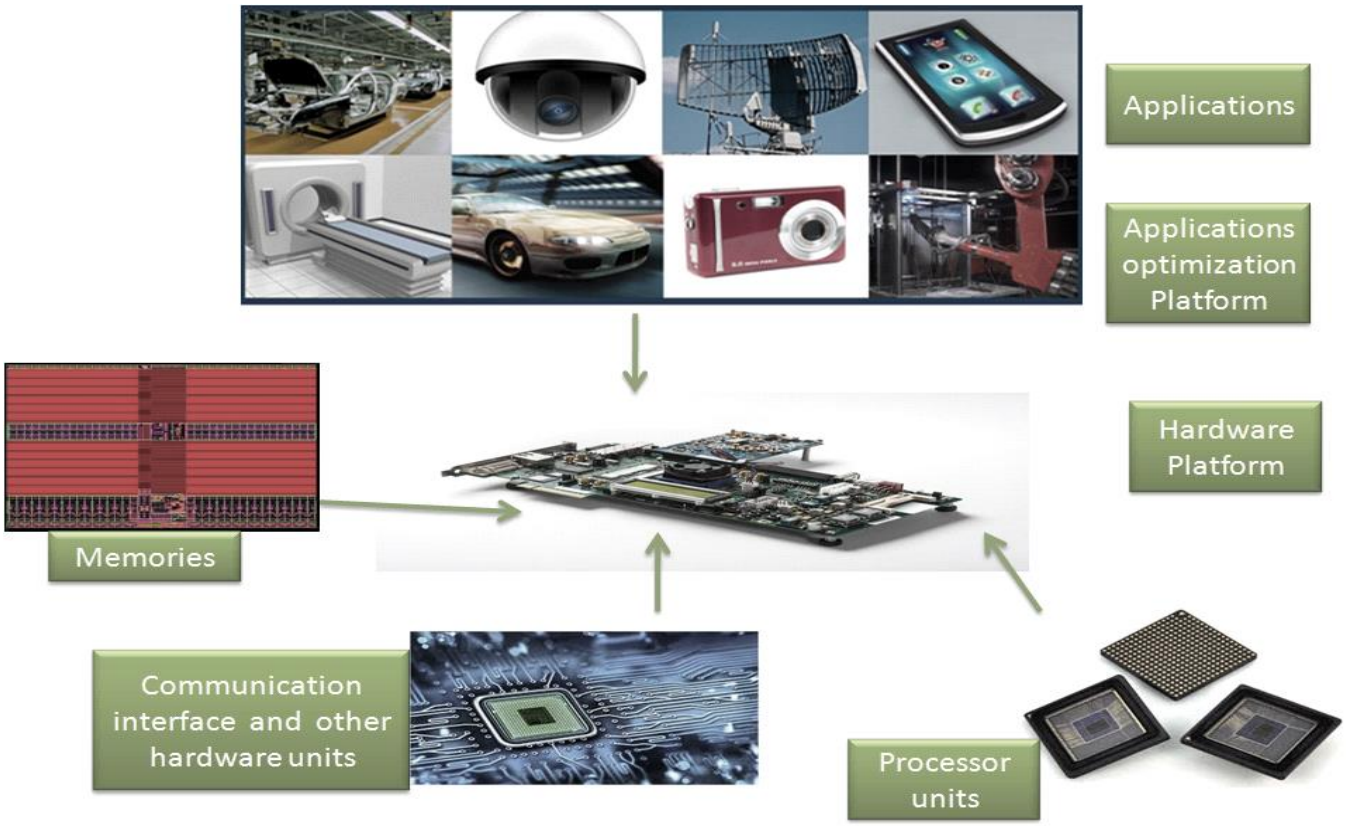


Fig. 1. Application Specific Requirements Based Embedded System Design

II. Application Specific Embedded Process Analysis

The basic process of embedded system is implemented with three basic mechanisms such as application compilation, synthesis and implementation, IP based integration and test and verification by specific simulator. By the help of this mechanism we implement the application based embedded systems design for low power embedded devices. In embedded system the HLS design mechanism Memory designer used high level language and implements behavioral specifications into register-transfer (RT) specifications by converting behavior on general-purpose processors to assembly code. The memory Designer also refines the register-transfer-level specification of a single-purpose processor into a logic specification and finally implements machine code for general-purpose processors and utilizes the gate-level net list. First Compilation/Synthesis process the designer specifies desired functionality in an abstract manner. A compiler translates the source language into its target machine language without having the option for generating intermediate code. Each new machine have a full native compiler is required [Fig. 2]. The Software compiler converts a sequential program to an assembly code, which is essentially a register-transfer code and a system synthesis tool converts an abstract system specification into a set of sequential programs on general and single-purpose Processors. A logic synthesis tool converts Boolean expressions into a connection of logic gates (called a net list). A register-transfer (RT) level synthesis tool converts finite-state machines and register-transfers into a data path of RT components and a controller of Boolean equations. A behavioral synthesis tool converts a sequential program into finite-state machines and register transfers.

Second **Libraries/IP based implementation** phase is used the logic-level library and it consists of layouts for gates and cells. The RT-level library may consist of layouts for RTL components, like registers, multiplexers, decoders, and functional units. A behavioral-level library may consist of embedded components, such as compression components, bus interfaces, display controllers, and even general- purpose processors. IP integration design is used to implement the memory or various peripheral devices and integrating the device according to our application requirements. Finally, a system-level library might consist of complete systems, solving particular problems, such as an interconnection of processors, memory with accompanying operating systems and programs to implement an interface.

Finally, **Test/Verification** phase we have analyzed the functionality of the design is correct or not and checked the mechanism with low abstraction levels to high abstraction levels. Simulation mechanism better utilizes the testing for correct functionality. The Logic level, gate- level simulators provides output signal timing waveforms with a given input signal waveform. And finally RTL level, hardware description language (HDL) simulators used to execute the RTL-level descriptions and provide output according to the given input waveforms. The behavioral level, HDL simulators used to simulate sequential programs and co-simulators connect HDL and processor simulators to enable hardware/software co-verification at the system level. Model simulator simulates the

initial system specification using an abstract computation model, that independently of any kind of processor technology and these simulators verify the correctness and completeness of the specification [Fig. 3].

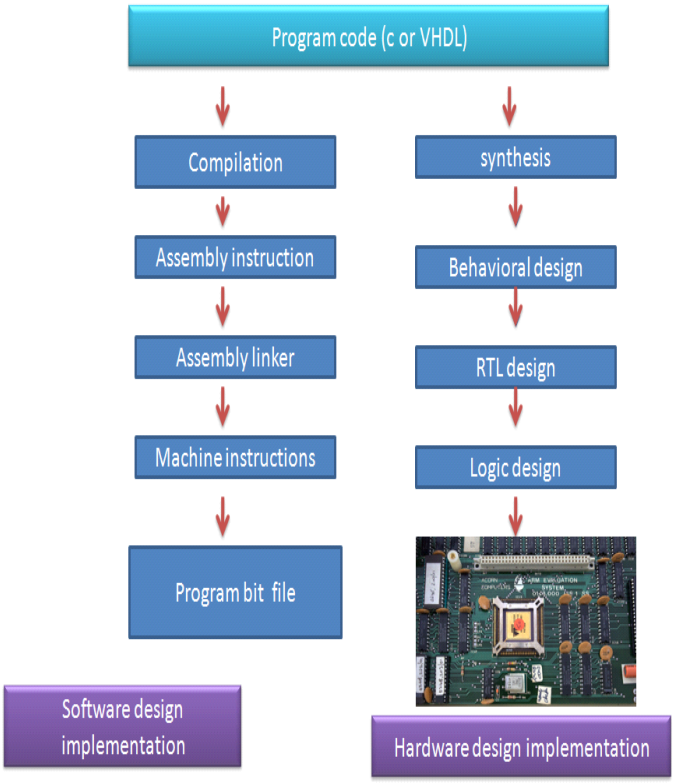


Fig. 2. Design process used in embedded system

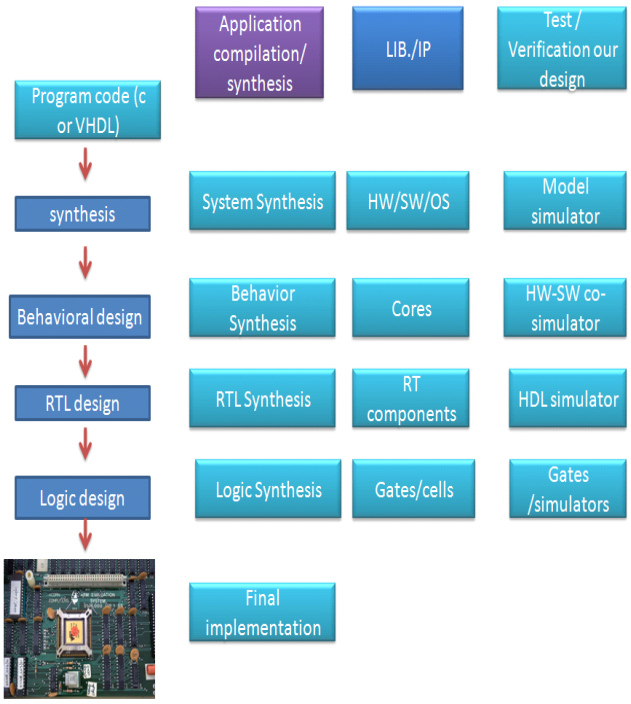


Fig. 3. Design process technology of embedded system

III. Appplication Specific Data Storage Analysis

Application specific computation frequencies analyzed according application behavior and their design complexity. Various standards application benchmarks used and analyzed the computation complexity with various critical conditions such as higher multiplication or lower multiplication. At First condition contains computation time O (n) loop used d n time's units and repeats a programming statement in n times. Second condition contains every iteration of the loop counter will be divided by 2 so computation designs as 2^4=16 words used. Third condition contains the nested loop used so computation designs complexity as O (n^2) and it represents a loop executing inside loops. Fourth condition contains operational computation complexity is O (n) loop independently of each other. Fifth condition contains computation complexity is O (n log n). Six condition computation complexity is 2^N complexity used due to loop multiply 2 so 2^N-1 possibility available and final condition have computation possibility in two part O (n) and O (n^2) available for memory allocation area. Critical benchmark application and their computation design complexity [Fig.4, Fig. 5, Fig. 6, Fig. 7 and Fig. 8]. Various critical application such as vision application, robotic memory design, mind mapping artificial neural network based application designs are implemented according to their computational complexity [Fig. 9 and Fig. 10].

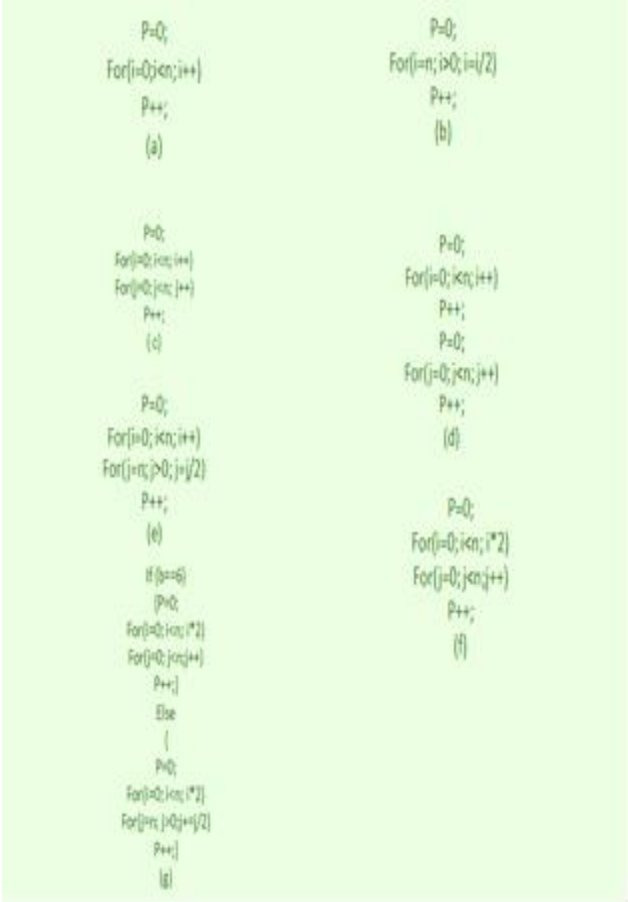


Fig. 4. Application specific operations frequency analysis with design complexity

```
*/
{
  int i;

  for ( i = 0; i < n; i++ )
  {
    x[i] = ( ( a + b ) + ( b - a ) * x[i] ) / 2.0;
  }
  for ( i = 0; i < n; i++ )
  {
    w[i] = ( b - a ) * w[i] / 2.0;
  }
```

Fig. 5. Patterson application design sections

```
cout << "CELLULAR_AUTOMATON:\n";
cout << "  C++ version.\n";

n = 80;
step_num = 80;

x = new char[n+2];
x_old = new char[n+2];

for ( i = 0; i <= n + 1; i++ )
{
  x[i] = ' ';
}
x[40] = '*';

for ( i = 1; i <= n; i++ )
{
  cout << x[i];
}
cout << "\n";

for ( j = 1; j <= step_num; j++ )
{
  for ( i = 0; i < n + 2; i++ )
  {
    x_old[i] = x[i];
  }
  for ( i = 1; i <= n; i++ )
```

Fig. 6. Cellular automation application section

```
for e = 1 : 17

  c = ( ( ( x - x0(e) ) * v11(e) ...
        + ( y - y0(e) ) * v12(e) ...
        + ( z - z0(e) ) * v13(e) ) / a1(e) ) .^2 ...
    + ( ( ( x - x0(e) ) * v21(e) ...
        + ( y - y0(e) ) * v22(e) ...
        + ( z - z0(e) ) * v23(e) ) / a2(e) ) .^2 ...
    + ( ( ( x - x0(e) ) * v31(e) ...
        + ( y - y0(e) ) * v32(e) ...
        + ( z - z0(e) ) * v33(e) ) / a3(e) ) .^2;

  i = find ( c <= 1.0 );

  f(i) = f(i) + g(e);

end

return
end
```

Fig. 7. 3D Shepp-Logan application design section

```
w = 1.0;

dc = ( double * ) malloc ( n * sizeof ( double ) );

for ( j = 0; j < n; j++ )
{
  dc[j] = 0.0;
}

while ( k < m )
{
  if ( k < m )
  {
    k = k + 1;
    for ( j = 0; j < n; j++ )
    {
      dc[j] = dc[j] + omega[k-1] * sin ( w * pi * x[j] );
    }
  }

  if ( k < m )
  {
    k = k + 1;
    for ( j = 0; j < n; j++ )
    {
      dc[j] = dc[j] + omega[k-1] * cos ( w * pi * x[j] );
    }
  }

  w = w + 1.0;

}
```

Fig. 8. Cycle reduction application design section

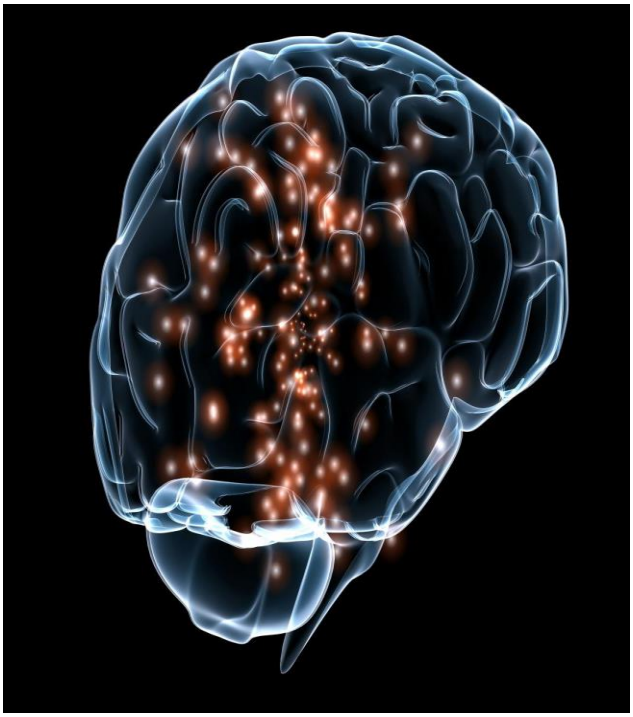


Fig. 9. Basic Mind mapping application section area

Fig. 10. Basic Embedded vision application section

IV. Data Storage Optimization in Memory System

Memory blocks are important concepts from both code generation and optimization point of view. Basic blocks play an important role in identifying variables (data), which are being used more than once in a single basic block. Memory blocks analysis with various schemes such as data flow, data allocation, data reusability, scheduling approaches and control flow design mechanism. When the first instruction is executed, then all the instructions in the same basic block will be executed in their sequence of appearance without losing the flow control of the program. The Blocking is another way of reordering iteration in a loop and greatly improves the locality of source code [Fig. 11 and Fig. 12]. Each memory element is implemented with the memory array design, architecture so block architecture design easily identify the data arrangements. Data reusability design is implemented by a References data storage mechanism. Matrix blocks are divided into sub blocks or sub matrixes and column implemented design effective reusable code design mechanism implements a cache hit condition [Fig. 16] [19].

Data stored in memory with the Serial execution mechanism of data elements depends upon row and column design architecture. Memory design is implemented with matrix level blocks architecture. In Matrix designs innermost loop reads and writes the same elements of z and use the row of x and column of y. Each block is designed according to the row & column accessing scheme [Fig. 12]. The Application based storage element is arranged in our block area and a single row is spread among only n/E element cache lines. When all data element is filling in the cache only n/E cache misses occur for a fixed value index and the entire total operation use n^2/E. If the cache is big enough that all n^2/E cache lines holding column Y can reside together in the cache, then no more cache misses [Fig. 14] occurred. Column index implementation technique implements the repositioning of memory data arrangements which reduces the cache misses or data cluster and it's easily serialize operational frequency. The total number of misses is depending upon $2n^2/E$, half for x and half of y. The Single processor will be computed n^2/E elements of Z; performing n^P/E where operation complexity p is changed according to application computation.

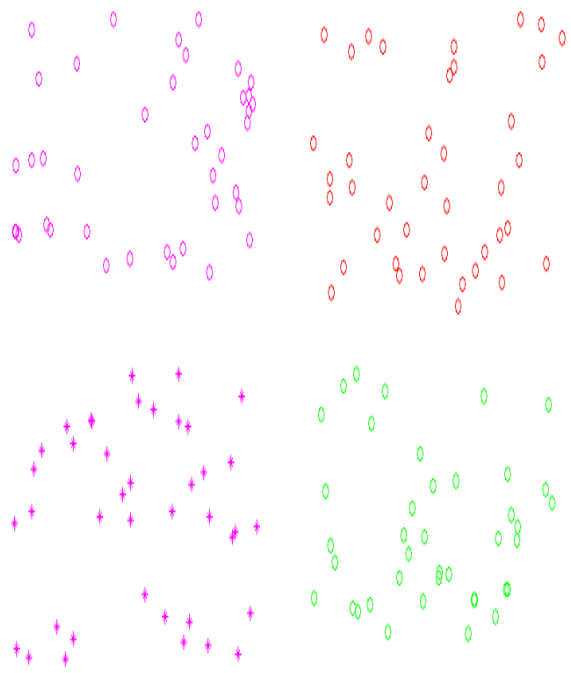


Fig. 11. Complex Cluster block analysis

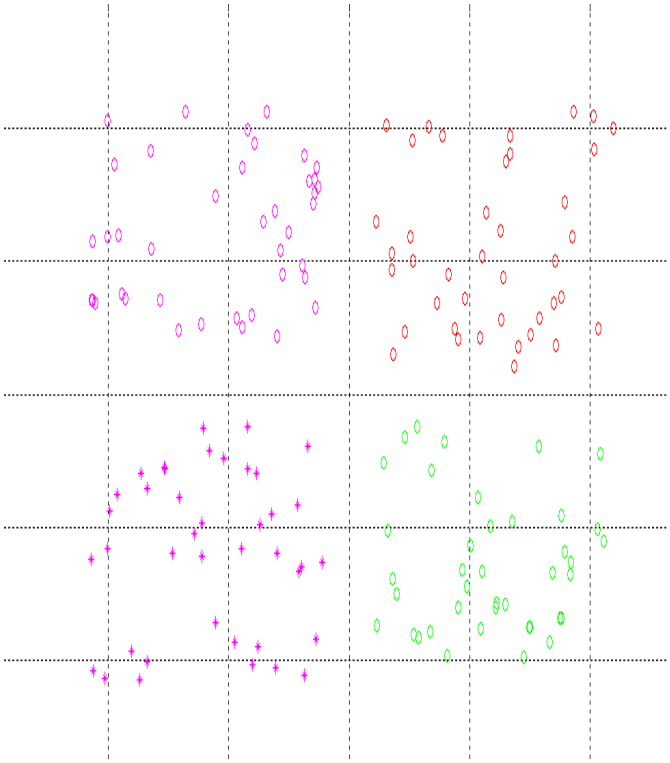


Fig. 12. Block filter mechanism which Increase data probability

A. Delay reduction design optimization

The critical path would be combination logic delay plus the logic circuits setup time, plus the clock output delay. The critical path analysis with various nodes based implementation. Complex memory structures have various critical sections. Various critical paths cell delay analyzed and combinational path delay is implemented with column based cell architecture. Application based column design implements cycle reduction and this design is used for data allocations which can implement data shifting and reduce the memory misses [Fig.16]. We have analyzed the performance based lower and higher frequency order based access time variations used in memory implementation mechanism [Fig. 13, Fig. 14 and Fig. 15]. The probability degree based access time pattern implements the critical section area. Higher critical section area has longer access time probability and it takes longer access time [Fig. 16]. Various approaches such as scheduling, allocation and binding pattern implements the access time and have a low probability frequency design which reduces the critical section area [Fig 17 and Fig. 18]. Node based critical section is implemented the high and lower order path for access time point of view and column implemented shortest path have lower access time path which have global impact in system performances.

```
    cout << "  Zero diagonal entry, index = " << i << "\n";
    exit ( 1 );
  }
}

if ( job == 0 )
{
  for ( it_num = 1; it_num <= it_max; it_num++ )
  {
    x[0] =   ( b[0]                  - a[2+0*3] * x[1]     ) / a[1+0*3];
    for ( i = 1; i < n-1; i++ )
    {
      x[i] = ( b[i] - a[0+i*3] * x[i-1] - a[2+i*3] * x[i+1] ) / a[1+i*3];
    }
    x[n-1] =   ( b[n-1] - a[0+(n-1)*3] * x[n-2]             ) / a[1+(n-1)*3];
  }
}
else
{
  for ( it_num = 1; it_num <= it_max; it_num++ )
  {
    x[0] =   ( b[0]                    - a[0+1*3] * x[1]     )
         / a[1+0*3];
    for ( i = 1; i < n-1; i++ )
    {
      x[i] = ( b[i] - a[2+(i-1)*3] * x[i-1] - a[0+(i+1)*3] * x[i+1] )
           / a[1+i*3];
    }
    x[n-1] =   ( b[n-1] - a[2+(n-2)*3] * x[n-2]                    )
           / a[1+(n-1)*3];
  }
}
```

Fig. 13. Diagonal matrix application design sections

```
k = seed / 127773;

seed = 16807 * ( seed - k * 127773 ) - k * 2836;

if ( seed < 0 )
{
  seed = seed + i4_huge;
}

r = ( float ) ( seed ) * 4.656612875E-10;
/
/  Scale R to lie between A-0.5 and B+0.5.
/
r = ( 1.0 - r ) * ( ( float ) a - 0.5 )
  +         r   * ( ( float ) b + 0.5 );
/
/  Use rounding to convert R to an integer between A and B.
/
value = round ( r );
/
/  Guarantee A <= VALUE <= B.
/
if ( value < a )
{
  value = a;
```

Fig. 14. Data allocation application design section

```
while ( 1 < il )
{
  ipnt = ipntp;
  ipntp = ipntp + il;
  il = il / 2;
  ndiv = ndiv * 2;

  for ( j = 0; j < nb; j++ )
  {
    ihaf = ipntp;
    for ( iful = ipnt + 2; iful <= ipntp; iful = iful + 2 )
    {
      ihaf = ihaf + 1;
      rhs[ihaf+j*(2*n+1)] = rhs[iful+j*(2*n+1)]
        - a_cr[2+(iful-1)*3] * rhs[iful-1+j*(2*n+1)]
        - a_cr[0+iful*3]     * rhs[iful+1+j*(2*n+1)];
    }
  }
}

for ( j = 0; j < nb; j++ )
{
  rhs[ihaf+j*(2*n+1)] = rhs[ihaf+j*(2*n+1)] * a_cr[1+ihaf*3];
}

ipnt = ipntp;

while ( 0 < ipnt )
{
  ipntp = ipnt;
  ndiv = ndiv / 2;
  il = n / ndiv;
  ipnt = ipnt - il;
```

Fig. 15. Cycle reduction application design section

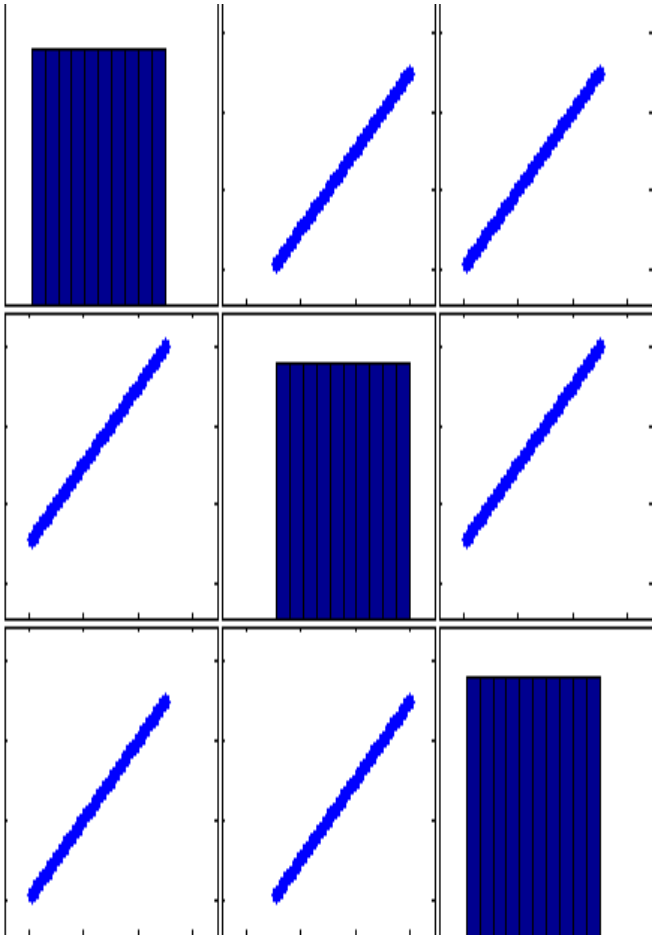


Fig. 16. Data filtering according to Diagonal design mechanism

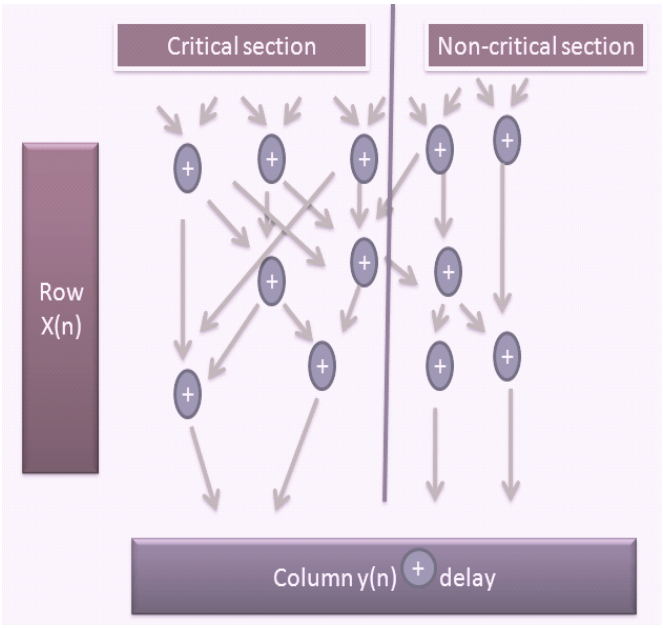


Fig. 17. Critical or non critical delay design section

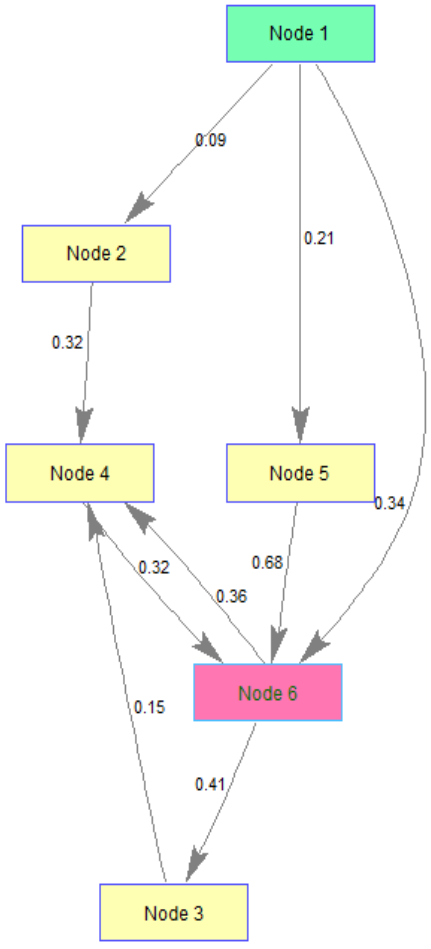


Fig. 18. Lower and higher delay point based critical section analyis

B. Scheduling approaches for reduction of memory space

Scheduling techniques are required to schedule the memory operations and operation scheduling effective determine the memory cost area. The scheduling algorithm will attempt to parallelize the operation to meet the timing constraints and scheduler mechanism will serialize the operation to meet the resource constraints [17]. Various scheduling problem implemented with different requirement such as time constraints, resource constraint, feasibly constrained [19, 20]. Memory operation scheduling implemented with three conditions such as FCFOP (First Come First Operational), LCLOP (Last Come Last Operational) and operational optimal degree based operational [Fig. 19, Fig. 20 and Fig. 21]. Scheduling approaches implement according to some conditions time, resource and feasible levels. The max no. of time step finds the cheapest schedule which satisfied the constraints. Lower resources find the fastest with satisfied the constraints [Fig. 22]. Feasible conditions decide if there exists a schedule which satisfied the constraints or not.

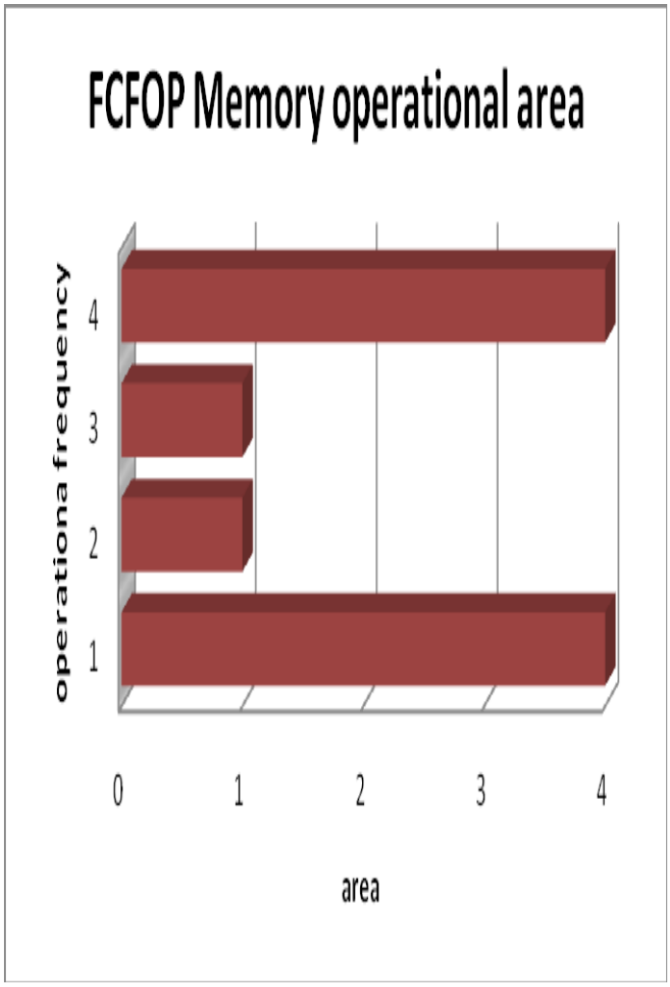


Fig. 19. FCFOP scheduling

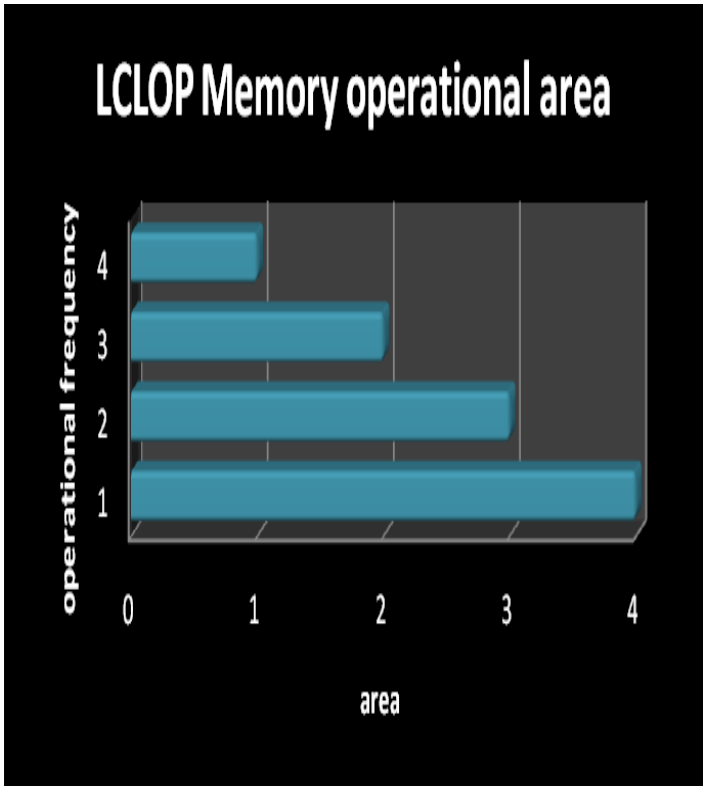


Fig. 20. LCLOP scheduling

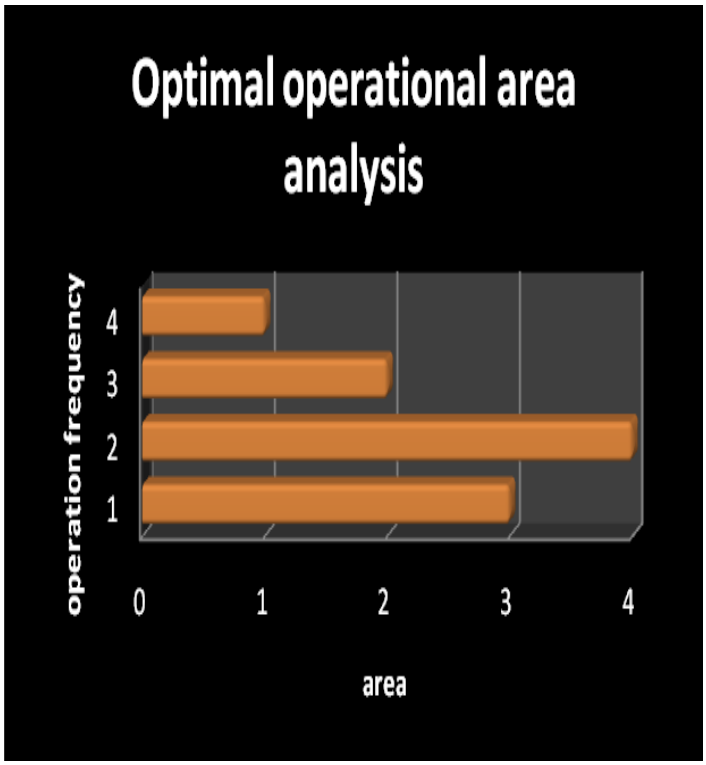


Fig. 21. Optimal design complexity based operational Scheduling

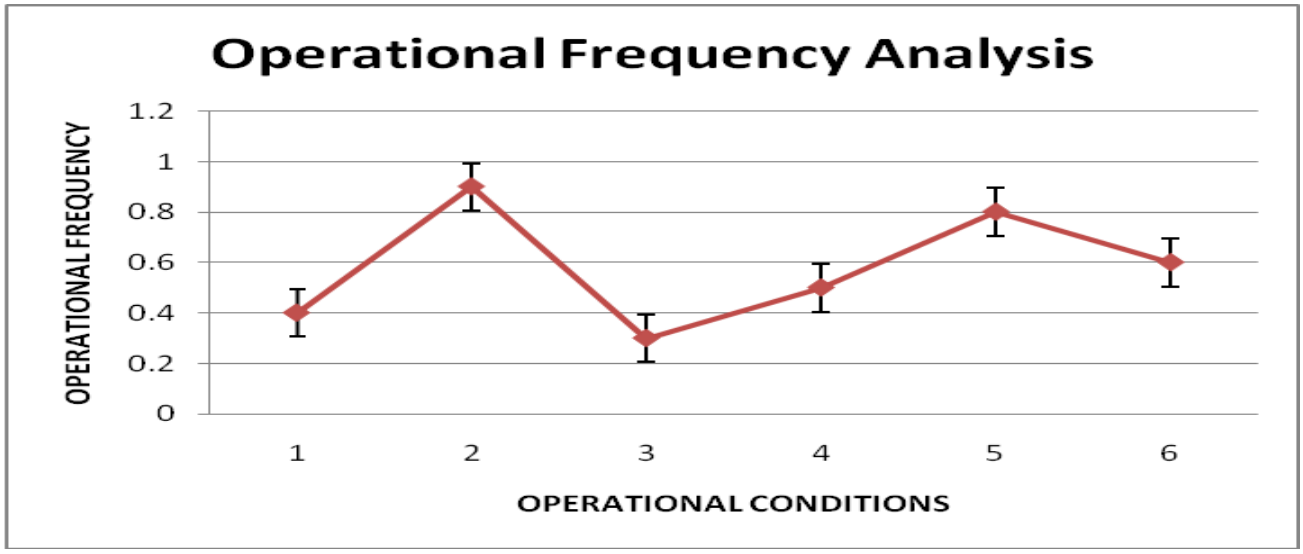


Fig. 22. Operational optimal frequency analysis

V. ASIP Behavior Optimization in Embedded System

Memory optimization techniques and performance area is determined by standards benchmark application. An application specific memory simulation analyzed by various simulators such as trace driven, cheetah, cache, ARM DS-5 etc. The advantages of SRAM used in programming technology so designer reuse the chip during prototyping and a system can be manufactured using in system programming.

In co-design technology effective memory performance area is analyzed by various simulators. The Co-design technology of ASIP used hardware and software implementation designs system to achieve an effective performance in the form of cycle count, low power consumption, latency and execution time [Fig. 23]. The source code profiling approach easily understands the application to guide the ASIP design methodology.

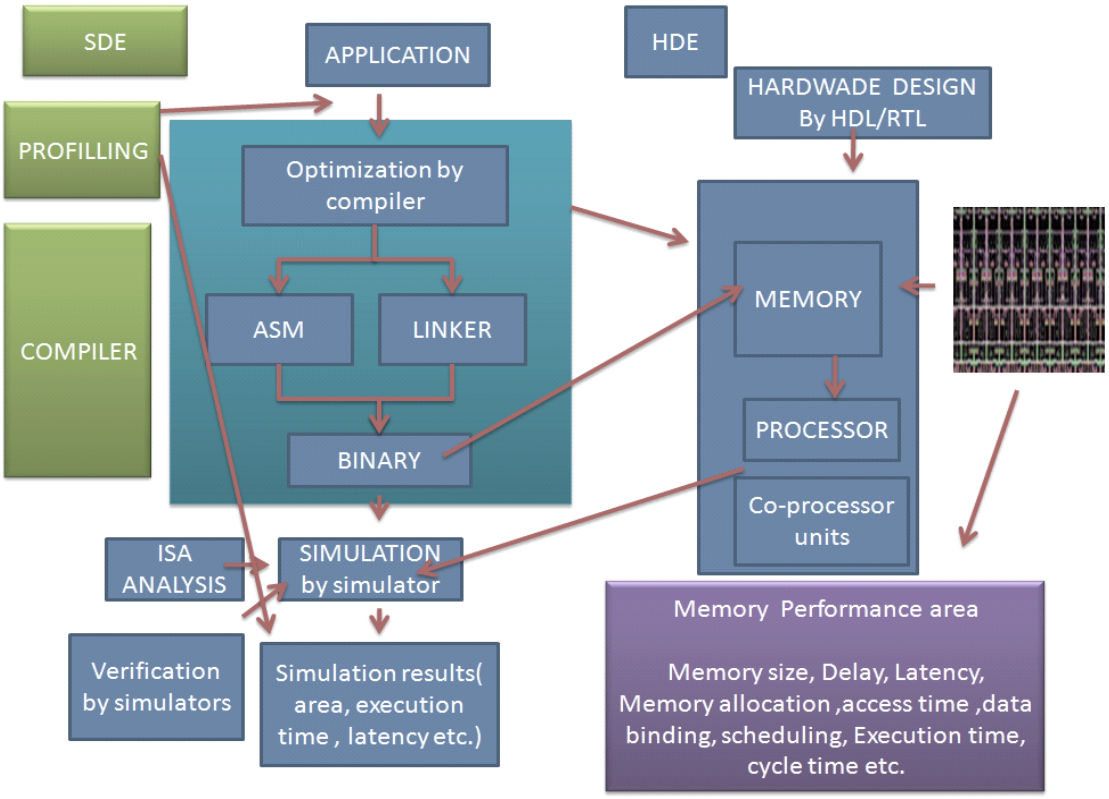


Fig. 23. Application specific Memory Integration and Performance Area

A. Application specific profiling and compilation Overview

Profiler have used to analyze the target source programs by collecting information on their execution based due to their data granularity scheme [10]. Profiler implements Pre-allocation of memory architecture and implements the execution time of application. A memory profiler used which implements dynamic profiling techniques to generate memory traces [10]. Memory object is computed load/store information for ASIP design mechanism. Micro-profiling approach also fills the gap between source level and instruction level profiler and implements speed and accuracy for ASIP design system [11]. LANCE [15] is mainly intended to facilitate C compiler design for embedded processors, so as to eliminate the need for time- consuming assembly programming [Fig. 24].Figure 24 shows the basic framework of LANCE profiler overview which requires profiling library, source code and instrumented binary file for profiling. Embedded processors for which LANCE based C compilers have been successfully built include both RISCs and DSPs design. The implementation of edge profiling, path profiling methods combines profiles with in the Low Level Virtual Machine [16] compiler infrastructure [Fig. 25]. A Codelets EXTRACTOR and RE player implements the code isolation. Codelet is basically designed for implementing, compiled, run and measure independently for the original application. The ISA design require an effectively for a fine grained profiling mechanism is based on C compiler mechanism.

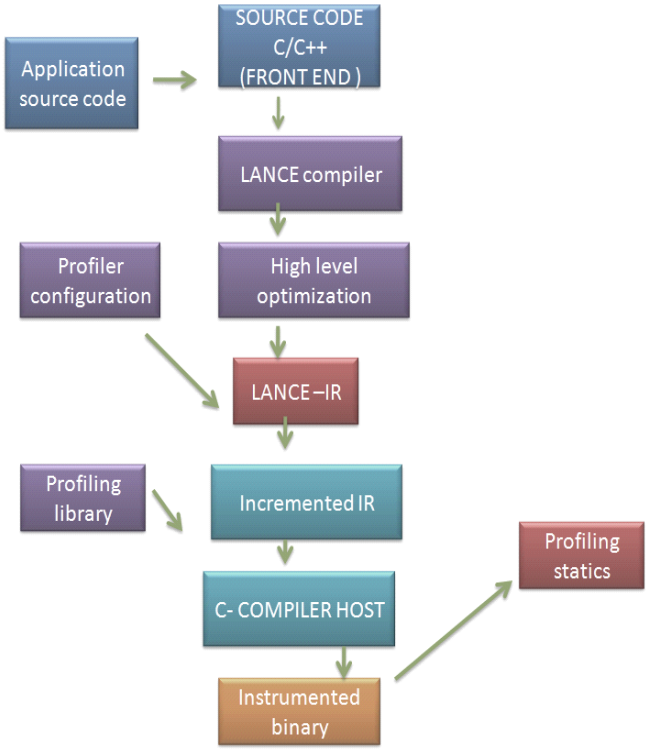


Fig. 24. LANCE Profiler in ASIP Embedded system [16]

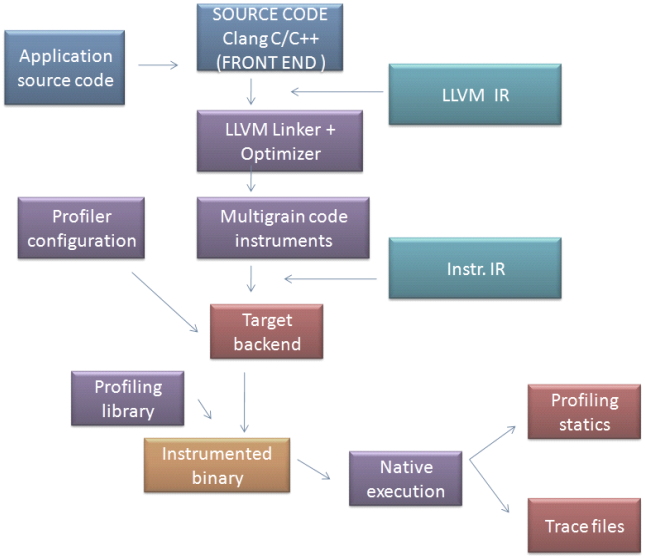


Fig. 25. LLVM based profiling analysis [17]

B. Application specific Latency prediction

Recently high level synthesis design is used efficient latency prediction techniques which implements the applications specific system performances and latency prediction design also used in clock cycle reduction mechanism or operational serialization. The number of time unit's clock cycles between initiations of stage is the latency between them. A latency of k means that the initiation are separated by k clock cycles. Any attempt two or more initiations to use the same stages at the same time they will cause a collision and collision must be avoided by scheduling a sequence initiations stages. In state diagram mechanism we have analyzed the function x from the initial stage (101010101110), only five outgoing transition are possible, corresponding to the five permissible latencies 10,8,6,4 and 1 in the initial collision vector. Similarly Free State (10101011), one reaches the same state offer three, five or seven shifts.

When condition is n+1 or greater, all the data transitions are redirected back to the initial states. A Collision can be implemented them by greedy cycles. Greedy cycles from the state diagram we can determine optimal latency cycles which result in the MAL. There are infinitely many latencies cycles, one can from state diagram, suppose that (1,12),(1,4,6,8,10,12),(4,6),(4,6,8)…… are legitimate cycles traced from the state diagram. As simple cycles are latency cycles in which each state appear only ones. Only (4),(6),(8),(6,8),(10,12) are simple cycles the cycles (6,12,10,12) are a complex cycle because of its travels these the states (101010101110) twice or more. Similarly (4,6,4,6,8,6) is not simple it repeats the state so we need greedy cycles is one whose edge are all made with minimum latencies from their respective starting states. The greedy cycles (1, 12) average latency is 6.5, which is lower than that of the simple cycle (10, 12) is 11[Fig. 26].

Greedy cycles have a constant latency which is equal the MAL (minimal average latencies points) for evaluating function X without causing collision the collision free scheduling approaches is thus reduced to finding greedy cycles from the sets of simple cycles. The greedy cycles yielding the MAL are the suitable choice for performance improvements. A latency sequence is a sequence of permissible forbidden latencies between the successive task initiations. A latency cycle is a latency sequence which repeats the same sequence indefinitely. Repeating of the cycles that reduces the collision between them and used the average latency that reduces the collision. Constant cycles contain is the latency cycles which contain only are latency value. The average latency cycles of a constant cycle are simple the latency itself. The target machine [RISC, CISC, and VLIW] can deploy more sophisticated instructions, which can have the capability to perform specific operations much efficiently.

If the target code can accommodate those instructions directly, that will not only improve the quality of code, but also yield more efficient results [Fig. 27]. Fixed point based latency optimal frequency optimized according various mechanism such as delay point and optimal operational frequency prediction mechanism. Operational serialize means how application computation complete the task with the least waste of time or least waste of hardware resources. Optimal condition is required to serialize the computational operations so resource reducible or operational optimal condition implements the latency design for ASIP system.

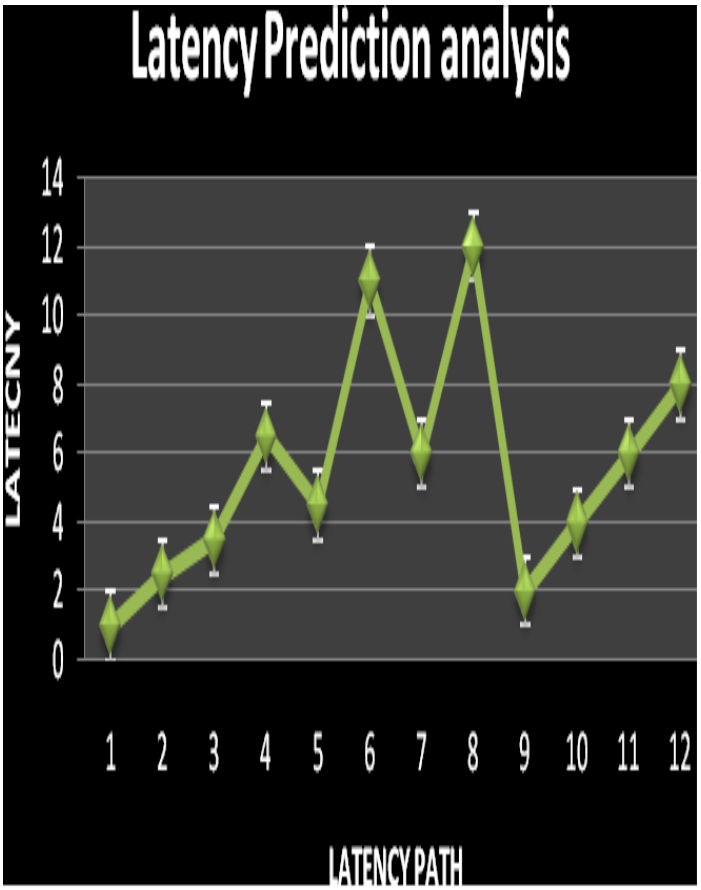


Fig. 26. Greedy cycle based LATENCY prediction analysis

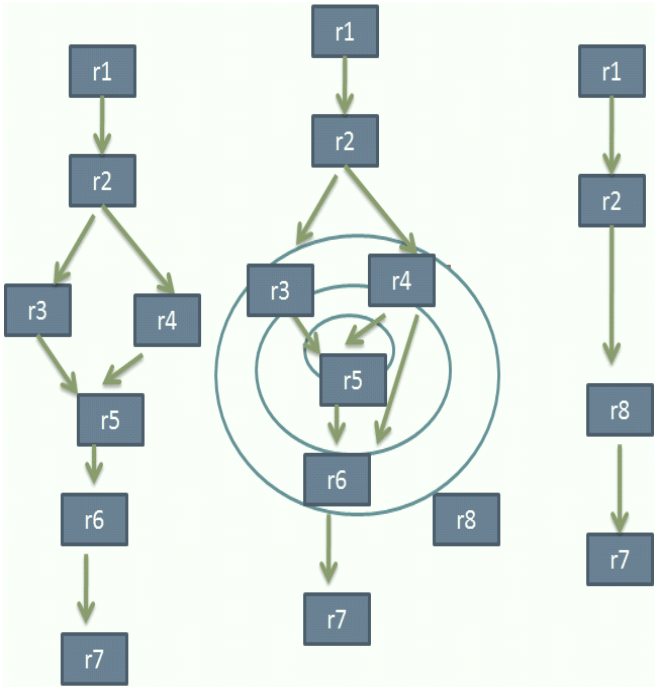


Fig. 27. Memory area implemented with resources Reducible flow mechanism and operational optimal condition

VI. Conclusion

Recently ASIP in our embedded system provides the benefits of flexibility and achieving excellent performances with low power consumption and ASIP also improves the functionality and design complexity with retargateable compiler technology. In real time embedded system designer implements the processor and memory architecture according to our application specific operational probability. ASIP system used the target machine can deploy more sophisticated instructions, which can have the capability to perform specific operations much efficiently for low power embedded system. Compilers and profiling mechanisms are also analyzed for ASIP and implements memory area reduction technique which improve the application execution performance. An effective cycle time, delay and scheduling prediction mechanism is used for memory implementation. An Efficient latency prediction technique is designed for operational serialization with the help of profiler and application specific computational complexity analyzed according to profiling execution delay time which is used in various high performance embedded devices.

References

[1] P. R. Panda, Nikhil D. Dutt "Data Memory Organization and Optimization in Application Specific Systems," In *Proceedings of the IEEE design & tests of Computers,(2001).*pp.56-58.

[2] M. K. Jain, M. Balakrishnan and A. Kumar "Integrated on-chip storage evaluation in ASIP synthesis," In *Proceedings of the 18th International conference on VLSI design (2005).*pp. 274-279. DOI**:** http://dx.doi.org/10.1109/ICVD.2005.112

[3] P. Meloni, S.Pomata, G. Tuveri, S. Secchi. L. Raffo, M. Lindwer. "Enabling fast ASIP design space exploration: An FPGA based runtime reconfigurable prototype," Hindawi Publication Cooperation, J. VLSI design (2012).

[4] Z. Prikryl, J.Kroustek, T. Hruska, D. Kolar. "Fast just in time translated simulator for ASIP," IEEE 14TH International symposium on Design and

diagnostics of electronic circuits and system (DDECS), 2011, pp.279-282.

[5] P. Meloni., S. Pomata, L. Raffo, M. Lindwer "Combining on-hardware prototyping and high level simulation for DSE of multi-ASIP system," IEEE Embedded Computer Systems (SAMOS), 2012, pp.310- 317.

[6] L. T. Clark, E. J. Hoffman, J. Miller, M. Biyani, Y. Liao, S. Strazdus, M. Morrow, K.E. Velarde and M. A. Yarch. "An Embedded 32-b Microprocessors core for low-power and high performance applications," IEEE J. of Solid-State Circuits 36(11), 2001, pp.1599-1608.

[7] E. Diken, R. Jordans, H. Corporaal "Build master: efficient ASIP architecture exploration through compilation and simulation result caching," IEEE 17th International Symposium on Design and Diagnostics of Electronic Circuits & Systems, 2014, pp. 83-88.

[8] E. Diken, R.Jordans. R. Corvino, H. Corporaal, F.A. Chies. 2014. "Construction and exploration of VLIW ASIPs with Heterogeneous vector-width," J. of Microprocessor and Microsystem. 2014. Vol.18, no. 8,pp.947-959.

[9] M. F. Jacome, G. D. Veeciana "Lower bound on latency of VLIW ASIP data paths," *International conference on computer aided design.IEEE*.1999, pp. 261-269.

[10] K. Karuri, R. Leupers, "Fine grained application source code profiling for ASIP design," *In the proceeding of 42nd design automation conferences.2005*.pp.329-334.

[11] X. Li, W. Zhou, D. Liu. "Application source code profiling for ASIP memory subsystem design," Procedia engineering, 2012, vol. 29, pp..3160-3164.

[12] A. Hoffmann. T. Kogel, A. Nohl. S. O. Brarun,O. Wahlen, A. Weiferink, and H. Meryr "A Novel Methodology for the Design of Application Specific Instruction-Set Processor Using a Machine Description Language". IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems (TCAD*)*, 20, 11 2001. pp.1338-1354.

[13] Z. Prikryl "Fast simulation of pipeline in ASIP simulator," IEEE International 14th workshop on microprocessor test and verification, 2014.pp.10-15.

[14] H. M. Hassan**,** K. Mohammed and A.F. Shalash. "Implementation of a reconfigurable ASIP for high throughput low power DFT/DCT/FIR engine," EURASIP J. on Embedded Systems, 2012.

[15] R. Leupers. "LANCE: A C Compiler Platform for Embedded Processors. Embedded System/Embedded Intelligence," Feb 2001.

[16] P.D. O. Castro, C. Akel, E. Petit, M. Popov, W. Jalby "CERE: LLVM based Codelet Extarctor and REplayer for Piecewise Benchmarking and Optimization," J. ACM Transactions on Architecture and Code Optimization (TACO), 2015. DOI**:** http://dx.doi.org/10.1145/2724717

[17] A. Mathur, M. Fujita, E. Clarke and P. Urard "Functional equivalence verification tools in High level synthesis flows," IEEE design and test of computer,vol. 26,no. 4 2009,pp. 88-95.

[18] P. J. Pingree, L. J. Scharenbroich, T. A. Werne and C. Hartzell "Implementing legacy-C algorithm in FPGA co-processors for performance accelerated smart payloads," *In proceeding of Aerospace conference*, 2008, pp.1-8.

[19] A. Putnam, S. Eggers, D. Bennett, E. Dellinger, J. Mason, H. Styles, P. Sundararajan and R.Witting "Performance and power of cache-based reconfigurable computing," *In Proceeding of ISCA' 09,* 2009. pp.395-405.

[20] J. Zhang, Z. Zhang., S. Zhou, M. Tan, X. Liu, X. Cheng, and J. Con.. Bit-level optimization for high level synthesis and FPGA-based acceleration. *In proceeding of FPGA'10,*pp. 59-68.

MAS based on a Fast and Robust FCM Algorithm for MR Brain Image Segmentation

Hanane Barrah
Laboratory of Innovative Technologies
National School of Applied Sciences
Tangier, Morocco

Abdeljabbar Cherkaoui
Laboratory of Innovative Technologies
National School of Applied Sciences
Tangier, Morocco

Driss Sarsri
Laboratory of Innovative Technologies
National School of Applied Sciences
Tangier, Morocco

***Abstract*—In the aim of providing sophisticated applications and getting benefits from the advantageous properties of agents, designing agent-based and multi-agent systems has become an important issue that received further consideration from many application domains. Towards the same goal, this work gathered three different research fields; image segmentation, fuzzy clustering and multi-agent systems (MAS); and furnished a MAS for MR brain image segmentation that is based on a fast and robust FCM (FRFCM) algorithm. The proposed MAS was tested, as well as the sequential version of the FRFCM algorithm and the standard FCM, on simulated and real normal brains. The experimental results were valuable in both segmentation accuracy and running times point of views.**

Keywords—agents; MAS; FCM; c-means algorithm; MRI images; image segmentation

I. INTRODUCTION

Agent technology has received a significant consideration from many application areas such as computer science, industry and medicine. Thus, developing theories and methods of designing agent-based and multi-agent systems has been adopted by several researchers. in fact, a variety of work has been published in this context. An overview of research and a historical context to the field were presented by N. R. Jennings et al. [1]. The authors concentrated on the interactions (cooperation, coordination and negotiation) within a MAS. Furthermore, they listed the first wave of agent based applications (industrial, commercial, entertainment and medical applications).

In their publication [2], M. Wooldridge and N. R. Jennings pointed out the three main elements needed to design and implement intelligent agents (theories, architectures and languages). They also examined some of the potential applications developed before 1995 (Air traffic control, patient care, believable agents ...).

In 1997, S. Franklin and A. Graesser [3] proposed an agent definition based on the autonomy concept to distinguish software agents from computer programs. Thus, they consider that all software agents are programs and the opposite is not true. The authors fostered their work by discussing two important points. The first one was about the agents classification, where they gave a natural classification of autonomous agents. The second point was an explanation of subagents and societies of agents.

In their review of industrial deployment of MAS, M. Pechoucek and V. Marik [4] presented a detailed list of potential applications of MASs associated to logistics, manufacturing control, production planning, space exploration and other application domains. In spite of the diversity of the applications depicted, the authors concluded that there is a gap between fundamental researchers and industrial users of agent technology.

F. Stonedahl et al. implemented the framework MAgICS as a coherent introductory computer science curriculum based on agent-based model (ABM) and MASs [5]. This framework is composed of nine models spanning seven computer science topics. The authors consider their framework as a starting point for future researches about reinventing introductory computer science education focused on MASs.

In the aim of providing agents able to bilaterally negotiate joint plans with humans, A. Fabregues and C. Sierra [6] proposed a modular software architecture based on an innovative search&negotiation method and which includes the BDI model (belief, desire and intention). This architecture is able to be extended by incorporating new components, which helps to build skillful agents.

As has been reported previously in literature. The agents and MASs' technology has had a wide range of application domains, this is owing to the beneficial properties of agents such as autonomy, social ability and reactivity [7]. Thus, this work used this recent technology as a solution to the MRI (Magnetic Resonance Imaging) image segmentation problem, which also has been in the center of interest of many researchers for many years. In fact, a wealth of methods have been developed to segment the MRI images [8]. To extract brain tumors, Eman Abdel-Maksoud et al. [9] integrated the k-means algorithm with its fuzzy version c-means [10], [11]; in order to get benefits from their advantages; and used the median filter as a pre and post-processing to remove noise. As a pre-processing method, the median filter presents two main problems. The first one is increasing the computational time, while the second one lies on the loss of some fine details [12] and which alters the clustering quality in a negative way. To get over this latter limitation and increase the efficiency of the c-means algorithm in presence of noise, several researchers improved it in many ways. the majority tried to include the filtering step in the clustering process by integrating spatial

information, while the rest tried to modify the dissimilarity measure [13]–[16].

This paper is mainly concentrated on designing a sophisticated MAS for MR brain image segmentation based on a fast and robust FCM algorithm (FRFCM) that includes the median filter into the clustering process. The main idea here, is to utilize the collective work of agents in order to segment the hole brain slices more accurately in a reasonable time.

The remainder of this paper is organized as follows: In section 2 are presented the key concepts of the agents and MASs' field. The FRFCM algorithm and the overall architecture of the proposed MAS are described in section 3. Section 4 is dedicated for some experimental results. A conclusion is presented in section 5.

II. Multi-Agent Systems : MAS

Given the degree of interest and the level of activity of the field of agents and MASs, in this section are presented the key concepts needed to design a convenient architecture for MR brain image segmentation.

A. Agents and Multi-agent Systems

According to various publications in the field of agents and MASs, there is no unique definition of an agent. Although, this lack of definition has not been a bottleneck in the development of this field. In this work, the definition of M. Wooldridge and N. R. Jennings [7] is adopted. Based on this definition, a MAS can be defined as a network of autonomous agents that interact with each other and with their environment to achieve a common goal.

The autonomy is a key concept that has been mentioned, explicitly or implicitly, by several definitions. It is the property that distinguishes an agent from an ordinary program.

B. Agent Architectures

Architecture is the organization of different elements within a system (agents and environment) and their relationships. Agent architectures is one of the key issues in this field. Indeed, the architecture design has become an interesting research subject of several researchers from different application fields. In literature, many agent architectures have been proposed, they can be roughly categorized into three types: *Reactive*, *deliberative* and *hybrid* architectures.

- Deliberative architectures. Are called also intelligent architectures. They contain a basic knowledge about the environment and can make decisions using a logical reasoning. The best-known deliberative architecture is the BDI architecture [17].
- Reactive architectures. In contrast to deliberative architectures, the reactive ones don't have any basic knowledge of the environment and they don't use any complex symbolic reasoning. In such architectures, the intelligent is seen as the result of interactions between the environment elements. The well-known reactive architecture is subsumption architecture, it was developed by Rodney Brooks in 1985 [18].
- Hybrid architectures. in order to get the best benefits of both deliberative and reactive architectures and to design more complex and sophisticated architectures, the hybrid architectures are more suitable. In fact, they try to combine the best aspects of the above architectures [1]. These architectures are also called layered architectures, because a such architecture may contain two (or more) layers: a deliberative one and a reactive one. The Touring Machines hybrid agent architecture is one of the best-known examples of these architectures, it was developed by Ferguson in 1992 [19].

C. Interactions within a MAS

In order to meet the purposes for which a MAS is designed, the agents must be able to interact with each other and with their environment. This interaction can be defined as an exchange of information between agents or between the environment and its agents. Generally, there are three main types of interactions: *cooperation*, *coordination* and *negotiation*.

- **Cooperation**. Corresponds to a collective work in order to achieve a common goal.
- **Coordination**. Aims to keep the coherence in the system, thus it seeks to organize the agent activities.
- **Negotiation.** Seeks to find an agreement that satisfied all the involved agents.

III. MAS based on a Fast and Robust FCM Algorithm

A. Fast and Robust FCM Algorithm: FRFCM

In the preceding section, an overview of the key concepts needed to design a convenient MAS is presented. In this section, we turn our attention to a different research area that is not less important than the agents and MASs' research field. It is the clustering problem. Indeed, the clustering is an essential step in several application domains such as data mining and image segmentation. It consists of grouping data into the most homogeneous groups as much as possible [20], [21]. In literature, several methods and techniques about the clustering problem have been developed [20], [21]. However, this work is mainly interested in the fuzzy clustering methods. The best-known fuzzy clustering algorithm is *c-means* [22], it creates fuzzy clusters by minimizing iteratively an objective function. The major drawback of this algorithm lies on the lack of any spatial information or constraints, which makes it sensitive to noise. In order to overcome this problem and faster the segmentation process, we proposed ; in an earlier work [23]; a fast and robust FCM (FRFCM) algorithm that is a combination of two powerful extensions of the standard FCM algorithm [24], [25].

The FRFCM algorithm takes as input the dataset $D = \{x_j \in \mathbb{R}\}_{j=1,\ldots,N}$, the number of clusters C, and minimizes iteratively the following objective function:

$$J(D,U,C)=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}.\left\|x_j-c_i\right\|^2+\alpha\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}.\left\|\bar{x}_j-c_i\right\|^2 \quad (1)$$

$\|.\|$ is the Euclidean distance, m is the fuzziness exponent and $U=[u_{ij}]$ is the fuzzy partition matrix that satisfies the following condition:

$$\left\{u_{ij}\in[0,1]\,\middle|\,\sum_{i=1}^{c}u_{ij}=1,\forall j \text{ and } 0<\sum_{j=1}^{N}u_{ij}<N,\forall i\right\}.$$

$\bar{x}_j$ is the median value of the neighbors within a specified window around $\bar{x}_j$.The parameter α controls the tradeoff between noise elimination and detail preserving. The minimization of the objective function presented in Eq. 1 is carried out by updating iteratively the fuzzy partition matrix and the cluster centers as follows:

$$u_{ij}=\frac{\left(\left\|x_j-c_i\right\|^2+\alpha\left\|\bar{x}_j-c_i\right\|^2\right)^{-\frac{1}{(m-1)}}}{\sum_{k=1}^{C}\left(\left\|x_j-c_k\right\|^2+\alpha\left\|\bar{x}_j-c_k\right\|^2\right)^{-\frac{1}{(m-1)}}} \quad (2)$$

$$c_i=\frac{\sum_{j=1}^{N}u_{ij}^{m}\left(x_j+\alpha\bar{x}_j\right)}{(1+\alpha)\sum_{j=1}^{N}u_{ij}^{m}} \quad (3)$$

To speed up the clustering process, this work took advantage from the suppressed version [24] of the standard FCM algorithm. The main idea behind this algorithm latter is prizing the biggest memberships and suppressing the others.

Let x_j be a pixel and u_{bj} be its degree of belongingness to the b^{th} cluster. If u_{bj} is the biggest value of all the clusters, then the membership degrees of x_j will be modified as follows:

$$u_{bj}^{*}=1-\gamma\sum_{i\neq b}u_{ij}=1-\gamma+\gamma u_{bj} \quad (4)$$

$$u_{ij}^{*}=\gamma u_{ij},\ i\neq b \quad (5)$$

Where $\gamma\in[0,1]$.

When γ gets closer to 0 (to 1 respectively), the algorithm becomes more hard (fuzzy respectively). Thus, the parameter γ balances between the fastness of the hard clustering and the good quality of the fuzzy clustering, in other words, a better selection of γ leads to a better clustering quality in a reasonable amount of time. It is always chosen equal to 0.5.

This latest modification has to be done immediately after updating the fuzzy partition matrix in order to bias the membership values to converge rapidly.

Algorithm steps

Step 0. Fix the clustering parameters (the converging error ε, the fuzziness exponent m and the number of clusters C) and initialize the clusters centers and the new parameters α and γ.

Step 1. Compute the median filtered image.

Step 2. Update the partition matrix using (Eq. 2).

Step 3. Modify the partition matrix using Eq. 4 and Eq. 5.

Step 4. Update the clusters centers using (Eq. 3).

Repeat steps 2-4 until the following criterion is satisfied:

$$\left\|C_{new}-C_{old}\right\|<\varepsilon$$

B. Agent Based FRFCM Algorithm

To segment a single slice of the MR brain, the FRFCM algorithm is sufficient, where it furnishes good results in a reasonable amount of time, this is owing to the used spatial information and the balance between hard and fuzzy clustering. To segment the hole slices of an MR brain, the FRFCM algorithm performs slowly because of the big size of data. To get over this problem and provide a sophisticated segmentation system, a multi-agent system based on the agents cooperative work to achieve the global segmentation of the brain is proposed.

It is known that the horizontal brain slices are roughly symmetrical in shape as well as in matter, where if a horizontal slice of a normal brain is split into four equal parts, each part contains necessarily the three brain tissues: white matter, gray matter and cerebrospinal fluid. This fruitful information is the origin of our idea. In fact, we proposed splitting each slice into four equal parts and segment them separately and in a parallel fashion, which can drastically reduce the processing time.

The proposed multi-agent system (Figure 1) takes as input a series of MRI images and return through its output the correspondent segmented images. It is composed of five agents; one master agent that takes control of the system and four slave agents that perform partial segmentations; sharing a memory space and communicating via exchanging ACL (Agent Communication Language) messages; in order to achieve the global segmentation.

The master agent is the first agent created in the platform. When it is created, it gets the data (brain slices) and makes it ready to be used in the shared memory, initializes the

clustering parameters and creates the slave agents. When all the clustering conditions are verified, the master agent notifies the slave agents in order to start their tasks and waits for them to reply.

Since a slave agent received the master's notification message, it performs the clustering process; using the FRFCM algorithm presented in the previous subsection. When it finishes segmenting its specific part of a given slice, it proceeds automatically to the next slice without waiting for the other agents to finish and when it is done; there is no more slices to process; it notifies the master agent that its mission is accomplished successfully.

When the master agent receives all the slave agents replies, it gathers their partial results in order to form the global one.

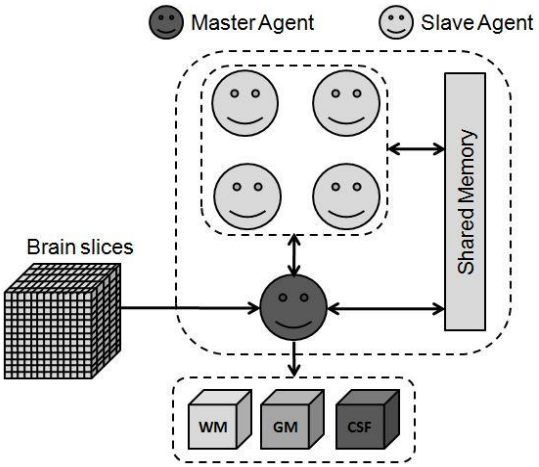


Fig. 1. MAS architecture

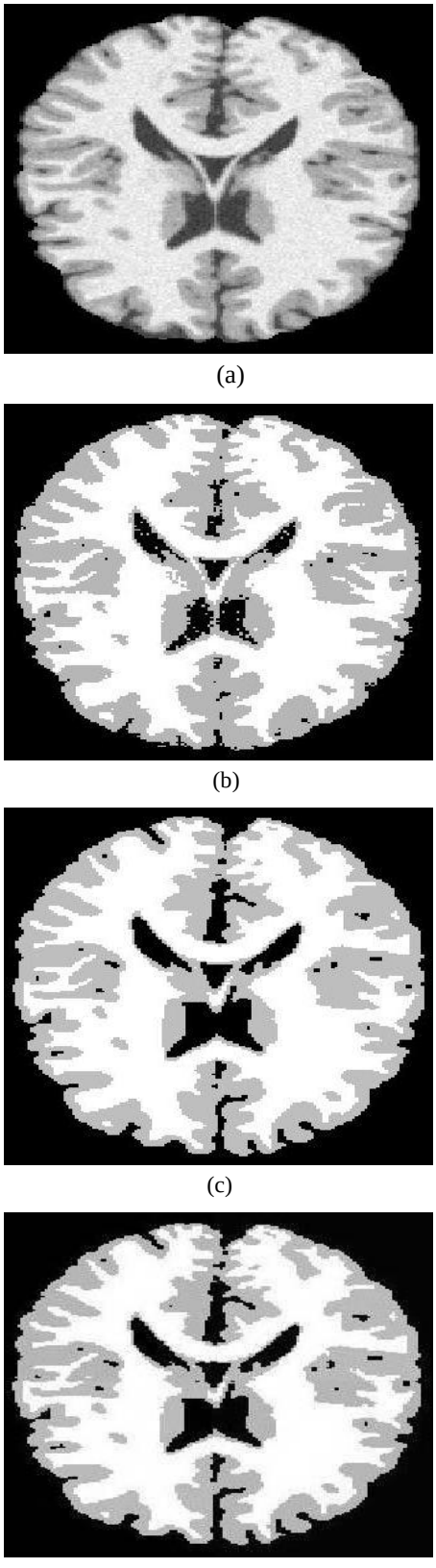
For simplification, the proposed MAS is going to be noted in the rest of this paper as FRFCM_MAS.

IV. Experimental results

In this section, the proposed MAS is compared with the sequential version of the FRFCM algorithm and the standard FCM in the segmentation accuracy and running times standpoints. In fact, the three approaches were tested on simulated [27] and real [28] normal brains. All the experiments were performed on an Intel Core i7 (4.4 GHz) machine. The proposed MAS was implemented on the JADE middleware [26]. And the clustering parameters were fixed as follows: $C = 3, \varepsilon = 10^{-8}$, $m = 2$, $\alpha = 4$ and $\gamma = 0.5$.

The parameter α is determined by experience; by seeking the interval $[1\ 10]$; for $\alpha \geq 4$ the changes of the segmentation results are negligible.

As FRFCM and FRFCM_MAS minimize the same objective function, their segmentation results are necessarily very close. To verify this conclusion and show the efficiency of the FRFCM_MAS system over the standard FCM algorithm, we run the three methods on the 90th horizontal slice of a normal brain [27] simulated with 3% of noise. The segmentation results are presented in Figure 2 and Table 1.



(a) (b) (c) (d)

Fig. 2. Segmentation results. (a) Original image. (b) FCM result. (c) FRFCM result. (d) FRFCM_MAS result

TABLE I. CLUSTERING RESULTS ON THE 90TH SLICE

Method	Misclassification Errors (%)	Running Times (s)
FCM	22.43	0.79
FRFCM	19.51	0.57
FRFCM_MAS	19.54	0. 342

From Figure 2, we notice that the FRFCM and the FRFCM_MAS outperformed the standard FCM algorithm; where they succeeded to some extent to handle noise and extract the most homogeneous regions, and their results are very close, which is also confirmed by the numerical results depicted in Table 1. In fact, the misclassification rate generated by the standard FCM is the biggest one, while the difference between those generated by the FRFCM and FRFCM_MAS is very small, this difference is due to the random initialization. Moreover, from Table 1, it is remarkable that the proposed MAS performed faster. Thus, the FRFCM_MAS combined between the robustness to noise and the fastness.

To show the strength of the proposed MAS against the sequential version of the FRFCM algorithm, seven experiments were performed on a real normal brain [28], in each experiment the number of the input images is increased. The results are summarized in Figure 3.

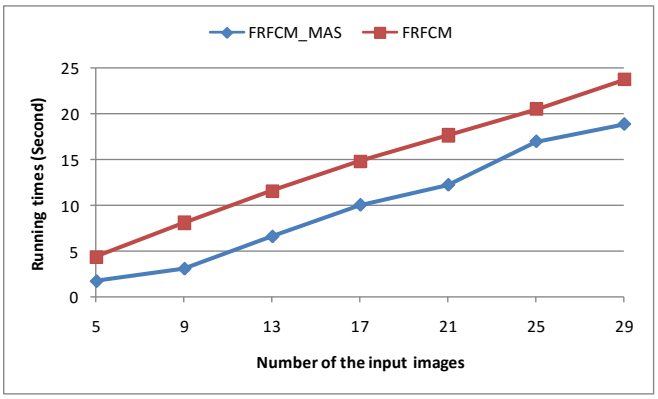


Fig. 3. The running times consumed by both our MAS and the sequential algorithm

From Figure 3, we note that as the number of the input images increases, both methods require much more time. In addition, all the running times performed by the FRFCM_MAS are smaller than those performed by the sequential version of the FRFCM. Which proves the fastness of the proposed method over the sequential version of the FRFCM algorithm.

V. CONCLUSION

In this work, the key concepts of the MASs' field have been used along with a fast and robust fuzzy clustering algorithm to design a MAS for segmenting the MR brain images. By testing this MAS as well as the sequential versions of the FRFCM and the FCM on simulated and real normal brains, it showed its robustness and fastness. For this special thanks, the future work will be focused on improving the FRFCM_MAS system in order to include other image processing techniques such as tumor extraction and volume estimation.

ACKNOWLEDGMENT

The authors would like to thank the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN) for making the MRI brain data sets freely available under the following grant numbers: P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

REFERENCES

[1] N. R. Jennings, K. Sycara, and M. Wooldridge, "A Roadmap of Agent Research and Development," Auton. Agents Multi-Agent Syst., vol. 1, no. 1, pp. 7–38, Jan. 1998.

[2] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," Knowl. Eng. Rev., vol. 10, no. 02, pp. 115–152, 1995.

[3] S. Franklin and A. Graesser, "Is It an agent, or just a program?: A taxonomy for autonomous agents," in Intelligent Agents III Agent Theories, Architectures, and Languages, J. P. Müller, M. J. Wooldridge, and N. R. Jennings, Eds. Springer Berlin Heidelberg, pp. 21–35, 1997.

[4] M. Pechoucek and V. Marik, "Review of industrial deployment of multi-agent systems," Gerstner Lab. Agent Technol. Group Dep. Cybern. Czech Tech. Univ. Prague Czech Repub. Rockwell Autom. Res. Cent. Prague Czech Repub., 2006.

[5] Stonedahl, Forrest, Michelle Wilkerson-Jerde, and Uri Wilensky. "MAgICS: Toward a Multi-Agent Introduction to Computer Science." Multi-Agent Systems for Education and Interactive Entertainment: Design, Use and Experience: Design, Use and Experience, pp. 1-25, 2010.

[6] A. Fabregues and C. Sierra, "HANA: A Human-Aware Negotiation Architecture," Decis. Support Syst., vol. 60, pp. 18–28, Apr. 2014.

[7] M. J. Wooldridge, "The logical modelling of computational multi-agent systems," Citeseer, 1992.

[8] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation 1," Annu. Rev. Biomed. Eng., vol. 2, no. 1, pp. 315–337, 2000.

[9] E. Abdel-Maksoud, M. Elmogy, and R. Al-Awadi, "Brain tumor segmentation based on a hybrid clustering technique," Egypt. Inform. J., vol. 16, no. 1, pp. 71–81, Mar. 2015.

[10] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Comput. Geosci., vol. 10, no. 2, pp. 191–203, 1984.

[11] M. C. Clark, L. O. Hall, D. B. Goldgof, L. P. Clarke, R. P. Velthuizen, and M. S. Silbiger, "MRI segmentation using fuzzy clustering techniques," Eng. Med. Biol. Mag. IEEE, vol. 13, no. 5, pp. 730–742, 1994.

[12] A. B. Hamza, P. L. Luque-Escamilla, J. Martínez-Aroza, and R. Román-Roldán, "Removing noise and preserving details with relaxed median filters," J. Math. Imaging Vis., vol. 11, no. 2, pp. 161–177, 1999.

[13] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data," Med. Imaging IEEE Trans. On, vol. 21, no. 3, pp. 193–199, 2002.

[14] J. Wang, J. Kong, Y. Lu, M. Qi, and B. Zhang, "A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints," Comput. Med. Imaging Graph., vol. 32, no. 8, pp. 685–698, Dec. 2008.

[15] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," Pattern Recognit., vol. 40, no. 3, pp. 825–838, 2007.

[16] Z. Ji, J. Liu, G. Cao, Q. Sun, and Q. Chen, "Robust spatially constrained fuzzy c-means algorithm for brain MR image segmentation," Pattern Recognit., vol. 47, no. 7, pp. 2454–2466, Jul. 2014.

[17] A. S. Rao, M. P. Georgeff, and others, "BDI Agents: From Theory to Practice.," in ICMAS, 1995, vol. 95, pp. 312–319.

[18] R. A. Brooks, "How to build complete creatures rather than isolated cognitive simulators," Archit. Intell., pp. 225–239, 1991.

[19] I. A. Ferguson, "TouringMachines: An architecture for dynamic, rational, mobile agents," University of Cambridge UK, 1992.

[20] D. Lam and D. C. Wunsch, "Chapter 20 - Clustering," in Academic Press Library in Signal Processing, vol. Volume 1, J. A. K. S., Rama Chellappa and Sergios Theodoridis Paulo S.R. Diniz, Ed. Elsevier, pp. 1115–1149, 2014.

[21] L. Rokach and O. Maimon, "Clustering methods," in Data mining and knowledge discovery handbook, Springer, pp. 321–352, 2005.

[22] S. Ghosh and S. K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," IJACSA, vol. 4, pp. 35–38, 2013.

[23] A. Cherkaoui and H. Barrah, "Fast Robust Fuzzy Clustering Algorithm for Grayscale Image Segmentation," in Xth International Conference: on Integrated Design and Production, Dec 2015.

[24] J.-L. Fan, W.-Z. Zhen, and W.-X. Xie, "Suppressed fuzzy c-means clustering algorithm," Pattern Recognit. Lett., vol. 24, no. 9–10, pp. 1607–1612, Jun. 2003.

[25] S. Chen and D. Zhang, "Robust Image Segmentation Using FCM With Spatial Constraints Based on New Kernel-Induced Distance Measure," IEEE Trans. Syst. Man Cybern. Part B Cybern., vol. 34, no. 4, pp. 1907–1916, Aug. 2004.

[26] F. L. Bellifemine, G. Caire, D. Greenwood, and Wiley InterScience (Online service), Developing multi-agent systems with JADE. Chichester, England; Hoboken, NJ: John Wiley, 2007.

[27] Brainweb, Simulated Brain Database. http://www.bic.mni.mcgill.ca/brainweb/

[28] The Open Access Series of Imaging Studies (OASIS). http://www.oasis-brains.org

Development of the System to Support Tourists' Excursion Behavior using Augmented Reality

Jiawen ZHOU
Graduate School Student,
Graduate School of Information Systems,
University of Electro-Communications
Tokyo, Japan

Kayoko YAMAMOTO
Associate Professor,
Graduate School of Informatics and Engineering
University of Electro-Communications
Tokyo, Japan

***Abstract*—The purpose of this study is to develop an information system (AR recommended GIS) to support tourists' excursion behavior by making the accumulating, sharing, and recommending of information concerning urban tourist spots possible. The conclusion of this study can be summarized into three points as shown below. (1) In order to support tourists' excursion behaviors by integrating SNS, Twitter, Web-GIS, recommendation system, and Smart Eyeglass, in addition to making the accumulating, sharing, and recommending of information regarding urban tourist spots possible, the AR recommended GIS was designed and developed. (2) 91% were between the age of 20-40 among the 91 users, and the ultimate number of submitted information was 161. In addition, concerning the operation using Smart Eyeglass, which was conducted with tourists in the Minato Mirai area, the total number of users were 34, age of users were spread out, and all users had no experience in using Smart Eyeglasses. (3) From the results of the Web questionnaire survey, the system is compatible for the collection method of tourist spot information for users, and is mainly used to collect tourist spot information using the viewing and recommendation functions. From the results of the access analysis using the log data form during the operation, the utilization method of the system with PCs and mobile information terminals were very similar. Additionally, as the system using AR Smart Eyeglass was rated extremely highly, it was evident that it is possible to support tourists' excursion behavior using PCs, mobile information terminals, and AR Smart Eyeglasses are possible.**

Keywords*—Augmented Reality; Web-GIS; Social Media; Recommendation System; AR recommended GIS; Tourists' Excursion Behavior

I. INTRODUCTION

In recent years, regarding Japan, which has progressed in advanced information development, various information has been transmitted by means of the internet. Regarding the tourism field also, the internet transmits various information, and is the main information source for planning sightseeing tours and searching for information about tour destinations. However, because of the volume and variety of information, it has become difficult for users to appropriately select and obtain necessary information on their own. Especially for urban tourist spots, because the amount of information submitted and made public is great compared to that of regional tourist spots, and this makes it difficult for those with little knowledge or good sense of locality to efficiently obtain vital information for sightseeing, an information system to help users obtain the appropriate information is necessary.

Because small and multifunction mobile information terminals have become widespread in recent years, many can use this conveniently in their daily lives. As many urban tourist spots in Japan are mostly focused in specific areas, walking to more than one tourist spot has become mainstream. Although sightseeing plans may be changed flexibly when on foot, it is inconvenient, and can even be dangerous, to have to search for information about tourist spots on a personal assistance device, while walking on unfamiliar streets. On the other hand, as one application example of virtual reality (VR), augmented reality (AR) has recently been drawing attention, and practical application has been advancing. When using AR technology, unlike VR, regarding the reality space surrounding the user, information accumulated in virtual space can be presented. Although this technology has been used before with devices such as cellular phones, AR Smart Eyeglass and other distinctive exclusive terminals, which has been designed to be convenient, has been recently released. Accordingly, concerning sightseeing tours on foot in urban tourist spots, introduction of such AR technology in addition to existing mobile information terminals has also been anticipated.

The purpose of this study, which is based on the background as shown above, is to develop an information system (AR recommended GIS) to support tourists' excursion behavior by making the accumulating, sharing, and recommending of information concerning urban tourist spots possible. More specifically, Web-GIS, SNS, and the recommendation system will be integrated to develop a system appropriate for three information terminals including PC, mobile information terminal, and AR Smart Eyeglass, in order for the system to be available in various situations.

II. RELATED WORK

This study is related to 4 research areas including (1) research related to the sightseeing tour support system, (2) research related to recommendation methods of tourists spots, (3) research related to social media GIS, and (4) research related to AR. Concerning the (1) sightseeing tour support system, Kurata (2012) [1] developed an automatic sightseeing course making system using Web-GIS and genetic algorithm. Sasaki et al. (2013) [2] developed a system that collects information regarding regional resources, and supports the tours of each user. Fujitsuka et al. (2014) [3] developed an

outing plan recommendation system using a pattern mining method that lists and extracts time series of sightseeing tours. Ueda et al. (2015) [4] developed a system to create posterior information from the action of users while sightseeing, and to support tourists' excursion behavior by sharing the aforementioned information as prior information to other users.

(2) Concerning the research related to recommendation methods of tourists spots, Uehara et al. (2012) [5] extracted tourists spot information from the Web, and proposed a system that recommends sightseeing locations by evaluating the similarities between tourist spots by means of several feature vectors. Batetel et al. (2012) [6] proposed a sightseeing location recommendation system using the multi-agent system. Shaw et al. (2012) [7] developed a system that presents a list of nearby tourist spots to users based on the location information and visit history of the user. In addition, research related to the recommendation of Point-of-Interest (POI) among research related to LSBN (Location-Based Social Networks) is also applicable to this research field. As a representative example, Ye et al. (2011) [8] , Liu et al. (2013) [9], and Chen et al. (2016) [10] proposed a POI recommendation method which focuses on the user's individual attributes. Yuan et al. (2013) [11] proposed a POI recommendation method which takes into consideration time and space information using check-in data regarding LBSN.

(3) Concerning the research related to social media GIS, Yanagisawa et al. (2011) [12] in addition to Nakahara et al. (2012) [13] developed an information sharing GIS with the purpose of accumulating and sharing information regarding the local community using Web-GIS, SNS and Wiki. Yamada et al. (2013) [14] and Okuma et al. (2013) [15] developed a social media GIS which reinforced the functions of social media included in the information sharing GIS as mentioned above. By using the systems of these prior researches as a base, Murakoshi et al. (2014) [16] in addition to Yamamoto et al. (2015) [17] developed a social media GIS to support the utilization of disaster information with the assumption that it would be continually used from normal times to disaster outbreak Times. Additionally, with the social media GIS as a base, Ikeda et al. (2014) [18] developed a social recommendation GIS in order to accumulate tourist spot information and recommend it according to the taste of each user.

(4) Concerning the research related to AR, Fujimoto et al. (2014) [19] developed a navigation system within buildings by means of arrow marks using AR. Tosa et al. (2013) [20] developed a coupon use purchasing support system using AR.

However, the prior researches listed above did not propose a system that integrated the Web-GIS, SNS, recommendation system, and AR technology. In this study, by integrating the above and developing a new system, originality as a system is made evident. Additionally, the main purposes of the systems of prior researches are the accumulation and sharing of information, but lack functions that can recommend necessary information for the support for tourists' excursion behavior. In addition, with the systems of prior researches, because the access from mobile information terminals at tourist spots were not taken into consideration, real-time tourists' excursion behavior cannot be supported in a safe or effective way. Therefore, this study will develop a system to support tourists' excursion behavior that values real-time information recommendations by means of the integration of SNS, Web-GIS, and the recommendation system. Additionally, by making smooth access to the system from AR Smart Eyeglasses and other various information terminals in addition to mobile information terminals possible, making sightseeing plans beforehand or searching for information at a tourist spot while touring has also become possible, and being able to obtain real-time information of tourist spots so conveniently shows the usefulness of the system.

III. Outline and Method of this Study

This study will be conducted according to the outline and method as follows: First, the design (Section IV) and development (Section V) of the AR recommendation GIS specifically for the purpose of this study will be conducted independently. Next, with the assumption that users are over 18, the operation test and operation of the AR recommendation GIS (Section VI), in addition to evaluation and extraction of points needing improvement (Section VII) will be conducted. Here, it will be assumed that each user will use the system for a period of one month, and the operation will be conducted after the operation test and operation test evaluation is completed. In addition, the evaluation of the system will be based on the results of the Web questionnaires for users and an access analysis using the log data during the operation that will be conducted, which will make it possible to extract points needing improvements in order to help users enjoy an even more safe and successful tourists' excursion behavior.

The center of Yokohama City, Kanagawa Prefecture, was selected as the operation area. The first reason for this selection is because Yokohama is a popular urban tourist spot that has many tourists, there is an abundance of information submitted and made public, it is difficult for tourists to efficiently obtain necessary information, there are various types of tourist spots available, and a variety of tourist spots can be recommended to each user according to their taste using the system. The second reason is because walking is the main form of transportation for tourists' excursion behavior in this area. Therefore, it is appropriate to use the system for supporting those sightseeing on foot. The third reason is that there is information on a total of 200 tourist spots in the same area collected by Ikeda et al. (2014) [18], and by accumulating such information to the system in advance, the disadvantage many independent information systems have, which is not having enough data at the initial stage of operation, will be solved, and the operation of the system can be expected to be more effective.

IV. System Design

A. System configuration

The system proposed in this study, as shown in Fig. 1, is made up of the combination of three applications including Web-GIS, SNS, and the recommendation system. Specifically, the management and visualization of submitted information on digital maps of Web-GIS, constraint of users by means of individually structured SNS, and sharing and exchanging of information between limited users have become possible,

which enable users to submit, view and evaluate information, while grasping geographical information related to tourist spot information on digital maps. In addition, by combining recommendation systems, information that matches the tastes of each user from the accumulated and shared information can be offered by means of the digital map. Thus, even when the system is operated for a long term and the amount of accumulated information becomes massive, it is possible to direct each user to the appropriate information, and it is anticipated that it will support the efficient obtaining of tourist spot information.

Additionally, the system can be used with three information terminals including PC, mobile information terminal, and AR Smart Eyeglass. PCs are used for prior research, AR Smart Eyeglasses are used while sightseeing on foot, and mobile information terminals are used when temporarily resting. By making use of these three information terminals, the system can be used for information searches anytime from the planning stage prior to the sightseeing trip to the actual sightseeing stage, and it can be anticipated that it will effectively and safely support tourists' excursion behavior.

Therefore, the usefulness of the system stated in Section II can be explained in details as shown below.

1) Amelioration of constraints related to information searches

As one scenario of a situation where information searches become restricted, if there is an information overload when various information is submitted/transmitted, a situation where users find it difficult to select and obtain necessary information efficiently can be assumed. Therefore, to ameliorate the constraints related to such information searches, the recommendation system will be combined with the system. By doing so, among the large amount of information, it will enable users to be appropriately directed to tourist spot information according to their taste in a short period of time.

2) Amelioration of spatial constraints taking security into consideration

As one scenario of a situation where spatial constraints taking into consideration security occurs, a situation where users need information in various stages of sightseeing can be assumed. More specifically, although many use PCs for prior research before the actual sightseeing trip, there are many situations where tourist spot information may be constantly required while sightseeing. Especially when sightseeing on foot, it is inconvenient to obtain information using mobile information terminals and using such a device while walking can also be dangerous. In addition, if the information on the PC and the mobile information terminal is not synchronized, this may also cause an inconvenience. In order to ameliorate the spatial constraints while taking safety into consideration, in addition to an interface that can be used on both PCs and mobile information terminals, by introducing a AR Smart Eyeglass to support sightseeing on foot and an application dedicated to it, the system can be used in quiet situations where the user is resting or in an active situation where the user is sightseeing on foot. Thus, users will not have to rely on spatial constraints taking into consideration security while sightseeing, and by means of recommendations and viewing of appropriate tourist spot information, users can receive support for excursion behaviors.

3) Amelioration of constraints related to continuous operation

Without being restricted by time or location by means of (1) and (2) as shown above, in order to create a lasting environment where submitting, viewing, and recommending tourist spot information is possible, a design for a system that can manage submitted information is necessary. In addition, if the system is to be made so that anyone can join and there is no structure that can manage submitted information, when inappropriate information is submitted, there is a possibility that operation to serve the purpose may become difficult. However, while the system will have combined management of submitted information by means of database, because users with ill will will be found by managing accounts through SNS, this will enable the operation to be long term.

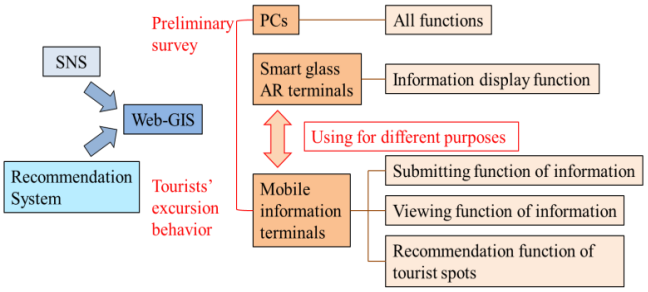


Fig. 1. System design of AR media GIS

B. Outline of system design

1) System structure

The social media GIS of this study will be developed by means of three servers including the Web server, database server, and GIS server. The Web server mainly handles the processing related to SNS and integrates each function by accessing the GIS server and the database server. SNS is implemented by JavaScript and PHP, while the recommendation system is implemented by PHP, and the database server is managed by MySQL, which accumulates collected information submitted online through SNS. The Web server and database server use the rental server from the information base center of the authors' university. The WindowsServer2008 of Microsoft is used as OS for the GIS server and ArcGISServer10.1of ESRI is used as a GIS server software. In addition, to enable the use with AR Smart Eyeglass, an Android application using the Smart Eyeglass SDK of Sony was developed.

2) Web-GIS

In this study, ArcGISServer10.1 of ESRI was used for Web-GIS and Mapple10000 of the Shobunsha Publication Mapple digital data SHAPE edition, including detailed road network data, was used for the GIS base map of the Web-GIS. The map that will be superposed with the digital map data is, among those offered by ESRI which is compatible with API of ArcGISServer10.1, the user interface of GoogleMaps, which was the most used in prior researches in relating fields to this study. Concerning the superposing of the Mapple10000 (SHAPE edition) and GoogleMaps, while GoogleMaps used

the new geodetic coordinates, Mapple10000 is based on the old geodetic coordinates. For this reason, using the ArcTky2Jgd, which was provided by ESRI as product support, Mapple10000 was changed to the new geodetic coordinates and changes were made in order to enable the input of information of operation areas using the ArcMap10.1.

3) SNS

In this study, an SNS will be selected as social media to be integrated with the Web-GIS and recommendation system, and will be individually designed and developed according to its purpose. This SNS, unlike other social media, is for the purpose of designing and developing the ideal system according to the purpose of use in addition to enabling detailed system development individually according to the characteristics of the operation area. In addition, by self-making SNS as mentioned in section IV-A, it will enable the information transmission of the system to have a two-way direction and to integrate with the recommendation system.

First, functions related to the registration and releasing of user information in addition to submitting, viewing, and recommending information were independently designed according to the purpose of this study. Next, because it is ideal that communication between users of the system be voluntary, without designing friend registration or community functions like other SNS, comment functions, evaluation function, and sub tag functions that can be updated freely were designed as a means of communication. The comment function will be used for communication between users and to provide supplementary information to existing submissions. The evaluation function will be used by users to evaluate tourist spots on a 5-grade scale. In this way, the average of evaluation points will be shown, and in addition to providing users with this information, it can also be used as a weighting of recommendations. Additionally, the evaluation function will have a ranking function listing tourist spots in the order of the average evaluation points. Concerning sub tag functions, users will be able to easily tag a tourist spot. Because the tag shows the characteristics of the tourist spot, it can be used as a reference when doing research prior to sightseeing. Tourist spots with the same tag can also be listed.

4) Recommendation system

The recommendation system has three methods including the collaboration recommendation, content-based recommendation, and knowledge-based recommendation (Jannach et al., 2012) [21]. The knowledge-based recommendation will be used for the system. The reason for this is that it can solve the cold-start problem, which is an issue concerning the difficulties of making appropriate recommendations for new users who have just started using the system, and recommending new items introduced to the system as recommendation subjects. Kamishima (2008) [22] pointed out that there will be no problem if, regarding the former person, it is a knowledge-based recommendation and the user directly writes their own user profile, and regarding the latter person, if there is an user profile with a content-based recommendation and knowledge-based recommendation, a recommendation will not be difficult with the help of the feature vectors even with new items.

In addition, because the system is for the general public, concerning the creating of user profiles based on the user preference information, a clear and intuitive question form is ideal. Therefore, question items must be made to be answered on a five grade scale in order to create a user profile vector. Similarly, with evaluation information of tourist spots, question items must be made to be answered on a five grade scale by contributors of new tourist spot information in order to create a feature vector of tourist spots. In this way, to deal with the cold-start problem mentioned above, the system will be set so that users must input evaluation information when submitting information on tourist spots, and regarding the operation manual distributed when promoting the use of the system, the input of evaluation information aside from explaining the main purpose was requested.

Based on the created user profiles and feature vectors of tourist spots, the similarity level was calculated by means of equation (1), and the tourist spots with high similarity levels will be recommended.

$$Sim_j = \frac{\sum_{i=1}^{n} U_i * S_{ij}}{\sqrt{\sum_{i=1}^{n}(U_i)^2} * \sqrt{\sum_{i=1}^{n}(S_{ij})^2}} \quad (1)$$

Sim_j: Degree of similarity

U_i: User preference information

S_{ij}: Tourist spot evaluation information

i: Question information number

j: Tourist spot number

5) Management of submitted information

As it was mentioned concerning the amelioration of constraints related to the continuous operation in Section IV-A, design for a system that can manage submitted information is required. Therefore, with the aim to keep the operation long term, although there will be no constraints for submitted information regarding all users when released, if it is determined by the manager that the submission was written by a user with ill will, or that the submitted information is not appropriate for the purpose of the system, the system design will allow the exercising of authority to delete an account or a submission. Specifically, a function that can enable combined management of submitted information by means of database will be installed.

V. System Development

A. System frontend

In this study, as shown in detail below, original functions will be implemented for users in addition to the accumulating, sharing, and recommending of tourist spot information.

1) Information submission function

When submitting sightseeing information, by clicking on the "submit tourist spot information" button on the top page, users will be directed to the submitting page. Items that can be submitted include, the "title", "main tag", "description", "evaluation points of the tourist spot", and the "location information". In addition, to provide more information of the tourist spot, users can add a sub tag and upload images. After inputting and selecting the required items, if users select the tourist spot by clicking on the digital map, the location will be

entered into MySQL and the submission will be complete after the user pushes the send button.

2) Information viewing function

On the viewing screen, tourist spots are divided into 8 categories according to the main tags, and by selecting a checkbox that each item has, the marker for tourist spots in the corresponding category can be listed on the map. By clicking the marker that the user is interested in, the title and image of the selected spot will appear in a bubble. By further clicking on the title, users will be directed to the detailed information viewing page of the selected tourist spot. On the detailed information viewing page, in addition to being able to view detailed information of tourist spots, users can evaluate and comment on the page. In this way, by enabling other users to comment on tourist spots, communication between users is promoted. Concerning the evaluation points, evaluation of tourist spots are converted to numerical values of a five grade scale, and by using the average evaluation point of multiple users, the evaluation and recommendations of tourist spots are weighted. In addition, if a tourist spot has sub tags, when users click the sub tag, they will be directed to a page where they can view other tourist spots with the same sub tag.

3) Information updating function

By clicking the information update button on the detailed information viewing page of tourist spots, users will be directed to the information update page of tourist spots. Concerning information updates, users other than the information contributor can easily add or correct information, all users can add sub tags, and edit the description. In addition, to prevent falsification that is done with ill intent, aside from locking the modifications of important information such as titles, the nicknames of editors will be recorded and disclosed on the detailed information viewing page.

4) Tourist spot recommendation function

If users would like a recommendation of tourist spots, by clicking the "tourist spots recommendation" button, users will be directed to the recommendation page. On the recommendation page, in order to improve recommendation accuracy, users can select a checkbox of tourist spot categories they would like to be recommended in. When the category is selected and sent, the top ten tourist spots according to the user's preference will be listed as the recommendation results. When users select a checkbox of a tourist spot they are interested in, the location of the spot will appear as a marker on the digital map. If the marker is clicked on and the user further clicks on the title appearing in the bubble, the user will be directed to the detailed information viewing page of the tourist spot.

5) Information display function of AR Smart Eyeglass

By entering the identification code displayed on the user information page using the Android app linked to the Smart Eyeglass, users can link their own account to the Smart Eyeglass. When sightseeing on foot, the Smart Eyeglass will present nearby tourist spot information based on the user's preference information and location information. More specifically, taking into consideration the effect on visibility, by sliding the controller, the device can switch tourist spots and display the top five spots according to the user's preference. In addition, it will display the "title", "main tag", "average evaluation points", and the "distance" of the recommended tourist spot. Additionally, by tapping the controller and selecting, the distance and direction indication to the tourist spot will be displayed in real time.

6) Information management function

Files of all user's submitted information and images are all stored on the system's database as data. The manager can manage users and check submitted information on a dedicated list screen, suspend accounts of users that have made inappropriate remarks or actions, and delete submitted information with a single click if an inappropriate submission was uploaded. Because managers are not required to confirm whether inappropriate information has been submitted within the system, the manager's burden is reduced.

B. System backend

1) Processing related to recommendations on the internet

In this study, using the knowledge-based recommendation method, the similarity level of each item is calculated with the backend and tourist spots are recommended. By registering preference information on the user information page, users can receive recommendations. More specifically, a feature vector is created with the registered preference information and the characteristics information of tourist spots, and the similarity level of user's preference information and each spot's evaluation information is calculated by means of formula (1) as shown in Section IV- B-4). In this way, the top ten tourist spots in order of the highest similarity levels are displayed as recommendation results.

2) Processing related to information display on AR Smart Eyeglass

Regarding AR Smart Eyeglass, in order for users to use it for sightseeing on foot, tourist spots that are within a walkable distance are displayed. The backend will extract nearby tourist spots with the location information of the user and display the top five spots in order of the highest similarity levels. Regarding the distance of extracting tourist spots, according to the study of the daily sphere by Ishihara et al. (2006) [23], 400m is generally referred to as walking distance for normal people, acknowledged as a nearby distance in everyday life, and is defined as a distance that would not be difficult to walk to. Therefore, the system has set 400m as the distance which tourist spots can be extracted. If there are less than five tourist spots within that distance for recommendations, the distance will be expanded to double and tourist spots will be extracted from within that distance. Additionally, if users are traveling, in order to lessen the load on the server which is accessed frequently in addition to taking into consideration the real time of information presentation, when users travel half the distance (200m) from the center point from where the tourist spots were extracted, the similarity level will be recalculated and the tourist spot list will be updated.

3) System interface

The system has four interfaces including the PC screen of the user (Fig. 2), mobile information terminal screens (Fig. 3), screen for AR Smart Eyeglass (Fig. 4), and PC screen of managers.

VI. Operation Tests and Operation

Following the operation process of TABLE I, the operation was conducted after the operation tests/operation test evaluation of AR media GIS which was designed and developed by means of this study.

A. Operation tests and operation test evaluation

Before the operation, 6 students in their 20s were selected and a two-week operation test was conducted. From the hearing survey results of the test subjects after the operation test, three points of improvement were found in areas including the location information of tourist spots on the detailed information viewing page, the information display method on an AR Smart Eyeglass, and the updating of tourist spot information displayed on Android apps, and the system was restructured regarding these points only.

No.	Description
1	User greeting
2	Display of user information
3	Display of ten latest pieces of submitted tourist spot information
4	Go to submitted information list and ranking page
5	Go to initial page (Sample information is displayed on digital map)
6	Go to user information change and registration page
7	Go to page for submitting tourist spot information
8	Go to page for viewing submitted tourist spot information
9	Go to page where tourist spots are recommended
10	Logout
11	Go to page for mobile information terminals
12	Marker legend

Fig. 2. C interface and description of functions

No.	Description
1	Go to initial page (Sample information is displayed on digital map)
2	Go to user information change and registration page
3	Go to page for submitting tourist spot information
4	Go to page for viewing submitted tourist spot information
5	Go to page where tourist spots are recommended
6	Logout
7	Go to page for mobile information terminals
8	Marker legend

Fig. 3. Mobile information terminal interface and description of functions

Fig. 4. Mobile information terminal interface and AR glasses terminal interface

B. Operations on the internet

1) Summary and results of the operation

Regardless of being inside or outside the operation area, utilization was promoted through Websites of authors' laboratory, and we received cooperation by the Tourism Department of Kanagawa Prefecture and Yokohama City, and the Yokohama Convention Bureau (Yokohama Tourism Association) in distributing pamphlets and operating manuals. When accessing the system for the first time, users must register user information including "user name", "email address", "age", "gender", and "greetings" on the initial registration screen. To take into consideration those who do not want their user information made public as a profile, users can choose to enter their real name or a nickname in the "user name" box. Additionally, users can choose whether to make their "age" and "gender" public or private. When logging in after completing the initial registration, users can operate the submitting, viewing and recommending screen. In addition, by registering preference information in "My Information", each user can receive recommendations according to their tastes.

TABLE II indicates the user classification during the two-month operation period, and Fig. 5 shows the transition of the total number of users during that time. The number of users gradually increased, and the total number was 91, with 44 being male users and 47 being female users. Users in their 20s were approximately 69%, users in their 30s and 40s were each approximately 13% and 9%, and the total users in their 20~40s were approximately 91%. As indicated in the 2015 White Paper on Telecommunications [24], the numbers shown above coincide with the fact that main users of general SNS are for the most part in their 20~40s.

2) Submitted information and use of the comment function/tag function

Fig. 5 also shows the transition of the total number of submitted items during the eight-week operation period as shown above, and it is evident that there is a significant increase in the number of submitted items from the fourth week. It can be assumed that this increase was caused as users became used to the use of the system, they actively started submitting information that they have or think is necessary around the middle of the operation period. In addition, in order to solve the cold-start problem as mentioned in Section IV-B-4), the 181 tourist spot information items which were collected by Ikeda et al. (2014) [18] were prepared as initial data. As there were a total of 162 submitted items during the operation period, a total of 343 items were accumulated to the system.

TABLE III shows the submitted information regarding tourist spots according to category, and although there were many submitted information during the operation period in categories such as restaurants/cafes (65 items, approximately 40%), shopping (29 items, approximately 18%), and famous spots/historical sites (23 items, approximately 14%), information was submitted in every category. In addition, almost all the submitted information had a related image submitted with it. Furthermore, with the various submissions of tourist spot information, information of tourist spots to recommend to users according to each preference, with the images attached to it, was accumulated in line with the purpose of the system.

During the operation period, the number of times the comment function was used on tourist spot information was 32, the number of times the evaluation function was used was 204, and the tag registration number was 26. From the above, among the communication methods used for the system including the comment function, evaluation function, and sub tag function, the evaluation function was the most used. The reason for this is that, while the comment function requires writing and the sub tag function requires thinking of a sub tag, the evaluation function only requires users to express their rating of tourist spots on a five grade scale.

C. Operation using the Smart Eyeglass

Regarding the Minato Mirai area situated in the center of the operation area, an operation using the Smart Eyeglass was conducted on December 18th targeting tourists. TABLE IV shows the user classification of the above, and the total number of users were 34 including 20 males and 14 females. When dividing users by age, although users in their 20s make up 41% of the total users which is the highest percentage, the age of users were scattered and all users had no experience using Smart Eyeglasses. After the operation, all users answered the Web questionnaire survey.

VII. EVALUATION

In this section, based on the questionnaire survey results as summarized in TABLE II and IV, the evaluation using the Web system and Smart glass will be conducted. Next, the evaluation based on the access analysis results using the log data of the operation will be conducted. In addition, based on the evaluation results of the above, solution strategies for this system will be extracted.

A. Evaluation based on the Web questionnaire survey results of the operation on the Web

1) Evaluation related to the use of the system

a) Evaluation related to the suitability of the collection method of tourist spot information

In order to evaluate the system's suitability for tourist spot information collection methods of users, users were asked whether they used the internet to collect tourist spot information. 32% answered "often", 23% answered "sometimes", 20% answered "depends", 21% answered "almost never", and 4% answered "never". In addition, regarding the question (multiple answers) of what information terminal users use to collect tourist spot information, 84% answered PC, and regarding mobile information terminals, 54% answered smartphone, and 14% answered tablet. In this way, from the fact that many users use two or more information terminals to collect information, information collection can be done efficiently by using each information terminal according to the situation. From the information above, it was revealed that a total of 75% of the answerers use the internet to collect tourist spot information, a total of 68% of the answered use mobile information terminals as an information terminal for information collection, and that using different

information terminals according to the situation can make information collection more efficient. Therefore, the system, which operates with the assumption that users use different information terminals, has a fitting design for tourist spot information collection methods of users.

TABLE I. OPERATION PROCESS OF THE SYSTEM

Process	Aim	Period	Specific details
1. Survey of present conditions	To understand efforts related to tourism in the region for operation (Yokohama City)	December 2014 - March 2015	- Survey of government measures and internet services - Interviews with government departments responsible, tourist associations, etc.
2. System configuration	Configure the system in detail to suit the region for operation	April - June 2015	- Define system requirements - System configuration - Create operation system
3. Operation test	Conduct the system operation test	July 2015	- Create and distribute pamphlets and operating instructions - System operation test
4. Evaluation of operation test	Reconfigure the system based on results of interviews with operation test participants	August - September 2015	- Evaluation using interviews - System reconfiguration - Amendment of pamphlets and operating instructions
5. Operation	Carry out actual operation of the system	October - November 2015	- Appeal for use of the system - Distribution of pamphlets and operating instructions - System operation management
6. Evaluation	Evaluate the system based on the results of Web questionnaires, and the results of access analysis which used log data during the period of actual operation	November - December 2015	- Evaluation using Web questionnaires, access analysis which used log data - Identification of measures for using the system even more effectively

TABLE II. OUTLINE OF USERS AND RESPONDENTS TO THE WEB QUESTIONNAIRE

	Aged 10 to 19	Twenties	Thirties	Forties	Fifties	Sixties and above	Total
Number of users (people)	3	63	12	8	2	3	91
Number of Web questionnaire respondents (people)	3	40	5	6	1	1	62
Valid response rate (%)	100.0	63.5	41.7	75.0	20.0	33.3	63.3

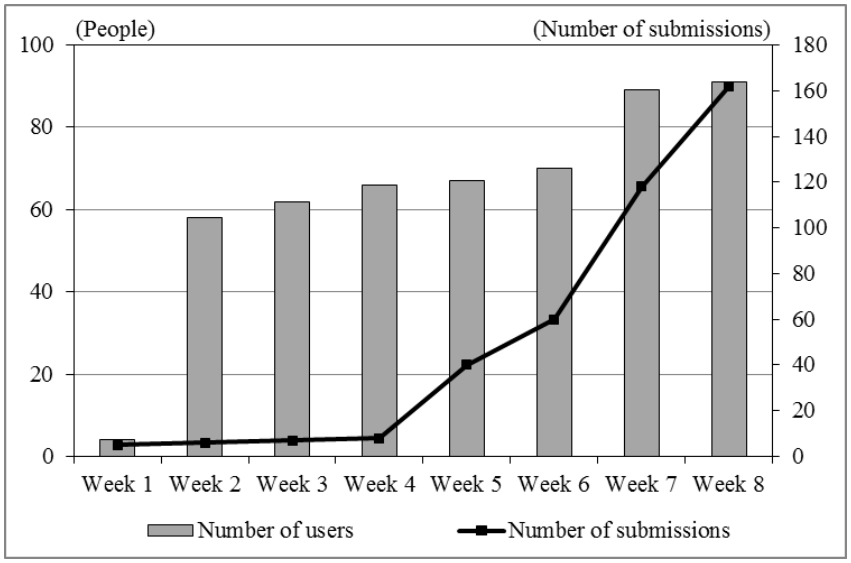


Fig. 5. Changes in the number of users and number of submissions during the operation period

TABLE III. SUBMISSIONS OF INFORMATION, CLASSIFIED BY TOURIST SPOT CATEGORY

Category	Number of submissions	Percentage (%)
Restaurants/Cafes	65	40.1
Other eating/drinking establishments	4	2.5
Noted places/Historic sites	23	14.2
Shopping	29	17.9
Theme parks/Parks	12	7.4
Art galleries/Museums	16	9.8
Scenery	2	1.2
Other	11	6.8
Total	162	100.0

TABLE IV. OUTLINE OF USERS AND RESPONDENTS TO THE WEB QUESTIONNAIRE

	Aged 10 to 19	Twenties	Thirties	Forties	Fifties	Sixties and above	Total
Number of users (people)	6	14	4	5	4	1	34

b) Evaluation related to the usefulness concerning sightseeing tour activities

In order to evaluate the usefulness of the system concerning the tourists' excursion behavior of the system, users were asked about the usefulness during the actual tourists' excursion behavior in the operation area. As a result, the system was rated highly as 70% answered "useful", 26% answered "somewhat useful", 4% answered "neither", and no user answered negatively. Therefore, it is anticipated that the system can be useful for users during tourists' excursion behavior.

2) Evaluation related to the original function of the system

a) Evaluation related to the use according to each function

In order to evaluate the use of the system according to each function, users were presented with each function, and were asked to choose the top 2 functions that were used the most. TABLE V shows the percentages of chosen individual functions and the different combination of functions, and the percentages were high as 27% chose the viewing function/recommendation function, 14% chose the recommendation function/evaluation function, 11% chose the viewing function, and there were no users who used the information updating function independently. In addition, the highest percentage was 53% for the viewing function, and because it is evident that half these users use the system combining the viewing and recommendation function, it can be assumed that the system is mostly used to obtain tourist spot information.

b) Detailed evaluation related to original functions

In order to evaluate the 4 functions that show the remarkable originality of the system, users were asked related questions. As shown in Fig. 6, 89% answered "suitable" or "somewhat suitable" regarding the suitability of the tourist spots information recommended. Therefore, calculating the similarity level based on the evaluation information of each tourist spot and the preference information registered in the "My information", the results of the recommended tourist spot information with high similarity levels with the preference information of users shows that the information was appropriate for the majority of users. The suitability of evaluation items regarding tourist spot had good results as 97% answered "suitable" or "somewhat suitable", and it can also be said that this has resulted in the high suitability of recommended tourist spots.

Regarding the usefulness of submitting tourist spot information, 89% answered "useful" or "somewhat useful", and it can be said the item setting, process, and display method when submitting tourist spot information in the system is appropriate. Concerning the usefulness of attaching sub tags to tourist spot, although 93% answered "useful" or "somewhat useful", 7% answered "not very useful". From this information, regarding users who are not used to sub tags, it can be assumed that such users considered the sub tags used in the system as not very useful.

B. Evaluation based on the Web questionnaire survey results of the operation using AR Smart Eyeglasses

In order to evaluate the operation using AR Smart Eyeglasses, first, users were asked questions regarding the accuracy and satisfaction level for information display of AR devices, in addition to security when using AR devices. As shown in Fig. 7, 82% answered "high" or "somewhat high" concerning the accuracy of information display on glasses, and 97% answered "high" or "somewhat high" concerning the satisfaction of information display on glasses. From the information above, it can be said that the information display on AR devices of the system is appropriate. In addition, regarding the safety when using AR terminals, 88% answered "high" or "somewhat high". However, 3% answered "somewhat low". The reason one answerer gave was that "as the screen is hard to see under strong lighting, users may not notice their surroundings because they're trying to focus on the screen".

Next, in the same way as Section VII.-A-1), in order to evaluate the usefulness regarding the tourists' excursion behavior of the system, users were asked whether the system was useful when actually on tourists' excursion behavior in the operation area. As a result, the system was highly evaluated as 77% answered "useful", 23% answered "somewhat useful", and there were no negative comments. Therefore, the system,

even if the AR Smart Eyeglass is used in the operation, can be assumed to be useful during tourists' excursion behavior.

C. Evaluation based on the access analysis results using the operation log data on the Web

1) Access analysis summary

In this study, by conducting an access analysis using the log data collected during the operation, an evaluation focused on the access numbers and access methods will be conducted. This study will integrate the Google Analytics API with the developed program in order to conduct the access analysis. Google Analytics is a free application provided by Google, and is often used as an analysis tool. Google Analytics can be used by entering the API within the program on each page of the site, and by doing so an access log can be obtained.

2) Evaluation based on the access analysis results

First, an analysis of users' access log was conducted concerning the operation in normal mode. The total session number was 415, 72% used a PC while 28% used a mobile information terminal as an information terminal in order to access the system. TABLE VI shows the number of visits to each function page, and the total number of visits was 2,846. Concerning visits from PCs, excluding the top page, although the tourist spot recommendation function page (8%) and viewing function page (7%) had many visits, the difference between each function page was not so great. Concerning visits from mobile information terminals, the same tendency as visits from PCs was seen. Therefore, comparing PCs and mobile information terminals, although there is a big difference regarding the number of visits of each function page, it can be said that the usage of the system is almost similar.

TABLE V. FUNCTIONS USED THE MOST FREQUENTLY IN THE SYSTEM (UP TO TWO SELECTED)

	Submission function	Viewing function	Updating function	Recommendation function	Evaluation function	Submission function & Viewing functions	Submission function & Updating functions	Submission function & Recommendation functions	Submission function & Evaluation function
Percentage (%)	5.4	10.7	0.0	7.1	5.4	7.1	7.1	3.7	1.8

Viewing function & Updating function	Viewing function & Recommendation function	Viewing function & Evaluation function	Updating function & Recommendation function	Updating function & Evaluation function	Recommendation function & Evaluation function
1.8	26.7	7.1	1.8	0.0	14.3

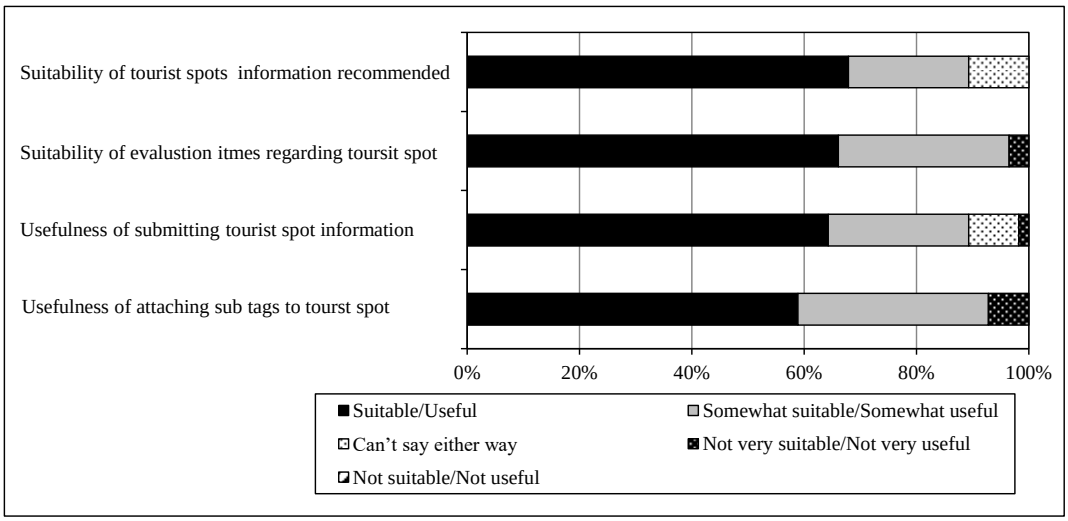


Fig. 6. Results for responses to four question items concerning original functions

Note: The response options shown on the right in the explanatory notes in the figure are for the two question items on usefulness – the usefulness of submitting tourist spot information, and the usefulness of attaching sub tags to tourist spots.

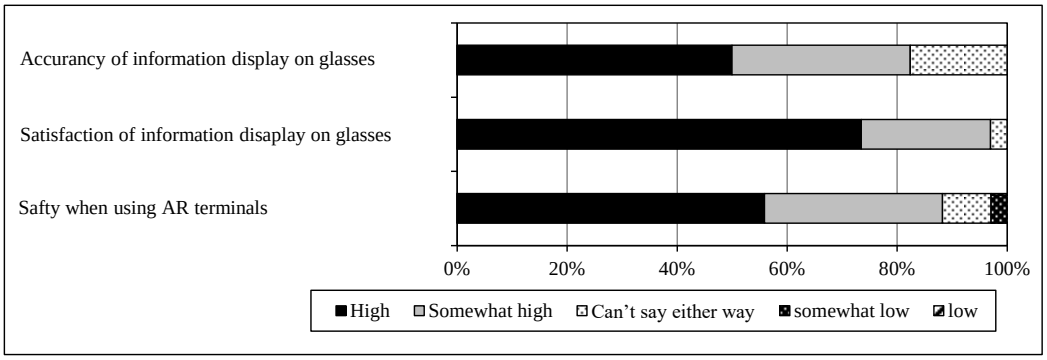


Fig. 7. Results for responses to accuracy and satisfaction of information display on glasses

TABLE VI. TOP FIVE VISITED PAGES, CLASSIFIED BY INFORMATION TERMINAL USED TO VISIT PAGE

PC			
Rank	Page name	Number of visits	Percentage (%)
1	Initial page	268	13.0
2	Viewing page	152	7.4
3	Recommendation page	161	7.8
4	Submission page	92	4.5
5	Updating page	75	3.6
Total		2,056	100.0
Mobile information terminals			
Rank	Page name	Number of visits	Percentage (%)
1	Initial page	88	11.1
2	Viewing page	52	6.6
3	Recommendation page	54	6.8
4	Submission page	29	3.6
5	Updating page	20	2.5
Total		790	100.0

D. Extracting solutions

From the evaluation results of this section, solutions for the system will be summarized as follows:

1) System function expansion

Regarding the AR Smart Eyeglass, a different function other than the automatic recommendation function will be implemented. More specifically, when tourist spots that users are interested in are saved with the system on the Web, users can view the information using AR Smart Eyeglass. In addition, by implementing the routing function with the AR Smart Eyeglass, not only will tourist spots nearby be recommended, but tourist spots that users are interested, even if it is a little far, will be recommended, thereby supporting sightseeing tour activities.

2) Improvement of the recommendation function

Recommendation method that takes into consideration the time series of tourists' excursion behavior will be implemented. By doing so, the system will support not only tourist spot recommendations but also tour plans. In addition, recommendations taking into consideration each user's age, group creation, and other detailed terms, will be conducted. By means of the above, users' tourists' excursion behaviors can be supported more effectively, and planning for tourists' excursion behavior will be made more useful.

VIII. CONCLUSION

The conclusion of this study can be summarized into three points as shown below.

1) In order to support tourists' excursion behaviors by integrating SNS, Twitter, Web-GIS, recommendation system, and Smart Eyeglass, in addition to making the accumulating, sharing, and recommending of information regarding urban tourist spots possible, the AR recommended GIS was designed and developed. This made the ameliorating of information search constraints, spatial constraints taking into consideration safety, and constraints related to continuous operation possible. In addition, the Minato Mirai area, situated in the center part of Yokohama City, Kanagawa Prefecture, was selected as the operation area, and the system details were developed after field surveys were conducted.

2) Because the operation was conducted over a period of 8 weeks, an operation test was conducted 2 weeks prior to the

operation, and the system was reconfigured by extracting points needing improvement. All subjects, whether inside or outside the operation area, were over the age of 18, and among the 91 users, 91% were between the age of 20-40. Additionally, the ultimate number of submitted information was 161. In addition, concerning the operation using Smart Eyeglass, which was conducted with tourists in the Minato Mirai area, the total number of users were 34, age of users were spread out, and all users had no experience in using Smart Eyeglasses.

3) From the results of the Web questionnaire survey given to users after the operation, the system, which sets using information terminals according to use as a premise, is compatible for the collection method of tourist spot information for users, and is mainly used to collect tourist spot information using the viewing and recommendation functions. From the results of the access analysis using the log data form during the operation, the utilization method of the system with PCs and mobile information terminals were very similar. Additionally, as the system using AR Smart Eyeglass was rated extremely highly, it was evident that it is possible to support tourists' excursion behavior using PCs, mobile information terminals, and AR Smart Eyeglasses are possible.

For research tasks in the future, points such as newly implementing functions proposed in Section VII-D in order to support tourists' excursion behaviors in a more safe and effective way, increasing usage performance by operating the system in other urban sightseeing areas, and improving the significance of use can be raised.

ACKNOWLEDGMENT

In the operation of the AR recommended GIS and the Web questionnaires of this study, enormous cooperation was received from those mainly in the Kanto region such as Kanagawa Prefecture and Tokyo Metropolis. We would like to take this opportunity to gratefully acknowledge them.

REFERENCES

[1] Y. Kurata, "Introducing a hot-start mechanism to a Web-based tour planner CT-Planner and Increasing its coverage areas", Papers and Proceedings of the Geographic Information Systems Association of Japan, Vol.21, CD-ROM, 2012.

[2] J. Sasaki, T. Uetake, M. Horikawa and M. Sugawara, "Development of personal sightseeing support system during long-term stay", Proceedings of 75th National Convention of IPSJ, pp.727-728, 2013.

[3] T. Fujitsuka, T. harada, H. Sato and K. Takadama, "Recommendation system for sightseeing plan using pattern mining to evaluate time series action", Proceedings of the Annual Conference on Society of Instrument and Control Enginnering 2014, SS12-10, pp.802-807, 2014.

[4] T. Ueda, R. Ooka, K. Kumano, H. Tarumi, T. Hayashi and M. Yaegashi, "Sightseeing support system to support generation / sharing of sightseeing information", The Special Interest Group Technical Reports of IPSJ: Information system and Social environment (IS), 2015-IS-131(4), pp.1-7, 2015.

[5] H. Uehara, K. Shimada and T. Endo, "Sightseeing location recommendation using tourism information on the Web", Technical Report of The Institute of Electronics, Information and Communication Engineers, NLC, "Natural langage Understanding and Models of Communication", Vol.112, No.367, pp.13-18, 2012.

[6] M. Batet, A. Moreno, D. Sánchez, D. Isern and A. Valls, "Turist@: Agent-based personalised recommendation of tourist activities", Expert Systems with Applications, Vol.39, No.8, pp.7319-7329, 2012.

[7] B. Shaw, J. Shea, S. Sinha and A. Hogue, "Learning to rank for spatiotemporal search", Proceedings of the sixth ACM international conference on Web search and data mining, pp.717-726, 2012.

[8] M. Ye, P. Yin, W. C. Lee and D. L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation", Proceedings of the 34th international ACM SIGIR conference on Research and Development in Information Retrieval, pp.325-334, 2011.

[9] X. Liu, Y. Liu, K. Aberer and C. Miao, "Personalized point-of-interest recommendation by mining users' preference transition", Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp.733-738, 2013.

[10] M. Chen, F. Li, G. Yu and D. Yang, "Extreme learning machine based point-of-interest recommendation in location-based social networks", Proceedings of ELM-2015, Vol. 2, pp. 249-261, 2016.

[11] Q. Yuan, G. Cong, Z. Ma, A. Sun and N. M. Thalmann, "Time-aware point-of-interest recommendation", Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.363-372, 2013.

[12] T. Yanagisawa and K. Yamamoto, "Study on information sharing GIS to accumulate local knowledge in local communities", Theory and Applications of GIS, Vol.20, No.1, pp.61-70, 2012.

[13] H. Nakahara, T. Yanagisawa and K. Yamamoto, "Study on a Web-GIS to support the communication of regional knowledge in eegional communities: Focusing on regional residents' experiential knowledge", Socio- Informatics, Vol.1, No.2, pp.77-92, 2012.

[14] S. Yamada and K. Yamamoto, "Development of Social Media GIS for information exchange between regions", International Journal of Advanced Computer Science and Applications, Vol.4, No.8, pp.62-73, 2013.

[15] T. Okuma and K. Yamamoto, "Study on a Social Media GIS to accumulate urban Disaster Information: Accumulation of Disaster Information during normal times for disaster reduction measures", Socio-Informatics, Vol.2, No.2, pp.49-65, 2013.

[16] T. Murakoshi and K. Yamamoto, "Study on a Social Media GIS to support the utilization of disaster information : For disaster reduction measures from normal times to disaster outbreak times", Socio-Informatics, Vol.3, No.1, pp.17-30, 2014.

[17] K. Yamamoto and S. Fujita, "Development of Social Media GIS to support information utilization from normal times to disaster outbreak times", International Journal of Advanced Computer Science and Applications, Vol.6, No.9, pp.1-14, 2015.

[18] T. Ikeda and K. Yamamoto, "Development of Social Recommendation GIS for tourist spots", International Journal of Advanced Computer Science and Applications, Vol.5, No.12, pp.8-21, 2014.

[19] A. Fujimoto, M. Niimi and H. Noda, "Development of navigation system for inside Buildings using Augumented Reality", The Institute of Electronics, Information and Communication Engineers: EMM, Multimedia Information Hiding and Enrichmen, Vol.113, No.480, pp.33-38, 2014.

[20] S. Tosa, S. Iwabuchi, S. Masuko and J. Tanaka, "Coupon use purchasing support system using the augmented reality to stimulate customers' interests", IPSJ Interaction 2013, 3EXB-36, pp.715-718, 2013.

[21] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, "Recommender systems: An introduction", Cambridge University Press, U.K., 2011.

[22] T. Kamishima, "Algorithms for recommender systems (2)", Transactions of Japanese Society of Artificial Intelligence, Vol.23, No.1, pp.89-103, 2008.

[23] H. Ishihara, T. Shimizu and Y. Izumi, "A basic study on the spheres of daily life", Urban Advance, No.45, pp.68-76, 2006.

[24] Ministry of Internal Affairs and Communications of Japan, "2015 White paper - Information and communications in Japan", Tokyo, 2

An Efficent Lossless Compression Scheme for ECG Signal

O. *El B'charri, R. Latif, A. Abenaou, A. Dliou, W. Jenkal
ESSI, National School of Applied Sciences
Ibn Zohr University
Agadir, Morocco

***Abstract*—Cardiac diseases constitute the main cause of mortality around the globe. For detection and identification of cardiac problems, it is very important to monitor the patient's heart activities for long periods during his normal daily life. The recorded signal that contains information about the condition of the heart called electrocardiogram (ECG). As a result, long recording of ECG signal amounts to huge data sizes. In this work, a robust lossless ECG data compression scheme for real-time applications is proposed. The developed algorithm has the advantages of lossy compression without introducing any distortion to the reconstructed signal. The ECG signals under test were taken from the PTB Diagnostic ECG Database. The compression procedure is simple and provides a high compression ratio compared to other lossless ECG compression methods. The compressed ECG data is generated as a text file. The decompression scheme has also been developed using the reverse logic and it is observed that there is no difference between original and reconstructed ECG signal.**

Keywords*—*ECG; lossless compression; data encoding; compression ratio

I. INTRODUCTION

In every year, according to an estimate given by the 2012 World Health Organization (WHO) statistics report, 56 million people died worldwide, of whom the heart diseases remained the leading cause of death throughout the world. The cardiovascular diseases killed 17.5 million people in 2012. That is 3 in every 10 deaths. Of these, 7.4 million people died of ischaemic heart disease and 6.7 million from stroke [1]. Some of these lives can be often saved if acute care by detecting the myocardial infarction early. So that those patients will have medical attention as soon as possible.

Electrocardiography is a fundamental part in both patient monitoring and cardiovascular assessment. It is an essential tool for investigating cardiac arrhythmias and is also useful in diagnosing cardiac disorders such as myocardial infarction [2]. It deals with the electrical activity of the central of the blood circulatory system, i.e., the heart. The contraction and relaxation of cardiac muscle result from the depolarization and repolarization of myocardial cells. These electrical changes produce currents that radiate through the surrounding tissue to the skin. The electrodes sense electrical currents when they are hooked up to the skin. The recorded currents are then transformed into waveforms called electrocardiogram (ECG).

To detect and identify cardiac problems, ECG signals should be continuously monitored typically over 24h period. Digitizing the ECG signals is performed at sampling rates ranging from 100 to 1000 Hz with a resolution of 8 or 12 bits per sample [3]. As a result, the data sizes of the produced ECG recording will enormously increase and fill up available storage space. Nowadays, storage space is relatively cheap. However, ECG data archives could easily become exceedingly large and expensive. Moreover, in mobile monitoring environments, compression is a fundamental tool to resolve and transmit physiological signals from the human body to any location, especially for real-time transmission applications.

The ECG compression methods are generally classified into two main groups, namely one-dimensional (1-D) and two-dimensional (2-D). 1-D ECG compression algorithms are the most widely employed in literature and can be further classified into four categories: direct time domain compression, model based compression, transformed domain compression and hybrid compression algorithms. In 2-D ECG compression algorithms, 1-D ECG signals are represented in 2-D then the transformation is applied on those 2-D representation.

The classification of ECG compression algorithms can also be observed by another angle, lossy and lossless compression. Although lossy compression has an important benefit of high Compression Ratio (CR), it introduces distortion into the original ECG signal and may lose some features that could be very important for future analysis of crucial patients. Considering lossless compression, it provides a moderate to high CR and maintains the original ECG signal away from any notable distortion..

Motivating by works of Mukhopadhyay et al. [4-6], which are based on voltages to text encoding, we propose a novel lossless ECG compression scheme that combines the advantages of the stating works. In [4], authors developed a lossless compression method that can preserve all the information in the reconstructed signal. However, the compression ratio has an unsatisfactory performance. The authors have also developed lossy compression algorithms in [5,6]. In [5], the entire signal values are quantified in the compression process while this quantification is only applied outside of QRS regions in [6]. The QRS regions are compressed using the same process as [4]. These lossy compression methods can achieve high CR but the major problem is that the reconstructed signal has a significant distortion that may even delete some important features in the ECG signal such as the T wave. Furthermore, from juridical and clinical point of view [7], it is strongly recommended to work on lossless ECG signal compression.

To overcome the aforementioned shortcomings, we developed a powerful lossless compression scheme that is able to reduce significantly the file size which provide a high compression ratio, all maintaining near-zero distortion in the reconstructed signal leading to an excellent quality score.

The paper is organized as follows. The proposed method of the ECG data compression and reconstruction is detailed in Section II. Section III presents the results of the proposed method of selected ECG records from the PTB diagnostic ECG database (PTB-DB). The performance analysis and comparison with other methods are also discussed. To conclude, some remarks and discussion are given in Section IV.

II. METHODOLOGY

The proposed algorithm is generally based on work done in [4], which achieves almost negligible distortion and a moderate good compression ratio. In this work, the compression ratio is highly improved by reducing the number of data sent to the output text file. Another improvement is made at the time of reconstruction to preserve the original signal reliably.

The compression scheme is divided into two main sections: data compression and data reconstruction. The key to this high compression lies in data encoding process and the rearrangement of the variables used in the output text file. The two ECG processing modules are described in the following of this section.

A. ECG data compression

The overall compression scheme is illustrated in figure 1. The boxes and the various shapes outlined in this figure are detailed step by step in the following of this subsection.

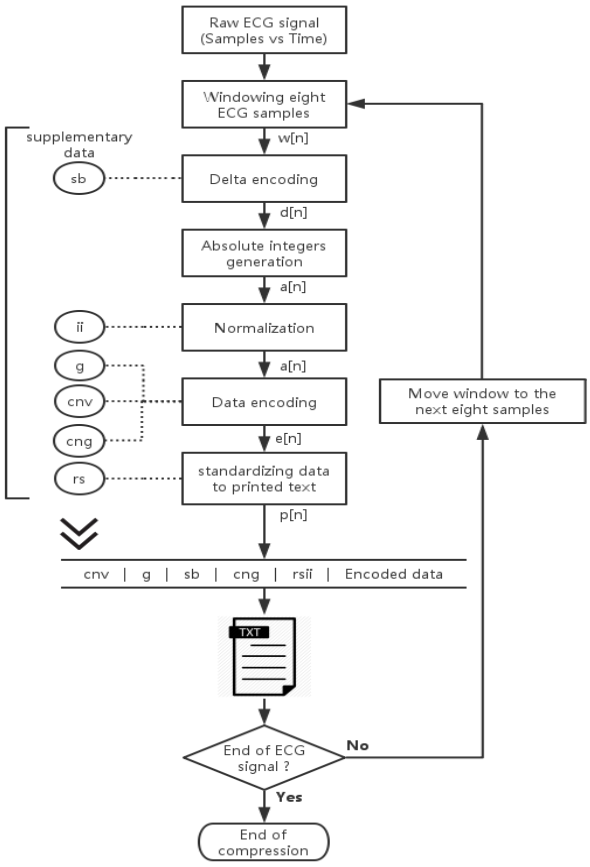


Fig. 1. Flowchart of the proposed scheme

1) Windowing eight ECG samples

The first step of the compression scheme is to take only ECG samples from the raw ECG signal file. The time axis is discarded since the sampling frequency is known. The ECG signal can be easily reconstructed by corresponding each sample to its equivalent instant of time. The whole compression procedure is applied on eight samples at a time until the end of the signal is reached.

2) Delta encoding

In order to get better compression, delta encoding is performed on those eight samples by subtracting two consecutive samples. The first value remained unchanged as described below.

$$d(0) = w(0)$$
$$for\ i = 1\ to\ 7$$
$$\quad d(i) = w(i) - w(i-1)$$
$$end$$

The motivation behind delta encoding is to get smaller number, which means concatenation possibility will increase in the later step of compression. Consequently, compression ratio will also increase.

3) Absolute integers generation

The sign of the each element of the array d[n] is verified. For those that contain positive numbers a binary zero (0) is made, and those containing negative numbers a binary (1) is made. The string obtained from those zeros and ones, which represents the sign of those eight samples, is converted to its equivalent decimal number and stored in a variable named ‘sb’. The absolute value of each sample is then multiplied by 1000 to avoid the fractional part since in standard ECG database; samples are recorded up-to three decimal points. This procedure is performed as follows:

$$for\ i = 0\ to\ 7$$
$$\quad if\ d(i) < 0$$
$$\quad\quad sb(i) = 1$$
$$\quad else$$
$$\quad\quad sb(i) = 0$$
$$\quad end$$
$$\quad a(i) = abs\{d(i)\} * 1000$$
$$end$$
$$(sb)_{10} \leftarrow (sb)_2$$

4) Normalization

In the delta encoding step, the first sample is kept unchanged. The problem starts when the ECG data begin with a high voltage (greater than |0.255|). As those samples after multiplying by 1000 become greater than 255 they cannot be printed in the output text file. Moreover, another problem has been observed at the time of dissociation in the decompression scheme that will be discussed in later section. So the first sample is chosen to be less than 200. To overcome this issue, the following instructions is used:

$$ii = a(0)/200$$
$$a(0) = a(0)\%200$$

Where 'ii' is assigned with the integer value from the division operation and a(0) is holding the reminder of this division.

5) Data encoding

In this section of the algorithm the main compression has occurred. To reduce the data size, these eight samples are minimized using the following logic:

$$if\ ([a(i) * 100] + a(i+1) < 255)$$
$$e(j) = [a(i) * 100] + a(i+1)$$
$$j = j + 1$$
$$else\ if\ (a(i) + [a(i+1) * 100] < 255)$$
$$e(j) = a(i) + [a(i+1) * 100]$$
$$j = j + 1$$
$$end$$

For the above-described encoding algorithm, the variable 'i' is initialized by zero and incremented by a factor of two.

The samples are encoded and stored in a new array as described above. The new array is constructed from the previous one by either concatenating or not two consecutive samples:

- Regular direction: If two consecutive samples undergoes the first condition, then these samples are concatenated in the regular direction.
- Opposite direction: If the first condition is not satisfied then we check that whether the second formula is verified or not. If so, then these consecutive samples are concatenated in the opposite direction.
- None: If concatenation is not possible in both direction then, those samples are left unchanged and stored separately in the new array.

The encoding step is achieved. However, in the decoding scheme we cannot identify what type of concatenation is carried out. To memorize the type of concatenation, instead of using three variables as done in [4], we take only one variable 'g'. This variable is taken as a binary number (one byte). Each 1 is inserted in its even bits corresponds to the regular direction concatenation (it is supposed that the most significant bit is an even bit), while a 1 is inserted in the odd bits corresponds to the opposite direction concatenation. If two consecutive even and odd bits equals to zero means that there was no concatenation.

It has been observed that some concatenated data lead to wrong samples at the decompression scheme. For an example purpose, suppose an encoded value like '209'. It could be concatenated from three case:

▪ Case 1	▪ Case 2	▪ Case 3
0 and 209	1 and 109	2 and 9

The first case is eliminated during the normalization section (the numbers in a[n] array are less than 200). To differentiate the two other cases, we take a variable named 'cnv' as a binary number to denote the second case in the fourth last bits of 'cnv' (the fourth most significant bits). The third case will be treated as a normal concatenation case.

To further clarify the data encoding step, let us take an array a[n] having sample values as shown in figure 2. In the following scheme, Reg and Opp denote regular direction and opposite direction, respectively.

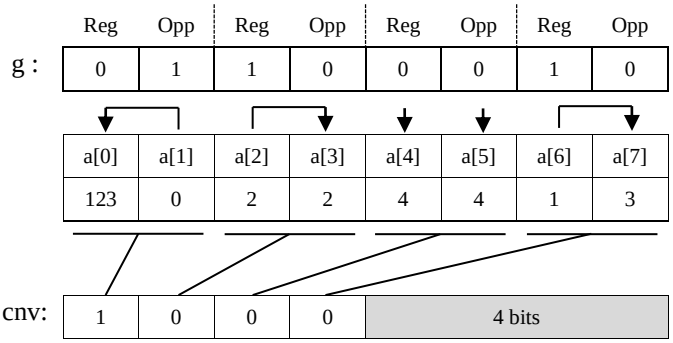


Fig. 2. An illustrative example of data encoding step

6) Standardizing data to printed text

The ECG samples are encoded in the range of text characters. However, some special characters are not stamped in the text file, which results to loss of data at the decompression stage. Those characters are 10 (line feed), 13 (carriage return), 26 (substitution) or 255 (blank). To avoid such problem, the numbers corresponding to those characters are substituted by some suitable numbers. To denote the modified numbers, it is necessary to add extra variables as described below.

- An extra variable named 'rs' is temporary taken which signifies whether the 'sb' is one of those special characters or not. If the variable 'sb' is changed by other number, 'rs' will be set to 1 (rs = 1) otherwise to 0 (rs = 0). Finally 'rs' is multiplied by 100 and is added with 'ii'. The result is stored in a new variable named 'rsii' which minimizes the 'rs' and 'ii' variables in one variable.
- A variable say 'cng' is taken as binary number to denote the positions of those special characters in e[n].
- There is a seldom probability that 'rsii', 'cng', and 'g' may contain any of those special characters. If it happens, some other suitable numbers will replace them. These changes will be denoted in the three first

unused bits of 'cnv' respectively (i.e. the three least significant bits). The 'cnv' variable will never be a critical number if we put a zero in its forth bit as described in figure 3.

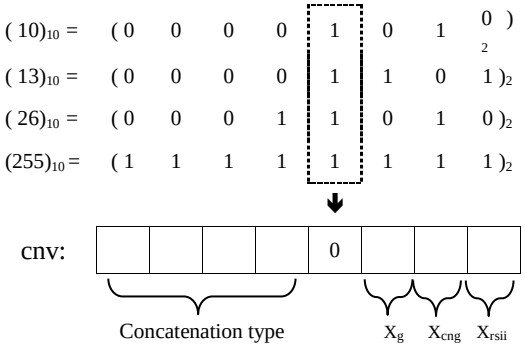


Fig. 3. The construction of the 'cnv' variable

Where X_i can be either 1 or 0 value depending if it was a special character or not.

Finally the encoded data along with all the necessary information (cnv, g, sb, cng, and rsii) are organized as a frame in the sequence shown in Figure 1 and are printed in their text code in a text file. The length of each frame may vary according to the concatenation possibilities. The order of these characters has a particular importance. Since from the variables 'cnv' and 'g' we can easily determine the length of each frame sent to the text file, the frames are sent tied to each other without any separation character. Following this procedure, we can minimize one separation character for each sent frame, which leads to a better compression.

Moving the window to the next eight samples, the whole compression scheme is repeated until the end of ECG signal is achieved.

B. ECG data decompression

Decompress the ECG data is necessary to retrieve the original signal at the receiving end. For this purpose, a decompression algorithm is also developed. Since the decompression scheme uses the same reverse logic described in figure 1, we briefly describe the procedure of the decompression. Some of the relevant steps are more detailed.

1) Extracting Frames from the text file

The first step of the decompression scheme is to identify the length of each frame from the continuous text contained in the compressed file. Here, the variable 'g' has an important feature apart from denoting the concatenation, it can also determines the length of the frame by calculating the sum of ones in 'g', then subtract this sum from 13. Where the value 13 represents the maximum length that can occupy a frame in the text file. This process is carried out by first checking the variable 'g' if it was a special character by testing on the third bit of the variable 'cnv' if it was 1. If it was the case, the original value of the variable 'g' is placed back. The decompression procedure is then applied on the extracted frame.

2) Replacement of the original numbers

Since the order of the variables is known in each frame, the modified data during subsection A.6 can be easily recovered by analyzing the variables 'cnv', 'cng' and 'rsii'.

3) Data decoding

The binary form of the variable 'g' gives the positions of concatenation. If it is found a one (1) in even bits (starting from the most significant bit) that represent the regular direction, while a one (1) in the odd bits represent the opposite direction. Or else, if two consecutive even and odd bits equal to zero (0), then there was no concatenation. These concatenated numbers are then dissociated according to the type of concatenation labeled in the four most significant bits of the variable 'cnv',

Now 'ii' is multiplied by 200 and added with a[0], the result is stored in a[0] position as the reverse was done during the 'Normalizing' subsection in the compression procedure.

As used in the 'Normalizing' subsection in the compression procedure, the variable 'ii' is multiplied by 200 and added to a[0]. The result is stored in the same a[0].

4) Signed values recovering

In this section, the variable 'sb' will be converted into its corresponding 8-bits equivalent. If any bit is '1' the corresponding positional element of d[n] array will be multiplied by (-1). The next step consist of dividing every number in this array by 1000.

5) Creating original ECG signal

To produce the original ECG signal, we have to generate ECG samples and the corresponding time axis. To get the original ECG samples, we use the reverse of delta encoding, e.i. each positional number is added with the previous value except the first one. Finally, each sample is stamped with its equivalent instant of time using the known sampling frequency.

Moving to the next frame, the whole decompression scheme is repeated until the end of the text file is achieved.

III. RESULTS AND DISCUSSION

A. Evaluation factor of the compression scheme

The criteria for testing the performance of the compression algorithms consist of three important components: compression measure, reconstruction error and computational complexity.

The computational complexity component is often attached to practical implementation consideration, which is recommended to be as simple as possible.

The Compression Ratio (CR) represents the ratio between the size of the file containing original and compressed signal, given by:

$$CR = \frac{B_0}{B_c} \quad (1)$$

Where B_0 is the total number of bits in the original file and B_c is the total number of bits in the compressed file.

The maximum absolute error is defined as the maximum element of sample-to-sample difference array:

$$E_{max} = Max(x_0(n) - x_r(n)) \quad (2)$$

Where $x_0(n)$ is the original signal, $x_r(n)$ is the reconstructed signal.

The definition of the error criterion for assessing the distortion of the reconstructed signal compared to the original one is of primary importance, the Percentage Root-mean-square Difference (PRD) measure defined as:

$$PRD(\%) = \sqrt{\frac{\sum_{n=1}^{N}\left(x_0(n) - x_r(n)\right)^2}{\sum_{n=1}^{N} x_0^2(n)}} \times 100 \quad (3)$$

Where N represents the window length, over which the PRD is calculated. In all scientific literature interested by ECG compression techniques, evaluation of this error is crucial and is the one commonly used. The clinical acceptability of the reconstructed signal is based on this criterion.

The normalized version of PRD is PRDN, which is independent of the signal mean value $\bar{x}$, is defined as:

$$PRDN(\%) = \sqrt{\frac{\sum_{n=1}^{N}\left(x_0(n) - x_r(n)\right)^2}{\sum_{n=1}^{N}(x_0(n) - \bar{x})^2}} \times 100 \quad (4)$$

One another evaluation factor given in (5), is Quality Score (QS). It quantifies the global performance of compression algorithm taken into account both the CR and PRD.

$$QS = \frac{CR}{PRD} \quad (5)$$

This all factors are evaluated on the proposed algorithm, and are given on their average value in Table 1.

The ECG data files under test are chosen from PTB diagnosis ECG database available under Physionet. To assess and compare the performance, the leads of every record were choosing the same as [4]. From table 1, we can see that using the proposed algorithm we can get average PRD of about 0.0092% and average CR of about 18.84:1. Since the developed method stands in the direct time domain compression and the compressed samples are encoded without any data truncation, the reconstructed signal from the decompression scheme is guaranteed to be the same as the original one without any distortion. Hence, we can have a zero PRD value. It should be noted that PRD, PRDN and E_{max} values depend strongly on the length of the original signal. If this length is a multiple of eight, these values is confirmed to be zero. Otherwise, the remaining ECG samples will not be compressed and will not be presented in the decompressed ECG signal which generate a non-zero PRD, PRDN and E_{max}. Regarding the compression ratio, we can observe that record n° S0305 provide the highest CR value, which means that the concatenating possibility was high compared to other signals.

To provide an overall estimate of the computational complexity of the algorithm, we performed an amount of simulation to each record then we calculate the average compression time of each signal. Compression time shows that this algorithm can be used for real-time systems.

Figure 4 shows the original (blue-a) and reconstructed (green-b) ECG signals of the proposed ECG compression scheme. Through visual inspection of this figure, It is obvious that the reconstructed signal is the same as the original ECG signal since the compression scheme performed in this work is lossless.

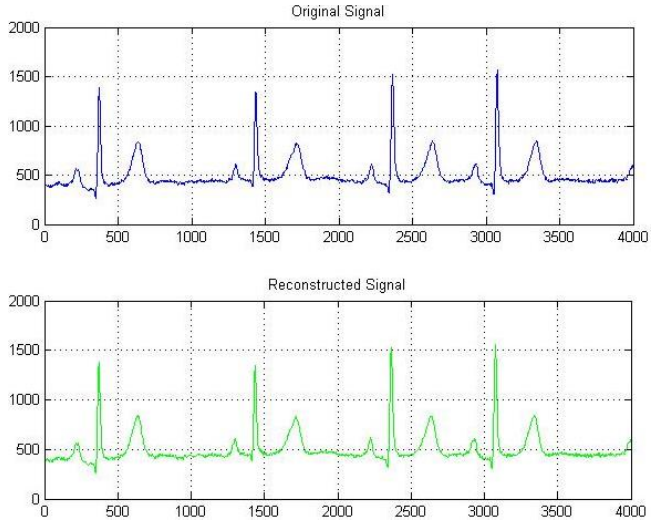


Fig. 4. Original (a) and Reconstructed (b) ECG signal. (File: S0305, Lead I, Duration: 4s)

TABLE I. THE PERFORMANCE MEASUREMENT OBTAINED FOR VARIOUS ECG FILES

Records	Performance evaluation contents *					
	Duration (s)	*PRD (%)*	*PRDN (%)*	*CR*	$\lvert Emax \rvert$	*Compression time (ms)*
S0022LRE	15	0.0000	0.0000	18.27	0	329.34
S0021ARE	15	0.0000	0.0000	18.86	0	417.67
S0015LRE	15	0.0000	0.0000	17.87	0	389.83
S0304	26.342	0.0082	0.0099	18.70	0.0003	717.00
S0305	23.770	0.0292	0.0811	19.76	0.0083	642.18
S0301	22.084	0.0181	0.0191	19.61	0.0028	613.74

(*) All the factors are represented in the average value

B. Performance comparaison

Comparison of some compression's evaluation factor of various methods with the proposed method is given in Table 2. It is divided into three parts namely lossy compression, near lossless compression and lossless compression.

The lossy compression methods are known by their high compression ratio but unfortunately, they may lose some

important information about the signal that can be noted in high values of PRD, while the near lossless compression algorithms give a moderate compression ratio, preserving the value of the PRD as low as possible. In the lossless compression algorithms, the importance is given to the value of PRD, which should be almost zero.

If lossy compression methods have a high compression rate, our method can be classified with these methods while maintaining almost zero distortion.

TABLE II. PERFORMANCE COMPARAISON OF VARIOUS ECG COMPRESSION ALGORITHMS

Algorithm	PRD (%)	CR	QS	PRDN (%)
Lossy compression				
m-AZTEC [8]	25.5	5.6	0.22	-
Mukhopadhyay et al. [5]	7.89	15.72	1.99	20.60
Mukhopadhyay et al. [6]	7.58	22.47	2.97	13.28
Near lossless compression				
USZZQ and Huffman coding of DSM [9]	2.73	11.06	4.05	-
SPIHIT [10]	1.18	8	6.78	-
Lossless compression				
JPEG2000 [11]	0.86	8	9.30	-
Fira and Goras [12]	0.61	12.74	20.89	48.38(max) -7.42(min)
SangJoon Lee et al. [13]	0.61	16.5	27.11	-
Mukhopadhyay et al. [5]	0.023	7.18	312.17	-
Proposed	**0.0092**	**18.84**	**2047**	**0.018**

Among the four lossless compression algorithms presented in table 2, the proposed method provides high CR (18.84), which is comparable with the CR of lossy compression methods generally known by their high compression ratio. The values of PRD and PRDN are negligible (almost zero) and are better than any other methods, wich leads to a very high QS (2047).

Our algorithm was based on the work of [4]. If we compare the results of our work with their results, we can found that the CR is considerably increased compared to the value of their CR. The value of the PRD is also improved since we treated some specific cases during the reconstruction of the signal.

IV. CONCLUSION

In this research, a high lossless compression scheme for ECG signals is proposed. The proposed algorithm was tested for the compression of normal and pathological types of cardiac beats ECG signals. This technique provides significant improvement in term of compression ratio compared to other lossless compression techniques, all ensuring a near-zero distortion that is notable, either through visual inspection, or the measured value of error loss. Consequently, the quality score that quantify the overall performance was far superior to the other compared algorithms. The proposed algorithm is simple and easy to implement. The output text file can be further compressed, using some standard text compression techniques. The compressed file can be either stored or transmitted over wireless network as text file for real time ECG analysis. The proposed scheme is suitable to use in portable and mobile ECG data monitoring system.

REFERENCES

[1] World Health Organization (WHO), Statistics 2012. Available online: http://www.who.int/mediacentre/factsheets/fs310/en/ (accessed on 21 July 2016).

[2] F. Morris, W. J. Brady and J. Camm, “ABC of Clinical Electrocardiography”, 2nd ed., Blackwell Publishing Ltd, 2008, pp. 1.

[3] M. S. Manikandan and S. Dandapat, “Wavelet threshold based TDL and TDR algorithms for real-time ECG signal compression”, Elsevier Ltd Biomedical Signal Processing and Control, vol. 3, 2008, pp. 44–66.

[4] S. K. Mukhopadhyay, S. Mitra, and M. Mitra, “A lossless ECG data compression technique using ASCII character encoding”, Computers and Electrical Engineering, vol. 37, 2011, pp. 486–497.

[5] S. K. Mukhopadhyay, S. Mitra, and M. Mitra, “An ECG signal compression technique using ASCII character encoding”, Measurement, vol. 45, 2012, pp. 1651–1660.

[6] S. K. Mukhopadhyay, M. Mitra, S. Mitra, “ECG signal compression using ASCII character encoding and transmission via SMS”, Biomedical Signal Processing and Control, vol. 8, 2013, pp. 354– 363.

[7] Koski A, “Lossless ECG encoding”, Computer Methods and Programs in Biomedicine, vol. 52, January 1997, pp. 23–33.

[8] V. Kumar, S.C. Saxena, V. K. Giri and D. Singh, “Improved modified AZTEC technique for ECG data compression: Effect of length of parabolic filter on reconstructed signal”, Computers and Electrical Engineering, vol. 31, issues 4-5, June-July 2005, pp. 334–344.

[9] M. S. Manikandan and S. Dandapat, “Wavelet threshold based ECG compression using USZZQ and Huffman coding of DSM”, Biomedical Signal Processing and Control, vol. 1, Issue 4, October 2006, pp. 261-270.

[10] Z. Lu, D. Y. Kim, W. A. Pearlman. “Wavelet compression of ECG signals by the set partitioning in hierarchical trees algorithm”, IEEE Trans. Biomedical Engineering, vol. 47, issue 7, July 2000, pp. 849–856.

[11] A. Bilgin, M. W. Marcellin and M. I. Altbach, “Compression of electrocardiogram signals using JPEG2000”, IEEE Trans. Consumer Electronics, vol. 49, issue 4, November 2003, pp. 833–840.

[12] Fira CM and Goras L. “An ECG signals compression method and its validation using NNs”, IEEE Trans. Biomed. Eng., vol. 55, no. 4, April 2008, pp. 1319–1326.

[13] S. J. Lee, J. Kim, and M. Lee, “A Real-Time ECG Data Compression and Transmission Algorithm for an e-Health Device”, IEEE Trans. Biomedical Engineering, vol. 58, issue 9, September 2011, pp. 2448-2455.

Albanian Sign Language (AlbSL) Number Recognition from Both Hand's Gestures Acquired by Kinect Sensors

Eriglen Gani
Department of Computer Science,
Faculty of Natural Sciences, University of Tirana
Tirana, Albania

Alda Kika
Department of Computer Science
Faculty of Natural Sciences, University of Tirana
Tirana, Albania

***Abstract*—Albanian Sign Language (AlbSL) is relatively new and until now there doesn't exist a system that is able to recognize Albanian signs by using natural user interfaces (NUI). The aim of this paper is to present a real-time gesture recognition system that is able to automatically recognize number signs for Albanian Sign Language, captured from signer's both hands. Kinect device is used to obtain data streams. Every pixel generated from Kinect device contains depth data information which is used to construct a depth map. Hands segmentation process is performed by applying a threshold constant to depth map. In order to differentiate signer's hands a K-means clustering algorithm is applied to partition pixels into two groups corresponding to each signer's hands. Centroid distance function is calculated in each hand after extracting hand's contour pixels. Fourier descrip-tors, derived form centroid distance is used as a hand shape representation. For each number gesture there are 15 Fourier descriptors coefficients generated which represent uniquely that gesture. Every input data is compared against training data set by calculating Euclidean distance, using Fourier coefficients. Sign with the lowest Euclidean distance is considered as a match. The system is able to recognize number signs captured from one hand or both hands. When both signer's hands are used, some of the methodology processes are executed in parallel in order to improve the overall performance. The proposed system achieves an accuracy of 91% and is able to process 55 frames per second.**

***Keywords*—Albanian Sign Language (AlbSL); Number Recognition; Microsoft Kinect; K-Means; Fourier Descriptors**

I. INTRODUCTION

Sign Language is very important for the inclusion of the hearing impaired persons in the society. They use sign language as natural way of communication. Every country has developed their own sign language and Albania has its own, which is relatively new [1], [2]. In all situations where deaf people are participating an interpreter is required which results in a non-effective method because it requires time and resources. An Albanian Sign Language recognition system would make possible the communication between hearing impaired persons and the hearing ones in a more effective way. Many countries have tried to develope a sign language translation system through natural user interfaces as for American Sign Language [3], Arabic Sign Language [4], Portugal Sign Language [5], Indian Sign Language [6] and many others.

Until now there doesn't exist a system that is able to recognize Albanian signs by using natural user interfaces (NUI). Body movements, head position, facial expressions and hands trajectory are used by hearing impaired persons to communicate with each other. Many work has been done to integrate some existing technologies to capture and translate signer's gestures, among them web cameras [7], data gloves [8] and Kinect sensors. Web cameras generate low quality of images and have an inability to capture other body parts. It is also hard to generalize the algorithms for web cameras due to many different shapes and colors of hands. Data gloves achieve high performance but are expensive and not a proper way to human-computer interaction perspective [2]. Kinect technology, launched by Microsoft, has many advantages as: provide color and depth data simultaneously, it is inexpensive, the body skeleton can be obtained easily and it is not effected by the light. Various researchers are using Microsoft Kinect sensor for sign language recognition as in [6], [9], [10]. We are trying to built a real-time, automatic system that is able to capture and translate numbers from 1 to 10 by using Microsoft Kinect sensors. It is used to obtain input data, K-Means algorithm to differentiate signer's hands and then Fourier descriptors algorithm to classify number gestures from 1 to 10.

Many researchers have tried to capture and translate number gestures by following different approaches. [11] and [12] use color camera to capture input gestures and then SVM (Support Vector Machine) and Fuzzy C-Means respectively to classify hand gestures.

[13] took another approach for number recognition. It used circle method to count the fingers which is scale, translation, and rotation invariant. Firstly the center of the hand is located and then the furthest pixel from the center is found. They both form the radius of the circle. An imaginary circle is drawn and the fingers intersecting with the circle have been counted. The method is limited to count the number of the fingers from 1 to 5 and cannot be extended to other numbers and signs.

Shape plays an important role in gesture categorization. Two approaches have been widely used for shape recognition and shape normalization including Fourier descriptors and HU moments. [14] conduced an experiment where Fourier descriptors and HU moments were compared in terms of

shape recognition accuracy. The result shows that Fourier descriptors were superior to HU moments in terms of shape recognition accuracy.

Fourier descriptors can be derived from different shape signature including complex coordinates, centroid distance, curvature signature and cumulative angular function. An evaluation is done in [15], which results that centroid distance is significantly better than other three signatures. The fact that centroid distance captures both local and global features makes it desirable as shape representation.

Section I gives a brief introduction and related work. The rest of the paper is organized as follows. Section II presents an overview of methodology and a brief description of each methodology's processes. Section III describes the experimental environment. Section IV presents the experiments and results. The paper is concluded in Section V by presenting the conclusions and further work.

II. METHODOLOGY

Figure 1 summarizes the methodology followed in number recognition captured from both signer's hands gestures:

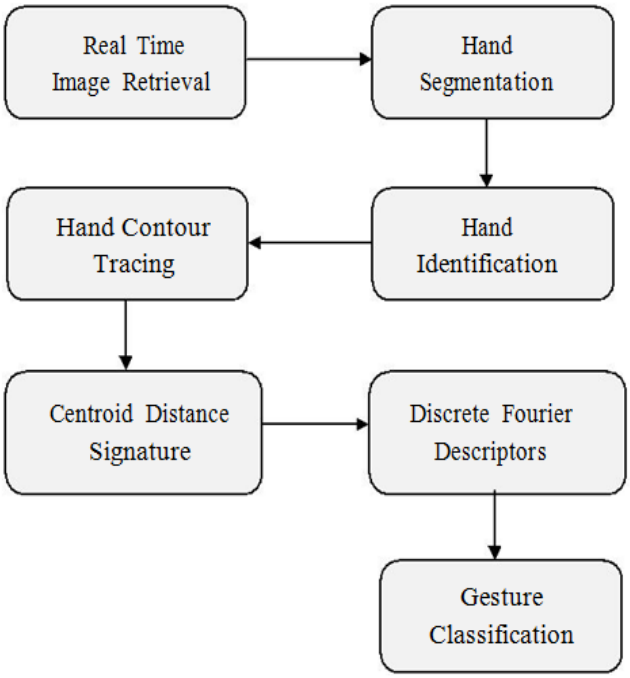


Fig. 1. Number recognition methodology

Microsoft Kinect is used to acquire hand gestures at rate of 30 frames per second. Attached to each frame it includes depth information at an interval of layers [0-4095] [16]. Each frame has dimensions of 640 x 480 (width x height) and includes a total of 307200 pixels.

Hand segmentation can be obtain by using two approaches illustrated at [17]. The first approach builds an histogram by including the number of white pixels grouped by depth layer. Histogram helps to understand the first object placed in front of Kinect. By applying a threshold constant, only the human hands are segmented. This approach does not function properly if an external object is placed between Kinect device and human body. To overcome this problem the second approach is focused only in body parts by using Kinect skeleton feature. It does not need to built an histogram because the skeleton feature excludes every object not part of human body. The first layers correspond always to user's hands. By applying again a threshold constant the human's hands are obtained.

In a 2D space the K-Means algorithm is applied in order to partition all pixels into two groups which corresponds to signer's hands. The K-means algorithm starts by placing K points (centroids) at random locations in 2D space. In our case only two centroid have been placed which correspond to user's hands. Each pixel, is assigned to a cluster with the nearest centroid, and then the new centroids are calculated as the mean of pixels assigned to it. The algorithm continues until no pixel change cluster membership. If the distance between two centroids is less then a constant, then they are merged into one.

After hand identification, the hand's contours pixels have to be extracted. A 8-connectivity algorithm [18] has been applied. The pixels that form the hand boundary are used as input data to centroid function.

Fourier descriptors are derived from shape signature. A compression between different shape signature is given at [15]. In our case centroid shape signature is used which is a one-dimensional function that represent two dimensional areas or boundaries. It is applied to pixels obtained from hand's shape boundary. Before applying Fourier transform the normalization process is performed. For matching purpose the training data set and input data must have the same number of points in the shape boundary. Number of points and points chosen affect the matching accuracy. Fast Fourier Transform (FFT) need a power-of-two [19] number of points. In our experiments 128 points have been chosen. Decreasing the number of points increases the computational results and decreases the accuracy of matching results. The chosen points must be distributed equally in all shape boundaries. There exist some methods including a)equal points sampling, b)equal angle sampling, and c)equal arclength sampling. In our experiment equal angle sampling is chosen.

Every input data must be compared against every sign in gesture dictionary by calculating Euclidean distances. The sign with the lowest Euclidean distance is considered as a match. The gesture dictionary is composed with ten number gestures representing numbers 1 to 10. Figure 2 visualizes number gestures dictionary, based on Albania Sign Language [1].

III. EXPERIMENT ENVIRONMENT

The environment where all tests are executed is composed of a Microsoft Kinect device and a DELL Notebook. Kinect device consist of an IR emitter, an RGB camera, an IR depth sensor, a microphone array and a tilt [20]. Kinect sensors are used to obtain skeleton stream and a depth map with a resolution of 640 x 480 at 30 FPS. DELL Notebook consists of a 64-bit architecture, a Windows 7 operating system, 4 GB of physical memory and an Intel Core i5-5200U processor 2.20GHz. System is developed using Microsoft C# programming language and Kinect for Windows SDK 1.8 library. Figure 3 visualizes the system environment.

IV. Experiment and Results

Experiments are based in two aspects: accuracy and computational latency. Firstly two data sets have been created, corresponding to training and testing data set. Each number gesture in both data sets contains 15 Fourier descriptors coefficients which are generated by running each of methodology processes.

Number Gesture	Meaning	Number Gesture	Meaning
	One		Six
	Two		Seven
	Three		Eight
	Four		Nine
	Five		Ten

Fig. 2. Number gestures dictionary

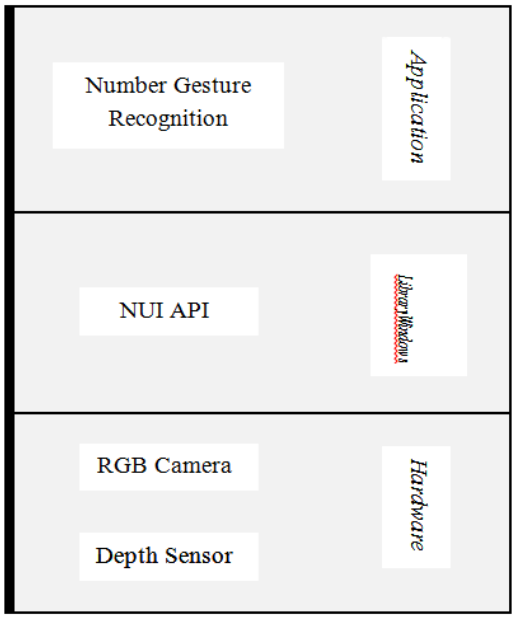


Fig. 3. System environment

Training data set contains in total 40 gestures only for numbers 1 to 5. Numbers 6 to 10, which are generated from both user's hands are calculated as sum of numbers recognized from each individual hand. There are 8 gestures for each number. To improve accuracy, they are taken from 4 different signers using their both hands. Testing data set contains a total of 400 gestures. There are 40 gesture for each number. They are captured from 4 different signers.

TABLE I. Testing Data Set Recognition Rate

Number Gesture	True Recognition	False Recognition
1	100%	0%
2	88%	12%
3	95%	5%
4	85%	15%
5	93%	7%
6	95%	5%
7	93%	7%
8	85%	15%
9	93%	7%
10	85%	15%
Average	91%	9%

Figure 4 shows the confusion matrix for number recognition system. Gestures 10, 8 and 4 have the lowest recognition rate. They are misclassfied as gestures 9, 7 and 5. Number 9 is the most ambiguous gesture even through having a high recognition rate of 93%. It is recognized as gestures 7, 8 and 10.

	1	2	3	4	5	6	7	8	9	10
1	40	0	0	0	0	0	0	0	0	0
2	3	35	2	0	0	0	0	0	0	0
3	0	0	38	0	0	2	0	0	0	0
4	0	0	1	34	4	0	0	0	0	0
5	0	0	0	2	37	1	0	0	0	0
6	0	0	0	0	1	38	0	0	0	1
7	0	0	0	0	0	2	37	0	0	1
8	0	0	0	0	0	0	6	34	0	0
9	0	0	0	0	0	0	1	1	37	1
10	0	0	0	0	0	0	0	1	5	34

Fig. 4. Confusion matrix

Firstly training data set elements are used as input data. The system receives 100% recognition rate. Secondly testing data set elements are used as input. Table I summarizes the results. Numbers 10, 8 and 4 have the lowest recognition rate while numbers 1, 3, and 6 have the highest recognition rate.

Computational latency is firstly calculated for each process running sequentially. Hand segmentation and K-Means calculation are the heaviest processes which take approximately 60% of total time. If all the processes are executed in sequentially order the system is able to calculate 48 frames per second. Microsoft Kinect is able to recognize number gestures in real-time since it generates frames at a rate of 30 frames per second. Table II summarizes the results. Time in milliseconds for each process is the average time of 50 experiments where single and both hands have been used.

Today's processors are implemented with a dozen of cores and this number is going to increase due to Moore's law [21]. Multi-core processors allow parallelism and multithread-ing. Since performance is an important factor in real-time recognition systems, computation latency is recalculated in situation where some level of parallelism is provided to some of processes. It is clear that some of the processes can be

TABLE II. COMPUTATION LATENCY CALCULATED SEQUENTIALLY

Processes	Time in milliseconds
Hand Segmentation	6.4381
K-Means Calculation	6.0661
Hand Contour Tracing	4.2646
Normalize Image (128 points)	0.0574
Centroid Distance Signature	0.0671
Discrete Fourier Description	3.7005
Gesture Classification	0.3477
Total	20.9415

processed in parallel. K-Means Calculation, Hand Contour Tracing, Normalize Image (128 points), Centroid Distance Sig-nature, Discrete Fourier Description and Gesture Classification can all be processed in parallel for each signer's hand. Then the results must be aggregated. Table III summarizes the results. Time in milliseconds for each process is the average time of 50 experiments where single and both hands have been used. The total time is improved by 2.9 milliseconds, and the system is able to process 55 frames per second.

TABLE III. COMPUTATION LATENCY CALCULATED FROM PARALLEL TASKS

Processes	Time in milliseconds
Hand Segmentation	6.4579
K-Means Calculation	5.6461
Hand Contour Tracing	3.2571
Normalize Image (128 points)	0.0434
Centroid Distance Signature	0.0503
Discrete Fourier Description	2.5015
Gesture Classification	0.1397
Total	18.0960

V. CONCLUSION AND FUTURE WORK

Aim of this paper is to propose an automatic, real-time solution for recognition of a limited set of numbers (1 to 10) obtained from signer's both hands, by using Microsoft Kinect. This system can be extended in the future to include more signs (including dactyls and not static signs) creating the first sign recognition system for Albanian Sign Language by using natural user interfaces (NUI). Input images are provided through Microsoft Kinect device. IR depth sensor is used to built a depth map. By using depth map the hand segmentation is performed. Skeleton feature of Kinect device can be used in hand segmentation process. In order to understand if gesture is formed from single hand or both hands a K-Means algorithm is applied. The system is able to recognize gestures capture from single hand or both hands by switching automatically. For each segmented hand the contour pixels are extracted. Each number gesture is represented by 15 Fourier descriptors coefficients which are based on centroid distance signature. In total, data set consists of 440 number gestures where 40 of them are used to form training data set. 400 number gestures are used to form testing data set. Every gesture in testing data set is compared against each gesture in training data set by using Euclidean distance. The gesture with minimum distance is considered as a match. Numbers 10, 8, and 4 have the lowest recognition rate. They are misclassified as gestures 9, 7 and 5. Number 9 is the most ambitious one. The system achieves an accuracy of 91%. Computation latency allows the system to be deployed in an image receiving technology that has an acquisition rate of less than 48 frames per second where no parallelism is applied and 55 frames per second where parallelism is applied. Microsoft Kinect can be part of this real-time recognition system since it's acquisition rate is 30 frames per second.

Future work consists of improving the overall system accu-racy by applying more reliable gesture data set and improving the execution time of the slowest processes. Future work will be extended to dactyls and other not static sign gestures.

REFERENCES

[1] ANAD, Gjuha e Shenjave Shqipe 1, ANAD, Ed. Shoqata Kombetare Shiptare e Njerezve qe nuk Degjojne, 2013.

[2] E. Gani and A. Kika, "Review on natural interfaces technologies for designing albanian sign language recognition system," The Third International Conference On: Research and Education Challenges Towards the Future, 2015.

[3] F. Ullah, "American sign language recognition system for hearing impaired people using cartesian genetic programming," in Automation, Robotics and Applications (ICARA), 2011 5th International Conference on. IEEE, 2011, pp. 96–99.

[4] N. R. Albelwi and Y. M. Alginahi, "Real-time arabic sign language (arsl) recognition," in International Conference on Communications and Information Technology (ICCIT 2012), Tunisia, 2012, pp. 497–501.

[5] P. Trindade and J. Lobo, "Distributed accelerometers for gesture recognition and visualization," in Technological Innovation for Sustainability. Springer, 2011, pp. 215–223.

[6] A. S. Ghotkar and G. K. Kharate, "Dynamic hand gesture recognition and novel sentence interpretation algorithm for indian sign language using microsoft kinect sensor," Journal of Pattern Recognition Research, vol. 1, pp. 24–38, 2015.

[7] S. Shruthi, K. Sona, and S. Kiran Kumar, "Classification on hand gesture recognition and translation from real time video using svm-knn," International Journal of Applied Engineering Research, vol. 11, no. 8, pp. 5414–5418, 2016.

[8] R. Rupasinghe, D. Ailapperuma, P. De Silva, A. Siriwardana, and B. Sudantha, "A portable tool for deaf and hearing impaired people."

[9] K. Stefanov and J. Beskow, "A kinect corpus of swedish sign language signs," in Proceedings of the 2013 Workshop on Multimodal Corpora: Beyond Audio and Video, 2013.

[10] H. V. Verma, E. Aggarwal, and S. Chandra, "Gesture recognition using kinect for sign language translation," in Image Information Processing (ICIIP), 2013 IEEE Second International Conference on. IEEE, 2013, pp. 96–100.

[11] J. Wachs, U. Kartoun, H. Stern, and Y. Edan, "Real-time hand gesture telerobotic system using fuzzy c-means clustering," in Automation Congress, 2002 Proceedings of the 5th Biannual World, vol. 13. IEEE, 2002, pp. 403–409.

[12] Y. Liu, Z. Gan, and Y. Sun, "Static hand gesture recognition and its application based on support vector machines," in Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08. Ninth ACIS International Conference on. IEEE, 2008, pp. 517–521.

[13] A. Malima, E. Ozgur¨ and M. C¸etin, "A fast algorithm for vision-based hand gesture recognition for robot control," in Signal Processing and Communications Applications, 2006 IEEE 14th. IEEE, 2006, pp. 1–4.

[14] S. Conseil, S. Bourennane, and L. Martin, "Comparison of fourier descriptors and hu moments for hand posture recognition," in Signal Processing Conference, 2007 15th European. IEEE, 2007, pp. 1960–1964.

[15] D. Zhang, G. Lu et al., "A comparative study of fourier descriptors for shape representation and retrieval," in Proc. 5th Asian Conference on Computer Vision. Citeseer, 2002.

[16] J. Webb and J. Ashley, Beginning Kinect Programming with the Microsoft Kinect SDK. Apress, 2012.

[17] E. Gani and A. Kika, "Identifikimi i dores nepermjet teknologjise microsoft kinect," Buletini i Shkencave te Natyres, vol. 20, pp. 82–90, 2015.

[18] T. Pavlidis, Algorithms for graphics and image processing. Springer Science & Business Media, 2012.

[19] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," Mathematics of computation, vol. 19, no. 90, pp. 297–301, 1965.

[20] MSDN, "Kinect for windows sensor components and specifcations," April 2016. [Online]. Available: https://msdn.microsoft.com/en-us/library/jj131033.aspx

[21] G. E. Moore, "Cramming more components onto integrated circuits," Readings in computer architecture, vol. 56, pp. 56–59, 2000.

A Collaborative Process of Decision Making in the Business Context based on Online Questionnaires

Rhizlane Seltani
TIMS Research Unit, LIROSA Laboratory, Faculty of Science
Abdelmalek Essaadi University
Tetuan, Morocco

Noura Aknin
TIMS Research Unit, LIROSA Laboratory, Faculty of Science
Abdelmalek Essaadi University
Tetuan, Morocco

Souad Amjad
TIMS Research Unit, LIROSA Laboratory, Faculty of Science
Abdelmalek Essaadi University
Tetuan, Morocco

Mohamed Chrayah
TIMS Research Unit, LIROSA Laboratory
ENSA, Abdelmalek Essaadi University
Tetuan, Morocco

Kamal Eddine El Kadiri
LIROSA Laboratory, Faculty of Science
Abdelmalek Essaadi University
Tetuan, Morocco

***Abstract*—This article is a component of a series of articles and scientific researches conducted by the research team which deals with the web 2.0 and its interactions with the different technology areas. During recent years, the emergence of the web 2.0 has revolutionized the world of new technologies, in particular the business intelligence field, providing businesses with new and innovative ways to make use of information in order to improve their overall performance. This article comes to consolidate the profit which can be taken from the new technologies of the web 2.0, especially blogs which constitute a valuable mean to gather exchanged information and results of the collaboration between users, by offering a new collaborative tool for decision making based on online questionnaires in order to exploit the collective intelligence which represents a very important source of significant data, and by adopting the SCAMMPERR method, a creative technique of stimulation of ideas and problem solving.**

This paper presents a practical innovation in the computing level and makes an impact on the economic and the organizational sides of the enterprise, by proposing a new methodology based on the SCAMMPERR technique and supported by the strengths of the web 2.0 to ensure a collaborative decision making. As a result, it provides relevant decisions which support the traditional decision support systems.

***Keywords*—Decision Making; Web 2.0; Blogs; Business Intelligence; SCAMMPERR Method; Online Questionnaire**

I. Introduction

Enterprises increasingly need to maintain and manage their competitiveness, enhance their market share, develop the loyalty of their customers, and optimize their processes and costs. To meet these needs, business intelligence was born.

Nowadays, organizations are more and more demanding and the needs are more important in terms of data constituting the basis of decisions, and the quality and the relevance of these data.

Web 2.0 is a major source of information and new technologies. Since its emergence, it has revolutionized the world of business intelligence as any other technology sector, offering new concepts and techniques as well as various sources of data, which influenced traditional methods of decision making which is becoming a difficult task for business leaders due to the increase of the number of companies and consequently, the competition becomes more and more hard, which makes the search for more innovative ways of decision making, based on adequate data, a primary case. A new way appears, exploit the tools of the web 2.0 and the variety of its resources to enrich the sources of data of the organization and as a result, improve the decision-making process. Among these resources, blogs constitute a precious way of collection of information results of the exchange and the collaboration between the internet users. The proposed solution enhances the advantage taken from this technology, by associating it to the SCAMMPERR method which is one of the most methodical and reasonable techniques of generation and stimulation of ideas as well as the resolution of problems, to provide a collaborative process of decision making.

This paper presents a new method of decision making in the business context. Its impact and its benefits concern the engineering level as well as the innovation management one. After the implementation of this solution, it provides a solid and a practical process, which allows organizations and businesses to monitor their systems and improve the process of the decision making with more meaningful and relevant decisions related to the enterprise issues and strategies in a short lapse of time, and consequently, boost and improve the overall performance of the organization.

The next section, gives a presentation of the web 2.0, followed by a section about the business intelligence and its limits. Then, we will introduce the SCAMMPERR method which constitutes the basis of this work to elaborate the collaborative process of decision making, called SCAMMPERR 2.0 and discussed in the main section of the paper. The two last sections are reserved to the modeling and the implementation of the process.

II. WEB 2.0

The term "Web 2.0" was diffused by Tim O' Reilly in 2004, more detailed later [1], to identify the participative web. What web 2.0 brings, are the progressive increase and the continuous evolution of technologies which allow more and more the participation of web users at the level of the creation of the web content. Improvements affect the material as well as the software.

Web 2.0 is a conjunction of technologies, business tactics and social skills [2], making it a social and a technological model at once and allowing to users to create web content and to follow the last updates of a website without visiting the web page source, and to developers, to quickly and easily create new web applications based on data, information and available services on the internet.

To ensure all these tasks, web 2.0 is based on a complex and diversified architecture [3], based on a permanent inclusion of approaches (providing storage, creation and diffusion capacities), such as: software server, messaging protocols, standards of navigation, content syndication and various client applications as plugins.

A website follows the web 2.0 approach if it is characterized by:

- Techniques of rich applications such as AJAX, a technique of web interface design, which allows the update without refreshing the web page [4].
- Content syndication through standard protocols: RDF, Atom and RSS, which ensures real-time diffusion of new websites information or blogs news.
- Classification by labeling to facilitate the search.
- Valid XHTML and micro-formats.
- Appropriate use of URL and REST architecture or XML web services.

Compared to web 1.0, web 2.0 has some strengths [5]: interactivity, participation and collaboration. The latter principle constitutes the basis of improving systems in different domains such as software engineering [6]. The internet user is no longer a spectator; he becomes an actor and an active contributor due to the emergence of new technologies allowing him to participate more and more in the creation of the web content.

- Social Networks: Any set of social entities, individuals or organizations, joined together by links, established through social interactions, can be considered as a social network. It is the small world which is based on the interactivity between the users and the community gathering around common points as values and passions. These last years, the social networks are more expanded and transform the curiosity to a global phenomenon [7]. Social networks are characterized by a large number of users and a variety of content and applications, such as tests and games.
- Blogs: Introduced by Justin Hall in 1994. In its simplest shape, a blog is a website with dated and published entries on the internet, according to the inverse chronological order [8]. This is a type of website that allows you to publish articles and all types of multimedia: images, videos and sounds. The owner of the blog can also post comments and answers the questions of the visitors who can comment and contact the blogger by e-mail. In blogs, the management is collaborative because all of users participate by their own contents and interventions.
- Wikis: Refer to dynamic websites containing pages which are editable by the web users, and represent collaborative writing spaces of varied information and an effective way of sharing knowledge. Wikis are rapid because reading and editing processes are combined. A common way of using wikis is to support planning meetings: a provisional agenda is set and the URL is distributed to the participants, who do not hesitate to comment or add their own elements [9].
- RSS Feeds: Really Simple Syndication is a manner of description of data, encoded in XML and constitutes a way of automatic distribution of information on a website, by receiving news headlines published on other websites, in real-time. Also, it allows other websites to republish simply the data, what is called the content syndication. RSS is not used only to display the news of blogs, but also for any kind of data regularly updated: weather report, availability of photos, etc [2].
- Podcasts and Videocasts: Podcasts and videocasts offer new means of distribution of digital content. A podcast is an audio file to which, people can subscribe and which can be afterward transferred to an audio player. If it is attached to a movie, it becomes a videocast. The podcasting remains an exceptional innovation in the publication of contents and largely based on the simplicity of use [10].

III. BUSINESS INTELLIGENCE

Business intelligence is a set of technologies of decision-making support within a company whose purpose is to allow executives, administrators and analysts to make better decisions more quickly [11].

In the 80s, the computerization continues, but some companies began to accumulate a lot of data, hence the birth of Data Centers. Only the IT department can create reports from the data sources to help analysts and managers to take decisions. However, the information search process involves a process of type: question - answer – question, that is why the IT specialists find themselves quickly overloaded. At the beginning of the 90s, the report generation software appears, but two side effects occurred further to the birth of reporting systems:

- Systems quickly become overloaded.
- Reporting systems provide "general public" reports.

In the 90s also, many concepts, tools, simple software: quick, independent of the production system, reliable and heterogeneous, appeared. Business intelligence is born. The

architecture of the traditional decision support system is essentially based on a data warehouse as it is shown in Fig. 1.

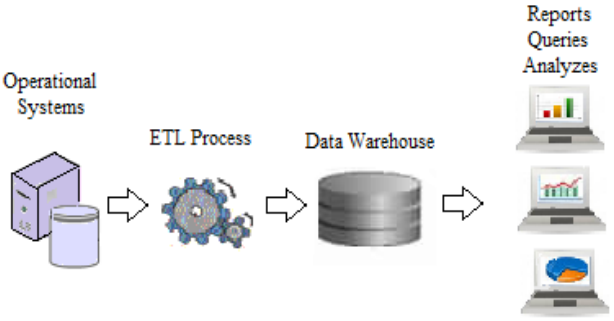


Fig. 1. Traditional business intelligence architecture

The business intelligence process includes the functions: collection, integration, distribution, presentation and administration. However, the traditional business intelligence has some limits:

- Introspective decision-making systems. Therefore, the strategic reach is very limited.
- Limit management of unstructured data.
- Techno centric approach focusing on technology.
- Very time-consuming tasks.

Also, due to the variety of people who express themselves on the web, data lack quality and relevance. Consequently, the integrity of the decision-making system is threatened. One of the major gaps is that the significant web 2.0 resources are not exploited in the ultimate ways in the field of decision-making.

That is why; and to solve problems already mentioned, it will be interesting to exploit the new inherent technologies of the web 2.0 as the collaboration, the interactivity, and the external data to enhance decision support systems with decisions based on a human centric approach and a collective intelligence reflected by the diverse collective applications, such as blogs, social networks, etc.

The collective intelligence is used in several disciplines. Here, it refers to diverse communities of people interacting to create clever outcomes [12]. So, the aim of this work is to participate to surpass the challenges facing the development of the business intelligence 2.0, which is a concept that raised a lot of questions of research to be exactly defined [13].

IV. SCAMMPERR METHOD

To guard their part of competitiveness, companies wish to seek ways of improving, in a continuous way, their products and services, which requires a high rate of imagination and innovation. Nevertheless, the creativity does not occur. It is a process that takes time and effort [14]. That is why; several researchers put a lot for the development of new techniques which fill this need, aiming to ensure the creative thinking and to solve problems for example: Hurson's productive thinking model, the six hats of critical thinking, the reversed brainstorming, etc; though, SCAMPER or SCAMMPERR is considered one of the structured, easiest, successful and most direct methods [15].

SCAMMPERR technique showed its effectiveness compared to the similar methods and constitutes a very rigorous and a powerful technique and at the same time, a very flexible and fast method, already implemented for experiments in important domains as education and engineering [16][17]. When talking about decision making, the most important goal is to save time and relevance, which justifies the choice of the method adopted in the process.

A. SCAMMPERR: Principle and Utility

SCAMMPERR refers to an associative method of creativity which gathers nine innovative techniques, its principle was proposed by Alex Osborn in 1953[18] and developed afterward by Bob Eberle [19] [20].

SCAMMPERR is a technique which provides a methodical and practical way of stimulation of the divergent thinking, the imagination, the originality, and the intuition [21] [22]. Each one of the letters in the SCAMMPERR acronym signifies an operation that can be applied to an idea, a concept, a project, a product or a service. The list of these operations is given in Table I [23].

TABLE I. SCAMMPERR OPERATORS

S	Substitute	components, materials, elements (ideas, people, features, services)
C	Combine	mix, combine with other ideas or services, add functions, elements or systems
A	Adapt	alter, change function, modify a part of an element, utilize a part of another element
M	Magnify	enlarge, make it enormous, higher, longer, add functions , features or additional capabilities
M	Modify	modify scale (increase or reduce it), shape (color, audio, …), attributes (texture, design, ...)
P	Put to another use	use it in a different context, identify more usages or advantages
E	Eliminate	delete elements, components, reduce, simplify, minimize
R	Rearrange	change the order, the sequence , interchange components, change patterns
R	Reverse	turn inside out, upside down, transpose, reverse usage

Each of the nine SCAMMPERR operators can refer to several questions, of which, the ones to adapt to a specific problem can be chosen, to generate answers which constitute new ideas. Some standard examples of these questions are presented in Table II [24].

TABLE II. SCAMMPERR QUESTIONS EXAMPLES

Operator	Questions
Substitute	Can we use something else instead of this product, object, service or process?
Combine	Can we combine anything to get something new or interesting?
Adapt	Does someone else have an answer that we can adapt to our situation?
Magnify/Add	Can we make it larger, add to it or extend it ?
Modify it	Can we change or modify it in some way?
Put it to some other use	How else could our product or process be used? Does it solve some other problem?
Eliminate something	Can we eliminate something to solve our problem?
Rearrange it	Must we rearrange the current order or sequence?
Reverse it	What if we reversed it? Did the exact opposite?

B. SCAMMPERR Process

The SCAMMPERR process is based on two essential steps:

- The identification of the idea, the problem or the subject, matter of the reflection.
- The formulation of questions related to the subject using the list of SCAMMPERR operators.

SCAMMPERR questions can be exploited through:

- Systematic exploration: consists in exploring an idea, a product or a service by using all the SCAMMPERR operators.
- Depth development: iterative use of a SCAMMPERR operator in particular to find new ideas.

In general, there are two important ways to develop the SCAMMPERR method [25]:

- Generate creative ideas from a problem or a topic: using SCAMMPERR for creativity and problem solving as shown in Fig. 2.

Fig. 2. Generation of ideas from a problem

- Apply SCAMMPERR on the results of a previous technique of stimulation of ideas: it aims to filter all the resulting ideas to focus on the best ones as it is shown in Fig. 3.

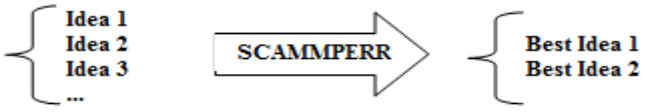


Fig. 3. Generation of ideas from the resulting ideas

In the context of the proposed process, the first method is noted SCAMMPERR1 and the second one is noted SCAMMPERR2. Here, the SCAMMPERR method is used as a way of stimulation, generation of ideas and resolution of problems launched by the company or the community.

V. SCAMMPERR 2.0: A COLLABORATIVE PROCESS OF DECISION MAKING BASED ON ONLINE QUESTIONNAIRES

SCAMMPERR 2.0 is a collaborative process, based on the good use of the web 2.0 tools. DBlog (Decisions Blog) is a blog which can be implemented on the web by the company, to present its problems and to question its marketing strategies, in order to take advantage of opinions and decisions of the customers and the community of the web, as well as to become aware of their needs and their interests, with the ultimate aim of facilitating, renovating and improving the process of the decision making. Questions on questionnaires follow the SCAMMPERR notation and depend on the nature of the treated issue or subject.

The acquisition of the opinions of users is made through an online questionnaire, appropriate to each problem or decision under process, given that the online questionnaire remains a very good way of inspection and evaluation. This online questionnaire comprises a set of questions related to the problem and following the SCAMMPERR1 reasoning, to ensure a methodical analysis. Answers (decisions of the users) undergo a SCAMMPERR2 treatment, before making the final decision. The general outline of this process is illustrated in Fig. 4.

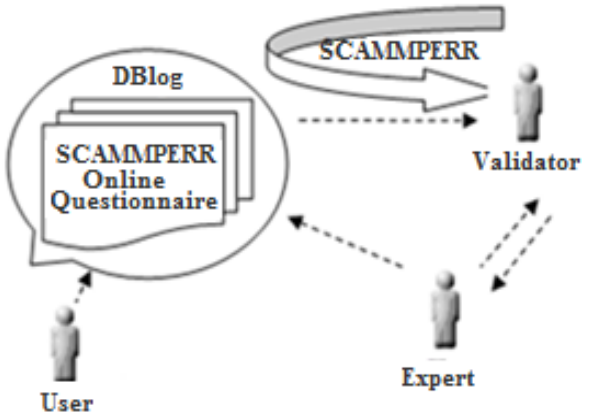


Fig. 4. General outline of the associative process

A. Categories of decisions

Decisions are classified into three main categories:

- Initial decision: the answer of a user for each SCAMMPERR question.
- Preliminary decision: decision of a validator after SCAMMPERR2.
- Final decision: decision of the expert after reviewing the decisions of the validators.

Every decision receives a score on 10 and follows the scale represented in Table III.

TABLE III. SCALE OF SCORES

Score	Signification
[1 – 3]	Low quality
[4 – 6]	Average quality
[7 – 8]	Good quality
[9 – 10]	Strategic Decision

B. Categories of users

Five categories of users interact in the system as shown in Table IV.

TABLE IV. CATEGORIES OF USERS

Actor	Signification	Eligibility
User	Simple user	<=5
User Plus	User	<=5 and client of the company
Validator	Validator of SCAMMPERR1 and actor of SCAMMPERR2	[6 – 9]
Validator plus	Validator of SCAMMPERR1 and actor of SCAMMPERR2	[6 – 9] and employee of the company
Expert	Monitor	= 10

The maximal value that the eligibility can reach is 10, given that: the maximal sum of the scores that can a user reach for each questionnaire is 90, with a score of 10 for each of the nine SCAMMPERR questions.

User eligibility is a parameter which reflects its decisional relevance. It is a factor that will be used to manage the promotion of users and it is expressed by the following equation:

$$\text{Eligibility} = \frac{\sum \text{Scores}}{\sum \text{Decisions}} \quad (1)$$

A degree of Influence is assigned to each user according to its category to designate the weight of its decisions as shown in Table V.

TABLE V. DEGREES OF INFLUENCE

Actor	dInf (Degree of Influence)
User	1
User Plus	2
Validator	4
Validator plus	6

C. Process of Collaborative Decision Making following SCAMMPERR

There are eight stages in the process:

- Step 1: One of the experts of the system develops the problem to solve or the decision to discuss. Then, he elaborates the associated questionnaire following SCAMMPERR notation. The user can start a process by proposing an idea; in this case the expert will take care of the rest.
- Step 2: The expert publishes the online questionnaire on the DBlog.
- Step 3: Users answer the questions of the online questionnaire.
- Step 4: The expert chooses three validators related to the topic of the problem.
- Step 5: The expert generates the matrix of the initial decisions.
- Step 6: The validators evaluate the initial decisions for the management and the promotion of users and apply SCAMMPERR2 to the matrix.
- Step 7: The validators make preliminary decisions.
- Step 8: The expert examines the preliminary decisions and makes the final decision.

The process is illustrated in Fig. 5.

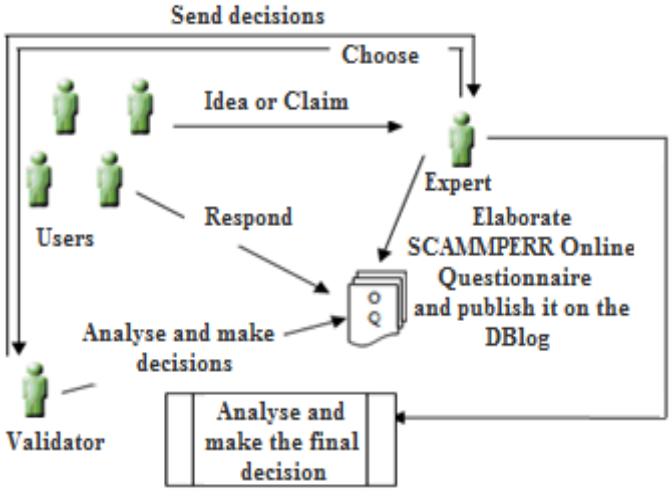


Fig. 5. General process of SCAMMPERR 2.0

When the SCAMMPERR questionnaire is online, users start to fill it. Once a considerable sum of filled copies is reached (as needed), the expert develops a matrix which comprises the diverse initial decisions.

The matrix takes the following form:

$$M(U,O) = \begin{pmatrix} d_{1S} & \cdots & d_{1R} \\ \vdots & \ddots & \vdots \\ d_{NS} & \cdots & d_{NR} \end{pmatrix} \quad (2)$$

Where: d_{UO} is the initial decision of the user U for the SCAMMPERR operator O.

The validators give a mark to initial decisions (a score going from 1 to 10). Afterward, they calculate the degree of relevance of every initial decision, based on the following formula:

$$dR(d_{UO}) = S(d_{UO}) * dInf(U) \qquad (3)$$

Where:

- dR (d_{UO}) is the degree of relevance of the decision d_{UO}
- S (d_{UO}) is the score assigned to the decision d_{UO}
- dInf (U) is the degree of influence of the user U

The degree of Influence of the user, the score of the decision and its degree of Relevance are three parameters which facilitate the application of the SCAMMPERR method by validators, which will allow making a first classification of the initial decisions before applying SCAMMPERR2, as well as for the expert during the evaluation of the preliminary decisions of validators.

After SCAMMPERR2, every validator extracts from the matrix of initial decisions, a SCAMMPERR vector containing nine decisions related to the SCAMMPERR operators.

$$v(V) = (d_{VS}, d_{VC}, d_{VA}, d_{VM}, d_{VM}, d_{VP}, d_{VE}, d_{VR}, d_{VR}) \quad (4)$$

The expert evaluates the decisions of the three validators according to their degree of Influence and the degree of Relevance of their decisions, and makes the final decision.

VI. MODELING

To model the system, the object modeling using UML (Unified Modeling Language) [26] is used. It proposes a rich set of different diagrams [27].

The modeling of the system comprises a use case diagram and a sequence diagram.

A. Use Case Diagram

Use case diagrams allow representing, in a simple way, the fundamental needs and the objectives of the system from an external point of view to it. The use case diagram is represented in Fig. 6.

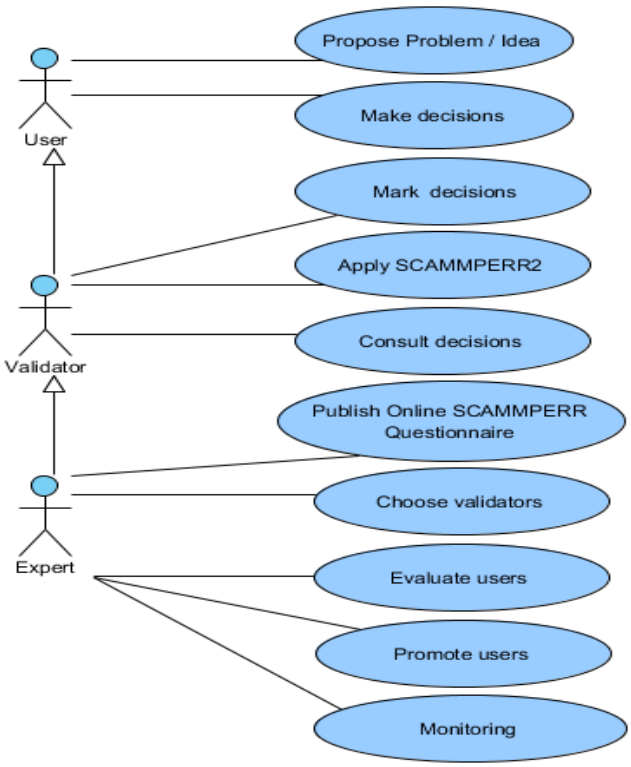


Fig. 6. Use case diagram of SCAMMPERR 2.0

The system has three main actors:

- User: a simple user or a user plus. He is the user of the internet platform, his main role is to make initial decisions by filling SCAMMPERR questionnaires posted by the company, as he can trigger the treatment of an issue or propose an idea.
- Validator: validator or validator plus, his mission is to evaluate the decisions of users by applying SCAMMPERR2 to initial decisions. Subsequently, he takes preliminary decisions.
- Expert: manages all the platform, develops the SCAMMPERR questionnaire, puts it online, collects the initial decisions and evaluates the preliminary ones to manage the promotion of the users and makes the final definitive decision. He is also the one who chooses the validators for each SCAMMPERR process.

B. Sequence Diagram

This diagram is mainly designed to represent the interactions between objects that communicate with each other by sending messages. The sequence diagram is represented in Fig. 7.

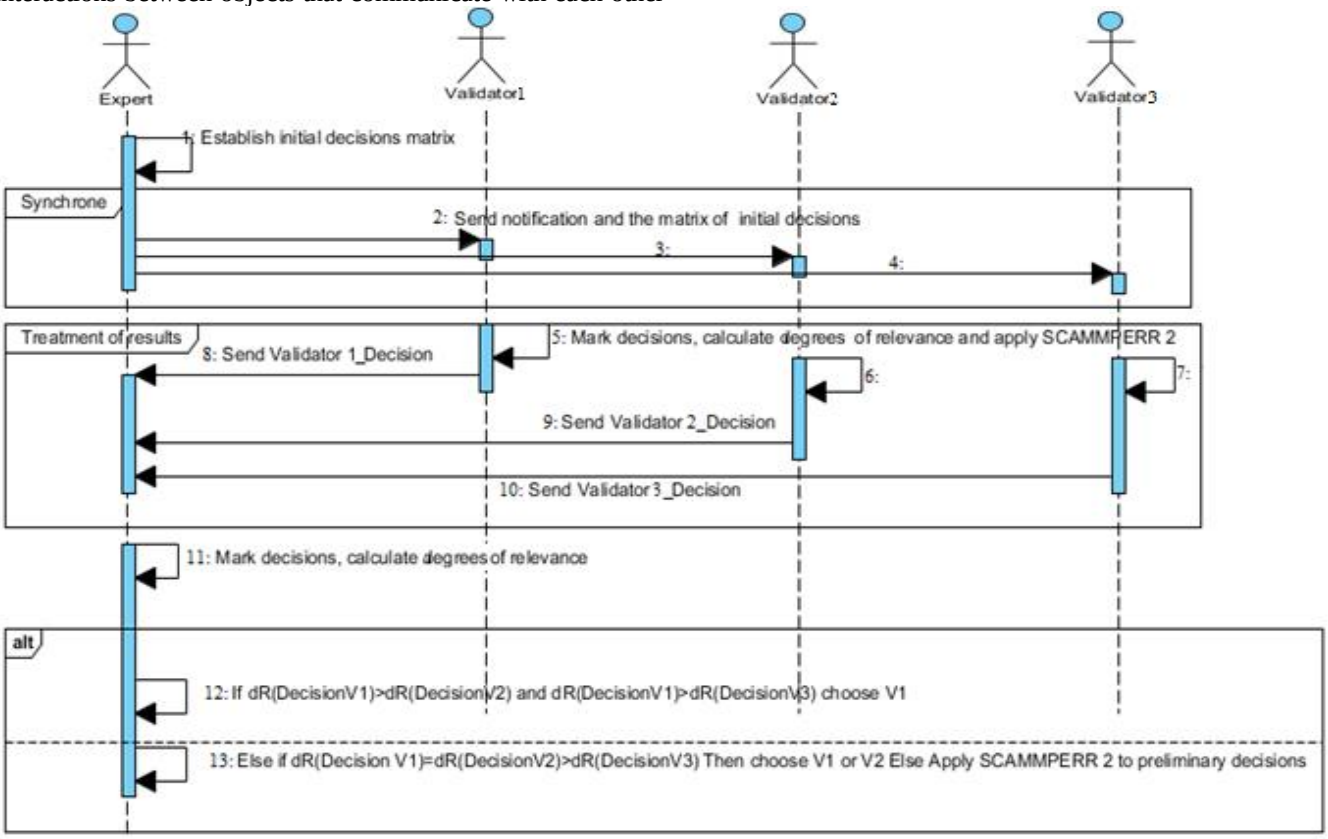


Fig. 7. Sequence diagram of the step of results treatment

After receiving the results, the expert establishes the matrix of the initial decisions, chooses three validators of the domain of the subject or the problem, and sends to them a notification of choice with information about the mission, accompanied by the matrix of the initial decisions. The validators assess decisions (nine decisions by user), then calculate their degrees of Relevance, according to which, the decisions are classified in decreasing order for every SCAMMPERR operator. These decisions will subsequently undergo a SCAMMPERR2 process, for example:

- Eliminate those with a low dR.
- Keep those which have a high dR as reference to which adapt (Adapt) those having a lower but a reasonable dR, or to mix them (Combine), with other decisions with a lower dR.
- The good decisions but which are badly positioned regarding to the handled subject, can be modified (Modify) or (Put to another use) for subsequent questionnaires.

Therefore, every validator makes a decision based on SCAMMPERR2, for every SCAMMPERR operator, which forms a vector of preliminary decisions to transmit to the expert, who evaluates the decisions of the three validators and calculates their dR before making a final decision, for example:

IF dR(Decision(V_1)) > dR(Decision(V_2)) and dR(Decision(V_1)) > dR(Decision(V_3)) THEN choose Decision(V_1)
ELSE IF dR(Decision(V_1)) = dR(Decision(V_2)) > dR(Decision(V_3)) THEN choose Decision(V_1) or Decision(V_2)
ELSE Apply SCAMMPERR2 to preliminary decisions

VII. USE CASE

A company wants to change its advertising slogan, or create a new one. An expert in the field is convened, he can be an employee of the enterprise or not.

The expert develops a SCAMMPERR questionnaire which corresponds to the problem and publishes it on the DBlog of the company which can be reached from the official site or from all the types of social platforms of the web 2.0.

The representation of the online questionnaire is illustrated in Fig. 8.

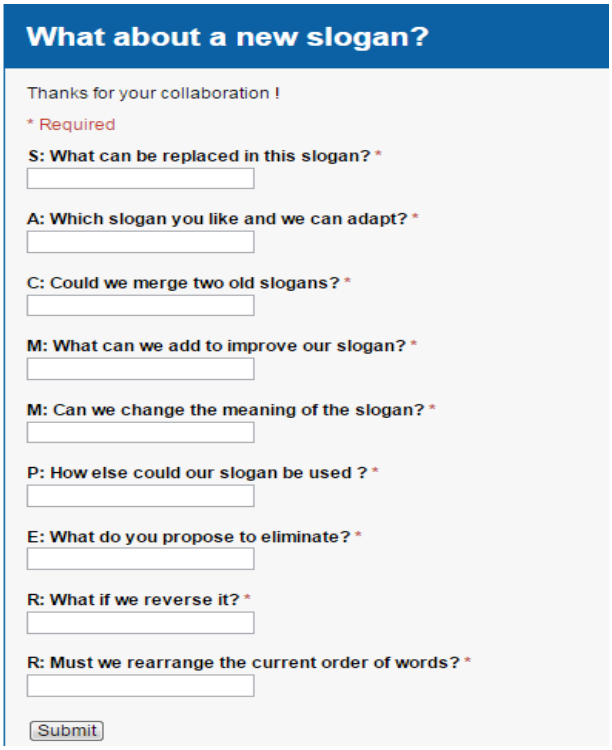


Fig. 8. A SCAMMPERR online questionnaire about a Slogan renovation process

After a certain period, the expert filters the results and creates a matrix of initial decisions. As for an example of 5 users, results are shown in Table VI.

TABLE VI. MATRIX OF INITIAL DECISIONS

User	Decisions
User 1	$(d_{1S}, d_{1C}, d_{1A}, d_{1M}, d_{1M}, d_{1P}, d_{1E}, d_{1R}, d_{1R})$
User 2	$(d_{2S}, d_{2C}, d_{2A}, d_{2M}, d_{2M}, d_{2P}, d_{2E}, d_{2R}, d_{2R})$
User 3	$(d_{3S}, d_{3C}, d_{3A}, d_{3M}, d_{3M}, d_{3P}, d_{3E}, d_{3R}, d_{3R})$
User 4	$(d_{4S}, d_{4C}, d_{4A}, d_{4M}, d_{4M}, d_{4P}, d_{4E}, d_{4R}, d_{4R})$
User 5	$(d_{5S}, d_{5C}, d_{5A}, d_{5M}, d_{5M}, d_{5P}, d_{5E}, d_{5R}, d_{5R})$

The corresponding matrix is:

$$M(U,O) = \begin{pmatrix} d_{1S} & \cdots & d_{1R} \\ \vdots & \ddots & \vdots \\ d_{5S} & \cdots & d_{5R} \end{pmatrix} \quad (5)$$

The matrix is transferred to the validators, who each, gives a mark for every decision and calculates its degree of Relevance taking into account the degree of Influence of each user.

As an example, the validator 1 establishes the results summarized in Table VII.

TABLE VII. SCORES AND DR OF INITIAL DECISIONS

User	dInf	Scores	dR(Decisions)
1	1	(5,6,7,9,7,7,8,5,7)	(5,6,7,9,7,7,8,5,7)
2	2	(6,4,6,6,2,7,5,7,8)	(12,8,12,12,4,14,10,14,16)
3	1	(7,8,6,9,5,8,8,8,9)	(7,8,6,9,5,8,8,8,9)
4	1	(1,3,4,4,2,1,2,3,3)	(1,3,4,4,2,1,2,3,3)
5	2	(9,9,9,7,9,9,9,8,9)	(18,18,18,14,18,18,18,16,18)

In the case of the substitution operation with the operator S, the obtained results are shown in Table VIII.

TABLE VIII. ACTIONS FOR THE SUBSTITUTION OPERATOR

User	dR of Substitute Decision	Actions
1	5	Eliminate
2	12	Mix with D(User5)
3	7	Save it if it is good for another purpose
4	1	Eliminate
5	18	Mix with D(User2)

The process is the same for all the operators, so a single decision is obtained for every operator. Thus, the result is a single SCAMMPERR vector by validator. Then, the expert assesses these three decisions and chooses the most relevant as definitive decision. Also, he can combine the two best decisions to obtain the final one, or apply SCAMMPERR2.

VIII. CONCLUSIONS AND FUTURE WORKS

Nowadays, the evolution of the web affects all areas and in particular, the approach of the decision making. In this paper, the objective of the proposed method is to find a way that promotes the combination of the web 2.0 and the business intelligence concepts by providing a new mechanism of decision making based on the integration of new technologies and tools of web 2.0.

This new system allows involving web users in the decision-making process of the enterprise, which generates a decision based on a collective intelligence strengthened by the use of a rigorous method of stimulation and generation of ideas. Thus, get more innovative and more relevant and fast decisions. The general process provides an independent and a flexible tool to generate significant decisions based on the exploitation of the web 2.0 data, especially through the social channels such as socials networks, blogs, etc. The process of the generation of decisions is characterized by a reduced time execution on demand and as needed. So, it can be executed at any time to get fast and relevant decisions. The strength of this

tool resides in the fact that it brings benefits on several levels, namely the technical, the economic and the organizational levels of the enterprise. So the decisions reached are more relevant, which helps in improving the overall performance of the organization.

As perspectives, the aim is to generalize the use of this tool by adapting it to other areas and other web 2.0 platforms. Also, to plan to extend the research by handling other components like semantics and integration.

REFERENCES

[1] T. O'Reilly, What is web 2.0. Sebastopol, CA: O'Reilly Media, Inc., 2009.

[2] J. Musser, T. O'Reilly, and O'Reilly Radar Team, Web 2.0 Principles and Best Practices. Sebastopol, CA: O'Reilly Media, Inc., 2006.

[3] J. Governor, D. Hinchcliffe, and D. Nickull, Web 2.0 Architectures: What entrepreneurs and information architects need to know. Sebastopol, CA: O'Reilly Media, Inc., 2009**.**

[4] B. Bacon, "Web 2.0 System Architecture Guidelines: Overview and Source Documentation," PIIM RESEARCH, published October 30, 2008.

[5] G. Solomon, and L. Schrum, Web 2.0: New Tools, New Schools. Interntl Soc Tech Educ., 2007.

[6] Z. Itahriouan, N. Aknin, A. Abtoy, and K. E. El Kadiri, "Building a Web-based IDE from Web 2. 0 perspective," *International Journal of Computer Applications*, *96*(22), 2014.

[7] R. Kumar, J. Novak, and A. Tomkins, "Structure and Evolution of Online Social Networks*,"* in *Link Mining: Models, Algorithms, and Applications*. New York: Philip S. Yu, Jiawei Han, and Christos Faloutsos, Ed. Springer, 2010, pp. 337-357.

[8] P. D. Duffy and A. Bruns, "The use of blogs, wikis and RSS in education: A conversation of possibilities," in *Proc. Online Learning and Teaching Conference*, 2006, pp. 31-38.

[9] B. Lamb, "Wide Open Spaces: Wikis, Ready or Not," *EDUCAUSE Review*, vol. 39, no. 5, pp. 36–49, Sep/Oct. 2004.

[10] E. Coghlan, D. Futey, J. Little, C. Lomas, D. Oblinger and C. Windham, "ELI Discovery Tool: Guide to Podcasting," *ELI Publications, EDUCAUSE*, 2007.

[11] S. Chaudhuri, U. Dayal and V. Narasayya, "An overview of Business intelligence technology," *Communications of the ACM*, vol. 54, no. 8, pp. 88-98, Aug. 2011.

[12] Z. Gill, "User-driven collaborative intelligence: social networks as crowdsourcing ecosystems," in *Proc. CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 2012, pp. 161-170.

[13] G. S. Nelson, "Business Intelligence 2.0: Are we there yet?," in *Proc. SAS Global Forum*, 2010.

[14] R. K. Conyne, ed. The Oxford Handbook of Group Counseling. Oxford University Press, 2010.

[15] R. Elmansy, "A Guide to the SCAMPER Technique for Creative Thinking," http://www.designorate.com/a-guide-to-the-scamper-technique-for-creative-thinking/, April 10, 2015, [visited on 30/04/2015].

[16] O.H. Kwon and K.S. Song, "Enhancement idea conception and creative expression in vocational high schools," in *Proc. EDULEARN13*, 2013, pp. 6425-6430.

[17] M. Luckie, "Scamper: a scalable and extensible packet prober for active measurement of the internet," in *Proc. 10th ACM SIGCOMM Conference on Internet measurement*, 2010, pp. 239-245.

[18] A.F. Osborn, Applied Imagination. Oxford: Scribner's, 1953.

[19] B. Eberle, Scamper on: For Creative Imagination Development. Cheltenham : Hawker Brownlow Education, 1990.

[20] B. Eberle, Scamper on: Games for Imagination Development. Texas: Prufrock Press Inc, 1996.

[21] R.E. Glen, Scamper for student creativity. Education Digest, 62(6), 1997, pp. 67–68.

[22] H. Hoover, "SCAMMPERR– The Series On Creativity," http://www.my-creativeteam.com/blog/scammperr-the-series-on-creativity/, February 21, 2013, [visited on 10/04/2015].

[23] M. Michalko. Thinkertoys: A handbook of creative-thinking techniques. New York: Ten Speed Press, 2010.

[24] J. Kilbride, Making Better. Kilbride Consulting, Inc., 2003

[25] "SCAMMPERR - As a creativity tool," http://www.trainingcoursematerial.com/free-training-articles/creativity-problem-solving-decision-making-and-lateral-thinking/scammperr-creativity-tool, 05 March 2014, [visited on 06/07/2014].

[26] C. Larman, Applying UML and Patterns: An Introduction to Object-Oriented Analysis and and Iterative Development. New Jersey: Prentice Hall, 2004.

[27] D. Ohst, M. Welle, and U. Kelter, "Differences between versions of UML diagrams," in *Proc. ACM SIGSOFT Software Engineering Notes*, 2003, pp. 227-236

Intelligent Sensor based Bayesian Neural Network for Combined Parameters and States Estimation of a Brushed DC Motor

Hacene MELLAH
Electrical Engineering Department,
Ferhat Abbas Sétif 1 University,
LAS laboratory, Sétif, Algeria

Kamel Eddine HEMSAS
Electrical Engineering Department,
Ferhat Abbas Sétif 1 University,
LAS laboratory, Sétif, Algeria

Rachid TALEB
Electrical Engineering Department,
Benbouali Hassiba University of
Chlef, Chlef, Algeria

***Abstract*—The objective of this paper is to develop an Artificial Neural Network (ANN) model to estimate simultaneously, parameters and state of a brushed DC machine. The proposed ANN estimator is novel in the sense that his estimates simultaneously temperature, speed and rotor resistance based only on the measurement of the voltage and current inputs. Many types of ANN estimators have been designed by a lot of researchers during the last two decades. Each type is designed for a specific application. The thermal behavior of the motor is very slow, which leads to large amounts of data sets. The standard ANN use often Multi-Layer Perceptron (MLP) with Levenberg-Marquardt Backpropagation (LMBP), among the limits of LMBP in the case of large number of data, so the use of MLP based on LMBP is no longer valid in our case. As solution, we propose the use of Cascade-Forward Neural Network (CFNN) based Bayesian Regulation backpropagation (BRBP). To test our estimator robustness a random white-Gaussian noise has been added to the sets. The proposed estimator is in our viewpoint accurate and robust.**

***Keywords*—DC motor; thermal modeling; state and parameter estimations; Bayesian regulation; backpropagation; cascade-forward neural network**

I. INTRODUCTION

We said that when we can measure a physical quantity, we know something about it, but when we cannot quantify it, our knowledge about it is very poor and insufficient, so without quantifying science does not advance.

The DC motor speed controllers frequently use feedback from a speed measuring device, such as a tachometer or an optical encoder [1,2], but this later, adds an additional cost and congestion throughout the installation [2,3], the problems related to the speed measurement are detailed in the [3].

The simplest estimation method is based on the steady-state voltage equation, where the speed is written as a function of armature voltage and current; the peaks due to converter especially in the transient state affect this speed and the link resistance-temperature is ignored on the other hand, it is the two major inconvenient of this method [1].

R. Welch Jr. *et all* [4] discuss the temperature effects on electrical and mechanical time constants, he prove that these time constants are not constant value, in addition the motor's electrical resistance and its back EMF are depend on temperature.

In [5-8] we find several methods about DC machine temperature measurement, but the problems of temperature measurement are more complicated and difficult to solve than the speed measurement problems, since, the rotor is in rotation. The temperature variation is strongly nonlinear depend on the load, the supply quality, the cooling conditions, the design and the environment conditions. Actually, the problems of armature temperature measurement are not totally resolved.

In literature [9-10], a finite element method (FEM) was usually used to obtain generally a 3D thermal distribution in all electrical machine point. The major advantage of this method is that is suitable to help a designer to optimize the cost, weight and cooling mode in the goal to increase the efficiency and motor's lifetime [10], generally, the FEM is hard to implement in real time both for the control or monitoring, on the other hand, this approach has an enormous resolution time.

According the literature [11-15], we can distinguish two types of electrical machines thermal modeling approaches:

The first one is thermal model-based approaches, this approach based to divide the machine into homogeneous components unscrewed in order to ensure each part has uniform thermal characteristics such as thermal capacitors, thermal resistances and heat transfer coefficients [11, 15]. The identification of the model is performed either by the finite element technique or by a high range of temperature measurement. These models are generally very detailed so, too complex for real time application [16], however, many researchers simplify this model for the real time applications [15, 17]. This approach is robust, unfortunately this model is not generalized and a few measurements are needed for each motor [11, 16].

The second one is the parameter-based approaches, this approach based to get the temperature from the online resistance estimation [12-14] or identified [18-19]. Therefore, the estimate temperature takes under consideration the thermal environmental conditions. This method can respond to changes in the cooling conditions, and is accurate, but it is generally too sensitive [20].

This work was supported in part by Electrical Engineering Department, Ferhat Abbas Setif1 University (UFAS1) and in other part by Algerian ministry of research and High education.

P.P. Acarnley *et all* [1] proposes an Extended Kalman Filter (EKF) is implemented to estimate both the speed and armature temperature, but the EKF has problems with the matrices initialization step for each machine, therefore, the risk of divergence is not very far and not forgotten its dependence of the mathematical model.

R. Pantonial *et all* [21] propose the using of EKF in two steps, the first one is in the steady-state used to estimate the electromechanical behavior, and the second one is a transient version used to estimate the thermal behavior. However, in this case, the system is decoupled and the temperature effect on the resistance is not into account for the steady-state model.

A new nonlinear estimation strategy is proposed in the recent paper in this field based on combining elements of the EKF with the smooth variable structure filter (SVSF) to estimate the stator winding resistance [22], in this research we find only a resistance estimation approach, also the link temperature-resistance is ignored, then this is the simplest estimator version.

M. Jabri *et all* use a fuzzy logic technic to estimate the field and armature resistance of DC series motor, this is an important problem in order to implement a robust closed loop control [23], in their newest version [24], present a comparative study between a Levenberg-Marquardt (LM) and LM with tuning Genetic Algorithms (GA) to adjust relaxation. However, in the two versions, only the resistance and the flux were estimated and the link temperature-resistance is ignored.

The most important electrical machine parameter is the winding temperature, the winding temperature affects both the machine's lifetime and accuracy of control, when the winding temperature is equal or superior to the supported winding insulation temperature, this critical temperature affect directly on the machine lifetime; thus, good knowledge of the thermal state of the machine is very important.

In this context, obtaining the temperature by brittle, expensive sensors and adds a congestion to the overall installation, without forgetting the problematic of the sensor placement, therefore, the sensor is not the right solution [16]. In addition, using a Kalman filter, which is difficult to stabilize and the problematic of covariance matrices choices, remains the two major inconveniences, we propose an intelligent universal estimator based on ANN.

The ANN widely used in different engineering domain, such as renewable energy [25], chemical [26], pharmaceutical [27] and mechanics[28], as well the ANN used in several engineering applications such as control [29], optimization [30], modeling [31] and condition monitoring [32]. In addition, the ANN can used alone [33] or mixed with other technic such as GA [34], Particle Swarm Optimization (PSO) [35] and Fuzzy Logic [36].

One of the most commonly phrased questions in neural computation techniques refers to the size of the network that provides the best results. Although various ‘‘hints and tips’’ like suggestions have been pointed out so far, there is still no clear answer to reply to this question [37,38]. This paper describes and applies an intelligent technique for combined speed, temperature and resistance estimation in a DC machine system.

The use of the proposed method for simultaneous estimation combines many advantages. We don't need to use the speed and temperature sensors, the armature temperature estimation may be used for thermal condition monitoring, and the estimate of speed can be used on speed drive process. The resistance estimation may be used in adaptive calculations in the goal to escape the maladjustment phenomenon of the control by parameter variations such as the PID gain correction. The proposed estimator is suitable both in the drive and in the thermal monitoring.

In section 2, a thermal model of DC motor is presented. In section 3, the DC motor model has been resolved and some simulation results have been presented. In section 4, the ANN topology and design steps have been introduced. In section 5, the simulation studies of ANN estimator is carried out to verify and validate the convergence, effectiveness and estimation quality.

II. Thermal Model of DC Motor

The model used in this paper is illustrated in [1].

A. Electrical equation

$$V_a = R_{a0}(1+\alpha\theta)i_a + l_a\frac{di_a}{dt} + k_e\omega \quad (1)$$

Where *Va* is armature voltage, R_{a0} is armature resistance at ambient temperature, α temperature coefficient of resistance, θ temperature above ambient, i_a armature current, l_a is armature inductance, *ke* is torque constant, and ω armature speed.

B. Mechanical equation

$$T = k_e i_a = b\omega + J\frac{d\omega}{dt} + T_l \quad (2)$$

Where *b* is viscous friction constant, *J* is total inertia and T_l is load torque.

C. Thermal equations

The thermal model is derived by considering the power dissipation and heat transfer [25]. The power dissipated by the armature current flowing through the armature resistance, which varies in proportion to the temperature can be represented by:

$$P_j = R_{a0}(1+\alpha\theta)i_a^2 \quad (3)$$

The iron loss is proportional to speed squared for constant excitation, this loss variation with speed in the armature body can represent by:

The iron loss is proportional to speed squared for constant excitation multiplied by the iron loss constant k_{ir}, this loss variation with speed in the armature body can represent by:

$$P_{ir} = k_{ir}\omega^2 \quad (4)$$

The power losses include contributions from copper losses and iron losses which frequency dependent:

$$P_l = R_{a0}(1+\alpha\theta)i_a^2 + k_{ir}\omega^2 \quad (5)$$

A simple representation of the assumed DC machine heat flow is given in Fig. 1. Heat flow from the DC motor is either directly to the cooling air with heat transfer coefficient *k*.

Fig. 1. Structure of thermal model of DC motor

The thermal power flow from the DC motor surface that is proportional to the difference temperature between the motor and the ambient air temperature, and the temperature variation in the armature which depends on the thermal capacity *H*.

$$P_l = k\,\theta + H\frac{d\theta}{dt} \quad (6)$$

The effect of the cooling fan is approximated by introducing a speed dependence of the thermal transfer coefficient k_T .

$$k = k_0(1+k_T\omega) \quad (7)$$

When K_o: thermal transfer coefficient at zero speed and is K_T thermal transfer coefficient with speed.

By arranging the previous eqs, we can write:

$$R_{a0}(1+\alpha\theta)i_a^2 + k_{ir}\omega^2 = k_0(1+k_T\omega)\theta + H\frac{d\theta}{dt} \quad (8)$$

The equations system can be written as:

$$\frac{di_a}{dt} = -\frac{R_{a0}(1+\alpha\theta)}{l_a}i_a - \frac{k_e}{l_a}\omega + \frac{1}{l_a}V_a$$

$$\frac{d\omega}{dt} = \frac{k_e}{J}i_a - \frac{b}{J}\omega - \frac{1}{J}T_l$$

$$\frac{d\theta}{dt} = \frac{R_{a0}(1+\alpha\theta)}{H}i_a^2 + \frac{k_{ir}}{H}\omega^2 - \frac{k_0(1+k_T\omega)}{H}\theta \quad (9)$$

III. Simulation Results

The resolution of the equations system (9) in Matlab/Simulink environment with the use of parameters from [1], we get the following results:

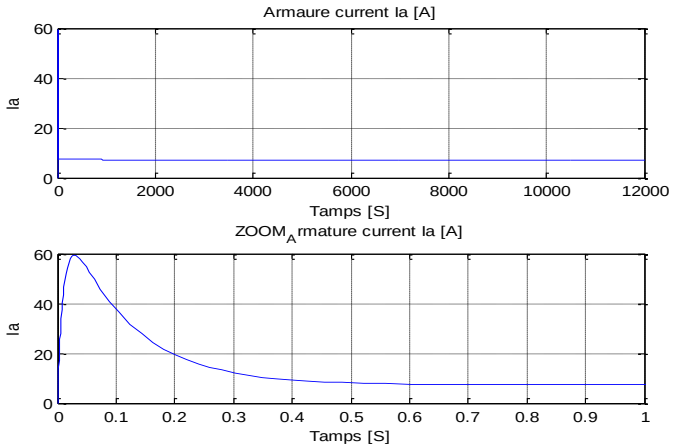


Fig. 2. Armature current

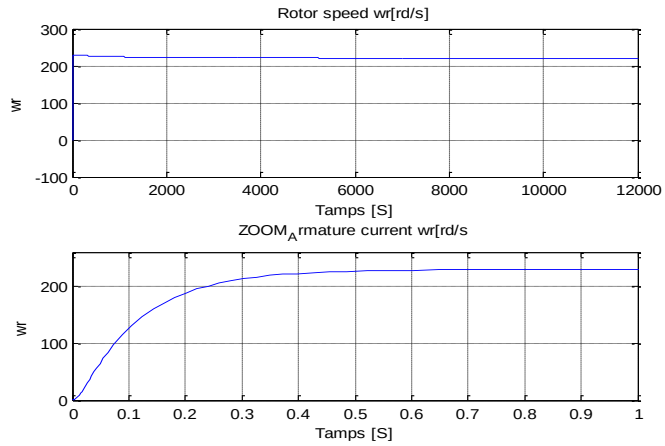


Fig. 3. Rotor speed

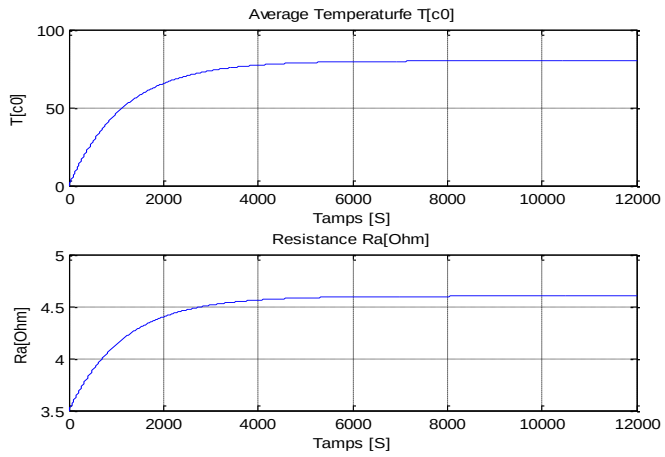


Fig. 4. Average temperature

The current curve variation is illustrated by Fig.2, we can see that in the transient stat the current reach 60A, but in the study state decrease by almost a factor of 10, the final value is 7.27A.

Fig.3 shows DC machine speed variation under load. Fig.4 shows armature average temperature in a brushed DC machine,

this temperature reaches 80 ^{0}C after 140 min, the armature resistance increase 31%.

IV. ANN ESTIMATOR

In this section, an ANN is used in tree steps in order to estimate the speed, temperature and resistance. In this section, we discuss the ANN design step, topology choice and the learning algorithms finally, we applicate the ANN to our study.

A. Types of ANN

Feed-Forward Neural Network (FFNN) is the simplest process neural network. Each subsequent layer in FFNN only has a weight coming from the previous layer. Due to the drawback of this topology structure, FFNN cannot solve some complex problems [38]. The convergence process is slow or even impossible to realize. To address these problems, a CFNN is proposed here.

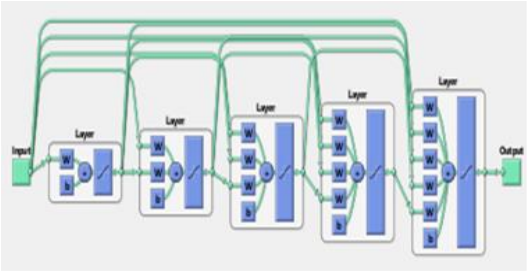


Fig. 5. The structure of the ANN used

CFNN are similar to FFNN, but include a connection from the input and every previous layer to following layers. As with FFNN, a two-or more layer cascade-network can learn any finite input-output relationship arbitrarily well given enough hidden neurons [38-29].

B. Data sets

We have create a Matlab program that breaks the input vector into three parts without losing the information of each part, to make the data obtained by simulation similar than the sensor data a random white-Gaussian noise signal has been added.

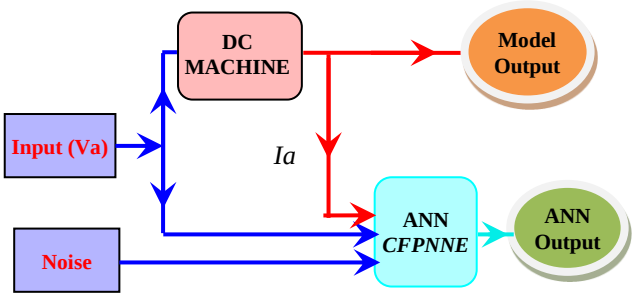


Fig. 6. ANN estimator schemes

This noise make the training very hard and requires a significant time, but the ANN is very trained and applicable on real time. so we have three sets: training, test and validation, each base part in the input vector of a well-defined percentage, 50% occupied by training set, 25% by the testing and 25% by validation set, this data was extracted from Fig. 6.

C. Training

LMBP is the default training function because it is very fast, but it requires a lot of memory to run [38-40]. In our case, we have a very large input vector so the problem of exceed memory is imposed.

We have created a Matlab program for optimize CFNNE performances, such as hidden layer number, number of neurons in each hidden layer, epochs number. For the activation functions, we try deferent functions, but the hyperbolic tangent sigmoid transfer function for the hidden layers and linear transfer function for the output are the best.

Fig. 7 shows ANN estimator used in the present paper.

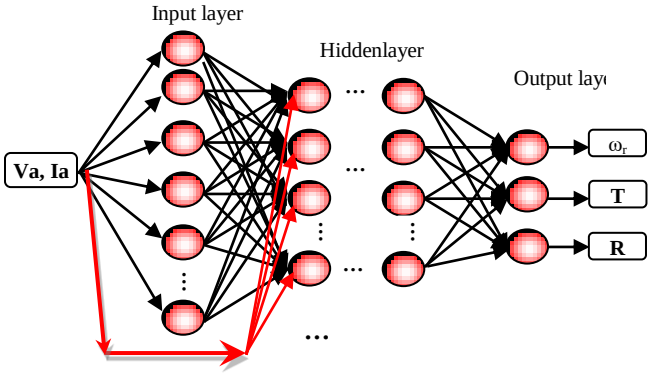


Fig. 7. ANN estimator used in the present paper

V. SIMULATION RESULTS

In this section we following the instructions discussed in the past section for obtained an optimized CFNNE, training step is the most important step to create any ANN, our optimized CFNNE is trained after 2000 epoch at the performance 1.6e-4.

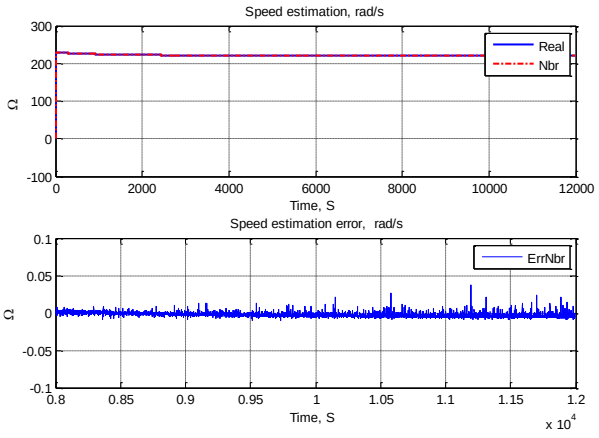


Fig. 8. Speed estimation by ANN

Fig.8 shows DC machine speed estimation and the corresponding estimation error at the testing step, in transient state we can see on the speed estimation error curve's a peak of 110 rad/s between the output of the model and the ANN output, the duration of this peak is 0.3s. In steady state our CFNNE give a good results with estimation error less than 0.04 rad/s that means less than 0.008%.

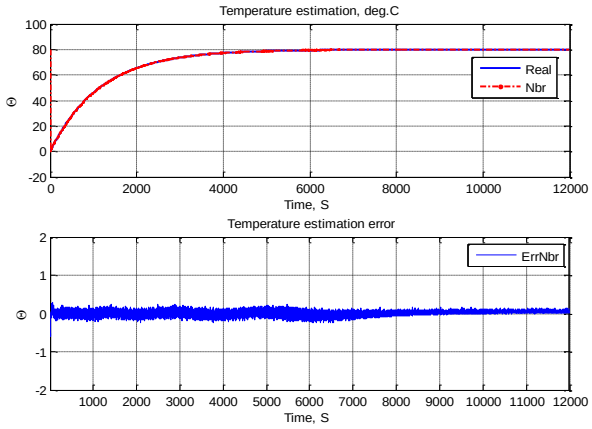


Fig. 9. Average temperature estimation by ANN

The temperature estimation is shown in Fig.9; the temperature value in the DC machine thermal steady state is approximately 80^0C, the corresponding error is less than 0.6^0C so, less than 0.75%.

The CFNNE can also estimate the resistance, this estimation is shown in Fig. 10 the estimation error is less than 0.004Ω.

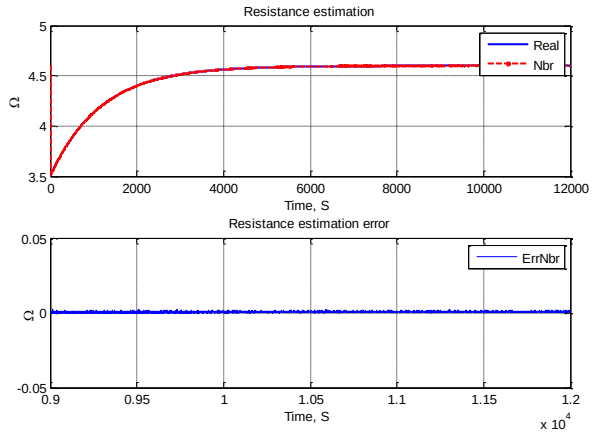


Fig. 10. Armature resistance estimation by ANN

VI. CONCLUSION

A thermal model of DC motor is presented and some results are discussed. The measurement problems and even the use of conventional estimators of speed, temperature and resistance were discussed; the simultaneous estimation of DC machine state variables and parameters not recognized in the literature, our goal is to simulate simultaneously the DC machine speed, temperature and resistance. The ANN makes it possible to achieve this goal, because it enables to estimate simultaneously the speed, temperature and resistance of a DC motor from only the knowledge of voltage and current. The specialized literature we give several ANN versions, according to the studied system characteristics the most suitable approach is CFNNE. The creation steps of CFNNE and the different data bases is discussed in Section IV, the addition of white Gaussian noise to the data set is very important, because if that which make the application in real time is possible and our ANN not be affected by current and voltage measurements noise. It can be seen that the network has worked with an acceptable error. The variable stat estimation may be used in condition monitoring or in robust control, the simulation results demonstrate that the new approach proposed in this paper is feasible.

REFERENCES

[1] P. P. Acarnley, J. K. Al-Tayie, Estimation of speed and armature temperature in a brushed DC drive using the extended Kalman filter, IEE Proc Electr. Power Appl., vol. 144, no. 1, pp. 13–20, Jan I997.

[2] E. Fiorucci, G. Bucci, F.Ciancetta, D. Gallo, C. Landi and M. Luiso, variable speed drive characterization: review of measurement techniques and future trends, Advances in Power Electronics, vol. 2013, pp.1–14, 2013.

[3] G. Bucci, C. Landi, Metrological characterization of a contactless smart thrust and speed sensor for linear induction motor testing, Instrumentation and Measurement, IEEE Transactions on , vol. 45, no.2, pp. 493 – 498, Apr 1996.

[4] R. J. Welch and G. W. Younkin, How Temperature Affects a Servomotor's Electrical and Mechanical Time Constants, Proc. IEEE Ind. Appl. Conference, vol. 2, pp. 1041–1046, 13-18 Oct. 2002.

[5] IEEE Recommended Practice for General Principles of Temperature Measurement as Applied to Electrical Apparatus, IEEE Std 119-1974,1974.

[6] T. Chunder, Temperature rise measurement in armature of a DC motor, under running conditions by telemetry, Proc. Sixth International Conference on Electrical Machines and Drives, pp. 44–48, 8-10 Sep 1993.

[7] L. Michalski, K. Eckersdorf, J. Kucharski, J. McGhee, Temperature Measurement, John Wiley & Sons Ltd, 2001.

[8] I. J. Aucamp, L .J. Grobler, Heating, ventilation and air conditioning management by means of indoor temperature measurements, Proc. 9th conference industrial and commercial use of energy (ICUE), pp. 1–4, 15-16 Aug, 2012.

[9] A. Cassat, C. Espanet and N. Wavre, BLDC Motor Stator and Rotor Iron Losses and Thermal Behavior Based on Lumped Schemes and 3-D FEM Analysis, IEEE Transactions on Industry Applications, vol. 39, no. 5, pp. 1314–1322, 2003.

[10] J. Le Besnerais, A. Fasquelle, M. Hecquet, J. Pellé, V. Lanfranchi, S. Harmand, P. Brochet and A. Randria, Multiphysics Modeling: Electro-Vibro-Acoustics and Heat Transfer of PWM-Fed Induction Machines, IEEE Transactions on Industrial Electronics, vol. 57, no. 4, pp. 1279–1287, 2010.

[11] R. Lazarevic, P. Radosavljevic, A. Osmokrovic, novel approach for temperature estimation in squirrel-cage induction motor without sensors, IEEE Transactions on Instrumentation and Measurement, vol. 48, no. 3, pp. 753–757, 1999.

[12] S. B. Lee, T. G. Habetler, R. G. Harley and D. J. Gritter, A stator and rotor resistance estimation technique for conductor temperature monitoring, Proc. IEEE Ind. Appl. Conference, vol. 1, pp. 381–387, 2000.

[13] S. B. Lee, T. G. Habetler, R. G. Harley and D. J. Gritter, An Evaluation of Model-Based Stator Resistance Estimation for Induction Motor Stator Winding Temperature Monitoring, IEEE Transactions on Energy Conversion, vol. 17, no. 1, pp. 7–15, 2002.

[14] S. B. Lee, T. G. Habetler, An Online Stator Winding Resistance Estimation Technique for Temperature Monitoring of Line-Connected Induction Machines, IEEE Transactions on Industry Applications, vol. 39, no. 3, pp. 685–694, 2003.

[15] K. D. Hurst, T.G. Habetler, A thermal monitoring and parameter tuning scheme for induction machines, Proc. IEEE Ind. Appl. Conference, IEEE-IAS Annu. Meeting, vol. 1, pp. 136–142, 1997.

[16] H. Mellah, K. E. Hemsas, Stochastic Estimation Methods for Induction Motor Transient Thermal Monitoring Under Non Linear Condition, Leonardo Journal of Sciences, vol. 11, pp. 95–108, 2012.

[17] J. F. Moreno, F. P. Hidalgo and M. D. Martinez, Realisation of tests to determine the parameters of the thermal model of an induction machine, IEE Proc Electr. Power Appl., vol. 148, no.5, pp. 393–397, 2001.

[18] R. Beguenane, M.E.H. Benbouzid, Induction motors thermal monitoring by means of rotor resistance identification, IEEE Transaction on Energy Conversion, vol. 14, no. 3, pp. 566-570, 1999.

[19] M.S.N. Saïd, M.E.H. Benbouzid, H–G Diagram Based Rotor Parameters Identification for Induction Motors Thermal Monitoring, IEEE Transactions on Energy Conversion, vol. 15, no. 1, pp. 14–18, 2000.

[20] Z. Gao, T. G. Habetler, R. G. Harley and R. S. Colby, An Adaptive Kalman Filtering Approach to Induction Machine Stator Winding Temperature Estimation Based on a Hybrid Thermal Model, Proc. IEEE Ind. Appl. Conference, IEEE-IAS Annu. Meeting, vol. 1, pp. 2–9, 2005.

[21] R. Pantonial, A. Kilantang and B. Buenaobra, Real time thermal estimation of a Brushed DC Motor by a steady-state Kalman filter algorithm in multi-rate sampling scheme, Proc TENCON 2012 IEEE Region 10 Conference, pp. 1–6, 19-22 Nov 2012.

[22] W. Zhang, S. G. Andrew and R.H. Saeid, Nonlinear Estimation of Stator Winding Resistance in a Brushless DC Motor, Proc American Control Conference (ACC), pp. 4699-4704, 17-19 June 2013.

[23] M. Jabri, I. Chouire and N.B. Braiek, Fuzzy Logic Parameter Estimation of an Electrical System, Proc. International Multi-Conference on Systems, Signals and Devices, pp.1–6, 2008.

[24] M. Jabri, A. Belgacem and Houssem Jerbi, Moving Horizon Parameter Estimation of Series Dc Motor Using Genetic Algorithm, Proc. International Multi-Conference on Systems, Signals and Devices, pp. 26–27, 2009.

[25] S. A. Kalogirou, Artificial neural networks in renewable energy systems applications: a review, Renewable and Sustainable Energy Reviews, vol. 5, no. 4, pp.373–401, 2001.

[26] E. Byvatov, U. Fechner, J. Sadowski and G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, Journal of Chemical and modeling, vol. 43, no. 6, pp. 1882–1889, 27 Sept, 2003

[27] S. Agatonovic-Kustrin, R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, Journal of pharmaceutical and Biomedical Analysis, vol. 22, no. 5, pp. 717–727, 2000.

[28] S. Ablameyko, L.Goras, M. Gorz and V. Piuri, Neural Networks for Instrumentation, Measurement and Related Industrial Applications, IOS Press, 2003.

[29] S. Haykin, Kalman filtering and neural networks, John Wiley & Sons, 2001.

[30] A. Cochocki, R. Unbehauen, Neural networks for optimization and signal processing. John Wiley & Sons, Inc, 1993.

[31] M. Y. Chow, Y. Tipsuwan, Neural plug-in motor coil thermal modeling, in Industrial Electronics Society, 2000. IECON 2000. 26th Annual Conference of the IEEE, vol.3, no., pp.1586–1591, 2000.

[32] L. P. Veelenturf, Analysis and applications of artificial neural networks, Prentice-Hall, Inc., 1995.

[33] M. Gupta, L. Jin and N. Homma, Static and dynamic neural networks: from fundamentals to advanced theory, John Wiley & Sons, 2004.

[34] L. C. Jain, N.M. Martin, Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications, vol. 4, CRC press. 1998.

[35] R. C. Eberhart, J. Kennedy, A New Optimizer Using Particle Swarm Theory, Proceedings of the Sixth International Symposium on Micro Machine and Human Science, MHS '95, vol.1, pp. 39–43. 1995.

[36] J.S.R. Jang, C.T. Sun and E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice-Hall, 1997.

[37] C. Dimoulas, G. Kalliris, G. Papanikolaou, V. Petridis and A. Kalampakas, Bowel-sound pattern analysis using wavelets and neural networks with application to long-term, unsupervised, Expert Systems with Applications, vol. 34, no. 1, pp. 26–41, 2008.

[38] B.M. Wilamowski, How to not get frustrated with neural networks, Proc. IEEE Int. Conf. Ind. Technol, pp. 5–11., 2011.

[39] Zhou Yao-ming, Meng Zhi-jun, Chen Xu-zhi and Wu Zhe, Helicopter Engine Performance Prediction based on Cascade-Forward Process Neural Network, IEEE Conference on Prognostics and Health Management (PHM), pp, 1–5, 18-21 June 2012.

[40] H. Demuth, M. Beale and M. Hagan, Neural Network Toolbox Users Guide, the MathWorks, Natrick, USA. 2009.

Social Computing: The Impact on Cultural Behavior

Naif Ali Almudawi
Department of Computer Science and Information System
Najran University
Najran, Saudi Arabia

***Abstract*—Social computing continues to become more and more popular and has impacted cultural behavior. While cultural behavior affects the way an individual do social computing, Hofstede's theory is still prevalent. The results of this literature review suggest that, at least for several cultural dimensions, some adjustments may be required to reflect current time and the recognition of the role of technology nowadays. Thus, today, social computing has evolved into continuous communication and interaction of many culturally diverse users.**

Keywords—social computing; Web 2.0; cultural behavior; culture; Power distance; Individualism vs. collectivism; masculinity vs. femininity; uncertainty; avoidance and time horizon

I. Introduction

A. Social computing

Social computing has been defined a number of ways by many different people, both scholars and practitioners. According to [9] social computing can also be defined as a communication that is computer-mediated and facilitates the interaction between how humans coordinate, collaborate and distribute news. Another more recent definition of social computing is that social computing is considered an area in the field of computer science that displays the connection of social behavior and computational systems [2]. Definition of social media, or Web 2.0 technologies, is Information and Communications Technology (ICT) that helps to advance both knowledge sharing and learning [18]. Social computing has the ability to promote a transition from a broadcast model of communication to a many-to-many model that allows individuals to converse and receive wisdom from others [20]. Social computing is interactive and collaborative behavior between technology and people. Personal computing is an individual user activity in that one user generally, commands computing. In social computing, the Internet allows users to interact through many mediums, including: Social media sites, Blogs and Wikis [69].

Organizations can use applications such as RSS feeds, podcasts, and blogging to rapidly push content to subscribers all over the world [3],[4],[17],[21]. Web 2.0, or social computing, could also be defined as a conceptual framework where a group of web-based tools could help users collaborate on tasks, interact in social networks, work and rework existing content, as well as share a host of information [6],[12]. All of these help in understanding what is meant by social computing; however, there are many types of social computing tools that are discussed in the literature review proper section.

Thus, social computing, often referred to as Web 2.0 or social media, is continuing to emerge as a new field of computing systems used for modeling social behavior through the use of software and technology. There are many different types of social computing technologies to include blogs, email, wikis, social networking, instant messaging, social bookmarking, and various others. Defined social computing as an enabler for people all around the world to communicate and share information instantly with a common interest or goal and with minimal costs [2],[3]. Some of the most important characteristics of social computing can be summarized as user-created content where users can control the data, unique sharing of content or media, the ability to tap into the intelligent minds of other users collectively, unique communication and collaborative environment, major emphasis on social networks, and an interactive, user-friendly interface [14].

B. Culture

Culture has been defined in a different way, which refers to the cumulative deposit of knowledge, understanding, principles, values, attitudes, religion, roles, concepts of the life, and possessions acquired by a group of people in the course of generations through group striving and individual [68].

Current research in Computer Information Systems (CIS) has examined the effect of culture in the adoption and use of different social computing technologies. However, research examining the impact of social computing on culture is more limited, albeit increasingly common in recent years. Search engine portals and e-commerce sites are universal internet destinations. Search engines assist in retrieving on-line information, regardless of the location or cultural background of the users. Most search engine technologies were originally developed in the United States, and therefore, intentionally or not, there are designs features embedded in these systems that reflect values that are characteristic of American culture.

Consequently, the cultural background of an individual affects on-line behavior. Nationality, a proxy for culture, has been reported to affect on-line behavior [16]. While the technology is identical everywhere, the way users perceive and use a particular technology varies. Some search engine users may select and emphasize using particular features of the search engine, while others may choose other features. For example, it has been observed that queries placed by users in the United States are likely to contain more operators (include, exclude, wildcard, etc.) than queries placed by users in Europe [53]. Furthermore, culture influences the perception that users develop about systems. Such is the case with:

- Social computing user perception of the effort required to use the search engine;
- Performance gain obtained when using a search engine; Other's perception of how the system should be utilized and;
- Perception of the individual about the support provided by the surrounding organization infrastructure to use the search engine.

These are all examples of variables that are influenced by a user's culture. System usage is also influenced by variables which include age, gender, and *0*experience [19]. Cultural background also influences the use of a system, as evidenced by the literature that examines system use and culture. According to [11] cultural background is composed of different dimensions, including individualism/collectivism, time orientation, power distance, masculinity/femininity, and uncertainty avoidance. These dimensions are described below in this literature review. This research seeks to understand how social computing impacts the variables identified by Hofstede.

The primary objective of this literature review is to investigate the impact of social computing on cultural behavior. In this review I highlight the prevalent position that social computing plays an impact on the cultural behavior of all human endeavors. Following this, I provide a brief definition of social computing and cultural behavior to justify the important effect that it has in all human endeavors.

II. Literature Review

Two areas of investigation make up the focus of this research: social computing and cultural behavior. Thus, the goal of this paper is to provide a review of previous work on both of these domains. We also look at the field of information retrieval, which precedes search engine research. The most significant frameworks proposed to explain cultural behaviors are reviewed, along with the seminal research that grounds this field. Computer information systems research using these well-accepted frameworks is also reviewed.

A. Overview of Social Computing Tools

Social computing, often referred to as social media or Web 2.0, has evolved greatly since 1966 with the ability to transfer Email messages between users on different computers [9]. Nowadays, there are various types of social computing tools that are used every day by organizations and individuals to include social networks, blogs or weblogs, wikis, instant messaging, and similar tools [7].

According to [1] a vast majority of these technologies are used to improve collaboration and communication efforts within most organizations. The transformation of the Internet with the introduction of social computing has been able to allow passive citizens because active content creators while also providing a greater sense of interactivity [10].

1) Social networks

Social networks are used all over the world to help people connect, meet, and share amongst each other. [8] Described social networking as the way people connect with one another through friendships, common interests, or ideas. Social networking applications can provide a collaborative work environment where individuals can share knowledge and ideas quickly and conveniently [13]. In addition, these can also allow one to quickly gather information about who they know and what they know in organizations [5]. These types of networks are said to exist because humans require social relationships with other humans for survivability.

Social networking sites are basically web-based services that allow individuals to do three things: (a) develop a public or semi-public profile within a system, (b) specify a list of other uses with whom they share a common connection, and (c) view and navigate through their list of connections and those made by others within a particular system [26] There are many applications social networks could be used for, such as a collaboration tool for education and as well as fighting crimes (World Future Society, 2010). Individuals typically create accounts on social networking sites to set up their personal profile. This profile showcases their profile to their online network of 'friends' or peers, many of whom they have pre-existing off-line relationships with. Through this initial network, individuals can then gain access to their friends' networks of friends, colleagues, and/or peers so that individuals are open to an array of diverse content through the weaker relationship ties [22]. Although these connections may vary from site to site, these social networking sites mainly help establish or maintain a means of communication through many networks.

Social networking sites are often used to catch up on personal information and current activities of those who have social ties. According to [23] users of social networking sites are usually readily disclose of private information for enjoyment and also for the convenience of establishing and maintaining friendships. They are not only used for social and playful uses, but also used as sources of information and productivity for those business-oriented social networking sites such as LinkedIn and Beehive [24].

Many users of social networking sites use these sites to connect with friends and colleagues they may have previously known [22]. 'Face to face' communication happens less often because of life's circumstances and the limited amount of free time available. Virtual communication is becoming increasingly popular because people spend a lot more time on the Internet with most of that time being spent on social networking sites. According to [25] a study conducted by blog.compete.com in 2011 revealed 75% of the time users spend on the Internet is being used for social networking. This shows just how much social networking has become a part of everyday life for many people.

There are multiple social networking tools that can be found on the Internet to include Facebook, Twitter, LinkedIn, Myspace, and much more. Although Twitter is a popular social networking tool used by many, Facebook stands out among the rest with over 1 billion users worldwide (Facebook, 2013). Those individuals that frequent Facebook tend to have a high level of trust in the site [27]. Many of these social networking sites are mostly centered around users so that the connections of the users could potentially reach larger

audiences with low costs.

2) Blogs and weblogs

Some of the most visible social computing applications are blogs. Blogs, which began in the late '90s, may be thought of as online journals in which individuals or small groups can publish. They are used to express opinions and share knowledge on any topic in a sequential format that is very similar to a personal diary. The archival, search, and categorization features in blogs help organize the content and retrieve specific information [30]. Those blogs that are popular attract many users that will engage in discussions thereby creating networks of blogs and online communities. Although some blogs are confined to personal expressions of a single person, others tend to stimulate reactions and comments from the readers. Because blogs can be used to convey different types of information, such as personal, public, commercial, and political, it has become a very effective communication tool that is constantly used over the Internet.

Blogs are fundamentally different from how they use to be, and many industries see them much differently from other industries. Research has shown that employees think that blogs are more effective than the traditional forms of communications such as emails or newsletters because they have the opportunity to comment, formulate ideas, and facilitate discussions publicly within an organization [34].

Some may see blogs as a good place to share knowledge, while others user blogs to be able to express themselves and feel a sense of empowerment. Blogging seems to make people more thoughtful and articulate observers of what's going on around them. Users can typically use a web browser to create conversations and reflections with respondents [47]. Blogs also offer the ability to do RSS feeds, which push new postings and reader comments to users automatically through syndicating and aggregating information [28]. According to [29] bloggers typically are motivated to publish information for various reasons to include self-presentation, relationship management, keeping up with trends, sharing information, storing information on the internet, entertainment, and for showing off. Many are able to take part in blogging because the software used to blog has become more technically advanced to allow web pages to be updated rapidly and easily.

3) Wikis

Wikis are another social computing approach used by many to manage web-based content or for collaborating with others. A wiki is a set of linked Web pages that are created incrementally by a group of collaborating users [30]. Wikis are similar to discussion forums and blogs in some ways because the most recent version reflects the cumulative contributions of all authors [32]. Wikis also allows users to see a history of changes, and if needed, it has the ability to revert pages to previous versions. A simplistic way of describing a wiki is that it is a "web based program that allows viewers of a page to change the content by editing the page online in a browser" [33].

Wiki, derived from the Hawaiian term Wikiwiki meaning "fast", was first developed in 1995 by Ward Cunningham to communicate specifications for software design [30]. Since it was developed, it has become an increasingly popular tool used by many for knowledge sharing and collaboration. One of the most visible instances of the wiki concept is Wikipedia, which provides primers on a wide variety of terms and names. Some key issues surrounding Wikipedia is the quality and credibility of the information being posted [36].

According to [67] warned that wiki users using the created web pages as a sole source of data could potentially diminish cognitive and affective learning outcomes that are assigned as a characteristic of wiki. A wiki can also be used as a collaboration tool. Many researchers have noticed the benefit of using wikis for student-to-student collaboration from elementary to graduate schools [64], [65].

The term "wiki", according to [37] generates roughly 436 million items on the Google search engine. More than 2.8 million English-language articles are in Wikipedia with more than 250 languages represented. The authors also determined that there were just below 13 million users of Wikipedia's English-language, which demonstrates just how important wikis are to users around the world. One of the main uses of a wiki, according to [39] is a content repository where wiki users can contribute their experiences and other content. They can also be used for organizational portals, for managing projects, and for creating a knowledge-base. Because of wikis are Internet-based, much of the content can either be extended within an organizational context or externally for customers and business partners.

Wikis can allow students the ability to share information interactively while fostering the vision of negotiated meaning, knowledge construction, and learner-to learner interaction [49]. Also reported how wikis and other social computing technologies could improve team collaboration, thereby enhancing learning among students. Explored the effect of using wikis on collaborate writing by using two writing tools (a wiki web site and MS Word) and three user modes (face-to-face, synchronous distribution, and asynchronous distribution). When comparing MS Word to the wiki web site, the authors found that the face-to-face collaborative writing sessions with wikis led to greater levels of participation. Wikis also produced documents with higher quality and provided greater satisfaction for the contributors. [37].

The private sector is increasingly engaging in the use of wikis to help influence business through innovative ideas and knowledge sharing [63]. According to [62] surveyed 168 corporate wiki users to determine how many are using wikis in a corporate setting. These experienced wiki users spent an average of 15 month contributing to company wikis and about 26 months contributing to wikis in general on average. The authors found that some of the most common activities wikis were used for included software development, e-learning, project management, posting of general information and knowledge management, communities of practice and user groups, ad-hoc collaboration, tech support, marketing and customer relationship management, resource management, and R&D [62]. The users indicated they benefited from corporate wikis because they enhanced reputation, made work easier, and helped the improvement of organizational processes. These benefits were more likely when wikis were used for those tasks requiring innovative solutions and when

the posted information was from credible sources.

4) Instant messaging

One of the most popular forms of social computing is instant messaging (IM). IM is a computer-based communication with fast transmission times that allows users to type messages to other users in a near-synchronous fashion [40] IM is a unique form of social computing because it allows immediate communication; however, it doesn't provide a lot of information about the user such as the profile pages that are involved with the social networking tool, Facebook. In addition to the immediate communication, many IM systems allow others to know users are currently logged in, how long they have been logged in, and if they are active or inactive. Some systems give users the ability to control who can see them online and also block those one may not want to communicate with. This type of social computing, could also be beneficial for those people that are geographically distant and prefer not to incur the financial expenses of face-to-face meetings.

Previous research indicated instant messaging is used in about 85% of enterprises in North America [50]. IM can function as a task-oriented, communication tool for users in the workplace, while also serving as an informal collaboration tool. Although there are still some organizations that have yet to find the benefit in using IM, there are many who have seen the value and are encouraging employees to use as a means for immediate communication in the enterprise. According to [38] investigated instant messaging to understand the determinants of collective intention, known as we-intention, which represents how someone may perceive a group of people that act as a unit. Based on the critical mass theory and social influence processes, the study's findings illustrated that critical mass influenced we-intention to use instant messaging indirectly and directly through two other factors known as group norm and social identity. The authors noted that understanding and recognizing the importance of collection intention can help managers advance their knowledge beyond that of the individual-based models that are greatly adopted in information systems research.

With the many types of social computing tools available, individuals can effectively collaborate and may change cultural behavior by simply being more innovative and creating an atmosphere that works for one`s needs. As technology continues to change, more cultures are starting to embrace the whole idea of social computing and are working to make that a part of everyday life.

B. Benefits and Challenges of Social Computing

Social computing could be both enriching and challenging for those who utilize these technologies. According to [35] students who have experience using social computing in the classroom typically accept the technology along with its emerging concepts, tactics, and course content available. The authors also explained how social computing could support peer learning. Students had no problems sharing what they learned and provided answers to questions that lessened the strain onfaculty resources. In addition students were also more comfortable asking their peers questions. Blackboard also has a messaging capability that allows students to submit assignments securely and provide a way for faculty to provide feedback in a timely fashion. A ccording to [66] developed a theory that looked at the relationship between emotional capital and internal social media use. Emotional capital was defined in this study as "the aggregate feelings of goodwill toward a company and the way it operates". The authors used comparative case studies and tested this theory using a survey. The findings indicated that executives who utilized social media to build emotional capital within employees were able to benefit in terms of an improvement in information flows, collaboration, lower turnover, and higher employee motivation.

Another potential benefit of social computing in the classroom is the ability of the faculty to manage the students. Recent research has proven how social computing technologies can benefit teaching and learning. Social computing can allow the tracking of student interactions through Blackboard, whichprovides a means for identifying those students who may be failing and to evaluate how the students are. [6]

Some other potential benefits of social computing include having a more flexible organization where employees or students could participate through contributing and providing feedback. Social computing could provide new styles of management where organizations allow the use of social computing for both work and personal use as it was often forbidden in the past. Also social computing could provide new ways to manage digital content by offering new ways of searching, managing, and effectively utilizing the information that is provided. Those organizations that are interested in maximizing the benefits of social computing should seek to integrate these systems with other systems that have similar purposes [44].

C. Use of culture behavior in Information systems Studies

An awareness of social computing and its impact on culture behavior is valuable to the understanding of how social computing technologies are used at the national, organizational, and group level and can have an effect in the implementation and use of social computing technologies [16]. First, finding an objective definition of culture has been an elusive task. In their ample review of culture, Note that there exist countless definitions, which relate to ideologies, beliefs, assumptions, shared values, collective will, norms, practices, symbols, language, rituals, myths, and other elements. Definitions come from multiple disciplines including psychology, sociology, anthropology, communication, linguistics, business, and others [16].

While these myriad of definitions exist, several authors agree that culture manifests itself at different levels. These authors agree that these values and assumptions form over time and are deeply embedded in individuals. In fact, these sets of values are acquired early on in life and generally transmitted by those surrounding an individual since infancy. Furthermore, these values and assumptions form a belief system that defines how individuals perceive and relate to each other and to the physical world, and how schemes and strategies are realized. While external circumstances may change during the life of an individual, this belief systems is

deeply rooted and likely to remain unchanged. In fact, this system is highly internalized by individuals, and it unconsciously influences all activities.

Note that social computing technology is not culturally neutral and "may come to symbolize a host of different values driven by underlying assumptions and their meaning, use, and consequences" [16]. Several definitions of culture have been used in cross-cultural studies in the computer information system literature. Three influential frameworks, those of [11], [51], [52] are cited repeatedly in social computing systems studies dealing with culture. Based on the strong empirical evidence provided, Hofstede's work went on to become ubiquitous within the social computing discipline. According to [51] proposed seven dimensions of culture; some of which overlap with those proposed by Hofstede. The other dimensions proposed dealt with variables not considered in [45] research, such as how individuals from different cultures perceive the world and their surroundings, how individuals from different cultures employ different strategies when thinking and deciding, and how rules and status impact relationships. Table 1 provides a short summary of conceptualization of culture.

TABLE I. TROMPENAARS DIMENSIONS OF CULTURE (ADAPTED)

Dimension	Definition
Universalism vs. Particularism	The extent to which rules and norms apply to everyone equally and the ability to make exceptions for some. Individuals in a society may apply rules and norms equally among all members, regardless of their position, status, or relationship, or may make special exclusions and adjustments for specific cases.
Analyzing vs. Integrating	Starting with the whole and decomposing into parts, or integrating the parts into the whole. Societies may tackle problems by taking a top down, or bottom up approach.
Individualism vs. Communitarianism	The rights and desires of the individual versus the rights and desires of the group. Individuals in a society may be willing (or not) to sacrifice personal goals for the goals of the group.
Inner-directed vs. outer-directed	The search for answers using thinking, intuition, and personal judgment, or to seek data in the outside world. In solving problems, a group may resort to their own insights, or to the physical world and empirical data.
Time as sequence vs. time as synchronization	Events happen in different time periods in a sequential fashion, or events may overlap and occur in parallel. In a society, every event and action is an individual unit that requires exclusive attention, or a individual or group could focus on many events and actions
Achieved status vs. ascribed status	Gaining status and recognition based on effort and performance, or by right Rank and standing is the result of either effort or performance, or it is inherited.
Equality vs. hierarchy	Equality among all members of the group, or ranks that distribute power. The distribution of power is a society may vary by concentrating authority on certain groups or distributing it among members.

Most computer information systems research dealing with cultural behavior will employ one of these frameworks, with Hofstede's dimensions of culture being the most prevalent [16]. Hofstede's dimensions of culture, as the most dominant framework, will be reviewed in the next section.

D. Hofstede's Culture Dimensions

There are multiple conceptualizations of culture. In this review we have presented those that are not only relevant, but have been widely used in computer information systems research. General agreement exists that the most commonly used definition of culture states that culture is "the collective programming of the mind, which distinguishes the members of one category of people from another" [11]. This programming extends from language and symbols to patterns and interactions. Hofstede's conceptualization of culture has been used extensively inside and outside of the field of computer information systems [16].

Hofstede's research involved more than 100,000 respondents from over 70 nationalities and more than 20 languages. The data collected resulted in the development of a model which includes five dimensions which can be used to measure national culture. Hofstede describes these dimensions as Power Distance (PD), Individualism versus Collectivism (IC), Masculinity versus Femininity (MC), Uncertainty Avoidance (UA), and Time Horizon (TH). These dimensions are summarized in Table 2.

TABLE II. HOFSTEDE'S (1980) CULTURAL DIMENSIONS (ADAPTED)

Dimension	Definition
Power distance	The degree to which the less powerful members of a society expect differences in the levels of power [hierarchical (authoritarian) or equalitarian (follower)]. The likelihood that an individual with less power (at a lower point in the hierarchy) can influence decisions made by those with more power (at a higher point in the hierarchy)
Individualism vs. collectivism	The extent to which people are expected to stand up for themselves, or act predominantly as a member of the group or organization. The willingness of an individual to sacrifice their own personal interests for the interests of the group and vice versa.
Masculinity vs. Femininity	The role overlaps that may exist among male and female members of a society.Masculine cultures value competitiveness, assertiveness, ambition, accumulation of wealth, and material possessions.Feminine cultures value relationships, quality of life, commitment, charity, compromise, and relationship building.
Uncertainty avoidance	How societies attempt to cope with anxiety by minimizing uncertainty.The level of risk taking and risk tolerance of a society. The strategies to minimize uncertainty include laws, rules and structures that limit outcomes
Time Horizon	Describes a society's time horizon and the willingness of individuals to sacrifice long-term goals for short-term goals and vice versa.

The national cultural dimensions presented by Hofstede have been used repeatedly in cross-cultural studies in many disciplines, including Computer Information System research

[16]. It is possibly the most cited and used work in the field of cross-cultural research [11]. These variables and dimensions, which distinguish cultures, are described below.

1) Power Distance

Cultural behavior affects the way decisions are made. While the studies reviewed did not examine the impact of power distance on search engine technology, several studies in information systems have linked power distance and participation in Group Decision Support Systems (GDSS). For example, explored whether the use of a GDSS would attenuate power distance. When using a GDSS, all users are presented at the same hierarchical level (organization-wise). If so, users may feel more comfortable expressing opinions. The effect of a GDSS would therefore be more pronounced in cultures with high power distance [56].

Power distances may also influence the process for selecting strategies to deal with complex problems and situations. In low power distance environments, assertive and control-oriented strategies take place more frequently [54]. In high power distance environments, assertive and control those who have a higher hierarchical status only take oriented strategies. In low power distance environments, any individual can propose strategies and take leadership, since decision making power is equal among members of a group.

2) Individualism versus Collectivism

Collectivist cultures tend to approach tasks, problems, and solutions as a group, sharing information in order to make decisions. Individuals from individualistic cultures prefer to undertake problems by themselves. Consequently, there is more shared meaning and common knowledge in an organization composed of collectivist members than in an organization composed of individualistic members. Based on this, we expect members of an individualistic culture to rely more on information systems to obtain information to make decisions than those of collectivist cultures, who gather/share information from/with each other [15].

In collectivist cultures, the amount of shared context or knowledge between participants in a dialogue is significantly higher than in individualistic cultures. In high context cultures, meaning is derived from the context of a communication exchange [42]. For collectivist cultures, where context is high, individuals share a vast array of information, which creates, shared knowledge while in low context communication is predominant in individualistic societies. High-context communication is prevalent in collectivistic cultures [42].

In high context cultures, implicit information is shared and the communication process relies on understanding the meaning of the verbal messages as well as interpreting cues such as tone of voice, body language, facial expressions, voice patterns, the use of silence, and past interactions. These cues, when understood, transmit information that would otherwise need to be encoded verbally. Participants of a conversation capture information from reading these cues from each other, which would be unnoticeable to those who do not share the same context.

While collectivist cultures are generally regarded as high context cultures, individualistic cultures can generally be classified as low context cultures. In these, individuals have limited shared knowledge, or assume a limited shared knowledge. Verbal messages are the primary communication medium. Other cues are not as important, and are sometimes blocked. Individuals in low context cultures generally opt for a reduced number of non-verbal cues since non-verbal cues could transmit equivocal messages due to the lack of common context. For these cultures all information needs to be communicated explicitly since there are few shared codes and symbols. When communicating, there is only one literal meaning to a message, and the meaning is not affected by occasional non-verbal cues that may be transmitted simultaneously [55].

The impact of technology adoption is moderated by culture, and individualism and collectivism have an impact. Individuals who come from collectivist cultures will provide information and seek approval from the members within their social boundaries, the “in-group”, and will discard those who are outside of the social boundaries, the “out-group”. Those who come from cultures characterized as individualist will give equal value to those in the in-group as to those in the out-group. The previously mentioned behavior has been reported in collaborative search environments, where those who were characterized as collectivist exchanged more information with their in-group. On the other side, those characterized as individualist did not give preference to any group. For collectivist cultures technology usage is perceived as a means to achieve organization among the group, with emphasis on the group. Individualistic cultures see technology as a means to achieve individual efficiency and decision making [61]. More specifically to on-line search behavior, members of collectivist cultures would find relevance ratings constructed from other users’ opinions more trustworthy than relevance ratings constructed with measures such as number of hits. Therefore, collectivist cultures are likely to value a search engine that presents relevance rating based on other’s opinions, and vice versa.

3) Masculinity versus Femininity

The level of masculinity or femininity of a culture has been linked to behavior in GDSS. Members of masculine cultures value recognition [48]. A GDSS meeting in which the anonymity feature is enabled will result in reduced participation from participant who reflects values associated with masculinity cultures [48] Furthermore, such an anonymous GDSS meeting will encourage masculine members to “free-ride”, while member who reflect values associated with low masculinity cultures will contribute to ensure the “well-being” of the group. Individuals from feminine cultures will also appreciate anonymity because they fell that this setting creates less conflict. In a different setting, individuals from masculine cultures tended to generate more conflict than individuals from cultures that are classified as low in masculinity. In addition, individuals from masculine cultures propose fewer conflict resolution strategies than other participants [60].

Furthermore, it has been reported that in some groups, time dominance, which is the time allocation obtained by contentious techniques such as raising the voice, is decreased since a GDSS system may be unable to transmit these cues

[58]. In such environments members of masculine cultures tended to participate less than members of feminine cultures. Based on the research described above, assertiveness and aggressiveness, which are values associated with masculine cultures, are difficult to convey in these media.

Another study analyzed web sites and their manifestation of masculine and feminine values [31] In this research, several websites were analyzed and masculine and feminine "signifiers" were found. Several masculine cultural values were identified such as strength, challenge performance, dominance, success, and leadership. The feminine values identified were sympathy for the weak, charity, relationship, commitment, sharing, and concern for life. The study found that those websites that were categorized as masculine generally contained numerical and statistical information and tables to describe events and facts. Masculine cultures tend to rely on factual information. The websites that were categorized as feminine generally resorted to intuition and feelings when describing events and facts [31]. Sites classified as masculine and feminine also used different tones to communicate, where words may be emphasized by using bold typefaces, and exclamation marks. This is in addition to the use of an assertive tone and challenging, sarcastic, and ironic comments to justify claims. Feminine websites were found to resort to explanation to justify a claim, and deferring explanation to experts, if necessary.

Feminine cultures value relationships. In those websites that were categorized as feminine, the language intended to build a relationship with the reader. Articles such as "you" were used often, as opposed to "one"

which was more common in websites that were categorized as masculine. In addition, imperatives, which show power and assertiveness, were more frequent in masculine rated web sites. The amount of dependence and fixation on technology by a culture is also a result of the level of masculinity/femininity. Masculine cultures tend to be more technology focused [46] Feminine cultures also value technology, but emphasis is placed on users and relationships. Masculine cultures may evaluate a technology by examining quantitative performance; while feminine cultures evaluate a technology by looking at the impact it has on its users and the workplace.

4) Uncertainty Avoidance

Uncertainty avoidance has been examined by Information System research. Technology adoption and diffusion has been linked to the uncertainty avoidance level of the culture. The adoption of certain technologies may take longer in some cultures, where users need to have certain assurances about a technology before the technology is widely adopted and standardized.

Hofstede determined that a culture with a high level of uncertainty avoidance generally prefer rules and structure, and enjoy having a higher degree of control. Individuals that are characterized as high in uncertaintyavoidance will require a larger number of searches to come to a conclusion (Wilson). Individuals that rate low on the uncertainty avoidance dimension will come to a conclusion with a lower number of search iterations. In addition, the risk profile of an individual can be weighed against the potential social impact of a decision.Individualistic cultures value risk taking and confrontation which may result in increasing personal benefits while harming the status of other [43].

5) Time Horizon

Culture influences an individual's acceptance of different time horizons or outcome expectations. An individual who comes from a short term oriented culture places more value on immediate results which are tangible. A higher value is given to any method or strategy that will provide immediate results. Efficiency is a key aspect of a process, and it is as important as the final result (Hofstede). Cultures that exhibit values of long term orientation uphold that perseverance, persistence, and thrif are necessary to achieve goals. Immediate satisfaction is not seen positively, since all future rewards should be the result of present effort. In contrast with short term oriented cultures, shortcuts are not acceptable, and may be considered dishonest.

In short term oriented cultures, the criteria used by an individual to evaluate the quality of a method, process, or service received will place more weight on delivery time. On the other hand, an individual who belongs to a long term oriented culture will not be concern with the time period required to complete a process or service, as long as the end result is what is desired.

An individual's time horizon has been evaluated in the context of on-line shopping, within the context of TAM, where a user's degree of time orientation moderated the relationship between trust and intention to use [59]. These results are significant because they suggest that in cultures that are long term oriented, trust is more important than perceived ease of use and perceived usefulness, within the TAM framework.

Time orientation has also been researched in the context of computer security. Long term horizon societies tend to have a different disposition and awareness in regards to potential threats to computer systems. Research has shown that in Asian cultures, which rate as long term oriented cultures, it would be more effective to describe long term benefits of an adequate computer security policy, than the immediate benefits [41].

III. SUMMARY

This paper provided a discussion of social computing and how it has been defined over the years. It also emphasized that described some of the social computing tools that are available and their potential uses. Additionally, This paper described a culture and cultural behavior with Hofstede's culture dimensions as conceptualizations of culture. In this paper concluded by presented those cultural dimensions that are not only relevant, but have been widely used in computer information systems research.

Social computing is an active area of research. In light of ongoing developments in on-line technology and new applications, many users switch from searching as application (e.g. Google) to searching as a function of an advanced, more complex system (e.g. Facebook's search function). This paradigm switch may require review of survey instrument in

future research. This literature review is based on Hofstede's model developed more than three decades ago. While the model has been repeatedly updated, changes may not have taken into consideration all newly developed technologies. In particular social computing technologies such as social networking applications, which re-defined the concept of personal computing and empower members of cultures that reflect collectivist values. Also, acceptance and use of new technologies will affect user behavior and consequently new or modified hypotheses will need to be developed. A potential area or future review may concentrate on a particular application of social coimuting such as search engines or social networking, and nclude a modified Hofstede's instrument to specifically address the advances of computation technology.

REFERENCES

[1] S. Andriole, "Business impact of Web 2.0 technologies." Communications of the Association for Computing Machinery, vol. 53, pp.67-79., 2010.

[2] M. Banan and A. Banan, "What about correlation between metrics and social computing?ComputerScience & Telecommunications. [On-line]. 6, pp. 47-55., 2009.

[3] N. Barnes and E. Matson, "Social media in the 2009 Inc. 500: New tools and new trends. University of MA Dartmouth, Center for Marketing Research.

[4] N. G. Barnes & E. Matson,"The Fortune 500 and social media: A longitudinal study of blogging and Twitter usage by America's largest companies. Unpublished research report.

[5] M. Brandel. Social networking goes corporate. ComputerWorld, 42(32), 24-27., 2008.

[6] L.Buffington, Creating and consuming Web 2.0 in art education. Computers in the Schools, 25(3), 303-313., 2008.

[7] J. Bughin, J.Manyika, and A. Miller. (2008). Building the Web2.0 enterprise.

[8] Coyle, and H. Vaughn, "Social networking: Communication revolution or evolution? " Bell Labs Technical Journal, 13(2), 13-17., 2008.

[9] Culley, (2006). Social computing. Retrieved from http://www.instructionaldesign.com.au/Academic/SocialComputing.htm

[10] M. Dadashzadeh, Social media in government: From eGovernment to eGovernance. Journal of Business & Economic Research, 8(11), 81-86., 2010.

[11] G. Hofstede, Culture Consequences Sage, Beverly Hills, CA., 1980.

[12] Jonassen, Howland, J., Marra, and D. Crismond, Meaningful learning with technology. Upper Saddle River, NJ: Pearson, Prentice Hall.2008.

[13] J. Kratzer, R. Leenders, nad j Van Engelen, The social network among engineering design teams and their creativity: A case study among teams in two product development programs. International Journal of Project Management, 28(5), 428-436. 2010.

[14] L. Lai, nad E. Turban, Groups formation and operations in the web 2.0 environment and social networks. Group Decision and Negotiation, 17(5), 387- 402., 2008.

[15] Leidner, H. Koch, and E Gonzalez, Assimilating Generation Y IT new hires into USAA's workforce: The role of an Enterprise 2.0 system. MIS Quarterly Executive, 9(4), 229-242.,2010.

[16] D. Leidner, and T. Kayworth, "Review: A Review of Culture in Information Systems Research: Toward a Theory of Information," Management information systems quarterly. (30:2) 2006, pp 357-399.

[17] Ramdani, nad T. RajwaniEnterprise 2.0: the case of British Telecom. Journal of Strategic Management Education, 6(2), 135-148., 2010.

[18] O. Serrat, Social media and the public sector. Washington, DC: Asian Development Bank. 2010.

[19] Venkatesh, J. Thong, and X. Xu, Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. MIS Quarterly, 36(1), 157-178.,2012.

[20] T. Williams, R. Williams, Adopting social media: Are we leaders, managers, or followers? Communication World, 25(4), 34-37.,2008.

[21] M. Zeisser, Unlocking the elusive potential of social networks. McKinsey Quarterly. 2010.

[22] N. Ellison, C. Steinfield and C. Lampe The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. Journal of Computer- Mediated Communication. Jul1;12(4):1143-68., 2007.

[23] H. Krasnova, S. Spiekermann,K. Koroleva andt Hildebrand , Online social networks: why we disclose. Journal of Information Technology. 2010 Jun 1;25(2):109-25.

[24] J. DiMicco, D. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M.Muller, November. Motivations for social networking at work. InProceedings of the 2008 ACM conference on Computer supported cooperative work (pp. 711-720).

[25] L. Chiş, M. Talpoş, "PROS AND CONS OF CORPORATE SOCIAL NETWORKING." Review of Management & Economic Engineering10, no. 2 , 2011.

[26] N. Ellison, and D. Boyd. ". Sociality through Social Network Sites." 2013.

[27] J. Fogel, and E. Nehmad, CInternet social network communities: Risk taking, trust, and privacy concerns. Computers in human behavior25(1), pp.153-160., 2009.

[28] T. Kidd, and I. Chen, Wired for learning: an educator's guide to web 2.0. IAP; 2009.

[29] Lee, S,Im, C. Taylor. Voluntary self-disclosure of information on the Internet: A multimethod study of the motivations and consequences of disclosing information on blogs. Psychology & Marketing. 2008 Jul 1;25(7):692-710.

[30] H. Du andC. Wagner, Learning with weblogs: An empirical investigation. InProceedings of the 38th Annual Hawaii International Conference on System Sciences 2005 Jan 3 (pp. 7b-7b). IEEE.

[31] Zahedi, W. Van Pelt, and M. Srite, Web documents' cultural masculinity and femininity. Journal of Management Information Systems, 2006, 23(1), pp.87-128.

[32] O. Arazy, I.Gellatly, S.Jang, and R. Patterson, Wiki deployment in corporate settings. IEEE Technology and Society Magazine, 28(2), pp.57-64.,2009

[33] Ebersbach, M.Glaser,R. Heigl, and A. Warta, Wiki: web collaboration. Springer Science & Business Media. 2008.

[34] Zhang ,Y. Zhu, and H. Hildebrandt, Enterprise Networking Web Sites and Organizational Communication in Australia. Business Communication Quarterly. 2009;72(1):114-9.

[35] Okoro, A. Hausman, and M. Washington. Social media and networking technologies: An analysis of collaborative work and team communication. Contemporary Issues in Education Research (Online). 2012 Jul 1;5(4):295.

[36] M. Parameswaran and A.Whinston, Research issues in social computing. Journal of the Association for Information Systems, 8(6), p.336., 2007.

[37] S. Chu, Y. Kim, Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. International journal of Advertising. 2011 Jan 1;30(1):47-75.

[38] X. Shen, C. Cheung, and M. Lee, Perceived critical mass and collective intention in social media-supported small group communication. International Journal of Information Management. 2013 Oct 31;33(5):707-15.

[39] R. Rubio, S. Martín, and S. Morán, Collaborative web learning tools: Wikis and blogs. Computer Applications in Engineering Education. 2010 Sep 1;18(3):502-11.

[40] B.Nardi, S.Whittaker, and E. Bradner, Interaction and outeraction: instant messaging in action. InProceedings of the 2000 ACM conference on Computer supported cooperative work 2000 Dec 1 (pp. 79-88). ACM.

[41] M. Schmidt, A. Johnston, K. Arnett, J. Chen, and A. Li ,cross-cultural comparison of us and chinese computer security awareness. Journal of

Global Information Management. 2008 Apr 1;16(2):91.

[42] W. Gudykunst,. and Y. Kim, Communication with Strangers: An Approach to International Communication. 2008.

[43] Adams G, Gercek-Swing B. Avoid or Fight Back? Cultural Differences in Responses to Conflict and the Role of Collectivism, Honor, and Enemy Perception Ceren Günsoy Susan E. Cross Ayse K. Uskul 2.

[44] J. Blackwell, J. Sheridan, K. Instone, D. Schwartz, S. Kogan, Design and adoption of social collaboration software within businesses. InCHI'09 Extended Abstracts on Human Factors in Computing Systems 2009 Apr 4 (pp. 2759-2762). ACM.

[45] G.Hofstede, Culture and organizations. International Studies of Management & Organization. 1980 Dec 1;10(4):15-41.

[46] Hasan, G.Ditsa, The impact of culture on the adoption of IT: An interpretive study. Journal of Global Information Management (JGIM). 1999 Jan 1;7(1):5-15.

[47] W. Richardson, Blogs, wikis, podcasts, and other powerful web tools for classrooms. Corwin Press, 2010.

[48] R. Robichaux, and N. Keesee, What can social workers do for warriors in transition?. US Army Medical Department Journal. 2008 Jan 1:25-7.

[49] C.Tu, M. Blocher, and G.Roberts, Constructs for Web 2.0 learning environments: A theatrical metaphor. Educational Media International. 2008 Dec 1;45(4):253-69.

[50] P.To, C. Liao, C. Chiang, M. Shih and C. Chang, An empirical investigation of the factors affecting the adoption of Instant Messaging in organizations. Computer Standards & Interfaces. 2008 Mar 31;30(3):148-56.

[51] P. Smith, S. Dugan, and F. Trompenaars, National culture and the values of organizational employees a dimensional analysis across 43 nations. Journal of cross-cultural psychology. 1996 Mar 1;27(2):231-64.

[52] W. Gudykunst, S.Ting-Toomey, and E.Chua, Culture and interpersonal communication. Sage Publications, Inc; 1988.

[53] S. Koshman, A. Spink ,and B. Jansen, Web searching on the Vivisimo search engine. Journal of the American Society for Information Science and Technology. 2006 Dec 1;57(14):1875-87.

[54] S. Singh, Cultural differences in, and influences on, consumers' propensity to adopt innovations. International Marketing Review. 2006 Mar 1;23(2):173-91.

[55] S. Singh, Cultural differences in, and influences on, consumers' propensity to adopt innovations. International Marketing Review. 2006 Mar 1;23(2):173-91.

[56] Smith, and I. Walker I. The Rocky Road from Comparisons to Actions. Improving intergroup relations: Building on the legacy of Thomas F. Pettigrew. 2008 May 30;2:227.

[57] Karahanna, J. Evaristo, and M. Srite, Levels of culture and individual behaviour: An integrative perspective. Advanced Topics in Global Information Management. 2006 Apr 30;5(1):30-50.

[58] Robichaux and R. Cooper, "Gss Participation: A Cultural Examination " Information & Management (33:6) 2008, pp 287-300.

[59] Yoon. "The Effects of National Culture Values on Consumer Acceptance of ECommerce: Online Shoppers in China," Information & Management (46:294-301) 2009.

[60] L. Tunga and M. Quaddus, "Cultural Differences Explaining the Differences in Results in Gss: Implications for the Next Decade " Decision Support Systems (33:2) 2005, pp 177-199.

[61] H. Cho, and J. Lee, "Collaborative Information Seeking in Intercultural Computer Mediated Communication Groups: Testing the Influence of Social Context Using Social Network Analysis," Communication Research. (35:4) 2008, pp 548-573.

[62] Majchrzak, C. Wagner, & D. Yates .Corporate wiki users: results of a survey. Proceedings of the 2006 international symposium on Wikis. Odense, Denmark. Retrieved from ACM Digital Library. 2006.

[63] Kane. A multimethod study of information quality in wiki collaboration. ACM Transactions on Management Information Systems, 2(1). 4.2011.

[64] Collier. Wiki technology in the classroom: Building collaboration skills. Journal of Nursing Education, 49(12), 718-718.

[65] Morgan & R. Smith,. A wiki for classroom writing. The Reading Teacher, 62(1), 80-82.2008.

[66] Q. Huy and A. Shipilov. The key to social media success within organizations. MIT Sloan Management Review. 2012 Oct 1;54(1):73.

[67] B. Alexander. Social networking in higher education. In R. N. Katz, (Ed.), Thetower and the cloud: Higher education in the age of cloud computing (pp. 197- 201). 2010.

[68] P. McFarlane. "Developing a culturally specific e-learning website." In2006 7th International Conference on Information Technology Based Higher Education and Training, pp. 473-480. IEEE, 2006.

[69] Dix, Human-computer interaction, pp. 1327-1331, Springer US,2009.

Improvisation of Security aspect of Steganographic System by applying RSA Algorithm

*Manoj Kumar Ramaiya
Research Scholar, Computer Engineering
Suresh Gyan Vihar University
Jaipur, India

**Dr. Dinesh Goyal
Professor & Principal, School of Engineering
Suresh Gyan Vihar University
Jaipur, India

Dr. Naveen Hemrajani
Professor & Head, Computer Engineering
JECRC University
Jaipur, India

***Abstract*—The applications accessing multimedia systems and content over the internet have grown extremely in the earlier few years. Moreover, several end users or intruders can simply use tools to synthesize and modify valuable information. The safety of information over unsafe communication channel has constantly been a primary concern in the consideration of researchers. It became one of the most important problems for information technology and essential to safeguard this valuable information during transmission. It is also important to determine where and how such a multimedia file is confidential. Thus, a need exists for emerging technology that helps to defend the integrity of information and protected the intellectual property privileges of owners. Various approaches are coming up to safeguard the data from unauthorized person.**

Steganography and Cryptography are two different techniques for security data over communication network. The primary purpose of Cryptography is to create message concept unintelligible or ciphertext might produce suspicious in the mind of opponents. On the other hand, Steganography implant secrete message in to a cover media and hides its existence. As a normal practice, data embedding is employed in communication, image, text or multimedia contents for the purpose of copyright, authentication and digital signature etc.

Both techniques provides the sufficient degree of security but are vulnerable to intruder's attacks when used over unsecure communication channel. Attempt to combines the two techniques i.e. Cryptography and Steganography, did results in security improvement. The existing steganographic algorithms primarily focus on embedding approach with less attention to pre-processing of data which offer flexibility, robustness and high security level. Our proposed model is based on Public key cryptosystem or RSA algorithms in which RSA algorithm is used for message encryption in encoding function and the resultant encrypted image is hidden into cover image employing Least Significant Bit (LSB) embedding method.

Keywords—Image Steganography; Cryptography; LSB insertion; Public key Cryptosystem; RSA algorithm

I. INTRODUCTION

While in multimedia communications, the need of privacy and confidentiality gains more and more significance, mostly in open networks like the Internet. In the era of worldwide electronic connectivity, of viruses and hackers, of electronic eavesdropping and electronic fraud, there is indeed a need to protect information from passing before curious eyes or, more importantly from falling into wrong hands. Thus, multimedia security is much to consider in distributing digital information safely.

The past three or four decade led to the wide spread transfer of data from one end to the other end of the world. The remarkable evolution of the internet also evolved and eased various E- Commerce applications. This demand the assurance of security of information. Further the communication between private parties demanding absolute privacy also necessitate the data transmission in modified or encoded mode.

In multimedia communication the necessity of privacy and confidentiality gains additional importance mainly in open, unsecure communication network like internet. Present era of universal connectivity, of viruses, intruders, eavesdropping and digital fraud need to safe-guard information from releasing into erroneous hand.

Cryptography techniques [5, 6] scramble a source message in to unintelligible form so it cannot be understood while steganography hides the message in to other media, so it cannot be perceived. The term steganography [2, 3] originates from the Greek Steganos which means "covered" and Grafia means "writing" i.e. Steganography means "covered writing "[4].

Cryptography and Steganography are extensively used in the field of information hiding [1] and has received attention from the businesses and academic world in the past. Former conceals the original data but latter conceal the very fact that data is hidden.

Public Key Cryptosystem

A different concept of achieving the same results as from digital signature [11, 12] and steganography is the asymmetric key crypto system [7, 10] using two key termed as public key and private key. In symmetric encryption, the key need to be communicated before at both senders and receiver. Also to make the digital authentication look analog to the current practice some sort of identification like signature need to be inserted. To fulfill the above requirements Diffie and Hellman proposed the most widely accepted and implemented principle in 1976 termed as Public Key Cryptosystem [8, 9].

In contrast to symmetric key encryption, asymmetric key cryptosystem employ one key for encryption and different but related key for decryption. To fulfill the security requirement the approach need to have the following characteristics:

* First Author ** Corresponding Author

1) The cryptographic algorithm be such that it is infeasible to find the decryption key if only the encryption key and cryptographic algorithm is available.

2) Either of key pair (two related keys) can be used for encryption and the other used for decryption.

Fig. 1. Basic Model of Public key Cryptosystem

The process of encryption be as follows:

1) Each user generate a pair of keys for the encryption and decryption of the massage.

2) Each user place one of the two key in public or in open domain, accessible to all. This key is termed the public key. On the other hand, the companion key is kept private with each of them, and is termed the private key.

The messages from A are encrypted using the B's public key. On receiving the message, B decrypt using his private key. Since the private key of B has been generated and kept safe by B himself, the message remains secure.

Here the private keys have been generated by each locally and never transmitted nor distributed, thus remains protected and secrete, and hence providing the requirement of security

The rest of the paper is organized as follows. Section 2, literature surveyed dealt with techniques involving purely combination of cryptographic method or steganography methods. Section 3 proposed highly secured system which combines both cryptography and steganography techniques in order to provide higher payload, more robust and secure. In Section 4, the proposed hybrid techniques were tested on various standard images set namely Cameraman, Lena and Baboon etc. The PSNR (Peak Signal to Noise Ratio) value, to evaluating the quality of reproduced image (cover image and stego image) qualitatively. Finally, Section 5 concludes this paper.

II. The Related Work

Considering the strength and weakness of steganography and cryptography, researchers tried to combining them in practice, so that the new method would simultaneously possess the advantages of steganography and cryptography while overcoming the respective shortcomings.

The literature surveyed dealt with techniques involving purely cryptographic method or steganography methods. Both of the techniques have shortcoming from the view point of a degree of security and robustness against attacks and efficiency and ease of implementation in terms of hardware and runtime.

Attempt to combines two techniques [21, 22] to ensure more secure encoding have been made. In the most of the cases, techniques involved works on plaintext and very rare attempt have been made to encode images.

The major techniques comprises cryptography and steganography detailed in the literature can be broadly be classified into five categories, four being in the special domain while others one encrypt in the transform domain. They are as follows:

1) Idea employing the two techniques in tandem. Shouchao Song et Al [15] suggested a protocol merging cryptography and steganography techniques based on LSB matching method and well developed Boolean function in stream cipher. The protocol accomplishes the encryption and hiding all at once resulting in less computation them all the existing methods. The LSB method is used for hiding the encrypted message in cover image.

2) Text encryption with Data Encryption Standard (DES) and LSB insertion Dhawal Seth et Al [14] combines cryptography and steganography, so as to ensure more security over insecure communication channel. DES cryptographic algorithm being used for encrypting the text message in conjunction with LSB substitution for embedding the encrypted message in the cover image.

3) The techniques proposes compressing the signal before encrypting and employing steganographic techniques. It also proposed use of hash function so as to generate a message authentication code by hashing the key. The resulting model is claimed to survive image manipulation and attacks. Khalil Challita and Hikmat Farhat [16] proposed multiple encryption. Embedding the encrypted text secret message in more than one cover objects.

4) For a highly secure communication Ankit Uppal et Al [17] proposed dual security method by combining the RC5 enhance algorithm for encrypting and JPEG LSB coding for steganography.

5) The techniques proposed by Dipti Kapoor Sarmah and Neha Bajpai [13] apply Advance Encryption Standard (AES) encryption techniques for secrete message. The encrypted message is embedded in the Discrete Coeficint Transform (DCT) of the cover image. The DCT of image is obtained and the coefficient is embedding in the image.

The slight variant in the combined techniques is proposed by Pye Pye Aung and Tun Min Naing [18], using the same AES algorithm for encryption. In the steganographic a part of encrypted message as a key is used to hide in DCT of a cover image.

III. Proposed Hybrid Model

Proposed steganographic model is based on RSA Algorithms is depicted in figure 2.

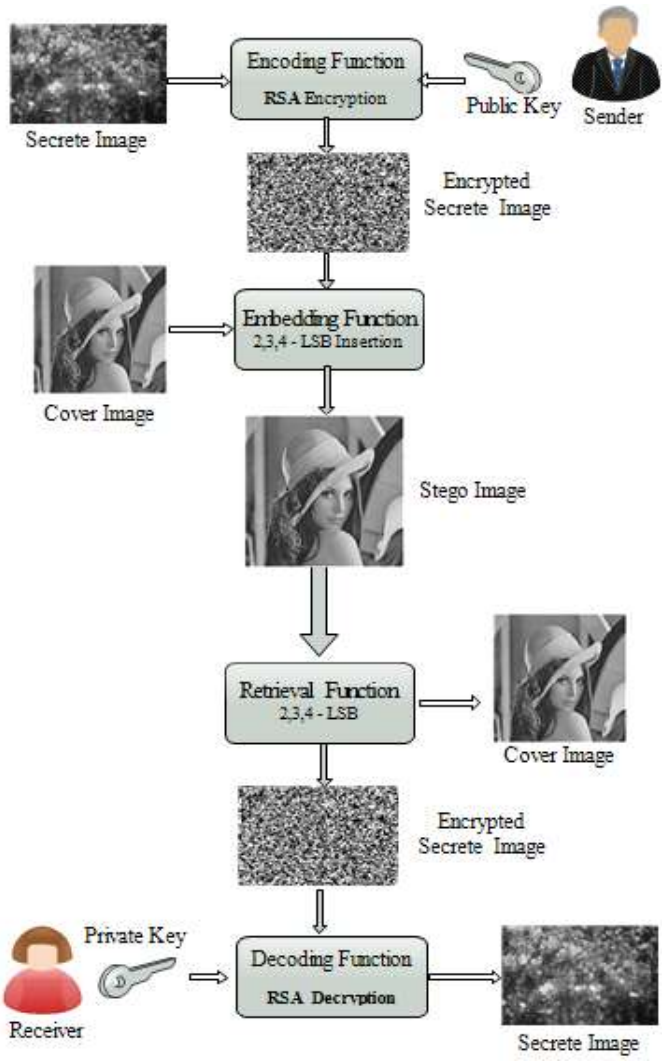


Fig. 2. Proposed Steganographic Model

A. RSA Encoding Function

First the secrete image is selected (e.g. of 256 × 256). The intensity value of each pixel of secrete image are converted from binary to decimal value. Now the first pixel values from secrete image is inputted to RSA encoding function as described below.

The RSA algorithm [22, 23] is implemented to encrypt input pixel value as follows:

1) Two prime number p and q are chosen such that they are the prime numbers.

2) $n = p \times q$ is calculated and made available to public.

3) e is chosen such that gcd $(\varphi(n), e) = 1$; $1 < e < \varphi(n)$ made public.

4) d is private and calculated as $d = e^{-1} \varphi(n)$.

Then the private key pair is (d ,n) and public key pair is (e ,n) .

The equivalent cipher value for first pixel is now calculated by using public key pair (e ,n)

$$C = M^{e} \bmod n$$

After the execution of RSA, first pixel value is now encrypted and this value are placed at first position by again convert it into decimal value. Now taking second pixel value convert it into decimal and inputted to RSA encoding function getting the second pixel encrypted value, likewise sequentially take pixel one by one, input to encoding function and obtain encrypted value of encrypted image.

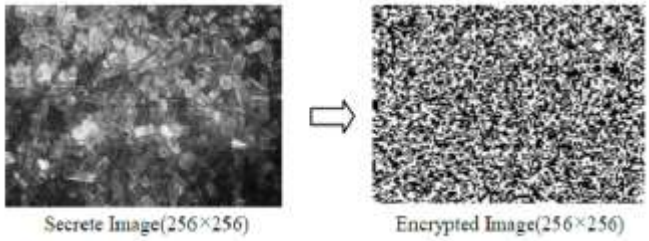


Fig. 3. Conversion of Secrete Image to Encrypted Secrete Image

B. Embedding Function using LSB Method

1) Bit Division: Taking the cipher encrypted image, the values are converted from decimal to binary.

The binary value of $(173)_{10} = (1\ 0\ 1\ 0\ 1\ 1\ 0\ 1)_2$

Next divide this 8 bit value into 4 part taking 2 bits in each. After bit division, value of b1 = 1 0 , b2 = 1 0 , b3 = 1 1 , b4 = 0 1 are getting.

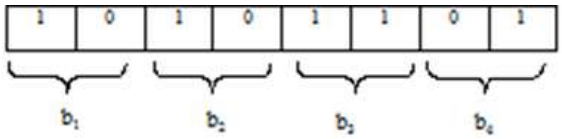


Fig. 4. Bit Division for LSB Embedding

2) Insertion of Bit value into the cover image: After receiving the values of b1, b2, b3, b4, these values are inserted into the cover image. The 2 bit LSB of the four consecutive pixels in cover image are replace. Taking the pixels one by one from the cover image, the 2 LSB bits are replaced by 10,10,11,01 respectively.

3) Formation of Stego Image: After receiving the new pixel value the stego image is formed by replacing these values at their original position. Likewise the pixels value one by one from encrypted secrete image and insertion into the cover image and replaced them. Result becomes the stego image.

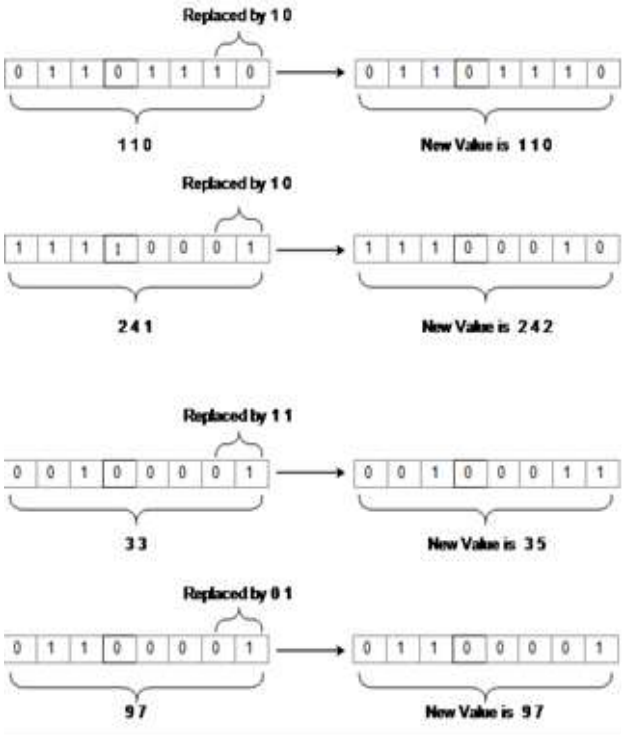


Fig. 5. Insertion of Bits into cover Image

C. Image Retrieval function

At the receiving end, decoding of stego image perform the following process:

1) Generate the 2 LSB bits from the stego Image: The pixels value are handled one by one from the stego image. Convert these pixel value from decimal to binary values and take 2 LSB bits from first four consecutive pixel values:

Similarly taking next three pixels. i.e. 242, 35, 97;

$(242)_{10}$ = $(1\ 1\ 1\ 1\ 0\ 0\ 1\ 0)_2$
$(35)_{10}$ = $(0\ 0\ 1\ 0\ 0\ 0\ 1\ 1)_2$
$(97)_{10}$ = $(0\ 1\ 1\ 0\ 0\ 0\ 0\ 1)_2$

Getting,

$b_1 = 1\ 0$; $b_2 = 1\ 0$; $b_3 = 1\ 1$; $b_4 = 0\ 1$;

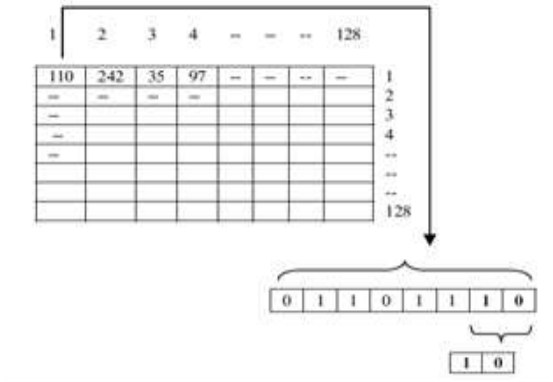


Fig. 6. LSB (2- bits) Extraction of Stego Image

2) Concatenation of bits: Now concatenating the input, the 8 bits of first pixel value of encrypted secrete image is acquired as

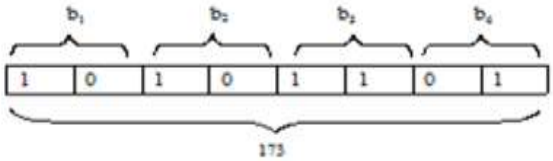


Fig. 7. Concatenation of Bits (Extracted)

3) Reformation of Encrypted Secrete Image: Now the generated value is placed into first position. Similarly taking the next four pixel value from stego image, the process is repetitive and the whole encrypted secrete image is recovered.

D. RSA Decoding Function

1) Creation of Secrete image: In decoding function the pixel value from the encrypted secrete image are again inputted to the RSA decoding function by using private key pair (d,n) to obtain pixel value of original secrete image as follows:

$M = C^d \bmod n$

After execution of decoding function for every pixel, the secrete image or original image is created.

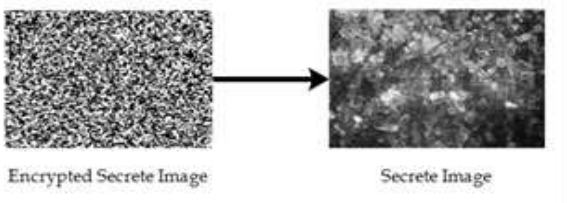


Fig. 8. Conversion of Encrypted Secrete Image to Secrete Image

IV. Results and Analysis

Proposed model using RSA algorithm is robust Steganography technique because without knowing the receiver secrete keys the extraction of secrete image from the stego image is impossible. Here the private keys have been generated by each user locally and never transmitted nor distributed via any transmission media , thus key remains protected and secrete, and hence system providing the requirement of security and authentication. Furthermore in embedding process quality of cover image is also not degrading due to variation in two LSB of each pixel which replicates only 0 – 3 difference in pixel value.

Moreover the proposed system is capable of not just scrambling data but it also changes the intensity of the pixels which contributes to the safety of the encryption.

TABLE I. Capacity and PSNR

Name of Image	Size (Pixel)	Capacity	PSNR In DB
Baboon.jpg	256× 256	25 %	44.23
Cameraman.jpg	256×2564	25 %	44.86
Lena.jpg	256× 256	25 %	44.48
pirate_gray.jpg	256× 256	25 %	44.36

V. CONCLUSION

In the proposed RSA based steganographic model is more secure as compare to the traditionally symmetric cryptosystem because in public key cryptosystem the private keys have been generated by each user locally and never transmitted nor distributed, thus no question of stealing or discloser of key, improves image quality and security compare to existing systems. Steganography, especially combined with the cryptography is a powerful tool which enables to communicate safely with the little computational overload in the system

REFERENCES

[1] N.F. Johnson, S. Jajodia, Exploring steganography: seeing the unseen, IEEE Computer 31 (2) (1998) pp.26–34.

[2] Ross J. Anderson, Fabien A.P. Petitcolas , "On The Limits of Steganography ", IEEE Journal of Selected Areas in Communications, 16(4):474-481, May 1998.

[3] N. Provos, P. Honeyman, "Hide and Seek: an Introduction to Steganography", IEEE Security and Privacy 1 (3) (2003) 32–44.

[4] J.C.Judge, "Steganography: past, present, future", SANS Institute publication, /http://www.sans.org/reading_room/whitepapers/stenganography/552.phpS, 2001.

[5] Lt. James Caldwell ,U.S. Air Force , "Steganography " , CROSSTALK The Journal of Defense Software Engineering , June 2003 , pp. 25 – 27 .

[6] N. Provos, P. Honeyman, "Hide and Seek: an Introduction to Steganography", IEEE Security and Privacy 1 (3) (2003) 32–44.

[7] Mohammed AbuTaha, Mousa Farajallah, Radwan Tahboub, Mohammad Odeh, "Survey Paper: Cryptography Is the Science of Information Security ", International Journal of Computer Science and Security (IJCSS), Volume (5): Issue (3): 2011, pp. 298 – 309.

[8] James L. Massey, "An Introduction to Contemporary Cryptology ", Proceedings of the IEEE, VOL. 76, NO. 5, MAY 1988, pp. 543 – 549.

[9] Alexander W. Dent, "Choosing key sizes for cryptography ", Information Security Technical Report, Vol. 15 No 1, 2010 Elsevier, pp. 21-27.

[10] Amin Daneshmand Malayeri , Jalal Abdollahi , "Modern Symmetric Cryptography methodologies and its applications " , IEEE Transactions On Information Theory, Vol. 97, No. 6, October 2009 , pp. 505 -509.

[11] Anthony T.S. Ho, Siu-Chung Tam, Kok-Beng Neo, Sim-Peng Thia , "Digital Steganography for Information Security", Internet Business99, 1999 - researchgate.net , pp. 1-9.

[12] Christoph Busch, Klara Nahrstedt, Ioannis Pitas, "Image Security", IEEE Jan – Feb 1999, pp. 16.

[13] Dipti Kapoor Sarmah, Neha Bajpai," Proposed System for Data Hiding Using Cryptographyand Steganography", International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010, pp. 7- 10.

[14] Dhawal Seth, L. Ramanathan, Abhishek Pandey, "Security Enhancement: Combining Cryptography and Steganography", International Journal of Computer Applications (0975 – 8887) Volume 9– No.11, November 2010, pp. 3-6.

[15] Shouchao Song, Jie Zhang, Xin Liao, Jiao Du, Qiaoyan Wen," A Novel Secure Communication Protocol Combining Steganography and Cryptography", Advanced in Control Engineering and Information Science , Procedia Engineering 15 (2011) , pp. 2767 – 2772

[16] Khalil Challita, Hikmat Farhat," Combining Steganography and Cryptography: New Directions", International Journal on New Computer Architectures and Their Applications (IJNCAA) 1(1): pp.199-208.

[17] Ankit Uppal, Rajni Sehgal, Renuka Ngapal, Aakash Gupta, "Merging Cryptography& Steganography Combination of Cryptography: Rc6 Enhanced Ciphering and Steganography: JPEG ", International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106, Volume-2, Issue-10, Oct.-2014, pp. 85-87.

[18] Pye Pye Aung, Tun Min Naing,"A Novel Secure Combination Technique of Steganography and Cryptography ", International Journal of Information Technology, Modeling and Computing (IJITMC) Vol. 2, No. 1, February 2014. Pp 55-62.

[19] Vijay Kumar Sharma, Vishal Shrivastava ," A Steganography Algorithm For Hiding Image In Image By Improved Lsb Substitution By Minimize Detection " , Journal of Theoretical and Applied Information Technology 15th February 2012. Vol. 36 No.1 © 2005 - 2012 JATIT & LLS. All rights reserved.

[20] Abhinav Gupta, Anurag Pandey, Himanshu Agarwal," Analysis on Comparison of Cryptography and Steganography over an Open Channel", MIT International Journal of Computer Science and Information Technology, Vol. 5, No. 1, January 2015, pp. 8-12

[21] Z. V. Patel, S. A. Gadhiya," A Survey Paper on Steganography and Cryptography ", RESEARCH HUB – International Multidisciplinary Research Journal (RHIMRJ), Volume-2, Issue-5, May-2015 .

[22] Tan, Wenxue, Wang,Xiping , Xi, Jinju , Pan,Meisen , "A mechanism of quantitating the security strength of RSA key " , Third IEEE International Symposium on Electronic Commerce and Security 2010 , pp. 357- 361.

[23] Rajan.S.Jamgekar, Geeta Shantanu Joshi, "File Encryption and Decryption Using Secure RSA", International Journal of Emerging science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-4, February 2013, pp. 11-14.

New Modified RLE Algorithms to Compress Grayscale Images with Lossy and Lossless Compression

Hassan K. Albahadily[1,2]
[1]Dept. of Telecommunication and Network Devices
BSUIR
Minsk, Belarus
[2]University of Mustansiriyah
Baghdad, Iraq

Viktar U. Tsviatkou
Dept. of Telecommunication and
Network Devices
BSUIR
Minsk,
Belarus

Alaa A. Jabbar Altaay
Dept. of Computer Sciences
University of Mustansiriyah
Baghdad, Iraq

Valery K. Kanapelka
Dept. of Telecommunication and Network Devices
BSUIR
Minsk, Belarus

***Abstract*—New modified RLE algorithms to compress grayscale images with lossy and lossless compression, depending on the probability of repetition of pixels in the image and the pixel values to reduce the size of the encoded data by sending bit 1 instead of the original value of the pixel if the pixel's value is repeated. The proposed algorithms achieved good reduction of encoded size as compared with other compression method that used to compare with our method and decrease encoding time by good ratio.**

Keywords—compression; Run Length Encoding; quantization

I. INTRODUCTION

Data files frequently contain the same character repeated many times in a row. Digitized signals can also have runs of the same value, indicating that the signal is not changing, also images and music. Run-length encoding is a simple method of compressing these types of files [1]. The basic idea of RLE is to remap a sequence of numbers into a sequence of pairs (Value, Run), where value represents the data in the input sequence and run represents the number of times that data is contiguously repeated. An example illustrating RLE for a binary sequence is shown in Fig.1 below

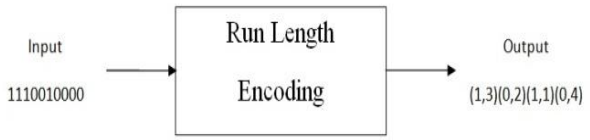


Fig. 1. Illustration of RLE for a binary input sequence

The goal of lossless image compression is to represent an image signal with the smallest possible number of bits without loss of any information, thereby speeding up transmission and minimizing storage requirements. While The goal of lossy compression is to achieve the best possible fidelity given an available communication or storage bit rate capacity or to minimize the number of bits representing the image signal subject to some allowable loss of information. In this way, a much greater reduction in bit rate can be attained as compared to lossless compression, which is necessary for enabling many realtime applications involving the handling and transmission of audiovisual information. In general, there is significant redundancy present in image signals. This redundancy is proportional to the amount of correlation among the image data samples [2].

The lossy compression of images is currently widely used algorithms like JPEG (Joint Photographic Experts Group) [3] and JPEG 2000 [4], which are based on entropy coding (Huffman and arithmetic), coefficients of discrete cosine and wavelet transforms. These algorithms are used to compress the image in hundreds of times with acceptable quality, but have a high computational complexity.

The lossless image compression algorithms are effective for encoding pixel values like LZW (Lempel Ziv Welch), GIF (Graphics Interchange Format), TIFF (Tagged Image File Format) [5,6], Deflate and LZMA (Lempel Ziv Markov chain algorithm) used in archive Zip, 7-zip [7], PPM (portable pixmap) and LZSS (Lempel Ziv Storer Szymanski) used in the archive Rar [7]. These algorithms are used to losseless compress grayscale images to about 2 times depending on the brightness distribution of pixels, but these techniques have a high computational complexity, and because the time and computational resources are limited we should use simpler algorithms for efficient coding like Run Length Encoding algorithm RLE[8], which is based on the character repeats. It can be used for lossless image compression with a small number of sharp luminance differences (animation, medical, segmented and quantized). In combination with pre-quantization image, RLE algorithm may be used for lossy compression. The disadvantage of this algorithm is the lack of

consideration of the probability of repetition of pixel values images.

The aim of this paper is to develop a lossless and lossy new compression algorithm based on the RLE method by using probability of pixel value repetition.

Our technique is implemented using MATLAB2012 on WINDOWS7 Operating System.

This study is organized as follows: Section I presented an introduction about the compression and RLE algorithm, Section II describes the Related work and some of recent modifications of the RLE compression method, Section III explain the suggested RLE algorithm and its modifications, Section IV explains the quantization which is necessary to make good compression ratio when using lossy compression, section V presented the experiment of suggested new algorithm and discuss the results and compare it with all others results, And Section VI presents conclusions of our work and suggestions for future studies.

II. Related Work

There are some recent modifications of the RLE compression method, some of these modifications focusing on the way of scanning pixels (row,coloumn,hilbert,…) [9], or on the bit depth of the runs of the repeated pixels and try to decrease the length of the bit reserved to represent the runs [10], or using a modified Entropy Coding to enhance RLE [11], or quantize image to decrease the number of values of the pixels and make them similar then using DCT to achieve a high compression ratio [11,12], or quantizing adjacent pixels which have small difference in the value of pixel [13], or negligee the non-repeated value [14]. All of these modifications are good to enhance RLE and achieving better results.

III. RLE Algorithm and its Modifications

A. RLE Algorithm

RLE algorithm has been chosen because it is mathematically simple and not complex so that we can achieve a high compression speed for the compression process.

The RLE algorithm based on counting the number of repetitions of the values of successive symbols and it is can be represent by the following diagram of structure coded data in Fig. 2 and the block diagram in Fig. 3

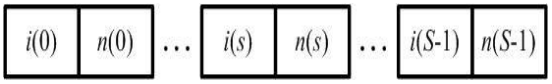


Fig. 2. Coded data structure

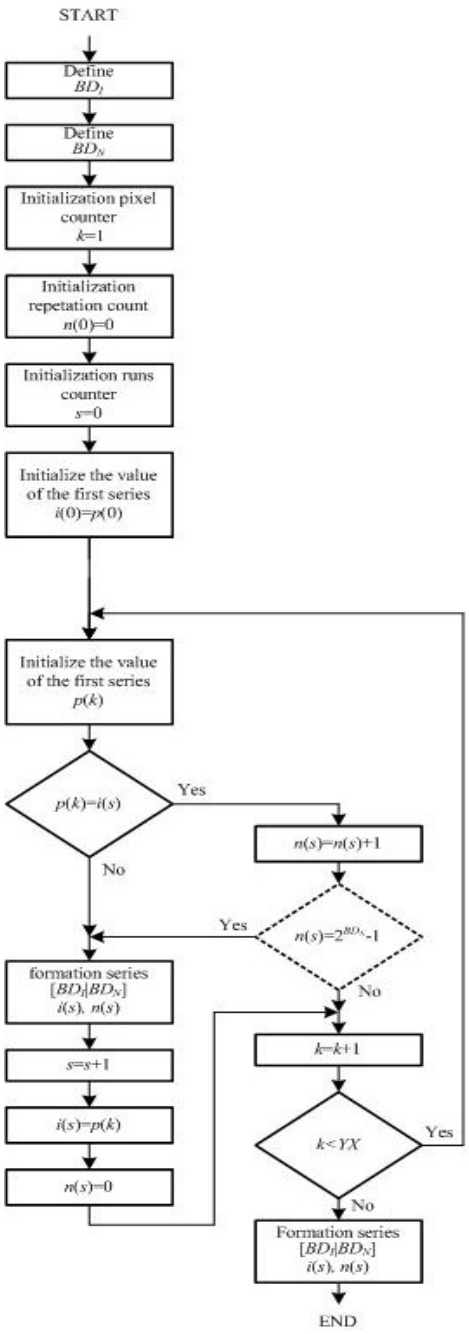


Fig. 3. RLE Block diagram

Modular implementation of RLE algorithm assumes that we have the accumulation of the encoded data for the analysis and selection of encoding parameters. The analysis is based on the table I/N length episodes in which one or more of the same value and consecutive symbols I put in correspondence number N as shown in table 1 below.

TABLE I. LENGTH SERIES

I	$i(0)$	$i(1)$	…	$i(s)$	…
N	$n(0)$	$n(1)$	…	$n(s)$	…

On the basis of run-length table determined bit depth BD_I and run length values BD_N using expressions:

$$BD_I = \left| \log_2 \left(\max \left(i(s) \right)_{(s=\overline{0,S-1})} \right) \right| \quad (1)$$

$$BD_N = \left| \log_2 \left(\max \left(n(s) \right)_{(s=\overline{0,S-1})} \right) \right| \quad (2)$$

Where $i(s)$ is encoded value with the first character of the run length of the table; $n(s)$ is the number of repetitions for i-th symbol (length of the series); S number of encoded symbols (the number of rows in Table 1).

Modular implementation of RLE algorithm provides a minimal amount of code. Its disadvantage is the delay in the coding required for the accumulation of data and the construction of the table of run lengths.

When the production implementation of the algorithm RLE table run length cannot be built, and count the number of characters it can be carried out as they become available. This can significantly improve the coding rate. The structure of the algorithm and the encoded data in the production implementation of RLE algorithm are the same as in the block. Values of the bit depth BD_I an image bit-depth and the lengths of series of values BD_N are selected independently from the incoming data. They can be selected with an excess or deficiency (the overflow $n(s)$ formed a new series $\{i(s), n(s)\}$). In some cases this can lead to an increase in the amount of code which is a disadvantage of RLE algorithm implementation. To improve the characteristics of the RLE algorithm, the production implementation can be through the formation of a dynamic table of run lengths and periodic updating of coding parameters.

When encoding image size $R_{I/N}$ (bit) code, CR compression ratio and computational complexity $C_{I/N}$ of RLE algorithm is determined using the following expressions:

$$R_{I/N} = S\left(BD_I + BD_N\right) \quad (3)$$

$$CR = 8YX / R_{I/N} \quad (4)$$

$$C_{I/N} = YX + 4S \quad (5)$$

where YX image size, determines the number of operations on the buffer and the formation of the table I/N size S records; 4S additional operations to find a maximum value for I (S operations), search for maximum values N (S operations), encoding and transmitting (2S operations).

B. RLE Algorithm I2BN

Suggested modification of RLE is I2BN algorithm to compress grayscale images based on the probability of repetition pixel values in rows. In the process of the algorithm builds a table N/P probability of repeats (table. 2), wherein each value of n the run length number placed such series $p_n(n)$.

TABLE II. PROBABILITY OF REPEATS

N	1	2	…	n	…
P	$p_n(1)$	$p_n(2)$	…	$p_n(n)$	…

For images characterized by a gradual decrease of the function $p_n(n)$ with increasing values n.

When encoding the length of Series the algorithm I2BN first formed character I. Then, if the character I is repeated, the formed bit $b1(s)=1$, otherwise $b1(s)=0$ (Repeat the first character). If the symbol I is repeated again, the bit generated $b2(s)=1$, otherwise $b2(s)=0$ (second repeat symbol). If the symbol I is repeated again, the character is formed $n(s)$, taking into account the number of repetitions (originally $n(s)=0$, if symbol I It is repeated again, $n(s)>0$). As a result, the series may be formed as:

$\{i(s), b1(s)=0\}$, $\{i(s), b1(s)=1, b2(s)=0\}$, $\{i(s), b1(s)=1, b2(s)=1, n(s)\}$.

The structure of RLE coded data can be represented by the following block diagram

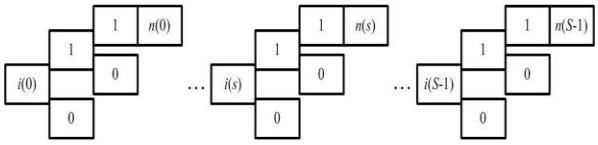


Fig. 4. Coded data structure of algorithm I2BN

The size $R_{I/2B/N}$ (Bit) code for the algorithm I2BN defined by the expression:

$$R_{I/2B/N} = S\left(BD_I + 1\right) + \sum_{s=0}^{S-1} b1(s) + BD_N \sum_{s=0}^{S-1} b2(s) \quad (6)$$

To determine the computational complexity of the expression

$$C_{I/2B/N} = YX + 4S + \sum_{s=0}^{S-1} b1(s) + \sum_{s=0}^{S-1} b2(s) \quad (7)$$

Equation (7) takes into account YX buffering operations and formation table I/N- run length size S records, S

Operations for searching for a maximum value I, S Operations for searching for a maximum value N, $2S+\sum_{s=0}^{S-1}b1(s)+\sum_{s=0}^{S-1}b2(s)$ operations for encoding and transmission.

C. RLE Algorithm I3BN

Another suggested modification of RLE is I3BN, which is differs from the algorithm I2BN by using an additional symbol $b3(s)$, which takes the value 1, if symbol I repeated for the third time in a row, and 0 if absent. The structure of coded data according to the algorithm I3BN is shown in the Fig.5 below

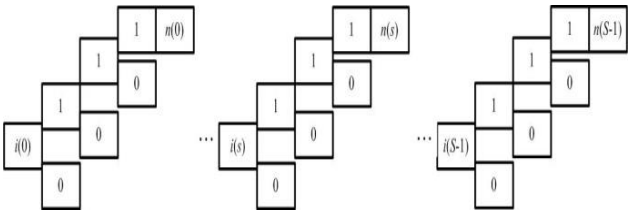


Fig. 5. Coded data structure of algorithm I3BN

Size of $R_{I/3B/N}$ (bit) code algorithm I3BN defined by the expression

$$R_{I/2B/N}=S(BD_I+1)+\sum_{s=0}^{S-1}b1(s)+\sum_{s=0}^{S-1}b2(s)+BD_N\sum_{s=0}^{S-1}b3(s) \quad (8)$$

The computational complexity of the algorithm I3BN estimated using the expression

$$C_{I/2B/N}=YX+4S+\sum_{s=0}^{S-1}b1(s)+\sum_{s=0}^{S-1}b2(s)+\sum_{s=0}^{S-1}b3(s) \quad (9)$$

Equation (9) accounts for buffering YX operations and forming a table I/N run length size records S, S maximum search operations on the values I, S maximum search operations on values N, $2S+\sum_{s=0}^{S-1}b1(s)+\sum_{s=0}^{S-1}b2(s)+\sum_{s=0}^{S-1}b3(s)$ operations for encoding and transmission.

D. Other modifications of RLE algorithm

In addition to I2BN, I3BN we developed algorithms characterized by the use of different numbers of additional characters to encode pixels repeat, all of these algorithms shown in table3 below

TABLE III. THE RLE MODIFIED ALGORITHMS

№	Algorithm	The peculiarity of the encoded data structure
1	I/S/N	variable size field run length
2	I/B/N	additional repeat symbol
3	I/B/S/N	additional character repeat and variable length fields run length
4	I/2B/N	Repeat two additional symbol
5	I/2B/S/N	Repeat two additional characters and a variable size field run length
6	I/3B/N	three additional characters repeat
7	I/3B/S/N	Repeat three additional characters and a variable size field run length
8	2I/N	encoded characters repeat
9	2I/S/N	repeat the encoded symbols and variable length fields run length
10	2I/B/N	Repeat encoded character and an additional character repeat
11	2I/B/S/N	repeat the encoded symbols, additional character repeat and variable length fields run length
12	2I/2B/N	encoded symbol is repeated and two additional symbol repetition
13	2I/2B/S/N	repeat the encoded symbols, two additional character repeat and variable-length field size series
14	2I/B/2N(L/R)	repeat the encoded symbols, additional characters, and repeat two-segment length field series

IV. IMAGE QUANTIZATION FOR LOSSY COMPRESSION IN THE SPATIAL DOMAIN

Methods of image lossy compression based on an efficient coding of transform coefficients with their pre-quantization. Quantization makes many lossy techniques determines the mainly compression ratio. Coding of transform coefficients can achieve the greatest compression ratios due to the concentration of the primary energy in a relatively small number of significant transform coefficients. However, the transformation itself requires substantial computing resources and time. Therefore, an urgent task is to develop an algorithm for using the quantization of the pixel values of the image and their subsequent efficient coding.

Determined quantized pixel neighborhood (left or one of the top three), the closest in value to the central pixel. We calculate modulus of the difference of the pixel values of all the values of neighboring pixels in the neighborhood. If these differences less than the threshold Δ_S, the central pixel in the neighborhood is set to the value of the quantized value of the pixel in the vicinity. If this condition is not met, then the central pixel keeps its value.

V. EXPERIMENTAL RESULTS

The test images shown in Fig. 6 below -which includes different grayscale images- has been used to test our algorithms

Fig. 6. Test images used in the experiment

A. Evaluating loseless Algorithms

We got the results shown in Table 4 below, which shows the code sizes obtained for RLE, Huffman, Zip, Rar and all modified algorithms.

TABLE IV. CODE SIZE (BYTE) FOR LOSSLESS COMPRESSION

Algorithm	Code size (byte) image		
	M1	M2	M3
RLE	18620	77896	242604
I/S/N	18739	72412	143162
I/B/N	21504	91397	303235
I/B/S/N	20957	81407	165901
I/2B/N	14694	57583	106871
I/2B/S/N	17689	68649	118177
I/3B/N	14577	56701	103512
I/3B/S/N	17687	68581	117997
2I/N	30714	120377	208857
2I/S/N	34484	135364	261921
2I/B/N	32599	127870	222123
2I/B/S/N	32826	128820	225640
2I/2B/N	32713	128345	223002
2I/2B/S/N	32825	128753	225460
2I/B/2N(L/R)	30046	118028	208692
Zip	13070	44603	76470
Rar	12505	43062	75887
Huffman	14361	58281	113091

The Fig 7 below shows these results

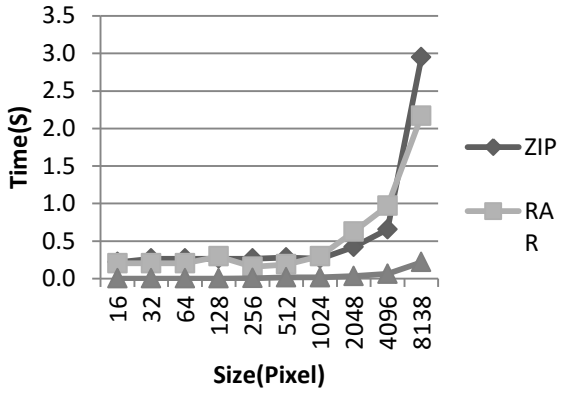


Fig. 7. Dependence of encoding time with image size for Zip, Rar and proposed RLE algorithms

From Table 4 we can see that the minimum amount of code provided by algorithms Rar, Zip and modified algorithm I/3B/N which shows about 26% of the worst result, but exceeds 57% and 8%, for RLE and Huffman respectively. Thus algorithms RLE, I/2B/N, I/3B/N provide comparison algorithms and Rar Zip decrease encoding time 31.2 and 53 times respectively, with an image size of 256×256 pixels, 16 and 18.5, 6 times, respectively, with image size 1024×1024 pixels, in 10 and 13.5 times respectively, when the image size 8192×8192 pixels.

B. Evaluating lossy Algorithms

Fig.6 shows the test images which used in the experiment, for which the coding efficiency analysis for lossy compression performed by RLE algorithm and its modifications with respect to the Zip, Rar, JPEG and JPEG2000

TABLE V. COMPRESSION TIME(SECOND) FOR LOSSY COMPRESSION

Algorithm	image compression time (Second)		
	M3	Lena	cameraman
RLE	0.00000085	0.00000241	0.00000085
I2/B/N	0.00000028	0.00000144	0.00000028
I3/B/N	0.00000142	0.000000962	0.00000028
RAR	0.249	0.219	0.300
ZIP	0.324	0.322	0.682
JPEG	0.073	0.200	0.116
JPEG 2000	0.102	0.182	0.190

For lossy compression based on RLE algorithm and its modifications I2BN and I3BN, Zip algorithms, Rar and Huffman algorithm used two-threshold progressive image quantization, we have performed a tests on the images above and we found that the probabilistic proposed RLE algorithms provides compression ratio up to 1.1-1.6 times, and 1.2-1.5 times as compared with Rar and Zip respectively. Also they are provides a ratio up to 1.4-2 times and 1.2-2.3 times according to RLE and Huffman respectively. The proposed algorithms provide a reduction in the mean square error MSE 3-10 times in comparison with JPEG 2000, and up to 1.2-18 times when the same proposed algorithm compared with JPEG.

Table 5 showing the compression time of lossy compression for the test images using different algorithms. From Table 5 the probabilistic proposed RLE algorithms provide a reduction in the compression time up to 1.7-6 times as compared with RLE, 170-1750 times as compared with the RAR, 230-2200 times as compared with ZIP, 50-400 times as compared with JPEG, 70-670 times as compared with the JPEG 2000.

VI. CONCLUSION AND FUTURE WORK

Results showed that these algorithms provide better performance in encoded image size compared with RLE and Huffman algorithms. The proposed lossless compression algorithms provide encoded image size reduction by 57% and 8% compared with RLE and Huffman algorithms, respectively; decrease encoding time 10-31 times and 13-53 times when changing the image size from 256×256 pixel to 8192×8192 pixel compared with Rar and Zip algorithms respectively. In lossy compression, the proposed algorithms provide improved image compression ratio up to 2 times in comparison with the algorithm RLE, and 2.3 times compared with the Huffman algorithm, reducing the mean square error (MSE) up to 10 times compared with the JPEG2000 compression algorithm when factor 2-5 time. also it is found that the proposed algorithms provide encoding time decreased up to 6 times compared with the RLE, 1750 times in comparison with RAR, 2200 times in comparison with ZIP, 400 times as compared with JPEG, and up to 670 times as compared with JPEG 2000.

For the future work it will be good to use the proposed algorithm with other modified algorithms especially to compress color images.

REFERENCES

[1] S. Smith, The Scientist and Enginer's Guide to Digital Signal Processing,2nd Edition, California Technical Publishing 1999.

[2] A. Bovik, The Essentia Guide to image Processing, Elsevier Inc. USA 2009.

[3] W. Pennebaker, J. Mitchell, JPEG Still Image Compression Standard, New York,Van Nostrand Reinhold, 1993.

[4] T. Ebrahimi, JPEG2000 still image coding versus other standards, Proc. of the SPIE, San Diego, CA, USA, Vol. 4115 July 2000.

[5] J. Miano, The formats and image compression algorithms in action, Publishing. Triumph, 2003.

[6] D. Salomon, Data compression: The Complete Reference, Springer-Verlag London Limited 2007.

[7] D. Vatolin, Data compression methods. Archiver device, image compression and video, Moscow Dialog-MIFI, 2003.

[8] S.W. Golomb, Run-Length Encoding, IEEE Transactions on Information Theory. July 1966.

[9] B. Karthikeyan, A Performance Analysis of Different Scanning Paths on Lossless Image Compression for Radiographic Welding Images, India, Journal of Scientific & Industrial Research Vol.73, April 2014.

[10] I. Made, Agus Dwi Suarjaya ,A New Algorithm for Data Compression Optimization, Udayana University Bali, Indonesia, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No.8, 2012

[11] R. Mala, M.Phil, S. Sarathadevi, A Lossless Image Compression using Modified Entropy Coding, International Journal on Recent and Innovation Trends in Computing and Communication, India, Vol. 2, Issue 8.

[12] S. Akhter and M. A. Haque, ECG COMPTRESSION USING RUN LENGTH ENCODING, Bangladesh University of Engineering and Technology, 18th European Signal Processing Conference Denmark, 2010.

[13] S. Joseph, A Novel Approach of Modified Run Length Encoding Scheme for High Speed Data Communication Application, International Journal of Science and Research (IJSR) Vol. 2 Issue 12, Dec 2013.

[14] M. VidyaSagar, J.S. Rose Victor, Modified Run Length Encoding Scheme for High Data Compression Rate, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol.2, Issue 12, Dec. 2013.

An Investigation and Comparison of Invasive Weed, Flower Pollination and Krill Evolutionary Algorithms

Marjan Abdeyazdan
Department of Computer Engineering
College of Electricity and Computer,
Mahshahr branch,
Islamic Azad University
Mahshahr, Iran

Samaneh Mehri Dehno
Department of Computer Engineering
College of Electricity and Computer,
Mahshahr branch,
Islamic Azad University
Mahshahr, Iran

Sayyed Hedayat Tarighinejad
Department of Computer Engineering
Yasooj Branch,
Islamic Azad University
Yasooj, Iran

***Abstract*—Being inspired by natural phenomena and available biological processes in the nature is one of the difficult methods of problem solving in computer sciences. Evolutionary methods are a set of algorithms that are inspired from the nature and are based on their evolutionary mechanisms. Unlike other optimizing methods of problem solving, evolutionary algorithms do not require any prerequisites and usually offer solutions very close to optimized answers. Based on their behavior, evolutionary algorithms are divided into two categories of biological processes based on plant behavior and animal behavior. Various evolutionary algorithms have been proposed so far to solve optimization problems, some of which include evolutionary algorithm of invasive weed and flower pollination algorithm that are inspired by plants and krill algorithm inspired by the animal algorithm of sea animals. In this paper, a comparison is made for the first time between the accuracy and rate of involvement in local optimization of these new evolutionary algorithms to identify the best algorithm in terms of efficiency. Results of various tests show that invasive weed algorithm is more efficient and accurate than flower pollination and krill algorithms.**

Keywords*—evolutionary algorithm; invasive weed algorithm; flower pollination algorithm; krill algorithm

I. INTRODUCTION

Evolution is a set of processes through which creatures have gradually learnt how to overcome the problems surrounding them and better interact with the environmental changes around them. In evolutionary processes, creatures that are more adapted to their environment are more likely to survive. Natural selection is considered one of the key terms in evolution and is defined as the process that creates different and various genes in animals over time and is known as one of the factors of formation of new species in nature. The environment surrounding these creatures can influence their characteristics and species that have become adapted to the environmental changes over time will continue living. Evolutionary algorithms are a set of algorithms inspired by laws of nature and Darwin's principles that play an important role in the form of an optimization problem in solving many practical issues of today. The complexity of real and practical issues will lead to the reduction of efficiency of traditional methods to solve such problems. The problem starts when nonlinear degree and the complexity of the atmosphere of problem solving are too difficult to be properly and efficiently solved by mathematical or gradient methods. A practical and optimization problem might be so complicated and non-linear that solving it using common methods might seem impossible to some extent. Using their dispersed population in problem solving environment, evolutionary algorithms have changed to one of the most efficient tools of such problem solving (Yang, 2010). Different evolutionary algorithms have been suggested so far for solving difficult and complicated problems. Unlike gradient-based methods, evolutionary algorithms are not deterministic and are inspired by random processes in nature. Evolutionary algorithms are a set of intelligent search algorithms that are able to search in the problem environment and be convergent with efficient answers with enough accuracy (Dasgupta and Michalewicz, 2013). Evolutionary algorithms perform based on different processes like genetics, evolution, ecosystem, swarm intelligence, etc.Charles Darwin has defined a set of fundamental laws for evolutionary rules that form the base of the science of Evolution. Evolutionary laws state that more adapted people are more likely to survive and continue their generation. Evolutionary behaviors are clear in all biological phenomena. For example, a specific kind of ringdove a kind of bird called Cuckoo) uses other birds' nests, whose eggs are similar to it, for laying eggs. In this reproduction behavior, the ringdove does not lie on its eggs and the victim bird takes care of all the eggs in its nest with the imagination that they are all its own. Ringdoves' behavior has not appeared overnight, but they have learnt over time that they can increase the probability of the survival of their generation by putting their eggs among other birds' eggs (Ghose et al., 2015). Evolutionary behaviors are clearly detectable among creatures like fireflies, dolphins, spiders, ants and bees. Evolutionary behaviors also exist among microscopic organisms like bacteria and body cells whose target is surviving and behaviors like being yokemates (Tripathy and Mishara, 2015). Plants' behaviors have also been formed reproducing. For example, bacteria have reached resistance against anti-biotic over time and this drug resistance is due to revolutionary based on evolution and natural selection. Getting flowers for reproduction or photosynthesis to produce sugar are some examples. Plants' behavior for surviving and adapting to their environment is one of the interesting evolutionary behaviors in nature. Plants compete with other plants to gain different resources such as water, soil and sunlight to survive. One of evolutionary algorithms that is formed based on plants behaviors and the competition between them is invasive weed algorithm (Mehrabian and Lucas, 2006). In this algorithm, any plant that is more adapted is more likely to survive by producing more seeds. Flower pollination algorithm is also an

evolutionary algorithm based on plant behavior that flowers' pollination is considered a vital issue in survival of flowers and consequently reproduction of the plant (Yang, 2012). Evolution-based behaviors are not just limited to animals and plants, but a set of behaviors in physical phenomena show rules and principles to create nature-based algorithms. For example, water drop behavior on river paths can be modeled using physics rules and principles and this behavior can be used today to solve the difficult problems. Being inspired by different behaviors of creatures like social interactions and human emotions can inspire formation of evolutionary methods.

Evolutionary algorithms attempt to use biological, social and natural processes to solve difficult problems and overcome available challenges like these phenomena. A wide range of evolutionary algorithms have been proposed today that indicate the significance of this computing branch in computer sciences. In this paper, the evolutionary algorithms of invasive weed and flower pollination as algorithms inspired by plants and krill algorithm as the evolutionary algorithm of animals (Alavi and Gandomi, 2012) will be investigated and in the following the accuracy and convergence of these evolutionary algorithms will be compared using a set of benchmark functions.

A. Invasive Weed Algorithm

Invasive weeds in a common and single definition are plants that are not the aim of farmers but they are growing on the farms. Any tree, bush, shrub or plant branch or leaf might be recognized as invasive weed. Invasive weeds are highly adaptive in growth and reproduction on farms. Their growth prevents from the growth of products and waste of resources such as soil, water and fertilizers. The life style of plants and invasive weeds follows a specific cycle of reproduction. Invasive weeds produce a specific number of seeds according to the properties and suitability of the plant and then the seeds are transformed into invasive weeds that compete with each other in absorbing water, sunlight, and soil and so on and only stronger and more adaptive plants survive. The interesting and yet simple mechanism of invasive weed reproduction has led to the invention of an algorithm based on these plants that is called Invasive Weed Optimization (IWO) (Mehrabian and Lucas, 2006). Invasive weed algorithm has been modeled based on the reproduction cycle of invasive weeds. In this algorithm, first plant seeds are scattered randomly in the problem space and the primary invasive weeds are formed by the growth of the seeds and each invasive weed creates a number of seeds around it depending on its fitness and these seeds grow and reproduce to compete with mother and other plants. Modeling plant behavior, the invasive weed algorithm attempts to solve the difficult and optimization problems. The general stages of this algorithm are as follows:

- Each plant scatters a number of seeds in the environment based on its fitness.
- The level of seed production around each plant is defined according to the fitness of that plant.
- Transmittal of seeds around parent plants with normal distribution.
- Seeds growth and production of new plans and investigation of plant fitness.
- Repetition of previous stages to reach proper convergence.

Invasive weed algorithm makes use of a multi-step and repetitive process to reach the desired answers to solve different optimization problems. This algorithm defines a fitness for each plant that can be defined based on the maximum or minimum of objective function.

First, each plant scatters a number of seeds in the environment based on its fitness. The number of seeds produced is defined according to the fitness of the plant. In figure (1), a linear relationship (the simplest pattern of seed production) is used to calculate the number of produced seeds around each plant. As it is shown in the figure below, the maximum number of seeds is produced by colony and invasive weed that had the most fitness. Similarly, the minimum number of seeds is produced by the invasive weed that had the least fitness (Mehrabian and Lucas, 2006).

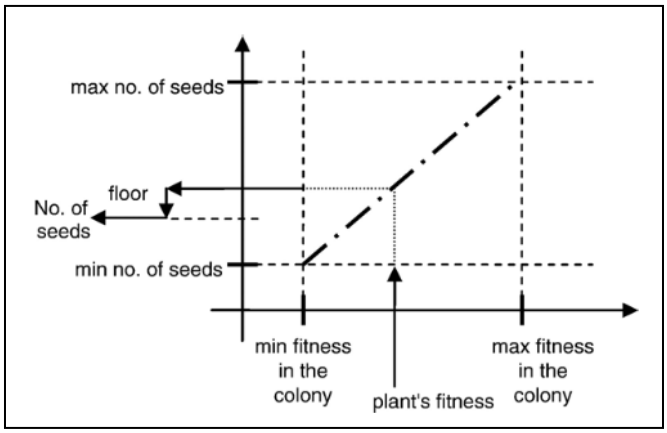


Fig. 1. Linear relationship between fitness and seed production of an invasive weed

In order to calculate the number of produced seeds by plant C, it suffices to write the line equation of figure (1) that passes through initial and final points as it is shown in equation (1).

Equation (1)

$$s = s_{min} + \frac{s_{max} - s_{min}}{fitness_{max} - fitness_{min}}(fitness - fitness_{min})$$

In this equation, smax, smin, fitnessmax and fitnessmin are maximum number of produced seeds, minimum number of produced seeds, maximum fitness of invasive weed and minimum fitness of invasive weed, respectively. S is the number of seeds produced and fitness is the level of fitness of an invasive weed that target seed has produced.

Transmittal of seeds around the parent plant is one of the significant stages of invasive weed algorithm. Seed transmittal can be considered based on a normal distribution because it can properly model distribution behavior of random and natural phenomena.

B. Flower Pollination Algorithm

There are about 250 thousand plant species and about 80 percent of them can produce flowers. It has been about 125 million years since the evolution of flowers in cretaceous period and imagining a world with plants without flowers is impossible. Flowers can be considered sexual organs of reproduction in flowering plants. Flowers play a crucial role in pollination to reproduce plants. Flower pollination is done in different methods like insects, birds or other animals. There are a number of plants that only let insects and animals do the pollination; that is, pollination in these plants takes place in a special and advanced way. About 90 percent of flowering plants need creatures like birds and insects for pollination and only 10 percent of them do not require insects or birds for pollination. Plants like willow and grass are only dependent on wind and rain for pollination. Therefore, they do not have petals to attract insects and their flower organs are simple. Pollination by insects such as honey bees is usually done only on one flower specie and they guarantee the reproduction of that flowering plant and increase the probability of its survival. Pollination of a flower by a specific insect with short memory that is able to learn only a limited amount will make the insect focus on a few number of flowers to find food and search less. On the other hand, searching for plants with good nectar for the insect might be time consuming and costly. Two types of pollination namely self-pollination and cross pollination can be observed in flowering plants. In self-pollination, pollens of a flower from a plant are placed on another flower of it and in cross pollination, pollens of one flower from one plant are placed on the flower of another plant. An example of self-pollination is in peach plant, in which the pollens of one flower are placed on another flower of the same tree. In global (cross) pollination, pollens of flowering plants are taken to different distances by insects. Considering the characteristics of pollination in flowering plants, the following four simple rules can be presented for modeling the algorithm (Yang, 2012):

1) Since cross pollination is done by insects flying pollens of flowers, it is considered as global pollination.

2) Self-pollination is considered a local pollination.

3) The probability of flower constancy is presented in a probability function of the similarity of the flower that has pollinated with this flower.

4) Selection of local pollination or global pollination of a flower is considered a probability in the p ϵ [0, 1] interval.

The global pollination of a flower is modeled in a mathematical equation in equation (2):

Equation (2):

$$\boldsymbol{x_i^{t+1} = x_i^t + L(x_i^t - g_*)}$$

In this equation, $\boldsymbol{x_i^t}$, $\boldsymbol{x_i^{t+1}}$, $\boldsymbol{g_*}$ and L are the ith place of pollen in the tth replication, the ith place of pollen in t+1th replication, the best place of pollen found so far and pollination power that show the direction and jumping of pollens, respectively. Pollen i or the solution vector of xi are the best solutions so far for t and $\boldsymbol{g_*}$ among all the current solutions of the current/repetition generation. Parameter L is pollination power which in fact is a step in equation (3) that has shown this probability distribution.

Equation (3):

$$\mathbf{L} \sim \frac{\boldsymbol{\lambda \Gamma(\lambda) sin(\frac{\pi\lambda}{2})}}{\boldsymbol{\pi}} \times \frac{\mathbf{1}}{\boldsymbol{s^{1+\lambda}}}, \boldsymbol{s > s_0 > 0}$$

In this equation, $\Gamma(\lambda)$ is the standard gamma function and the appropriate value for this function is $\lambda = 1.5$. Local pollination or self-pollination of the flowers can also be defined using equation (4).

Equation (4):

$$\boldsymbol{x_i^{t+1} = x_i^t + \varepsilon(x_j^t - x_k^t)}$$

In this equation, xtj and xtk are two different groups that are formed by similar flowers.

Krill ALGORITHM

FORMATION of categories and groups between sea animals is not merely a random phenomenon and many studies have been carried out on it (Gharavian et al., 2013). Living in groups or herds allows sea creatures to confuse attackers and on the other hand, leads them towards food sources. Many mathematical models have been proposed to describe the behavior of creatures that live in groups or herds (Wang et al, 2014). One of the creatures that live in great groups is Antarctic Krill that are sometimes referred to as Free Sea Krill. These creatures are capable of creating a group with 10 to 100 meters of radius in a short period of time and they can even join other groups and create even bigger groups. Each one of these groups can travel the sea or ocean in parallel. Many creatures like seals, penguins and birds attack krill. The purpose of attackers is to scatter these creatures from their group or herd and hunt them easily. Results of experimental cstudies have shown that hunting krill that are not in the herd is easy, while when the krill are moving in groups, the attackers will be confused due to krill' parallel motions and consequently the probability of hunting them decreases. When the group or herd of krill is attacked by attackers, their herd is destroyed and they tend to join the closes herd or group to decrease the probability of being hunted. In general, creating a herd or group by krill is a multi-purpose process, whose two purposes are stated bellow (Gandomi and Alavi, 2012).

- Increasing krill density
- Reaching food

Krill algorithm is a metaheuristic algorithm based on the behavior of increasing density and searching for food that is modeled to solve optimization problems. Krill algorithm tries to guide these creatures to places with higher density and more food. The objective function in krill algorithm is modeled in an area using krill density and amount of food. I krill algorithm, the global optimized points are the ones that include high density of krill and high amount of food that krill are finally guided there. Three main behaviors are considered for the motion of krill in krill algorithm (Gandomi and Alavi, 2012):

- Motion included by other krill individuals
- Foraging activity

- Random diffusion

Each krill moving in searching spaces of the problem is stated and modeled based on a lagrangian model that is shown in equation (5)

Equation (5):

$$\frac{dX_i}{dt} = N_i + F_i + D_i$$

Where, Ni, Fi and Di are the motion of ith krill to other individual krill, motion of ith krill for foraging and random motion of ith krill in searching spaces of problem. In this equation, Xi is the location vector of ith krill and $\frac{dX_i}{dt}$ is the speed of the ith krill at tth time. Individual krill can move according to equation (6). In fact, the effect of repulsion and attraction of each krill on other specific krill is modeled in this equation (Gandomi and Alavi, 2012).

Equation (6):

$$N_i^{new} = N^{max}\alpha_i + \omega_n N_i^{old}$$

Where, N_i^{new} is the speed and motion of the new i krill in problem space.

N^{max}: The maximum possible speed of an individual krill and experimentally it is commonly measured to be $N^{max} = 0.01 ms^{-1}$

N_i^{old}: The previous speed and motion of ith individual krill in problem space

α_i: Direction of ith individual krill in problem space

ω_n: Inertia and weight of ith individual krill in problem space that is commonly a random number between zero and one.

α_i: Local effect of ith individual krill that shows the angle of krill motion for local search.

Foraging motion of krill refers to the state in which krill move towards points with more food. In fact, minimization of the distance between the krill and food is an objective function for krill in foraging food. Krill use two important factors of food location and previous experience about the food location in foraging food. Motion vector of ith krill is defined in equation (7).

Equation (7):

$$F_i = V_f . \beta_i + \omega_f . F_i^{old}$$

Where,

V_f: Foraging speed and is usually considered to be $V_f = 0.02 ms^{-1}$

ω_f: Inertia weight in foraging food.

F_i^{old}: Last foraging motion value

F_i: Speed vector that is being used in foraging right now

Apart from motions influenced by other krill and foraging motions, each krill has another random motion that is called physical diffusion or random motion. The random motion of each krill can be shown in maximum speed value and a random vector like equation (8).

Equation (8):

$$D_i = D^{max}\delta$$

Where,

D^{max} and δ are maximum speed and random value between [-1, +1]. The appropriate value for maximum speed has been determined to be $D^{max} \in [0.002 , 0.01]$ (Gandomi and Alavi, 2012).

Having modeled the triple motions of the krill, the direction of krill motions can be determined in period from t to t + Δt using equation (9).

Equation (9):

$$X_i(t + \Delta t) = X_i(t) + \Delta t \frac{dX_i}{dt}$$

II. FINDINGS

One of the methods of evaluating evolutionary algorithm efficiency is using mathematical evaluative functions, four of which are presented in table (1) along with the formulas and target range (Moré and Wild, 2009).

TABLE I. EVALUATIVE FUNCTIONS USED IN THE RESEARCH

Name of evaluative function	*Mathematical formulas*	*Rang*
Sphere	$f(x) = \sum_{i=1}^{D} x_i^2$	$(-10, 10)^D$
Rastrigin	$f(x) = \sum_{i=1}^{D} [x_i^2 - 10\cos(2\pi x_i) + 10]$	$(-5, 5)^D$
Griewank	$f(x) = \frac{1}{4000}\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} \cos(\frac{x_i}{\sqrt{i}}) + 1$	$(-10, 10)^D$
Rosenbrock	$f(x) = \sum_{i=1}^{D-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	$(-10, 10)^D$

Sphere and Rosenbrock evaluative functions have global minimums and Griewank and Rastrigin evaluative functions have local minimums apart from having global minimums.

Sphere and Rosenbrock have simpler forms compared to Rastrigin and Griewank evaluative functions. Complexity of an evaluative function indicates that it is a more difficult evaluative criterion. The efficiency of evolutionary algorithms

is higher when they are more convergent with global minimum than local minimum and also when they find the global minimum more accurately in problem space. Invasive weed, flower pollination and krill evolutionary algorithms are carefully assessed and compared in this part and results obtained from 50 different tests for the three algorithms of invasive weed, flower pollination and krill have been applied on Sphere, Rastrigin, Griewank and Rosenbrock evaluative functions with initial populations of 40 and replications of 50, 40, 30, 20 and 100, respectively. The chart in figure (2) shows the convergence of these evolutionary algorithms for each replication. As the convergence chart of the three algorithms show, the invasive weed algorithm is faster optimized to the desired answers than flower pollination and krill algorithms. On the other hand, the level of convergence of flower pollination algorithm is better than that of Antarctic krill algorithm.

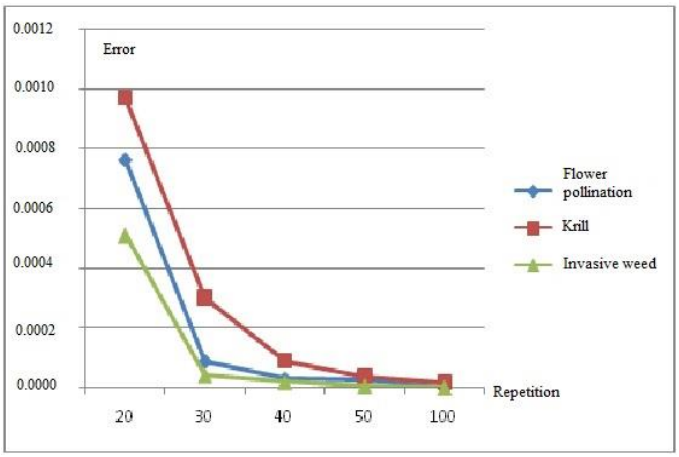


Fig. 2. Comparison of convergence of invasive weed, flower pollination and krill algorithms

Rastrigin and Griewank evaluative functions have local optimizations and there is the possibility that an evolutionary algorithm be convergent to the local optimization instead of being convergent to global optimization. The reduction in convergence rate to local optimizations is one of the most significant indices of a good and accurate evolutionary algorithm. One of the criteria of comparison between evolutionary algorithms is the tendency to convergence to global optimization answers and being away from local optimizations. For example, invasive weed algorithm can properly calculate the global optimization at $x = 0.004$, $y = 0.000$, $f_{min} = 0.005$ in Rastrigin evaluative function. While krill algorithm has been converged to a local optimization at $x = 0.9952$, $y = 1.9898$, $f_{min} = 4.9748$. the percentage of being involved in local optimization for Grienwank evaluative function by invasive weed, flower pollination and krill algorithms is %11, %15 and %18, respectively which shows that compared to flower pollination and krill algorithms, invasive weed algorithm has less tendency to be converged to local optimization.

III. Discussion and Conclusion

Optimization problems include a wide range of applications, especially in industrial designing, planning, timing, and etc. Real optimization problems are usually non-linear and complicated, such that they are not solved by common mathematical methods, gradient and numerical calculations. One of the effective methods of optimization problem solving is using evolutionary algorithms inspired by the nature. In this paper, three evolutionary algorithms of invasive weed, flower pollination and krill, which are less known, were studied in terms of accuracy and convergence. Test results show that invasive weed algorithm has a better efficiency than flower pollination and krill algorithms. Moreover, Mehrabian and Lucas (2006) also showed that invasive weed algorithm has better convergence compared to particles, genetics, differential difference and frog jump algorithms.

Considering the proper convergence of invasive weed algorithm compared to other evolutionary algorithms, the future paper aims at presenting an improved version of this algorithm that can also pollinate so that it will increase in accuracy and convergence.

References

[1] Yang, X. S. (2010). Engineering optimization: an introduction with metaheuristic applications. John Wiley & Sons.

[2] Dasgupta, D., & Michalewicz, Z. (Eds.). (2013). Evolutionary algorithms in engineering applications. Springer Science & Business Media.

[3] Ghose, R., Das, T., Saha, A., Das, T., & Chattopadhyay, S. P. (2015, October). Cuckoo search algorithm for speech recognition. In Computing and Communication (IEMCON), 2015 International Conference and Workshop on(pp. 1-5). IEEE.

[4] Tripathy, M., & Mishra, S. (2015). Coordinated tuning of PSS and TCSC to improve Hopf Bifurcation margin in multimachine power system by a modified Bacteria Foraging Algorithm. International journal of electrical power & energy systems, 66, 97-109.

[5] Mehrabian, A. R., & Lucas, C. (2006). A novel numerical optimization algorithm inspired from weed colonization. Ecological informatics, 1(4), 355-366.

[6] Yang, X. S. (2012). Flower pollination algorithm for global optimization. InUnconventional computation and natural computation (pp. 240-249). Springer Berlin Heidelberg.

[7] Gandomi, A. H., & Alavi, A. H. (2012). Krill herd: a new bio-inspired optimization algorithm. Communications in Nonlinear Science and Numerical Simulation, 17(12), 4831-4845.

[8] Gharavian, L., Yaghoobi, M., & Keshavarzian, P. (2013, April). Combination of krill herd algorithm with chaos theory in global optimization problems. In AI & Robotics and 5th RoboCup Iran Open International Symposium (RIOS), 2013 3rd Joint Conference of (pp. 1-6). IEEE

[9] ang, G. G., Gandomi, A. H., & Alavi, A. H. (2014). Stud krill herd algorithm. Neurocomputing, 128, 363-370.

[10] Moré, J. J., & Wild, S. M. (2009). Benchmarking derivative-free optimization algorithms. SIAM Journal on Optimization, 20(1), 172-191.

Mobile Forensic Images and Videos Signature Pattern Matching using M-Aho-Corasick

Yusoof Mohammed Hasheem, Kamaruddin Malik Mohamad, Ahmed Nur Elmi Abdi, Rashid Naseem
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia

***Abstract*—Mobile forensics is an exciting new field of research. An increasing number of Open source and commercial digital forensics tools are focusing on less time during digital forensic examination. There is a major issue affecting some mobile forensic tools that allow the tools to spend much time during the forensic examination. It is caused by implementation of poor file searching algorithms by some forensic tool developers. This research is focusing on reducing the time taken to search for a file by proposing a novel, multi-pattern signature matching algorithm called M-Aho-Corasick which is adapted from the original Aho-Corasick algorithm. Experiments are conducted on five different datasets which one of the data sets is obtained from Digital Forensic Research Workshop (DFRWS 2010). Comparisons are made between M-Aho-Corasick using M_Triage with Dec0de, Lifter, XRY, and Xaver. The result shows that M-Aho-Corasick using M_Triage has reduced the searching time by 75% as compared to Dec0de, 36% as compared to Lifter, 28% as compared to XRY, and 71% as compared to Xaver. Thus, M-Aho-Corasick using M_Triage tool is more efficient than Dec0de, Lifter, XRY, and Xaver in avoiding the extraction of high number of false positive results.**

Keywords—mobile forensics; Images; Videos; M-Aho-Corasick; (File Signature Pattern Matching)

I. INTRODUCTION

In the last few decades, Digital forensic (DF) plays a paramount part not entirely in availing in cracking cases against mobile phone malefactions like drug dealing, child trafficking, and arms trade. Mobile phone capabilities increase in public presentation, recollection capability and multimedia functionality turning phones into data pools that can fortify a wide range of personal information [1]. Nowadays mobile phone, personal digital assistant (PDA) and the Internet are widely accepted around the world were mobile phone became a component of our quotidian life activities due to rapid development in mobile phone technology. Mobile phone becomes personal and was habituated to avail in multimedia and personal task [2]. However, data held on mobile contrivances can be utilizable and paramount to law enforcement agencies when carrying an investigation in either civil or malefactor transactions. There are two ways of recuperating digital evidence, traditional data instauration, and file carving. The Traditional data integration is a customary technique applied to retrieve digital information where the metadata or file allocation table subsists. While on the other hand file carving was introduced to give assistance in malefactor cases where the traditional data recuperation techniques cannot be worked out. In this paper a new technique called images and videos signature pattern matching using M-Aho-Corasick is proposed to efficiently search for images and videos file from damaged mobile phone using M_Triage tool.

A. M_Aho-Crasick

One of the main components in M_Triage tool that efficiently search for images and videos utilizing multi-pattern signature matching is the M-Aho-Corasick algorithm. The algorithm is habituated and modified from the pristine algorithm kenned as Aho-Corasick, where the failure links function is abstracted and superseded with a signature database which contains all the pattern and file structure that pertains to investigator stored in it [3]. M-Aho-Corasick algorithm has remained constructed utilized for the set of patterns $D=C_1,C_2,...,C_K$ of total length $n=|n1|+|n2|+...+|nK|$. Entirely the patterns of interest have the same signature. The algorithm search for patterns in such a way that: The algorithm crisscross if a pattern *P* of length *m* is a subpattern in *O(m)* time. The algorithm discovers the main subsistence of the patterns *P 1 ,...,P q* of total length *m* as subpattern in *O(m)* time. The algorithm additionally discovers all *z* existences of the patterns *P 1 ,...,P q* of total length *m* as subpattern in *O(m+z)* time [4].

B. M_Aho-Crasick Algorithm

During the probing in M-Aho-Corasick algorithm, a file pattern like JPEG, 3gp and MP4 are probing by building their signature database, then followed by building the block tree and integrating pattern ID's into the tree utilizing automation. The signature are probing predicated on finite state machines (FSM) and if the pattern is probing within the dump file, the pattern will be compared with the once in the signature database for identifying the file of interest. If the signature is matched, then *"go to"* function will be called to mark the address of the valid signature and peregrinate to the next block. If the signature of interest is not found, then skip the block to the next block. Fig. 1 shows the probing pattern process utilizing M-Aho-Corasick algorithm.

Step 1 Start.

Step 2 Build the block tree and adding pattern ID's into the tree using automation.

Step 2 Search pattern within the dump file using finite state machines (FSM).

Step 4 Compared pattern with the once in the signature database for identifying the file of interest.

Step 5 If the signature is matched, then call "go to" function to mark the address of the valid signature.

Step 6 Move to the next block.

Step 7 If the signature of interest is not found, then skip the block to the next block.

Fig. 1. M-Aho-Corasick algorithm

C. M_Aho-Crasick Implementation

During the implementation an accepted algorithm known as Aho-Corasick (AC) is adapted [3]. The adapted algorithm has a clear implementation and understandable codebase which is illustrated in Fig 2 known as M_Aho- Corasick. Rudimentary, M-Aho-Corasick takes a set of finite pattern file signature as an array and an input file signature and outputs the details on the patterns matched such as their positions in the input signature. Nevertheless, M-Aho-Corasick takes both the input file signature and the set of patterns from reading the files; hence, both sets of the signature can be given as files.

In M-Aho-Corasick design, the algorithm reconstituted the state machines without the failure link transitions and as shown in Fig 3 for image files search and Fig 4 for multimedia file search. The algorithm transmuted the probing method with respect to the failure fewer transitions

```
public patternSearchResult[] FindAll(pattern)
{
  ArrayList ret=new ArrayList();
  TreeNode ptr=_root;
  int index=0;
  while(index<pattern.Length)
{
    TreeNode trans=null;
    while(trans==null)
{
      trans=ptr.GetTransition(pattern [index]);
      if (ptr==_root) break;
      if (trans==null) ptr=ptr.skipblock;
}
    if (trans!=null) ptr=trans;
    foreach(pattern found in ptr.Results)
      ret.Add(new patternSearchResult(index-found.Length+1,found));
    index++;
}
  return (patternSearchResult[])ret.ToArray(typeof(PatternSearchResult));
}
```

Fig. 2. 2 M-Aho-Corasick implementation

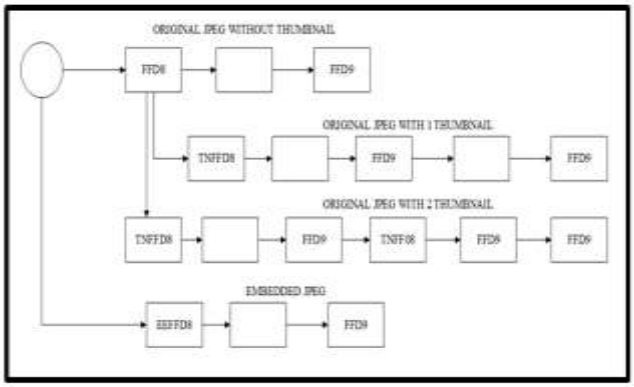


Fig. 3. Images pattern search

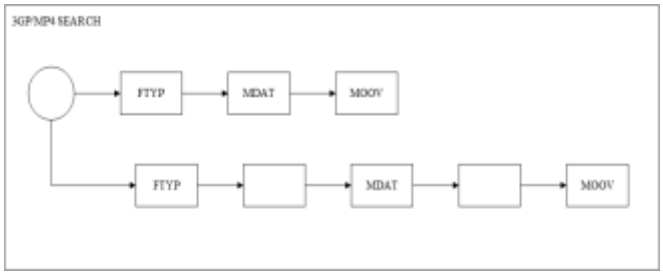


Fig. 4. 3GP/MP4 pattern search

II. Related Work

As it became necessary to find more efficient file signature matching algorithms and their implementations, a significant number of researchers have been being carried out in this area [3]. This literature grants some research work achieved on file signature matching algorithms.

Among the researchers one adapts the original Aho-Corasick and modified it to be known as Parallel Failure less Aho-Corasick (PFAC) implementation, all failure transitions are abstracted from the state machine, to enable the (PFAC) algorithm to probe for a string file in parallel.

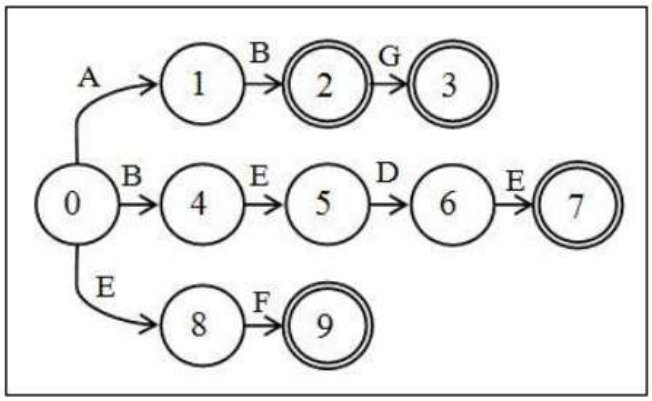


Fig. 5. Parallel Failure less Aho-Corasick (PFAC) [3].

Fig. 5 illustrates the modified state machine for the PFAC implementation for the same four patterns, ABG, BEDE and they used EF in their last example. The four patterns are the target probed files.

In 2011, another researcher [5] developed a mobile forensic triage tool called Dec0de, they acclimates techniques from natural language processing. They propose an efficient and flexible utilization of probabilistic finite state machines (PFSMs) to encode typical data structures. They utilize the engendered PFSMs along with a classic dynamic programming algorithm to find the maximum likelihood parse of the phone's recollection. DEC0DE uses Viterbi Algorithm twice. First, it passes the filtered byte stream to Viterbi with the Field PFSM as input. The output of the first pass is the most likely sequence of generic fields associated with the byte stream. That field sequence is then inputted to Viterbi along with the Record PFSM for a second pass [5]. They refer to these two phases as field-level and record-level inference, respectively. This type of file searching is time consuming because the process doesn't perform in parallel as Aho-Corasick. Furthermore, in 2014 the researcher proposed another tool called LIFTER [6] with the intention to improve the searching performance of their previous tool known as Dec0de by applying the new technique called initial ranking and relevance feedback.

The implement LIFTR's early classification orders single pages that have a file system sheltered; consummately other pages will have a zero quality score after initial ranking. Throughout the pertinence feedback stage, the implement LIFTR endeavors to increment the initial ranking by utilizing investigator feedback [6]. This sort of file searching is also time consuming because the performance of the system depends on the investigator initial ranking and feedback.

Bulk Extractor is another digital triage forensic tool implements by [7]. Information from digital evidence files like credit card numbers, email addresses, and URLs are extracted correctly using the bulk extractor command-line tool. The tool extracts evidence from the raw disk images. The disk image is split into pages and one or more scanners are used to process the pages after is being split for triage examination [7].

III. EXPERIMENTATION

This section discusses the experimental setup for M-Aho-Coriasick using M_Triage tool. During the experiment, additional datasets from Digital Forensic Reseach Workshop (DFRWS 2010) which is purposely created to solve the research problems regarding efficient file search. M_Triage is developed to efficiently search for valid address book, call logs, SMS, images, and videos.

A. Dataset Preparation

As mention earlier, Dataset from Digital Forensic Reseach Workshop (DFRWS 2010). And the once extracted using JTAG are chosen as the input to validate the output of the proposed technique. However, due to this flexibility of recovering any leads that might connect Monsieur Victor [8] to other individuals, companies, or bank accounts that are involved in his international arms business, valid address-book, call logs, SMS, images and videos, parameters are considered in this experiment as data of interest to validate M_Triage tool. Fig 6. presents the five data set and their total number of files which is used for the experiment.

DATASETPHONE_A	NUMBER OF FILE
ADRESS BOOK	11
CALL LOG	5
SMS	1
IMAGES	0
VIDEOS	0
TOTAL FILES	17

DATASETPHONE_B	NUMBER OF FILE
ADRESS BOOK	134
CALL LOG	20
SMS	12
IMAGES	293
VIDEOS	4
TOTAL FILES	463

DATASETPHONE_C	NUMBER OF FILE
ADRESS BOOK	358
CALL LOG	90
SMS	10
IMAGES	5
VIDEOS	0
TOTAL FILES	463

DATASETPHONE_D	NUMBER OF FILE
ADRESS BOOK	698
CALL LOG	120
SMS	16
IMAGES	70
VIDEOS	5
TOTAL FILES	909

DATASETPHONE_E	NUMBER OF FILE
ADRESS BOOK	1350
CALL LOG	107
SMS	41
IMAGES	471
VIDEOS	8
TOTAL FILES	1977

Fig. 6. Total number of files in Phone, A, B, C, D and E

B. Experiment on Dataset

- Datase Phone A

M_Triage processed datasetphoneA in 0.21 seconds while de0de in 0.50 seconds, Lifter in 0.30 seconds, XRY in 0.29 seconds while Xaver in 1 minute 20 second.

- Datase Phone B

M_Triage processed datasetphoneB in 1 minutes 10 seconds while de0de in 4 minutes 12 seconds, Lifter in 2 minutes 20 seconds, XRY in 2 minutes 17 seconds while Xaver in 5 minutes 23 seconds.

- Datase Phone C

M_Triage processed datasetphoneC in 4 minutes 20 seconds while De0de in 12 minutes 10 seconds, Lifter in 6 minutes, XRY in 5 minutes 30 seconds while Xaver in 15 minutes 40 second.

- Datase Phone D

M_Triage processed datasetphoneD in 6 minutes 57 seconds while de0de in 24 minutes 30 seconds, Lifter in 10 minutes 50 seconds, XRY in 8 minutes 20 seconds while Xaver in 25 minutes 47 seconds.

- Datase Phone E

M_Triage processed datasetphoneE in 13 minutes 40 seconds while de0de in 60 minutes, Lifter in 21 minutes, XRY in 19 minutes 35 seconds while Xaver in 41 minutes 30 seconds.

In order to justify the experiment refer to Fig. 7 and 8 for the result.

Universiti Tun Hussein Onn Malaysia/ Oricc

IV. Results and Discussion

This Section discusses the final result of the M-Aho-Corasick using M_Triage tool for performing efficient file search. A dataset from DFRWS 2010 and another four dataset are used for the experiment. The result is discussed in this section. The test of the experiment is performed in the context of mobile forensics. The examination is conducted on a set of real objects smartphone and future mobile phone, with functional and operational characteristics also different from each other.

Fig 7 and 8 present the time taken by all the recovery tools used during the experiment. Each tool is run ten different times using each data set. For each runs a time, taken value is obtained by each tool.

		Phone A	Phone B	Phone C	Phone D	Phone E	
	Size (MB)	2MB	66MB	98.8	124	400	Average Time
M	M-Triage	0.21	1.10	4.20	6.57	13.40	5.10
D	Dec0de	0.58	4.12	12.10	24.30	60.00	20.22
L	Lifter	0.30	2.20	6.00	10.50	21.00	8.00
X	Xry	0.29	2.17	5.30	8.20	19.35	7.06
Xa	Xarver	1.20	5.23	15.40	25.47	41.30	17.72
D-M		0.37	3.02	7.90	17.73	46.60	15.12
L-M		0.09	1.10	1.80	3.93	7.60	2.90
X-M		0.08	1.07	1.10	1.63	5.95	1.97
Xa-M		0.99	4.13	11.20	18.90	27.90	12.62
Improvement in Percentage	Dec0de	64	73	65	73	78	75
	Lifter	30	50	30	37	36	36
	Xry	28	49	21	20	31	28
	Xarver	83	79	73	74	68	71

Fig. 7. Computational time comparison for all tools and all data set

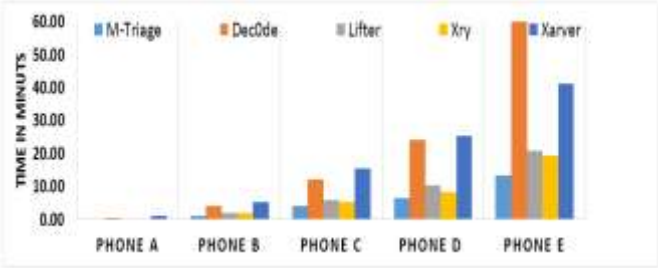


Fig. 8. The summary of computational time comparison between various forensic tools in graph

The result reported the average values of these ten runs, where Pim stands for percentage improvement and ET time stands for existing tools, also MT time stands for M_Triage time.

$$P_{im}MT(Existing\ Tool) = \frac{ET\ Time - MT\ Time}{ET\ Time} * 100 \qquad (1)$$

Furthermore, to calculate the percentage improvement, equation (1) is used as ETtime-MTtime divide by ETtime multiply by 100. For instance, in Fig 7 PhoneA has the size of 2.26MB, and M_Triage complete the execution time in 0.21 seconds while Dec0de completed in 0.58, to calculate the percentage improvement 0.58 minus 0.21 is equal to 0.37 then divide by 0.58 and multiply by 100, the result shows that in PhoneA size 2.26MB, 64% is achieved over Dec0de by M_Triage. The calculation steps apply to all data set used in the experiment.

V. Conclusions

This paper addresses the issue of searching a file during Digital forensic examination. A novel, multi-pattern signature matching algorithm called M-Aho-Corasick is developed in M_Triage to address such problem. Experiments are conducted on five different datasets which one of the data sets is obtained from DFRWS 2010. Comparisons are made between M-Aho-Corasick using M_Triage with Dec0de, Lifter, XRY, and Xaver. The result shows that M-Aho-Corasick using M_Triage has reduced the searching time by 75% as compared to Dec0de, 36% as compared to Lifter, 28% as compared to XRY, and 71% as compared to Xaver (refer to Fig. 7). Thus, this shows that M-Aho-Corasick using M_Triage is much more stable for searching a file during the forensic examination.

Acknowledgment

The authors would like to acknowledge Universiti Tun Hussein Onn Malaysia (UTHM), for providing financial support (Vot U061) towards this research.

References

[1] K. Curran, A. Robinson, S. Peacocke, and S. Cassidy, "Mobile Phone Forensic Analysis," vol. 2, no. 2, 2010.

[2] N. A. Abdullah, R. Ibrahim, and K. M. Mohamad, "An IMPROVE file carver of intertwined jpeg images using X-mykarve," UNIVERSITY TUN HUSSEIN ONN MALAYSIA, 2014.

[3] S. Arudchutha, T. Nishanthy, and R. G. Ragel, "String matching with multicore CPUs: Performing better with the Aho-Corasick algorithm," 2013 IEEE 8th Int. Conf. Ind. Inf. Syst. ICIIS 2013 - Conf. Proc., pp. 231–236, 2013.

[4] L. Benuskova, "Lecture 4 : Exact string searching algorithms," http://marknelson.us/1996/08/01/suffix-trees/, vol. 2, no. 1, p. 6, 2012.

[5] R. J. Walls, E. Learned-miller, and B. N. Levine, "Forensic Triage for Mobile Phones with DEC0DE," 2011.

[6] R. J. Walls and B. N. Levine, "Efficient Smart Phone Forensics Based on Relevance Feedback," 2014.

[7] S. L. Garfinkel, "Digital media triage with bulk data analysis and bulk-extractor," Comput. Secur., vol. 32, pp. 56–72, 2013.

[8] J. Blokhuis and A. Puppe, "DFRWS Challenge 2010 - Mobile forensics," 2010.

Visual Knowledge Generation from Data Mining Patterns for Decision-Making

Jihed Elouni
METS Micro Electro Thermal Systems
University of Sfax, National School of Engineers (ENIS)
Sfax, Tunisia

Mounir Ben Ayed
REGIM: REsearch Groups in Intelligent Machines
University of Sfax, National School of Engineers (ENIS)
Sfax, Tunisia

Hela Ltifi
REGIM: REsearch Groups in Intelligent Machines
University of Sfax, National School of Engineers (ENIS)
Sfax, Tunisia

Mohamed Masmoudi
METS Micro Electro Thermal Systems
University of Sfax, National School of Engineers (ENIS)
Sfax, Tunisia

***Abstract*—The visual data mining based decision support systems had already been recognized in literature. It allows users analysing large information spaces to support complex decision-making. Prior research provides frameworks focused on simply representing extracted patterns. In this paper, we present a new model for visually generating knowledge from these patterns and communicating it for intelligent decision-making. To prove the practicality of the proposed model, it was applied in the medical field to fight against nosocomial infections in the intensive care units.**

Keywords—Knowledge; patterns; visualization; data mining; Decision Support Systems

I. INTRODUCTION

Decision Support Systems (DSS) are interactive information Systems intended to assist decision-makers to use data, models and knowledge to solve structured or unstructured problems [7].Currently, decision-making is becoming more complex and dynamic. To cope with this increasing complexity, datamining becomes an interesting element for the enhancement of decision support quality [14]. It extends the decision support possibilities by analysing the raw data to extract new actionable insights, interesting patterns and hidden relationships in data.

Data mining for decision-makingis increasingly applied in many fields especially in those based on the treatment of large data quantities in complex and dynamic environments. This solution achieved positive results, but requires instead the integration of other tools to better attract the attention of users and then facilitate the decision-making. Recently, visualization techniques taking place in such systems provide interactive visual tools for more effective decision-making [15].

Information visualization can be applied to visualize raw data, data mining algorithm or extracted patterns[21]. Representing patterns in a visual form is considered as insufficient to transfer knowledge for decision-making [10]. In this context, we address the knowledge visualization from automatic extracted patterns. The importance of knowledge visualization, as a strong sub-discipline of knowledge management, had already been recognized in literature [4][5]. It examines the use of visual representations to improve the creation and transfer of knowledge between two or more users. It refers to all graphic means that can be used to build, communicate complex ideas, create, transform and communicate knowledge.

In this context, we propose a knowledge visualization model allowing users (i.e. decision makers) to identify, preview and interpret the knowledge behind the data mining patterns representation and integrate it into decision-making process. To validate this proposal, our proposed model was applied to the fight against nosocomial infections in the hospital intensive care units.

This paper is organized into 5 sections. In Section 2, a little state of art about our theoretical background concerning the data mining based DSS and the knowledge visualization is addressed. Our knowledge visualization proposal is described in section 3. Section 4, is dedicated to the discussion of the application of our proposal to develop a knowledge visualization tool. Finally, we present the evaluation of the model then we present our conclusions and future outlook.

II. THEORETICAL CONTEXT

A. Visual data mining for decision-making

Decision support systems (DSS) are interactive computer systems that are designed to help decision makers using data and models to identify problem, solve it and make appropriate decisions. Their mission is to improve effectiveness, rather than the efficiency of decisions [7]. The amounts of information available today are becoming increasingly large and complex [1] [2]. In this context, the mining of these data to extract useful information has emerged in order to improve decision-making process: *data mining based DSS*.

Several research works proposed the integration of data mining technology into decision support systems[13]. They confirm that the combination of data mining tools and decision support systems improved decision-making quality. In fact, this integration can provide solutions for new problems that are not addressed before. It can help to create new approaches to

problem solving, by allowing the merging of expert knowledge and automatically discovered knowledge [13].

Data mining is the central phase of the knowledge discovery process visible in figure 1. It is an iterative process that takes place after a series of operations. It begins by data selection according to the analysis domain. Data pre-treatment steps (data cleaning and transformation) occur before the data mining itself. Pre-treatment is about access to selected data in order to build specific data corpus. To this corpus, it is question of applying a data mining algorithm to extract interesting patterns. These patterns will be evaluated and interpreted to verify their quality. Validated patterns will be integrated as knowledge for the decision-making.

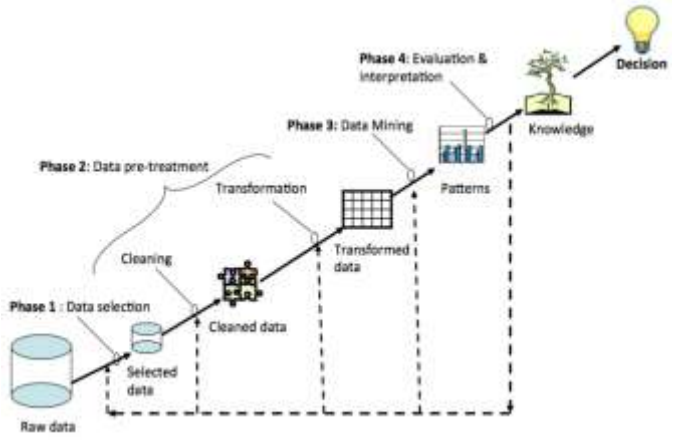


Fig. 1. knowledeg discovery process(Ltifi et al. 2013)

Visualization in data mining made important advances in literature, where several studies proved successes in assisting decision-makers exploring large complex data sets and making decisions: visual data mining based DSS[21]. The power of visual data mining comes from coupling: (1) the computational and data storage capabilities on the machine, and (2) the perceptive skills and the cognitive reasoning of the human. Visual data mining process span from human analytic tasks using domain knowledge to automatic tasks using data mining algorithms. Such visual analytic follows a specific process [13]. Figure2shows the different stages in this process.

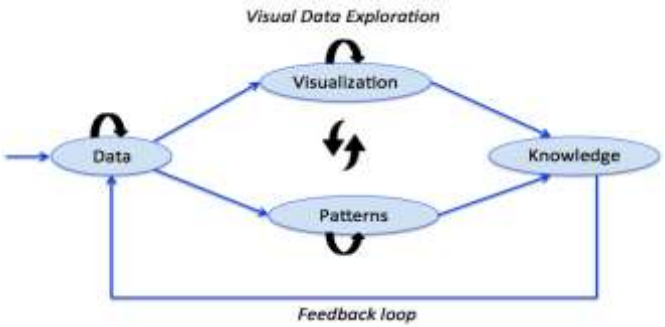


Fig. 2. Knowledege visualization framework[13]

The visual analytics process gives the authority to the user to guide the analysis methods during the execution of his/her tasks from raw data to decision. It provides collaboration between automatic patterns extraction and visualization (of data and patterns) across different abstraction levels. The feedback in this process allows the evaluation and rapid improvement of the visual patterns and eventually the improvement of the knowledge and generated decisions.

While this process allows structured analysis to find new patterns and gain insights into the decision problem domain, it does not support knowledge building and communication. To cope with this need, knowledge visualization would be valuable.

B. Knowledge Visualization

Knowledge Visualization can be considered as a new field of research[4][5]. The goal of this field is the creation and the transfer of knowledge by visualizations [18].Some definitions of knowledge visualization exist in literature. For Tergan (2006), “Knowledge visualization is a field of study, that investigates the power of visual formats to represent knowledge. It aims at supporting cognitive processes in generating, representing, structuring retrieving sharing and using knowledge” [4] (p.168).According to Burkhard "Knowledge Visualization ”[...] examines the use of visual representations to improve the transfer and creation of knowledge between at least two persons.” [4][5].

Burkhard[5] proposed a framework based on four perspectives to guide the knowledge visualization (cf. Fig. 3): the function type, the knowledge type, the recipient, and the visualization type.

- **The function type perspective**: the objective is to specify the aim that should be achieved. It can be a coordination, attention, recall, motivation, elaboration or new insight.
- **The knowledge type perspective**: defines the useful type of knowledge, which should be transferred. The knowledge can be declarative (to Know-What the facts are pertinent), procedural (to Know-How the things are made), experimental (to Know-Why the things happen), orientation focused (to Know- Where the information can be found) and finally individual (to Know-Who are the experts).
- **The recipient type perspective**: concerns the target group that can be individuals, groups, organizations or networks.
- **The visualization type perspective**: concerns the types of visualization. Burkhard[4] defines the seven visualization types relatively to the common visualization categories of architects (Sketch, diagram, image, map, object, interactive visualization and story).

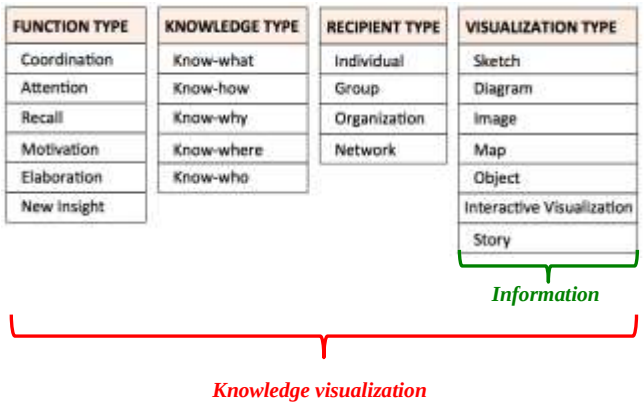


Fig. 3. Knowledege visualization framework[4]

Knowledge Visualization combines findings from various disciplines, particularly information visualisation. Table 1distinguishes between the two fields (information and knowledge visualization) according to different perspectives (Table 1).

TABLE I. KNOWLEDGE VISUALIZATION VS INFORMATION VISUALIZATION

Perspective	*Knowledge Visualization*	*Information visualization*
Objectives	Uses visual representations to improve the transfer and the creation of knowledge	Uses computer applications to get new insights
Content	Knowledge types like experiences, insights, social structures	Explicit data like facts and numbers
Recipients	Individuals or groups	Individuals
Contribution	Solution-oriented: apply new and traditional visualization problems to solve predominant problems	Innovation-oriented: create technical methods
Phases	1. Cognition 2. Perception 3. Communication	1. Information architecture 2. Design 3. Interaction

This paper aims to establish knowledge visualization modelling for visual data-mining based DSS. The model is based on the data mining reasoning, the Burkhard framework and the information visualization pipeline.

III. VISUAL KNOWLEDGE GENERATION FOR DECISION-MAKING

To propose a knowledge generation model, we have to take into account two things: (1) the knowledge visualization framework of Burkhard [4] does not take into account the type of extracted data mining patterns, and (2) the passage from these patterns to knowledge is inconspicuous. As visualization demonstrated successes in helping domain experts in visual analytics, we propose to apply to visually generate knowledge from patterns.

A. Knowledge Visualization Framework

In the context of visual data mining based DSS, the knowledge visualization framework must take into accountthe extracted patterns types. In fact, visualizing decision tree is different to visualizing clusters. An adapted knowledge visualization framework is presented by the figure 4.

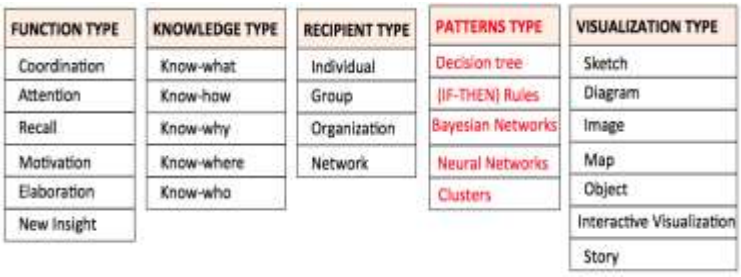


Fig. 4. Adapted knowledge visualization framework

The added patterns type perspective concerns the type of the extracted models by the data mining algorithms. We summarized them into five kind of patterns:

*1) Decision tree:*as its name suggests, it a technique for decision support that divides a population of individuals into homogeneous groups according discriminating attributes based on a fixed and known objective. It allows to issue predictions based on known data on the problem by reducing, level by level, domain solutions[3]. It is a method which has the advantage of being readable for analysts and to determine the discriminating couples from a very large number of attributes and values.

*2) (IF-THEN) rules:*set of rules where each one implies certain relationships of association between a set of objects in a database. Generally, the rules are propositions of the form "if premise then conclusion," noted premise → conclusion. They have the advantage of representing explicit knowledge (unlike connectionist models, for example), and are also the predominant model for many artificial intelligence applications [12].

3) Bayesian Networks: it is a causal graph oriented and acyclic to represent random variables with their dependencies. It shows the distribution of conditional probabilities of a set of variables[9][19]. Its nodes represent random variables and its arcs represent dependencies between these variables. Because of their ability to represent uncertain knowledge, Bayesian networks play an increasingly important in many medical applications[6] [16].

4) Neural Networks: it is a computational model in which the schematic operation was inspired from the functioning of biological neurons. During the learning phase, network learns by adjusting the weights to be able to predict the correct class label input tuples. Neural networks are very powerful to draw inaccurate data and can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques[3].

5) Clusters: it is a statistical method of data analysis that aims to bring together a set of data into different homogeneous groups. Each subset groups elements with common characteristics that match the criteria of proximity. The goal of clustering algorithms is to get subsets most distinct possible.

Distance measurement is a key element for the quality of the clustering algorithm[3].

After defining our adapated framework for visually generating knowledge. Following, we present the knowledge visualization model.

B. Knowledge Visualization Model

1) Cognition

The cognition phase of the model begins when the decision maker must take decisions. Once the code component of the knowledge exploration generates the concerned patterns, a request is sent to the patterns visualization element to visualize this patterns. The decision maker can chooses to browse, manage and structure the relevant models from which associated knowledge will also be visualized interactively. These measures aim to reduce the cognitive load of the decision makers in their challenges and situations cognition.

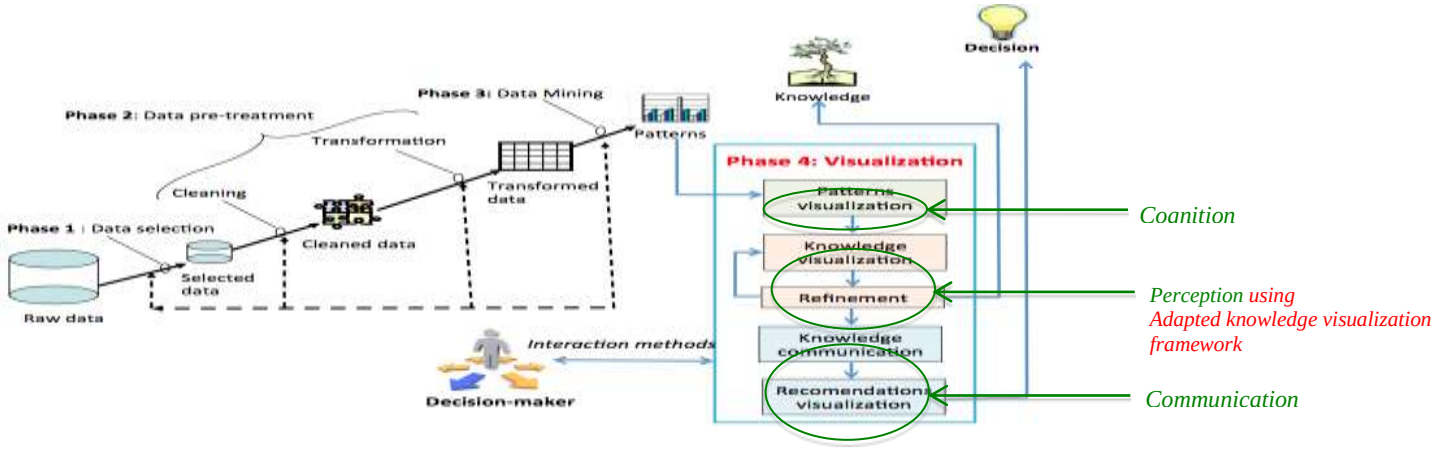


Fig. 5. adapted knowledge visualization framework

2) Perception

The interactive visual analysis helps to generate visual representations intended to be perceived and interpreted by the visual system of the user, which itself controls the display to compare the performances produced with its interpretation of the decision phenomen. So, the decision-maker can visualize and filter the knowledge displayed as needed, refine the visual design, if necessary, according to the perception of executives and knowledge of building a flexible manner.

a) Knowledge Visualization

The goal of this step is the translation of the patterns in a simple, natural and useful representation and description to the decision maker [15]. In this step we will rely on the principle of Shneiderman [20]to view the displayed information: (1) the representation of an overview of patterns, (2) extraction of points of interest and filtering irrelevant points. The display is interactive to allow manipulation of visual elements to easily identify areas where knowledge exists.

b) Refinement

Once the previous step is completed, the decision maker can perform refinement tasks to improve results of the knowledge visualization. During the interactive refinement step, decision maker can benefit from the capabilities of knowledge processing to choose the terms that accurately represent his/her need for knowledge.

3) Communication

We base this phase on the tasks introduced by Shneiderman (Shneiderman 1996) to communicate the knowledge and the decision recommendations. The principle is to provide details and send recommendations at any time. The objective of our visualization process is to:

a) provide decision makers with knowledge about data mining models; this knowledge must be integrated to generate any concrete recommendations

b) interactively communicating knowledge.

In the following section, we present the application of our model in the medical field.

IV. Medical Application

The goal of the work is to apply our proposal to design and develop a visual data mining based DSS in the medical field. The application must contribute to a better analysis, interpretation and knowledge generation from data mining patterns for medical decision-making. We aim aso to analyze the ability of a visual representation to produce changes in the decision making activity. The DSS to developaims to the fight against Nosocomial Infections (NI) in the hospital Intensive Care Unit(ICU).The purpose is to solve the problem of decision on the occurrence of NI during hospitalization of a patient that can weaken or delay his/her treatment. By preventing the occurrence of aNI every day during the patient's hospitalization period in the ICU. The objectif of the work concerns the visual generation of relevant knowledge extracted by a specific data mining technique, which is the association rules mining, for good analysis and better understanding of the patient's condition and to acquire useful knowledge for decision support.

A. Association rules mining

For analysis and extraction of large quantities of valuable knowledge, it becomes increasingly important to develop powerful tools. In our work we chose to work with the association rules as a data mining technique. Support and

confidence are the most known measures for the evaluation of association rule.

1) Dynamic Association rules used in our work

Considering the temporal nature of data to view, we try to improve conventional representations of association rules by adding the time factor for a visualization technique of dynamic association rules. We represent some of dynamic association used in our work:

- If Artificial Ventilation and Trachealintubation then probability of NI in 28 days.
- If Urinary Sonde and Perf.IntraVein then probability of IN in 7 days.
- If Trachealintubation et Urinary Sonde then probability of NI in 6 days.
- If Perf.IntraVein and Artificial Ventilation then probability of NI in 7 days.
- If Trachealintubation and Artificial Ventilation then probability of NI in 6 days.

It is true that the textual representation is easily comprehensible, the cognitive effort exerted to interpret a significant number of rules or patterns extracted remains high. And since a picture is more significant than a thousand words, hence we propose to explore the knowledge generated by the association rules in an interactive visual space. We will present in the next section, the application of our proposed visualization model to visualize knowledge associated with generated patterns.

2) Knowledge generation model application

Table 2 presents the proposed model application for the Dynamic Association Rules extracted patterns.

TABLE II. THE MODEL APPLICATION

	Phase	***Brief description***
Cognition		To view the generated patterns (acts, conclusion, time and metric) we present three main dimensions: Items, Time, Support and trust. Taking an action at a time t is expressed by the intersection of the X axis and the Y axis (2D array). Each cell of the matrix represents a rule, we use the space used by this case to represent. Metrics (Support and Trust) are shown in the bottom of the matrix, and their sizes are proportional to their values. We chose to combine the matrix visualization and 3D histogram to visualize patterns. We are interested in improving the visualization of association rules to help increase system performance while reducing cognitive load exerted by the user. This visualization technique aims to: a. Assist the decision maker in its reasoning and amplify cognition, b. Produce graphs reflecting changes over time association rules using graphical objects.
Perception		The visual representation used "Histogram interactive 3D" is an abstract graphic to explore and interpret the relationships and trends between different patterns generated by association rules over time. The applied knowledge visualization framework is presented by the figure 6. The decision maker can interact with the visualization component with several ways: c. Collect knowledge in relation to the rules and previous knowledge stored in the knowledge base, d. Visually interpret the differences or similarities between the rules over time.
Communication		By selecting a specific date in the course of time, the rules are displayed as interactive rectangles with different colors and the decision maker can interact with them (cf. figure 7). It can perform: e. Interactive filtering on the rules and relationships in each viewing area to facilitate research, and improve the display if the number of actions is important, f. Zooms to switch from an overview (containing all items) to a more detailed view or more specific (selected items). g. Dynamic filtering for displaying different views of the histogram or by rotation, zoom, or stretching.

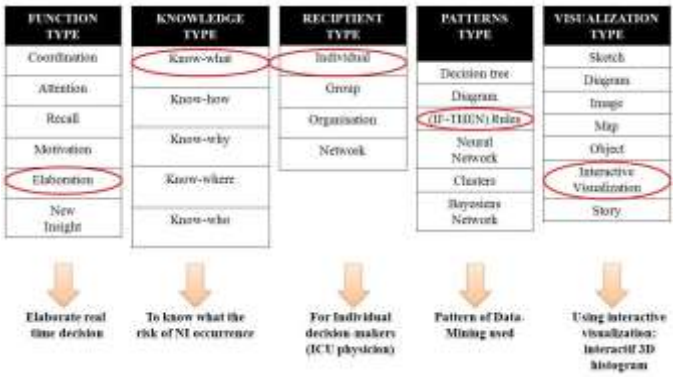


Fig. 6. Knowledge visualization framework

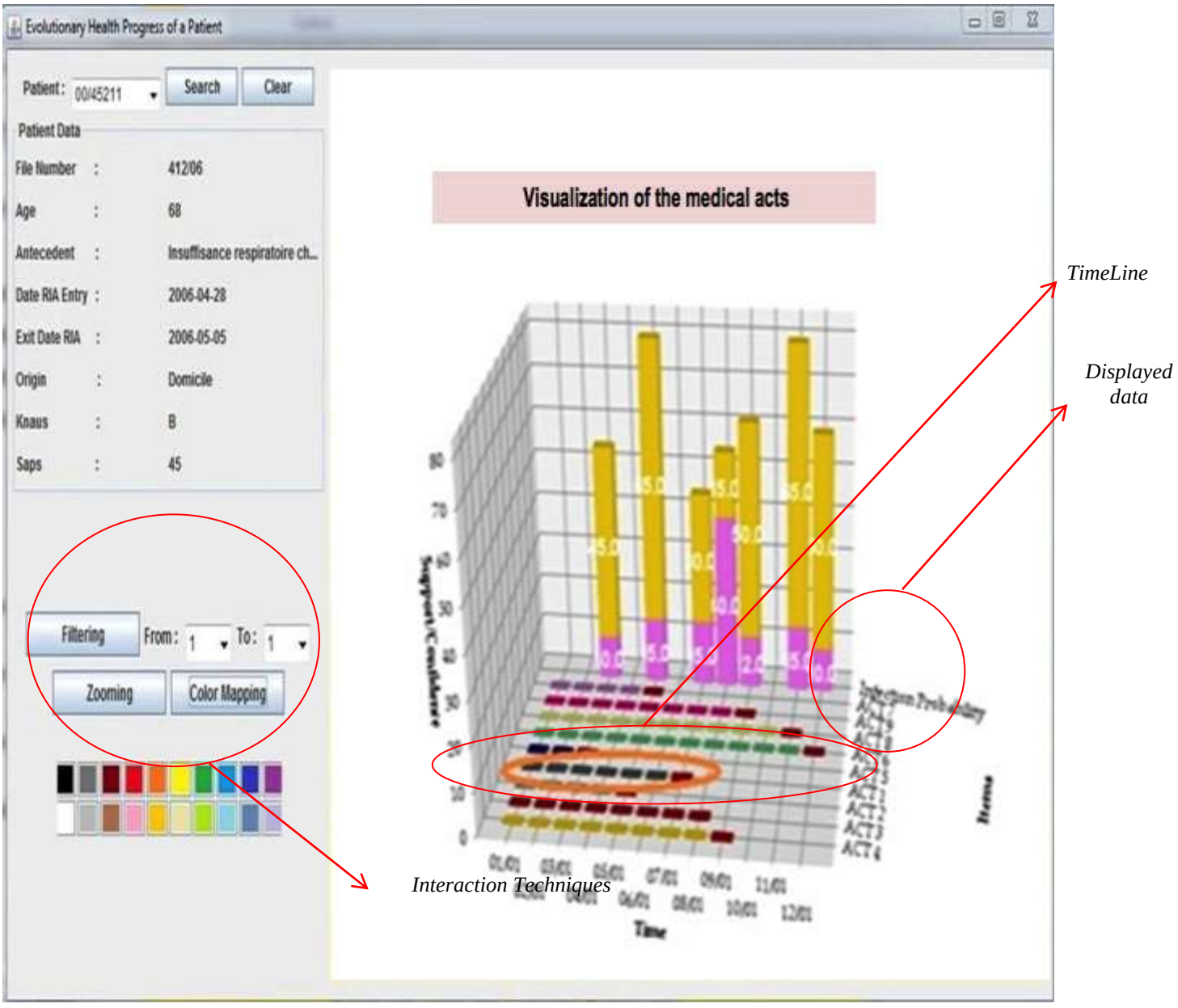


Fig. 7. Interactive visual representation

The Figure 2 presents the interactive visual representation allowing representing probabilistic patterns extracted by the RDA to visually generate knowledge.

V. Evaluation and Discussion

The evaluation of a visualization system consists of the analysis of its behaviour and its use of resources at its disposal. Knowing these resources, it is possible to verify the behaviour application that must be reliable and meet user needs. We are interested in our work to both conventional evaluation dimensions in the Human-Computer Interaction field that are "usability" and "utility".

A. Usability evaluation

the most widely used method for evaluating a display interface is conducting a user study. The principle of this technique is to provide a questionnaire for users to assess and rate the display interface through a set of questions. The list of proposed assessment criteria is divided into three categories that are related to: the user ("Who is it?"), his/her task ("What does he/she want? What for?"), and finally the system (including the characteristics of all the results obtained). This assessment questionnaire is presented in Table 3: the responses of representative users are defined by: Excellent (1) Good (2) Acceptable (3) and Bad (4).

TABLE III. Usability Questionnaire

N	Questions	E	G	A	B
	User				
Q 1	The terms employed are familiar to the users and it concerning the task?		**X**		
Q 2	Is it easy to manipulate the desired environment?		**X**		
	Tasks				
Q 3	There are any messages that inform about the success of the tasks performed?			**X**	
Q 4	The running time is it short?		**X**		
Q 5	Is it easy to find the resources to select the information?	**X**			
	System				
Q 6	The system response time is it short?	**X**			
Q 7	Does the use of interfaces is easy, clear and guided?	**X**			
Q 8	If necessary a previous knowledge, the time taken to back up system is it short?			**X**	
Q 9	The interface is customizable and / or adaptive?	**X**			
Q 10	If mishandled, is it possible to go back? There have some error messages?	**X**			
Q 11	Is it easy to quickly orient you to the documents related to the topic of research			**X**	
Q 12	The quality and the performance of the interface are satisfactory.		**X**		

From these results, we generated the histogram below:

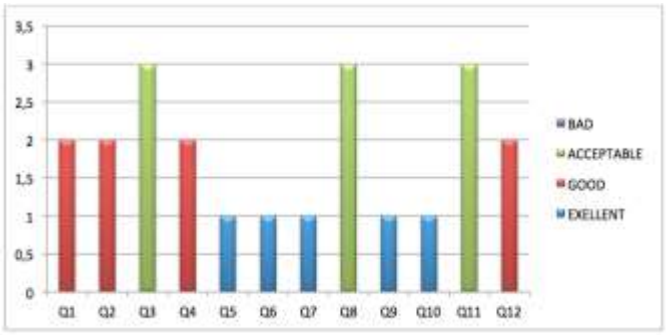


Fig. 8. Usability evaluation results

B. Utility evaluation

The tool application is considered effective if it can visually generate good prediction results. For the evaluation of the performance of our tool, we used a test database that contains 58cases (patients). We got the results given by the following confusion matrix.

TABLE IV. The Prediction Results Provided by the DAR

Observed values / *Predicted values*	**Yes**	**No**	**Total**
Yes	15	8	23
No	5	30	38
Total	20	38	58

From the prediction results obtained by the DAR structure, we found that the classification rate was correct to 0.77, which is interesting. The evaluation results of our system are encouraging. We noticed that users are generally satisfied with the proposed application.

VI. Conclusion

Data mining technology plays an important role in uncovering hidden and interesting patterns. Actually the discovering of models and relationships in data extends the possibilities to support decision-making. The objective of our work is to investigate how decision makers obtain knowledge from the extracted patterns representation. The aim was to design an efficient visual tool to articulate the knowledge produced by these patterns for providing actionable recommendations to make best decisions.

The introduced knowledge visualization model occurs on three main phases: cognition, perception and communication. Each phase contains some successive tasks to improve the visualization results and to help user to better make appropriate and good decisions.

We have applied our model for the daily fight against NI in the ICU. We have used the DAR classification results to calculate NI occurrence probability. These patterns are displayed using interactive 3D histogram representation to be visually and interactively interpreted for obtaining related knowledge. The utility and usability evaluation of this visualization prototype provided good results and reflected the feasibility of our proposed model.

In future work, we plan to apply the model to design other visualization techniques to represent patterns and build knowledge, and to use multi-Agent architecture to extract and visualize knowledge.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of this research by grants from the ARUB program under the jurisdiction of the General Direction of Scientific Research (DGRST) (Tunisia). Thanks are also due to all the ICU staff of HabibBourguiba Teaching Hospital for their interest in the project and all the time they spent helping us design, use and evaluate our system.

REFERENCES

[1] Bohn R.and Short J., How much information? 2009: Report on American consumers, 2010.

[2] Bohn R., Short J., and Baru C., How much information? 2010: Report on enterprise server information, January 2011.

[3] Boulila W. Extraction de connaissances spatio-temporelles incertaines pour la prediction de changements en imagerie satellitale. Thesis. Co-University of Manouba,2012.

[4] Burkhard R.A., Learning from architects: the difference between knowledge visualization and information visualization. In IV '04: Proceedings of the Information Visualisation, Eighth International Conference, pages 519–524, Washington, DC, USA, 2004. IEEE Computer Society.

[5] Burkhard, R.A., Towards a framework and a model for knowledge visualization: synergies between information and knowledge visualization, in Knowledge and Information Visualization, Berlin/Heidelberg: Springer, vol. 3426, pp. 238–255, 2005.

[6] Burnside E.S., Rubin D.L., Fine J.P., Shachter R.D., Sisney G.A., Leung, W.K. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. Radiology, vol. 240(3), pp. 666-673, 2006.

[7] Chen K.C., Decision Support System for Tourism Development: System Dynamic Approach, Journal of Computer Information Systems, 45, 1, pp. 104-112, 2004.

[8] Chittaro L., Information visualization and its application to medicine, Artificial Intelligence in Medicine, vol. 22, no. 2, pp. 81–88, 2001.

[9] Darwich A., A Differential Approach to Inference in Bayesian Networks, 2001.

[10] Elouni J., Ltifi H., Ben Ayed M., Knowledge visualization model for intelligent dynamic decision-making, Volume 420 of the series Advances in Intelligent Systems and Computing, 2015.Springer, pp. 223-235.

[11] HilbertM., LópezP., The world's technological capacity to store, communicate, and compute information, Science, Vol. 332 (6025), pp. 60-65, 2011.

[12] Julien B, Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association,Thesis, University of Nantes l'École Polytechnique, France, 2005.

[13] D. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer, et al. « Visual Analytics: Definition, Process and Challenges ». A. Kerren and J.T. Stasko and J.-D. Fekete and C. North. *Information Visualization - Human-Centered Issues and Perspectives*, Springer, pp.154-175, 2008, LNCS.

[14] Khademolqorani S., hamadani AZ, 2013, An adjusted decision support system through data minig and multiple criteria decision making, SocBehavSci, 73: 388-395.

[15] Ltifi H., Ben Mohamed E. and Ben Ayed M., 2016, Interactive visual KDD based temporal Decision Support System, Information Visualization, 15 (1): 31-50.

[16] Lucas P.J.F., de Bruijn N.C., Schurink K., Hoepelman I.M. *A Probabilistic and decision-theoretic approach to the management of infectious disease at the ICU*.Artificial Intelligence in Medicine, vol. 19(3), pp. 251–279, 2000.

[17] MladenićD., LavračN., BohanecM., MoyleS., *Data Mining and Decision Support: Integration and Collaboratio*, Dordrecht, Kluwer Academic Publishers,2002

[18] Meyer R., Knowledge Visualization. the Media Informatics Advanced Seminar on Information Visualization, 2009

[19] Murphy K.P., Dynamic Bayesian Networks: Representation, Inference and Learning, University of California, Berkeley Fall 2002.

[20] Shneiderman, B., The eyes have it: a task by data type taxonomy for information visualizations, in Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, CO, USA: IEEE, Los Alamitos, CA, United States, pp. 336–343, (1996).

[21] Simoff S.J., Böhlen M.H. and Mazeika A., 2008, Visual Data Mining, Theory, Techniques and Tools for Visual Analytics, Lecture Notes in Computer Science 4404, Springer Berlin Heidelberg.

[22] TerganS.-O., KellerT., and BurkardR.A., Integrating knowledge and information: digital concept maps as a bridging technology. Information Visualization, 5(3), pp.167–174, 2006.

A Novel Design of Miniaturaized Patch Antenna Using Different Substrates for S-Band and C-Band Applications

Saad Hassan Kiani
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

Sharyar Shafeeq
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

Khalid Mahmood
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

Mehre Munir
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

Khalil Muhammad Khan
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

***Abstract*—In advance communication technology, patch antennas are widely exploit due to their inexpensive and light weighted structure. This paper presents a novel design of miniaturized multiband patch antenna using different substrates frequently used in patch antennas. Various substrates such as Teflon, Roger 5880, Bakelite and Air are in use to achieve better gain and directivity. The proposed miniaturized multiband patch antenna contains 2 substrates where one substrate is FR4 (fixed and lossy) and the other substrates are changed to observe gain, directivity and return loss. Coaxial probe serving mode is presented in this paper. This serving mode is a contacting arrangement for patch, in which the outer conductor is linked to ground plane and the inner conductor of the coaxial connector spreads through dielectric and is bonded to the radiating patch. The proposed antenna can be used for various S-band and C-band applications.**

Keywords—substrates; microstrip; return loss; directivity; miniaturized; Impdence bandwidth

I. INTRODUCTION

Micro strip antennas are widely used in communication devices for various applications such as radars, satellites and mobile phones etc. They are low cost, small structured and easily fabricated. Considerable amount of approaches have been develop for reducing the size of antenna but the central concern with the reduced area of an antenna is its minor gain. Some methods here are discussed below.

With the use of synthetic magnetic conductor area of antenna was reduced but with the result of lowered gain [1]. As gain was significantly improved as compared to magnetic conductor using split ring resonators but the reverse outcome was that reduce size was approximately equal to 10% [2]. The size reduction was achieved above 20% by way of Koch Fractal shape but with few repetitions gain starts diminishing [3].

In this paper we have analyzed a fractal pi shaped dual substrate patch antenna where one substrate is FR4 (fixed and lossy) and second is a variable one as a double substrate antenna shows good return loss [4]. Stage of substrate material is an considerable job in designing a patch antenna, as the restrictions regarding micro strip antenna such as negligible gain, poor efficiency and directivity can be mitigated by deciding on proper substrate materials, because performance factors of patch antenna like bandwidth, gain ,radiation pattern are linked up to permittivity of substrate material .[5][6].

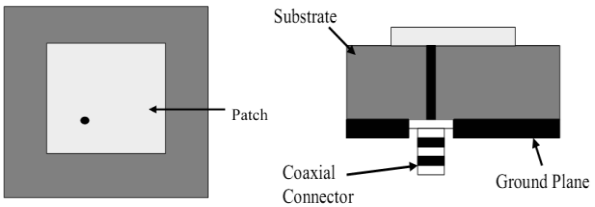
To the best of our knowledge, no literature review on miniaturized double substrate selection is available so we are proposing a novel design of miniaturized patch antenna substrate selection for S-Band and C-Band applications.

This paper is organized as follows:

Section I deals with introductions, section II deals with antenna design and methodology whereas section III deals with results and discussion. Concluding remarks are given in section IV.

II. ANTENNA DESIGN

The designing of antenna is done stage by stage in a convenient way. The basic shape is shown in figure 1(a and b) [5-6].



(a). (b)

Fig. 1. (a) Top view of Micro strip patch antenna (b) Side view of patch antenna (Coaxial Cable Feed)

A. Substrate Selectivity

Parameters of patch antenna like bandwidth, radiation efficiency are linked up with substrate material used for patch so setting up substrate material is prime assignment in designing patch. Thickness of substrate significantly impacts upon parameters.

According to Coulomb's la w:

$$E = \frac{q1\ q2}{4\Pi\ \varepsilon r\ r^2} \quad (1)$$

Where q1 and q2 are the charge bodies and $\boldsymbol{\varepsilon r}$ is the permittivity of free space and r is the distance.

Electric field (E) is inversely proportional to the relative permittivity εr by increasing or decreasing relative permittivity of the various substrates change in electric field can be observed [4].The substrates used in this paper have the following dielectric constants.

TABLE I. DIFFERENT TYPES OF SUBSTRATES

Substrate	Dielectric Constant
RO5880	2.2
Air	1.00
Bakelite	4.78
Teflon	2.1

B. Width of the Patch

With the help of following equation, width of the patch is calculated. [4].

$$W = \frac{c}{2\ f_0\sqrt{\frac{(\varepsilon r + 1)}{2}}} \quad (2)$$

Where c = $3x10^8$ m/s.

f_0 =Resonant Frequency

εr is the permittivity of a substrate.

C. Length of the Patch

With the help of following equation, length of the patch is calculated [4].

$$L = L(eff) - 2\Delta L \quad (3)$$

Where

$$L(eff) = \frac{c}{2f_0\sqrt{E(reff)}} \quad (4)$$

And

$$\varepsilon_{(reff)} = \frac{\varepsilon r + 1}{2} + \frac{\varepsilon r - 1}{4}\left(1 + \frac{12h}{W}\right)^{-1/2} \quad (5)$$

By calculating the length and width of the patch and ground, in this paper a 4 GHz antenna is designed with coaxial feeding technique. After miniaturizing and reducing its size up to 79.12%.This was achieved by using combination of U-shaped and L-shaped slots on the ground plane and pi-shaped slot on the double fractal patch of antenna with shorting pin between patch and ground [7] As a result antenna produced multiband response with a high gain and sufficient impedance bandwidth for each band [9]. We have considered thickness of the first and second substrate constant (2mm) but we have changed permittivity of second substrate one by one by changing various material like, RogerRT5880, Bakelite, and Teflon and Air.

For these substrates relative permittivity is $\boldsymbol{\varepsilon r} < 5$.

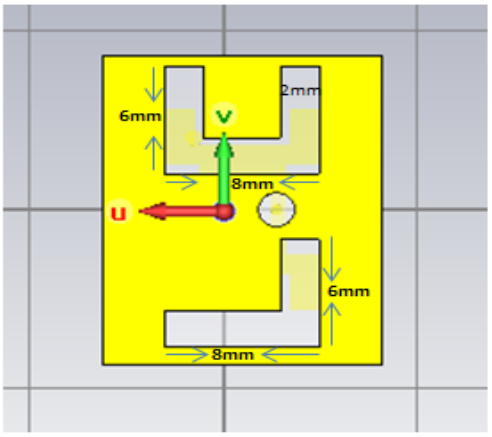


Fig. 2. Back view of Dual Fractal Patched Pi Antenna

TABLE II. DIMENSIONS OF PI SHAPED ANTENNA

Parameters	Values in MM
Length of Patch, LP	16.95
Width of Patch, WG	22.47
Length of Ground, LP	28.95
Width of Ground, WP	34.47
Slot Length, SL	5.00
Slot Width, SW	2.00
Pi Slot Length, PL	4.00
Pi Slot Width, PW	1.00
Height of Patch	0.035
Height of Ground, HG	0.8
Height of Substrate, HS	2.00
Horizontal U and L Slot Length, HUL&HLL	8.00
Horizontal U and L Slot Width, HUW &HLW	2.00
Vertical U and L Slot Length, VUL&VLL	8.00
Vertical U and L Slot Width, VUW &VLL	2.00

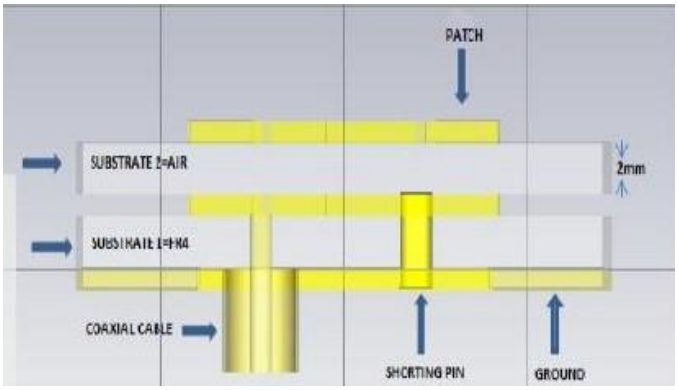


Fig. 3. Front View of Dual Fractal Patched Pi Antenna

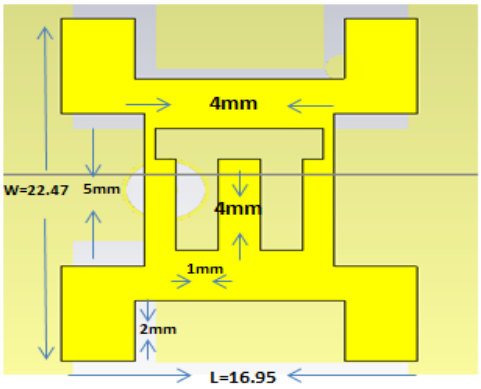


Fig. 4. Bottom View of Dual Fractal Patched Pi Antenna

The ground plane is shown in Fig II or in other terms back view of antenna. Approximately 35mm and 29mm are antenna width and length dimensions. 8mm U-shape and 2mm L-shape holes on ground plane reduced antenna size up to quite satisfactory level. In Fig III we clearly see the overall structure of proposed patch comprising of 2 substrates where lower substrate is FR4 and upper substrate is Air and feeding coaxial probe with shortening pin connecting 2 patches. AIR has a dielectric constant of 1.00 and its thickness is taken to be 2mm.

Current circulation on patch of antenna is shown on Fig IV. As marked from figure, it is clear that on ends of Pi-shape slot, current concentration is supplementary as compared to rest of the patch. Introducing slots increases current distribution path, increasing electrical length and resulting in resonating frequencies shifting downward direction hence revealing several band response. As compared to traditional antenna with familiarized resonance frequency, the overall surface of antenna is reduced to excessive level.

D. Substrate Varying

In this section, we have used second substrate of fractal pi shape antenna and observed the gain and directivity. The First substrate to replace air is RO5880 having permittivity of 2.2.

After R05880 substrate was changed from R05880 to Bakelite having permittivity of 4.78.similarly Bakelite substrate was changed from Bakelite to Teflon having permittivity of 2.1. Various changes were seen in gain and directivity as well as return loss.

III. Results and Disscussions

Satisfactory results were achieved as individual factors like radiation pattern, gain, and return loss were examined for several regularities of interest. The obtained results achieved are discussed below.

A. Air

With Air substrate, antenna showed very good results as shown in table III, by showing multi-level response at different resonant frequencies.

TABLE III. Air Substrates Results

Resonant Frequency	Return Loss	Gain	Directivity
2.603	-31.74dB	3.08dB	4.32dBi
3.119	-26.84dB	4.03dB	5.25dBi
3.5	-15.91dB	2.26	4.2dBi
4.21	-18.945dB	5.47dB	6.71dBi
6.16	-12.89dB	0.145dB	3.72dBi

The minimum return loss occurred at resonant frequency of 3.00 GHz which was -31.74dB having gain of 3.08dB and directivity of 4.32dBi.

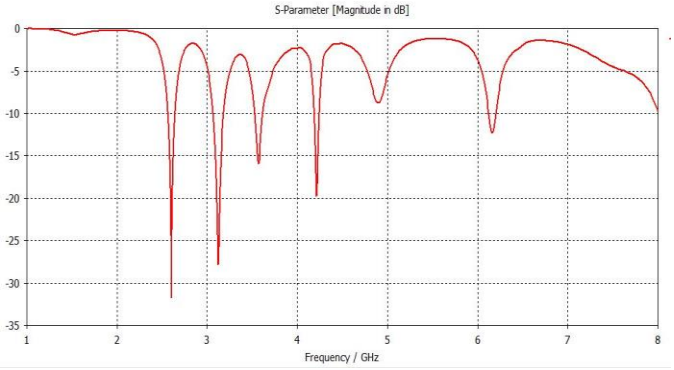


Fig. 5. Return Loss of Air

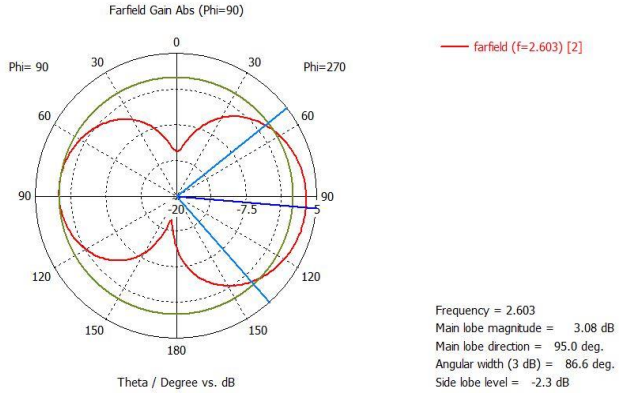


Fig. 6. 1D plot of Gain at Resonant Frequency 0f 2.603 GHz

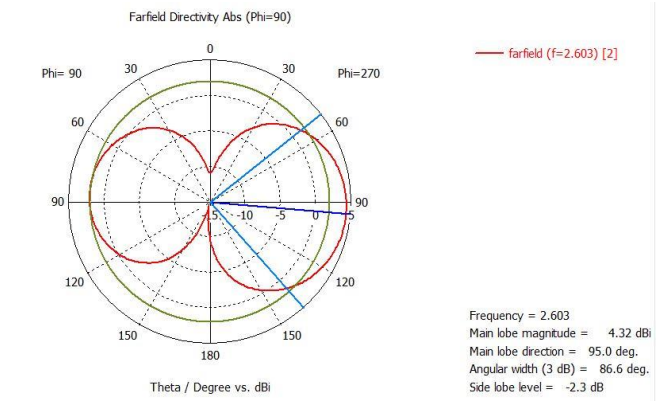


Fig. 7. 1D plot of Directivity at Resonant Frequency 0f 3.00 GHz

In 2.603GHz the level of main lobe is 3.08dB, direction is 95 degrees and angular width is 86.2 degrees with side lobe level of -2.3dB.

B. Teflon

With Teflon antenna showed good and satisfactory results showing multi-level response at different resonant frequencies.

TABLE IV. TEFLON SUBSTRATE RESULTS

Resonant Frequency	Return Loss	Gain	Directivity
2.50	-25.3dB	2.97dB	4.23dBi
3.00	-35.7dB	4.11dB	5.39dBi
3.70	-11.48dB	4.49dB	6.49dBi
5.89	-13.32dB	0.0078dB	3.26dBi

The minimum return loss occurred at resonant frequency of 3.00 GHz which was -35.7dB having gain of 4.11dB and directivity of 5.39dBi.

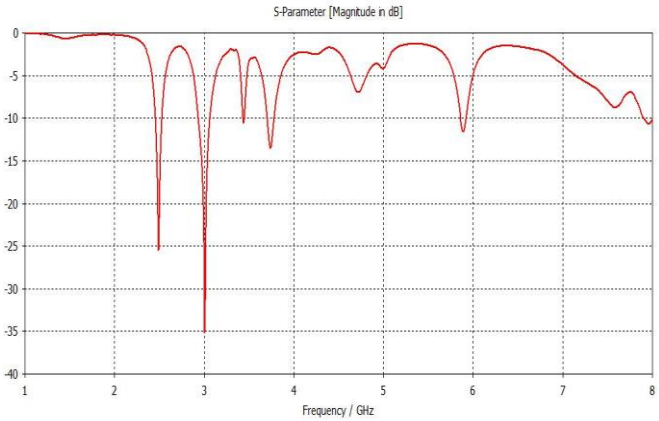


Fig. 8. Return Loss Plot of Teflon

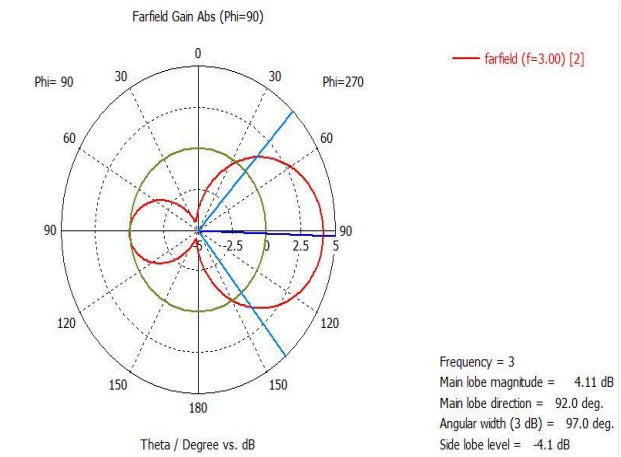


Fig. 9. 1D plot of Gain at Resonant Frequency 0f 3.00 GHz

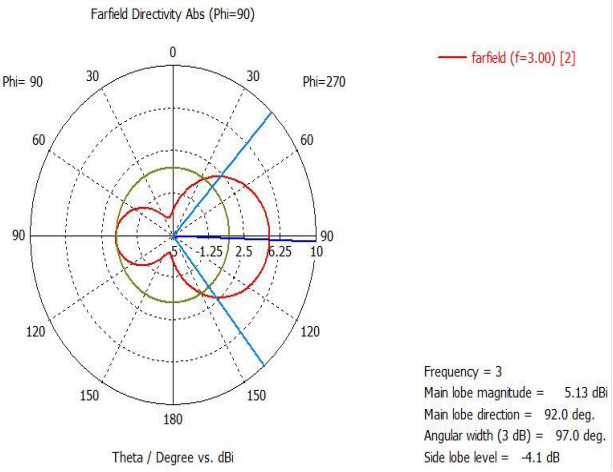


Fig. 10. 1D plot of Directivity at Resonant Frequency of 3.00 GHZ

In 3.00GHz the level of main lobe is 5.13dB, direction is 92 degrees and angular width is 97 degrees with side lobe level of -4.1dB.

C. Bakelite

With Bakelite antenna showed quite good and satisfactory results by showing multi-level response at different resonant frequencies.

TABLE V. BAKELITE SUBSTRATE RESULTS

Resonant Frequency	Return Loss	Gain	Directivity
2.30	-12dB	2.45dB	4.22dBi
2.73	-13dB	3.16dB	4.88dBi
3.45	-12.42dB	4.09dB	5.62dBi
6.58	-16dB	0.736dBi	3.81dBi
7.74	-15dB	5.39dB	6.44dBi

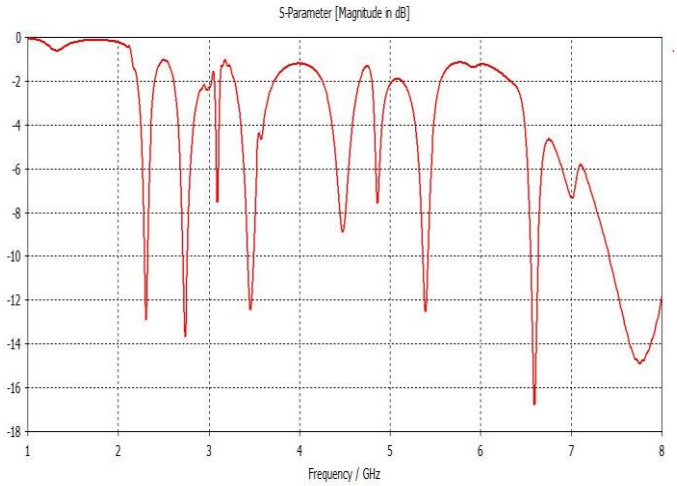


Fig. 11. Return loss of Bakelite

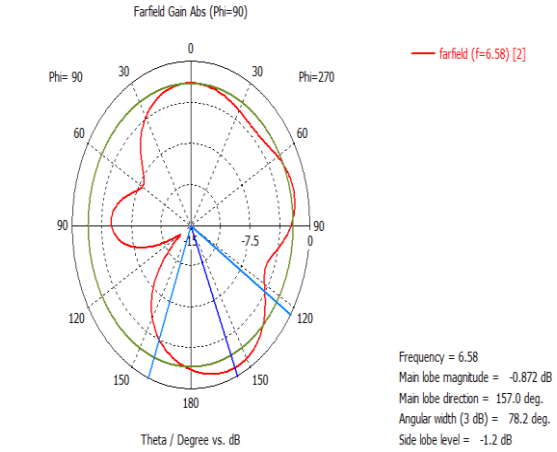


Fig. 12. 1D plot of Gain at Resonant Frequency of 6.58 GHZ

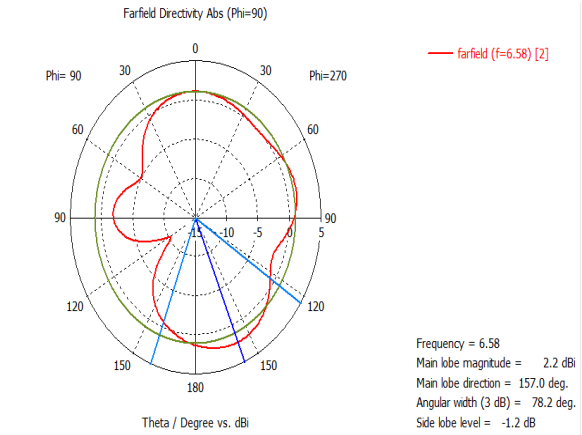


Fig. 13. 1D plot of Directivity at Resonant Frequency of 6.58 GHZ

In 6.58GHz the level of main lobe is 0.872dB, direction 157 deg, angular width is 78.2 deg with Side lobe level of -1.2dB.

D. Roggers5880

With R05880 antenna showed quite good and satisfactory results by showing multi-level response at different resonant frequencies.

TABLE VI. RO5880 SUBSTRATE RESULTS

Resonant Frequency	Return Loss	Gain	Directivity
2.48	-22.7dB	2.95dB	4.21dBi
2.98	-27.7dB	4.05dB	5.11dBi

The minimum Return Loss occurred at resonant frequency of 2.98 GHz which was -27.7dB having gain of 2.95dB and directivity of 4.21dBi.

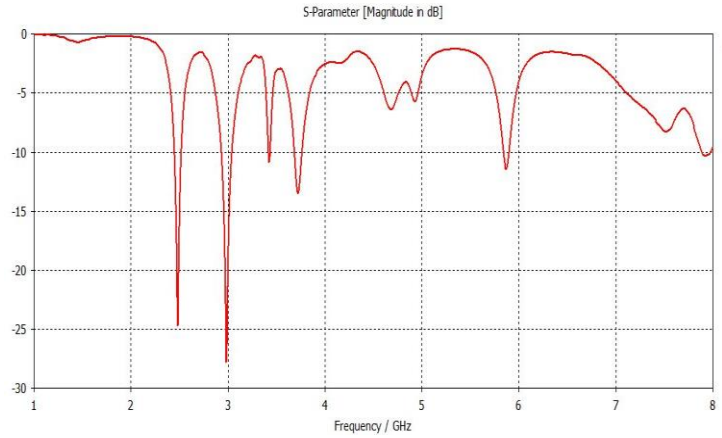


Fig. 14. Return loss of RO5880

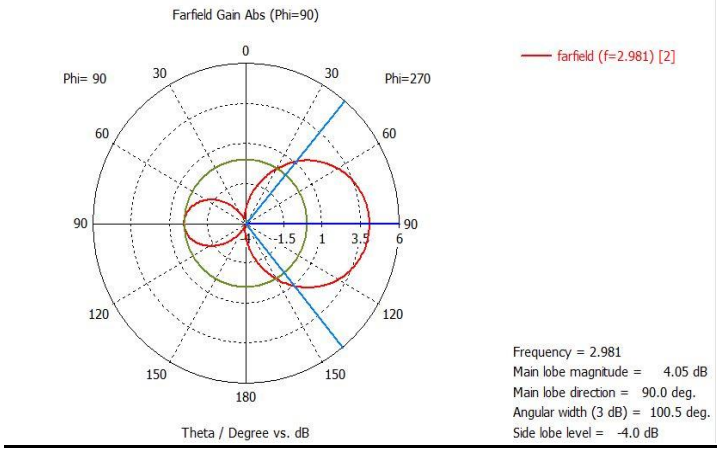


Fig. 15. 1D plot of Gain at Resonant Frequency of 2.98 GHZ

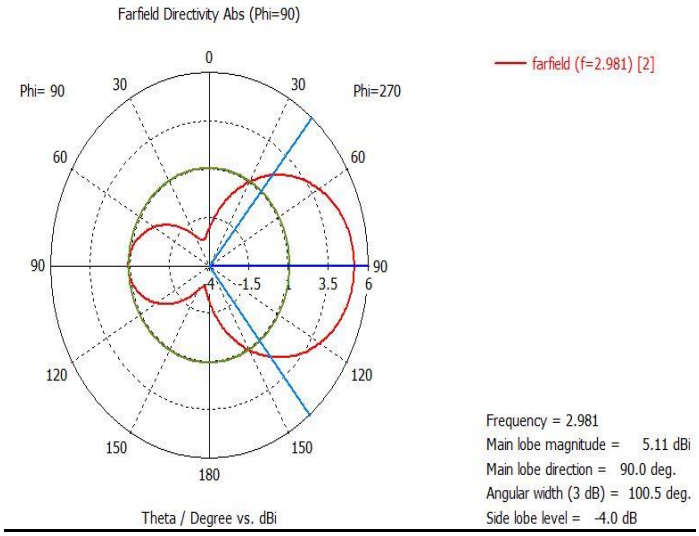


Fig. 16. 1D plot of Directivity at Resonant Frequency of 2.98GHz

In 2.981GHz the level of main lobe is 5.11dB, direction 90 deg with angular width of 100.5 deg and Side lobe level of -4.0dB.

IV. CONCLUSION

This paper presents a novel design of miniaturized multiband patch antenna using different substrates frequently used in patch antennas. The proposed miniaturized multiband patch antenna contains 2 substrates where one substrate is FR4 (fixed and lossy) and the other substrates are changed to observe gain, directivity and return loss. In this paper coaxial probe feeding technique is also used. The proposed antenna can be used for various S-band and C-band applications such as mobile communication, WIMAX, WLAN and Vehicular communication etc.

REFERENCES

[1] Rahamdani and A. Munir, "Microstrip patch antenna minaturaization using artifical magnetic conductor" in Telecommunication Systems, Services and Applications (TSSA), 2011 6th International conference on 2011,pp.219-223

[2] M. M. Bait-Suwailam and H. M. Al-Rizzo, "Size reduction of microstrip patch antennas using slotted Complementary Split-Ring Resonators," in Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on, 2013, pp.528-531,

[3] S. S. Gaikwad, et al., "Size miniaturized fractal antenna for 2.5GHz application," in Electrical, Electronics and Computer Science (SCEECS), 2012 IEEE Students' Conference on, 2012, pp. 1-4.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] C. A. Balanis, Antenna Theory: Analysis and Design, 3rd ed.NewYork,NY,USA:Wiley,2005.

[5] Abbaspour, M. and H. R. Hassani, "Wideband star shaped microstrip p atch antenna," Progress In Electromagnetics Research Letters, Vol. 1, pp.61–68, 2008.

[6] Ansari, J. A. and R. B. Ram, "Broadband stacked U-slot microstrip p atch antenna," Progress In Electromagnetics Research Letters, Vol. 4, pp.17–24, 2008.

[7] M. M. Bait-Suwailam and H. M. Al-Rizzo, "Size reduction of microstrip patch antennas using slotted Complementary Split-Ring Resonators," in Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on, 2013, pp. 528-531.

[8] H. Oraizi and S. Hedayati, "Miniaturization of Microstrip Antennas by the Novel Application of the Giuseppe Peano Fractal Geometries," Antennas and Propagation, IEEE Transactions on, vol. 60, pp. 3559-3567, 2012.

[9] John D Kraus ,Ronald J Marhafka and Ahmad S khan"Antenna and Wave Prop agation" Text Book.

[10] Y. Cheng-Chi, et al., "A compact antenna based on metamaterial for WiMAX," in Microwave Conference, 2008. APMC 2008. Asia-Pacific, 2008, pp. 1-4.

[11] D. Sievenpiper, H. P. Hsu, J. Schaffner, G. Tangonan, R. Garcia, and S. Ontiveros, "Low profile, four sector diversity antenna on high impedance ground plane," Electron. Lett. vol. 36, pp. 1343 1345, 2000.

[12] Zhang, X. Yang, "Study of a slit cut on a microstrip antenna and its applications," Microwave and Optical Technology Letters, vol.18, no.4, pp.297- 300, 1998.

Impact of Elliptical Holes Filled with Ethanol on Confinement Loss and Dispersion in Photonic Crystal Fibers

Khemiri Kheareddine
Sys'com laboratory ENIT
University of Tunis El Manar
Tunisia

Ezzedine Tahar
Sys'com laboratory ENIT
University of Tunis El Manar
Tunisia

Houria Rezig
Sys'com laboratory ENIT
University of Tunis El Manar
Tunisia

***Abstract*—To get a confinement loss value, the weakest possible We have interest to optimize an optical fiber our structure has a cladding which is formed by holes in silica. The geometry of the holes is special because they have an elliptical shape and oriented with some angle. The introduction of ethanol in the holes, the omission of some rings allows us to have values very close to zero for the confinement loss. In this paper, we have designed an ultraflat dispersion PCF. We notice that the zero dispersion can be in the range from 1000 nm to 1650 nm and has the value of 0±0.14ps/nm/km.**

Keywords—confinement loss; dispersion; doped Photonic Crystal Fiber; ethanol-filled holes; elliptical holes; FDTD

I. INTRODUCTION

The request for high bit rate is increasing in recent years. To meet this spectacular rise, optical fiber was used as a transmission medium. These fibers have a wide bandwidth and can carry high bit rates.

Despite this characteristic, these fibers have their limits due to the modal or chromatic dispersion. This makes then research to have not ceased and the researchers arrived in recent years to put on the telecommunications market the most reliable support that's the photonic crystal fiber [1-2] that meets a requirement in adjustable dispersion, unimodal character and high bit rate

PCF fiber presents a clear example of these media.The properties of these fibers such as dispersion who can be adjusted by the parameters of the fiber.

We are interested to fibers formed by a doped silica core which present a difference Δn between the core and the cladding [3-4]. This photonic cladding it is composed of holes filled with ethanol in the silica. These past years the research highlighted in this field and many researchers has touch in this axis [5-6] such as the study of photonics sensors [7-8]. With some arrangement this structure may be used for telecom applications.

Many parameters can influence the optical properties of such fibers as to know the number of rings surrounding the core, the core diameter and that of the holes denoted d, the distance between the centers of two adjacent holes noted (pitch). We can also reduce losses of guide by increasing the number of rings of holes [9]. The guiding of the light within these fibers is made by total internal reflection [10].

In previous work the effects of geometric deformations and arrangements of holes around the doped silica core on the confinement losses was stadied [11-12]. These geometric distortions can be introduced during the manufacturing. Although the fiber manufacturing techniques are very well controlled, geometric distortion effects and the arrangement of the holes around the core of the PCF on the confinement loss still remained to explore and to study.

Unlike conventional photonics fibers which are formed of air holes in silica our design structure is formed by a cladding which consists of ethanol-filled holes with some elliptical holes. Each ellipse possesses a major and a minor radius; we are interested on the effect of of the orientation angle and the omission of some rings on the confinement loss and dispersion and the effect of temperature after the introduction of ethanol into the holes.

An important characteristic of fiber is the dispersion .The Chromatic dispersion determines the transmission capacity of an optical communication system. This is an extension in the time limit the transmission rate as it forces to increase the time between two pulses. We stady the dispersion of our structure in a range of 800 nm to 1850 nm

II. THEORY

Since our design structure is in an XZ plane and since our numerical method is based on a temporal and spatial discretization of Maxwell equations [13-14]. The Y direction is laid as infinite and The propagation is along Z . All these hypotheses allow us to remove all derivatives from Maxwell's equations and divided them into two independent sets of equations which are The Transverse Magnetic (TM) and Transverse Electric (TE)

$$\min imum(\Delta x, \Delta y, \Delta z) \le \frac{\lambda_{\min}}{10 n_{\max}} \quad (1)$$

$$\Delta t \le \frac{1}{\gamma \sqrt{\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta z)^2}}} \quad (2)$$

the maximum value of the refractive index of the calculation area is represented by n_{max}.

A mode is an adaptation of the light to the guide, a guided mode has its own velocity and an effective index. The modes represent a solution of Maxwell's equations. For some modes, light is well confined in the core. The determination of the imaginary part of the complex effective index allows us to calculate the confinement losses [15].

$$CL(dB/m) = \frac{2 \times \pi \times 20 \times \mathrm{Im}(n_{eff})}{\lambda \times \ln(10)} \quad (3)$$

Chromatic dispersion is obtained from the effective index calculated on a spectral band, it is expressed by the following relationship [16]:

$$D_c = -\frac{\lambda}{c}\frac{d^2 n_{eff}}{d\lambda^2} \quad (4)$$

Where λ is the operating wavelength and c is the the velocity of light in vacuum

To calculate chromatic dispersion we can use Taylor expansion:

$$\frac{d^2 n_{eff}}{d\lambda^2}\Big|_{\lambda=\lambda_0} \approx \frac{1}{24(\Delta\lambda)^2}(-2n_{eff}(\lambda_0 + 2\Delta\lambda) + 32n_{eff}(\lambda_0 + \Delta\lambda) - 60n_{eff}(\lambda_0) + 32n_{eff}(\lambda_0 - \Delta\lambda) - 2n_{eff}(\lambda_0 - 2\Delta\lambda)) \quad (5)$$

Δλ is equal to 20 nm

III. Modeling and Analysis

We took a conventional structure formed by a doped silica core ($\Delta n = 2\ 10^{-2}$) and surrounded by five rings of holes in silica index n = 1.45. The holes are filled with ethanol whose index is a function of the location temperature [7].

The introduction of ethanol into the holes instead of air is specially designed for photonic fiber sensor [17]. In our article we will try to see the impact of the temperature of the places where is fibers installed on the confinement loss and dispersion for a telecom application λ=1,55µm

The ethanol index varies according to

$$n = n_0 - \alpha(T - T_0) \quad (6)$$

n_0 refractive index of ethanol at T_0. For T_0 = 20°, n_0 = 1.36048 and $\alpha = 3.94 \times 10^{-4}$

The core of our structure has 3 µm for radius, the pitch Λ= 4.5 µm. Holes filled with ethanol have an index value n = 1.358 (T = 25 °)

The structures is studied using the OPTIFDTD 8 software It is based on Finite Difference Time Domain method (FDTD). From the simulation we can see the profiles of the calculated mode; the result is shown in Figure 1.

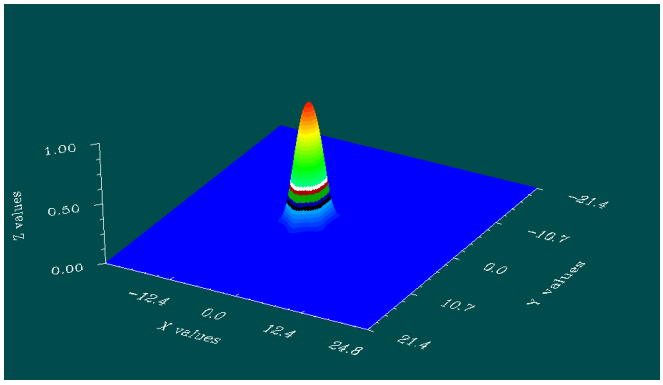


Fig. 1. 3D distribution of the fundamental mode of structure

We stady too structure designed (e) and (f) shown in fig.2.

(e): the core radius 3µm holes index 1.358; first rings of holes with 1.5µm radius; the second ring with q ratio = q_1=0.75 oriented 135 °with omission of the third rings.

(f): the core radius 3µm holes index 1.358; first rings of holes with 1.5µm radius; the second ring with q ratio = q_2= 0.75 oriented 135 ° with omission of the third rings.

with q_1= r /R and q_2= 2r /2R; R=0.6µm and r = 0.45 µm

The elliptical air holes have a major radius (R) and a minor radius (r) like is shown in Figure 3.

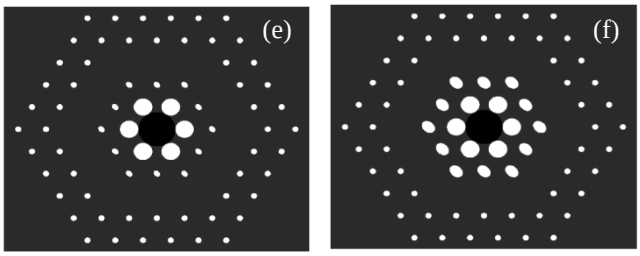


Fig. 2. cross section view of studied structure

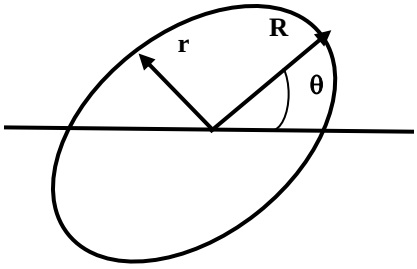


Fig. 3. Major radius, minor radius, and orientation angle

We can also see the behaviour of the effective refractive index in Figure 4.

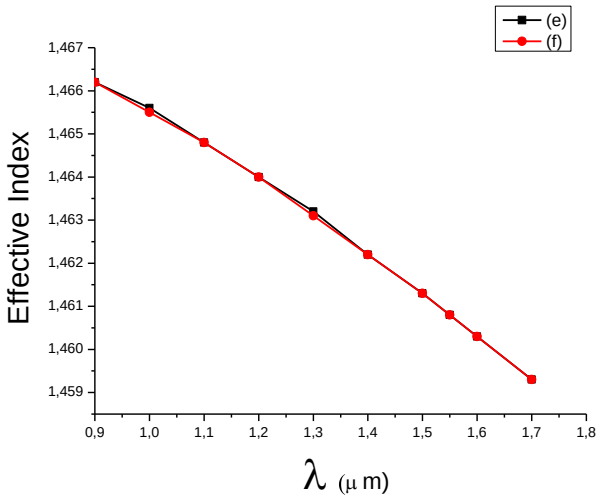


Fig. 4. variation of effective index with the wavelength λ for the studied structures

For the confinement loss we can see clearly in Figure 5 that the structure (f) in very close values in 0 from the value of λ = 1,4μm. It is a very interesting result which proves that such structure can be used for telecom applications in wavelengths λ = 1,55μm. although the number of ring of holes is reduced around the doped core. This represents a great technological interest in the manufacture of such a fiber.

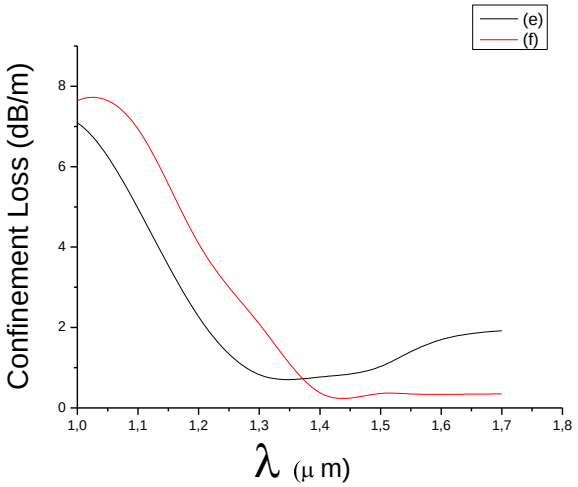


Fig. 5. Confinement loss plot according to wavelength

Generally optical fibers are put in places where there is a medium temperature variation principally made in military and the industrial machining.

We wanted to see the impact of the temperature on the value of the confinement loss of our held structure.

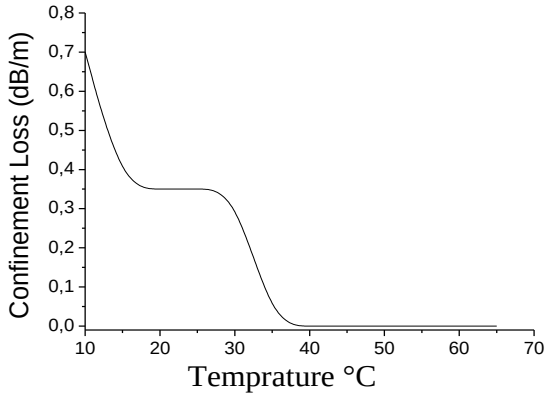


Fig. 6. Confinement loss of structure (f) as a function of temperature

We were able to notice that the value of the confinement loss is zero from 40°, as we shown in Figure 6.Thereafter we were interested in the value of the dispersion.

Realizing near-zero ultra-flat dispersion for a wide band wavelength is a major realisation of the PCFs engineering. For dispersion management Ultra-flat near-zero dispersion profile is helpful but also for achieving novel applications like high gain,broadband parametric amplification.

To realize the ultra-flat near-zero dispersion we fill holes with a liquide in this work we use ethanol

The simulation allowed us to determine the behavior of the dispersion as a function of wavelength, as we shown in Figure 7. We can notice that for the value of 1000 nm to 1650 nm dispersion value is 0±0.14ps / nm / km. This result is very important for transmissions.

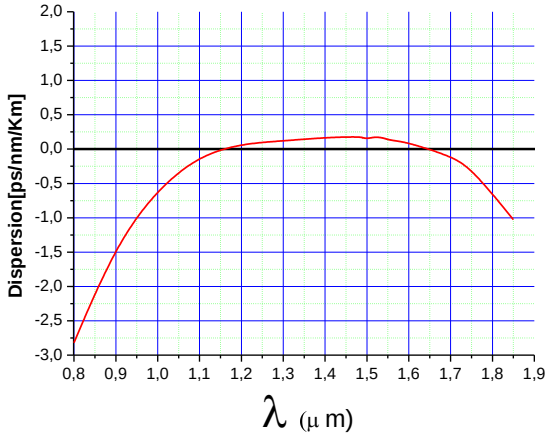


Fig. 7. Chromatic dispersion of structure (f) as a function of wavelength

IV. Conclusion

FDTD is used to simulate and analyse confinement loss and dispersion of a doped Photonic Crystal Fiber. A new structure is introduced, in our study we took a structure which has a cladding which is formed of holes in silica. The geometry of the holes is special because they have the elliptical shape and

oriented with an angle of 135 °. The orientation of the holes of the first ring has a great effect on the value of the confinement loss in a manner that reduces the confinement loss values. The introduction of liquid such as ethanol in the holes, the omission of some rings allows us to have values very close to zero for the confinement loss.

The Chromatic dispersion determines the transmission capacity of an optical communication system.

We notice an ultraflat dispersion that has a value of $D = 0 \pm 0.14$ps/nm/km which can be in the range from 1000 nm to 1650 nm.

REFERENCES

[1] Knight J. C., Birks T. A., Cregan R. F., Russell P. S. J. et de Sandro J.-P. *Electronics Letters*.**1998**,34 , 1347-1348.

[2] J. C. Knight, T. A. Birks, P. St. J. Russell, and D. M. Atkin .*Optics Letters*. **1996**, 21,1547-1549.

[3] B. Nagaraju, M.C. Paul , M. Pal b, A. Pal , R.K. Varshney , B.P. Pal , S.K. Bhadra ,G. Monnom, B. Dussardier . *Optics Communications*. **2009**,282, 2335–2338

[4] R. F. Cregan, J. C. Knight, P. St. J. Russell, and P. J. Roberts. *JOURNAL OF LIGHTWAVE TECHNOLOGY*. **1999**, 17, 2138 - 2141.

[5] Shen, G.-F., X.-M. Zhang, H. Chi, and X.-F. Jin. *Progress In Electromagnetics Research*.**2008**,80, 307-320.

[6] Nozhat, N. and N. Granpayeh. *Progress In Electromagnetics Research*.**2009**, 99, 225-244.

[7] Yongqin Yu, Xuejin Li, Xueming Hong, Yuanlong Deng, Kuiyan Song, Youfu Geng, Huifeng Wei, and Weijun Tong. *OPTICS EXPRESS*.**2010**,18, 15383-15388.

[8] Ginu Rajan, Manjusha Ramakrishnan, Yuliya Semenova, Karolina Milenko, Piotr Lesiak, Andrzej W. Domanski, Tomasz R. Wolinski, Member, *IEEE*, and Gerald Farrell *IEEE SENSORS JOURNAL*.**2012**,12, 39 - 43.

[9] Kunimasa Saitoh, Member, IEEE, and Masanori Koshiba, Fellow. *IEEE PHOTONICS TECHNOLOGY LETTERS*.**2003**,15, 236 - 238.

[10] J.C. Knight, T.A. Birks, P.St.J Russel, and D.M. Atkin. *Optics Letters*.**1996**,21,1547-1549,with errata.**1997**, 22,484-485.

[11] khémiri khéareddine,Ezzedine Tahar and Houria Rezig. *25th International Conference on Microelectronics (ICM)*.**2013**,1-4.

[12] Khémiri khéareddine, Tahar Ezzedine, Houria Rezig. *International Journal of Computer and Information Technology*. **2015**,4.

[13] OptiFDTD Technical Background and tutorials, version 7.0, Optiwave, Inc.

[14] D. H. Choi and W. J. Hoefer. *IEEE Transactions on Microwave Theory and Techniques*.**1986**,.34, 1464–1470.

[15] H.F.Taylor. *Journal of Lightwave Technology*.**1984**,2,617- 628.

[16] Mourad Zghal, Rim Cherif. Optical Engineering.**2007**,46,12

[17] TaoHu, YongZhao, YongliangZhu, QiWang. *Optics Communications*.**2013**,309,6–8.

Improving the Recognition of Heart Murmur

Magd Ahmed Kotb
Professor of Pediatrics
Department of Pediatrics
Pediatric Hepatology Unit
Faculty of Medicine, Cairo University
Cairo, Egypt

Hesham Nabih Elmahdy
Professor and Chairman of Information Technology Department
Faculty of Computers and Information Cairo University
Cairo, Egypt

Fatma El Zahraa Mostafa
Professor of Pediatrics
Department of Pediatrics
Pediatric Cardiology Unit
Faculty of Medicine, Cairo University
Cairo, Egypt

Mona El Falaki
Professor of Pediatrics
Department of Pediatrics
Head of Pediatric Allergy and Pulmonology Unit
Faculty of Medicine, Cairo University
Cairo, Egypt

Christine William Shaker
Lecturer of Pediatrics
Department of Pediatrics
Pediatric Allergy and Pulmonology Unit
Faculty of Medicine, Cairo University
Cairo, Egypt

Mohamed Ahmed Refaey
Lecturer of Information Technology
Information Technology Department
Faculty of Computers and Information Cairo University
Cairo, Egypt

Khaled W Y Rjoob
Department of Information Technology
Faculty of Computers and Information Cairo University,
Cairo, Egypt

Abstract*—Diagnosis of congenital cardiac defects is challenging, with some being diagnosed during pregnancy while others are diagnosed after birth or later on during childhood. Prompt diagnosis allows early intervention and best prognosis. Contemporary diagnosis relies upon the history, clinical examination, pulse oximetery, chest X-ray, electrocardiogram (ECG), echocardiography (ECHO), computed tomography (CT) and cardiac catheterization. These diagnostic modalities reliable upon recording electrical activity or sound waves or upon radiation. Yet, congenital heart diseases are still liable to misdiagnosis because of level of operator expertise and other multiple factors. In an attempt to minimize effect of operator expertise this paper built a classification model for heart murmur recognition using Hidden Markov Model (HMM). This paper used Mel Frequency Cepestral coefficient (MFCC) as a feature and 13 MFCC coefficients. The machine learning model built by studying 1069 different heart sounds covering normal heart sounds, ventricular septal defect (VSD), mitral regurgitation (MR), aortic stenosis (AS), aortic regurgitation (AR), patent ductus arteriosus (PDA), pulmonary regurgitation (PR), and pulmonary stenosis (PS). MFCC feature used to extract feature matrix for each type of heart sounds after separation according to amplitude threshold. The frequency of normal heart sound (range= 1Hz to 139Hz) was specific without overlap with any of the studied defects (ranged= 156-556Hz). The frequency ranges for each of these defects was typical without overlap according to examined heart area (aortic, pulmonary, tricuspid and mitral area). The overall correct classification rate (CCR) using this model was 96% and sensitivity 98%. This model has great potential for prompt screening and specific defect detection. Effect of cardiac contractility, cardiomegaly or cardiac electrical activity on this novel detection system needs to be verified in future works.

Keywords—component; Hidden Markov Model (HMM); Heart Murmur; Mel Frequency Cepestral Coefficient MFCC; Systolic Murmur; Diastolic Murmur; Auscultation Area; ventricular septaldefect (VSD); mitral stenosis (MS); mitral regurgitation (MR); aortic stenosis (AS); aortic regurgitation (AR); patent ductusarteriosus (PDA); pulmonary regurgitation (PR); pulmonary stenosis (PS); Electrocardiogram(ECG); Echocardiography(ECHO); Computed Tomography(CT); Correct Classification Rate(CCR); Artificial Neural Network(ANN); Back Propagation Neural Network (BPNN); Empirical Mode Decomposition (EMD); Support Vector Machines(SVM); Adaptive Neuro-Fuzzy Inference System (ANFIS); MATRIXLABORATORY (MATLAB); Radial Basis Function (RBF)

I. Introduction

Murmur detection is the cornerstone of diagnosis of congenital heart diseases [1,2]. Efficient detection and delineation of murmurs is important to achieve diagnosis [3]. Research in heart murmur recognition is divided into two domains; (1) heart murmur recognition and (2) suggested method for more accurate murmur recognition. Most of studies in the first domain focused on recognition of mitral regurgitation (MR), mitral stenosis (MS), aortic regurgitation (AR), aortic stenosis (AS), pulmonary stenosis (PS) and normal heart sound [4]. Accurate murmur recognition was reported to vary according to used method, where artificial neural network (ANN) based murmur classification achieved accuracy of 48.5% with recorded signal

and 85% with simulated sound. Researchers built a databank with 110 sounds from 28 patients with feature vector extraction from spectrogram using average single cycle. For model testing they used 7 examples for normal sound, 4 examples for aortic stenosis and 4 examples for aortic regurgitation [5]. Back propagation neural network (BPNN) and Hidden Markov model (HMM) were also employed in murmur recognition, with extraction based upon Mel Frequency Cepestral coefficient (MFCC) as a feature. The BPNN overall CCR was reported to be 82.8% and HMM model murmur sounds overall CCR was 94.2% [6]. The recognition using HMM with empirical mode decomposition (EMD) and MFCC yielded overall accuracy equal 98.9% [7]. Other algorithms ANN with back propagation techniques, support vector machines (SVM), ANN with radial basis function and Adaptive Neuro-Fuzzy Inference System (ANFIS) classifiers were also used to recognize four types of murmur aortic regurgitation, aortic stenosis, mitral regurgitation and mitral stenosis with 90% accuracy [8]. SVM also used in heart murmur recognition based on feature extraction including four feature sets, each feature set covered specific domain. They used 3 domains (time domain, frequency domain and statistical domain). They have sensitivity range (86%-100%) [9]. Some research papers suggested new method for feature extraction in presence of murmur; they extracted feature from different features in phonocardiogram (PCG). Each heart signal represented by feature vector contains 7 variables (maximum value amplitude, sum of positive area, absolute sum of negative area, variance, shanon energy, bispectrum and winger bispectrum) [10]. This research aimed at building a novel model with high CCR [11] for detection of the normal heart and with high CCR for murmur recognition using Hidden Markov Model (HMM) and the open source matrix laboratory (MATLAB) as a programming language to build model from scratch. The paper is structured in five main sections: first section subjects and methods, in the second section statistical analysis, in the third section results, in the fourth section discussion and in the final section conclusions.

II. Subjects and Methods

A. Subjects

This research studied 1069 records of heart sounds. The study commenced by April 2015 and ended by November 2015. The records belonged to normal and structurally abnormal hearts. The 1069 records belonged to 824 children whose diagnoses were confirmed by echocardiography devices (SIEMENS acuson CV70 and Vivid S5) and other diagnostic modalities according to clinical decision.

B. Methods

a) Heart Model Creation

Heartbeats were recorded at 16-bit accuracy and 44100 Hz sampling frequency and stored as **wav** format. This research studied 605 heart sounds to build the model, of them 177 (29.3%) were records of normal hearts and 428 (70.7%) were records of structurally abnormal hearts. The structural heart abnormalities studied included VSD, PDA, MR, PS, PR, AR and AS. The records were generated from the known the auscultation areas (Mitral Area, Tricuspid Area, Pulmonary Area and Aortic Area) as shown in figure 1.

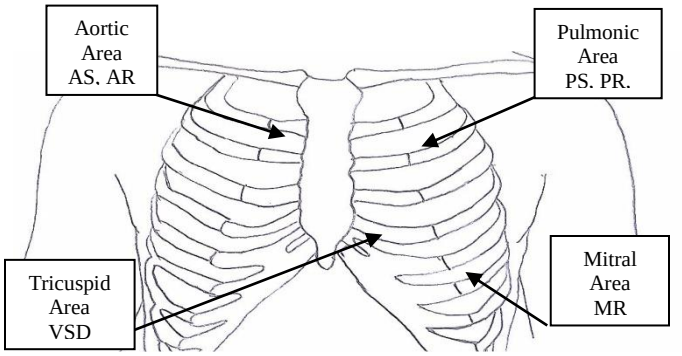


Fig. 1. Auscultation Areas

Heart sounds S1 and S2 were separated from other sounds depending on specific (0.014A amplitude threshold used to separate murmur from the original signal) threshold. Accordingly heart sounds were separated from overlapping

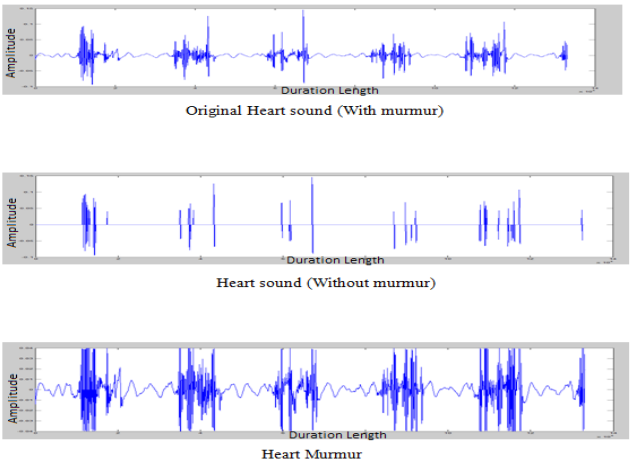


Fig. 2. Heart Murmur Separation

murmurs as shown in figure 2. MFCC feature used to extract feature matrix for each type of heart sounds. MFCC computation display is shown in figure 3. In MFCC computation 13 cepestral coefficients used for each type of heart sound to delineate clearly normal heart sounds frequency.

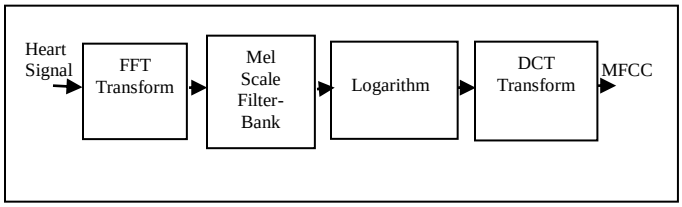


Fig. 3. MFFC Computation Steps

MFCCs computed as follow:

1) Initialize MFCC parameter analysis: frame duration 100 ms, analysis frame shift 99, pre-emphasis coefficient 0.97, number of filter-bank channels 20, number of cepstral coefficients 13, cepstral sine lifter parameter 22, lower frequency limit 130 Hz, upper frequency limit 500 Hz.

2) Preemphasis filtering:

$$y[n] = x[n] - a.x[n-1] \quad (1)$$

3) Framing and windowing signal. Window size=100, frame shift=99.And we applied hamming window to keep the continuity of the first point and the last point in each frame.

4) Compute fast fourier transform FFT using built in function fft.

5) Apply triangular filter-bank on mel-scale using trifbank function.

6) Apply filter-bank to unique part of the magnitude.

7) DCT matrix computation to eliminate discontinuities.

8) Compute DCT of the log filter-bank FBE. And keep the first 13 DCT coefficients.

Then the heart sounds classified according to HMM model as follows:

1) HMM trained using MFCC feature matrix.

2) Baum-Welch used in HMM to produce new parameter estimates that have equal or greater likelihood of having generated the training data.

3) Viterbi algorithm used to determine the best state sequence that maximizes the probability of generation of the observation sequence (each feature matrix represented one observation).

4) Forward-backward algorithm used to calculate the probability.

5) The heart model isolated HMM model for each auscultation area related murmurs. Auscultation area are divided into 4 areas to increase HMM model accuracy. Figure 4 shows model processing.

6) A heart model guided created by anatomic auscultation areas, to sense frequencies and designate origin of structural abnormality to overcome limitations of frequencies overlap.

7) Frequencies classified as low (1Hz-139Hz) and high (156Hz-556Hz). Structurally normal hearts frequencies were encountered in the low range but never in high range, yet the opposite was not correct, as we encountered low and high frequencies from mild cases of valvular defects. Thus any low frequency was subjected to amplification one fold before designation.

8) Detected signals classified by machine learning into nominal characters denoting specific structural defects.

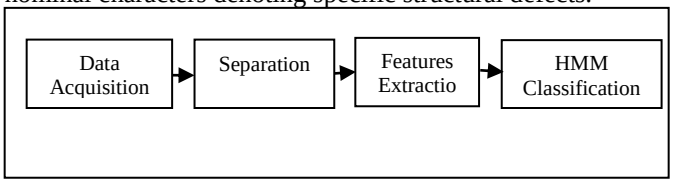


Fig. 4. Heart Model Processing

b) Heart Model Validation:

The created heart machine learning model based on HMM and MFCC used to recognize 464 blinded heart sounds. All validation studies were compared to diagnoses derived from standard echocardiography and other imaging studies.

III. Statistical Analysis

Statistical analysis was conducted using Statistical Package for Social Sciences version 15 (SPSS, Chicago, Ill). Simple frequency, cross-tabulation, descriptive analysis, and tests of significance (t test for parametric data and $\chi 2$ tests for nonparametric numbers N5) were used. Studied heart sounds validation was computed according to CCR [11]. Sensitivity testing was computed for created heart model.

IV. Results

The study was conducted on 1069 records that belonged to 824 children whose diagnoses were confirmed by echocardiography and other diagnostic modalities according to clinical decision. Their ages ranged from 1 week to 14 years (mean±standard deviation=2years 9.5months ±6 months. Of them 458 (55.6%) were males. Table 1 shows different diagnoses of the studied sound records.

A. Heart Model Creation:

605 heart sounds were studied to build the model, of them 177 (29.3%) were records of normal hearts and 428 (70.7%) were records of abnormal heart sounds. The structural heart abnormalities studied included VSD, PDA, MR, PS, PR and AS.

Table 2 demonstrates the frequencies of recognized sounds of the aforementioned structural valvular defects, and clearly separates frequencies from structurally normal heart from those of structurally abnormal hearts. We detected general overlap of frequencies between the aforementioned abnormalities, yet this overlap was not recorded on any specific auscultation area according to figure 1.

B. Heart Model Validation:

This paper used 464 heart signals covering 100 normal hearts, 62 PS, 56 PR, 46 MR, 64 VSD, 72 PDA, 61 AS and 3 AR to test heart sound recognition using machine learning model based on HMM and MFCC.

TABLE I. Diagnosis of Studied Recorded Signal

diagnoses	**Total Number of Sound Records**		**Machine Learning Model Based on HMM and MFCC Creation**		**Machine Learning Model Based on HMM and MFCC Validation**	
	N	**%**	**N**	**%**	**N**	**%**
PR	146	13.7	90	14.9	56	12
PS	142	13.3	80	13.2	62	13
PDA	152	14.2	80	13.2	72	15.5
VSD	154	14.4	90	14.9	64	13.9
AS	94	8.8	33	5.5	61	13
AR	3	0.2	0	-	3	0.6
MR	101	9.4	55	9.1	46	10
NO	277	26	177	29.2	100	22
Total	1069	100%	605	100	464	100

HMM: Hidden Markov Model, **N**=Total number of heart sounds.

Table 3 demonstrates ages of studied records according to ages.

TABLE II. HEART SIGNAL FREQUENCY RANGE IN HZ

Class	N	%	Frequency Range Min	Max
Normal	177	29.2	1	139
Abnormal	431	70.8	156	556
Tricuspid Area				
VSD	90	14.9	156	164
Mitral Area				
MR	55	9	158	162
Pulmonary Area				
PDA	80	13.2	157	167
PS	80	13.2	156	556
PR	90	14.9	156	200
Aortic Area				
AS	33	5.5	157	176
AR	3	0.1	156	160

Min: Minimum frequency, **Max**: Maximum frequency.

This paper evaluated machine learning heart model based on HMM and MFCC according to sensitivity as shown in table 4 and CCR in table 5.

TABLE III. STUDIED RECORDS ACCORDING TO AGES OF CHILDREN

	Mean age (years)	**SD**(months)	**t -test p =**
Heart Model Creation Group	2.8	7.9	0.00072
Heart Model Validation Group	2.82	7.8	

SD: Standard Deviation.

Finally, mean CCR of machine learning model based on HMM and MFCC was 96% and overall sensitivity was 98%. Machine learning model based on HMM and MFCC training time was 15 seconds and testing time was 3 seconds.

TABLE IV. HEART MODEL IMAGING DETECTED SENSITIVITY

	Machine Learning Model Based on HMM and MFCC detected	**ECHO detected**	**Machine Learning Model Based on HMM and MFCC Sensitivity (TP/(TOP+ FN)) %**	**ECHO detected Sensitivity (TP/(TOP+ FN)) %**
VSD	62	64	96.6	100
PS	62	62	100	100
PDA (Greater than 0.3mm)	72	72	100	100
PR	56	56	100	100
MR	44	46	95.5	100
AS	61	61	100	100
Normal	100	100	100	100

ECHO detected: Echocardiography detected, **TP**= True Positive, **TOP**=Total Positive, **FN**=False Negative.

TABLE V. HEART MODEL CCR

Heart Signal Frequency Range in Hz Min	Max	**Heart Signal**	**Cycles**	**Machine Learning Model Based on HMM and MFCC Interpretation**	**Machine Learning Model Based on HMM and MFCC CCR %**
156	164	VSD	64	62	97
156	556	PS	62	59	95
157	167	PDA	72	67	93
156	200	PR	56	52	93
158	162	MR	46	44	96
157	176	AS	61	61	100
1	139	Normal	100	100	100

CCR: Correct Classification Rate, **Min**: Minimum frequency, **Max**: Maximum frequency.

TABLE VI. COMPARISON BETWEEN MACHINE LEARNING HEART MODEL BASED ON HMM AND OTHER MODELS

Ref	**Sound**	**Sensor Type**	**Data Bank**	**Sensor Position**	**Method**	**Results**
Strunic et al., 2007 [5]	AS,AR	Simulator	110	Appropriate Auscultation Area	ANN	Up to 85±7.4% accuracy, 95±6% sensitivity
Zhong et al., 2013 [6]	MR, MS,AS, AR, PS	Not determined	600	Appropriate Auscultation Area	BPNN ,HMM and MFCC	HMM accuracy 94.5% BPNN accuracy 82.5%
Jimenez et al., 2014 [7]	Not determined	Welch Allynr Meditron model	400	Appropriate Auscultation Area	HMM and MFCC combined with statistical moment (EMD)	Accuracy 98.9% and 98.6% sensitivity
Devi et al., 2013 [8]	AS,AR MR, MS.	Not determined	Not determined	Appropriate Auscultation Area	ANN,BPNN,SVM, ANN with RBF ANFIS.	90% and above accuracy
Machine Learning Model Based on HMM and MFCC	VSD, MR, PDA, PS,PR, AS and Normal	hands-free tie-clip electrets (real heart sounds)	1069	Appropriate Auscultation Area	Machine learning Based on HMM and MFCC	CCR 96% And 98% sensitivity

ANN: Artificial Neural Network, **BPNN**: Back Propagation Neural Network**, HMM**: Hidden Markov Model, **MFCC**: Mel Frequency Cepestral coefficient, **EMD**: Empirical Mode Decomposition, **SVM**: Support Vector Machine, **RBF**: Radial Basis Function, **ANFIS**: Adaptive Neuro-Fuzzy Inference System.

V. DISCUSSION

A machine learning model based on HMM and MFCC, covered normal heart sounds and abnormal heart sounds including AR, VSD, MR, PS, PDA and AS. The machine learning model based on HMM and MFCC achieved 98% sensitivity and overall CCR =96%.

This work supports that the HMM as a classifier and MFCC as a feature matrix are widely used for heart sounds classifications, as they have demonstrated their effectiveness, especially if mixtures of features from different domains were employed [12]. The MFCC 13 features coefficients allowed reduction of calculation time and memory that will impact cost of recognition model. Another important point in favor of HMM in heart sound recognition is the ability of easy update.

This machine learning model successfully recognized other types of murmur as VSD, PDA, and PR which were not recognized by others. Table 6 compares all previously reported studies and types of murmur recognized [5-8].

This study comprised the largest reported databank size of real heart sounds (1069 heart sounds), of them 464 heart sounds were for testing and 605 heart sounds were for training. This research did not study simulated heart sounds, while all previous reports used simulated sounds. This research need to emphasize that simulated heart sounds models were not validated against real heart sounds thus the reported accuracy of systems based on simulated heart sounds should be cautiously interpreted [5-8]. The accuracy of this machine learning heart model for recognition of heart sounds has future implications in heart sound recognition using simpler devices compared to the more complex operator dependent ECHO machines, and promises new role in clinical education. Heart sound recognition using HMM model shortcomings is the difficulty of recognizing some mild cases of MR. The number of AR cases were limited thus This paper need to study more cases to enhance machine learning recognition. The study did not address effect of heart contractility, heart rate, conduction defect, hypertrophy and size on accuracy of heart sound recognition.

This paper aim to study effect of combining heart rate sensors with machine learning model on recognition ability and on time of training and computational complexity in future works.

VI. CONCLUSION

The machine learning model based on HMM as a classifier and 13 MFCC elements and real heart sounds is effective in recognizing VSD, MR, PS, PR, PDA AS and AR. It relies upon separation of murmur from original heart signal using amplitude threshold. It achieved 98% sensitivity and 96% CCR. Real heart sounds recognition sensitivity result is better than simulated heart sounds. Each heart sound should be recorded from specific auscultation area. Heart machine learning model may have the potential to assist clinicians for more accurate diagnosis. This paper used amplitude threshold to separate murmur from original heart sound.

REFERENCES

[1] Trivedi N, Levy D, Tarsa M, Anton T, Hartney C, Wolfson T, Pretorius DH. Congenital cardiac anomalies: prenatal readings versus neonatal outcomes. J Ultrasound Med 31(3), pp:389-99, 2012.

[2] Rajakumar K, Weisse M, Rosas A, Gunel E, Pyles L, Neal WA, Balian A, Einzig S. Comparative study of clinical evaluation of heart murmurs by general pediatricians and pediatric cardiologists. Clin Pediatr (Phila). 38(9), pp:511-8,1999.

[3] Asprey DP. Evaluation of children with heart murmurs. Lippincotts Prim Care Pract. 2(5), pp:505-13, 1998.

[4] D Kumar, P Carvalho, M Antunes , J Henriques ,M Maldonado, R Schmidt, J Habetha, "Wavelet Transform and Simplicity Based HeartMurmur Segmentation",The Proceedings of The Computers in Cardiology Conference, pp:173:172, Valencia, Spain, 17-20 Sep 2006.

[5] Strunic SL, Rios-Gutierrez F, Flores RA,Nordehn G,Burns S. Detection and Classification of Cardiac Murmurs Using Segmentation Techniques and Artificial Neural Networks. The Proceedings of Computational Intelligence and Data Mining IEEE Symposium on Conference, pp:397-404, Honolulu, USA, 1 March-5 April 2007.

[6] Zhong L, Wan J, Huang Z, Cao G, Xiao B. Heart Murmur Recognition Based on Hidden Markov Model. Journal of Signal and Information Processing.4, pp:140-144, 2013.

[7] Jimenez JA, Becerra MA, Delgado-Trejos E. Heart Murmur Detection Using Ensemble Empirical Mode Decomposition and Derivations of The Mel-Frequency Cepstral Coefficients on 4-Area Phonocardiographic Signals. The Proceedings of The Computing Cardiology Conference,pp:493-496, Cambridge, USA, 7-10 September 2014.

[8] Devi A, Misal A. A Survey on Classifiers Used in Heart Valve Disease Detection. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2, pp: 609-614, 2013.

[9] D. Kumar, P. Carvalho, M. Antunes, R.P.Paiva, J.Henriques," Heart Murmur Classification With Feature Selection", The Proceedings of The Annual International Conference of The IEEE Engineering in Medicine and Biology Society, DOI:10.1109/IEMBS.2010.5625940, pp:4566-4569, Buenos Aires, Argentina, 31August-4 September2 2010.

[10] Amir Mohammad Amiri and Giuliano Armano,"Segmentation and Feature Extraction of Heart Murmurs in Newborns", Journal of Life Sciences and Technologies Vol 1, pp 107:111, 2013.

[11] Classification accuracy. Centre for Bioscience: The Higher Education Academy. Clustering and Classification methods for Biologists. http://www.alanfielding.co.uk/multivar/accuracy.htm. Accessed on 22/5/2016.

[12] Leng S, Tan RS, Chai KT, Wang C, Ghista D, Zhong L. The electronic stethoscope. Biomed Eng Online. 10, pp:14:66, 2015.

Investigative Behavioral Intention to Knowledge Acceptance and Motivation in Cloud Computing Applications

Case study: Group of students from Jordanian universities

Sundus A. Hamoodi
Department of Business Networking and Systems Management
Philadelphia University
Amman –Jordan

***Abstract*—Recently the number of Cloud Computing users in educational institutions has increased. Students have the chance to access various applications and this gives the opportunity to take advantage of those applications. This study examined the behavioral Intention toward Cloud Computing Applications and evaluated the acceptance of these Applications. Participants population consisted of 110 students from different Jordanian universities. The results showed that Performance Expectancy, Effort Expectancy, Attitude toward using Technology, Social Influence, Self-Efficacy, Attention and Relevance have different levels of correlation with Behavioral Intention in Cloud Computing Applications, and there was no correlation between Anxiety and Behavioral Intention in Cloud Computing Applications.**

Keywords—Cloud computing; formatting; ARCS model; UTAUT model

I. INTRODUCTION

Cloud Computing is considered as the most modern technique which can be used in education today in order to deliver services that can help students and learners to accomplish their tasks in an effective and efficient way.

Many researchers defined and described cloud computing in different ways. The researchers shared a number of facts about Cloud Computing that convert Information Technology from product to service. A number of researches clarified that "The cloud can provide exactly the same technologies as "traditional" IT infrastructure, the main difference, as mentioned previously, is that each of these technologies is provided as a service" [7]. Yuvaraj gave an example of Cloud Computing tools that may help users in different categories, "There are various cloud based tools for reference service needs of the libraries such as cloud-based video services (e.g. YouTube, TeacherTub), information collection services (e.g Google forms) and files sharing services (e.g. Dropbox)" [23].

Cloud Computing transfers the processing and storage to the cloud that saves more cost by providing more storage space, software license, and hardware device maintenance. Many features can be offered by Cloud Computing applications such as availability of the storage space that can be reached at any time at any place. According to Al Mourad and Hussain, "Cloud services deliver compute, storage, software, applications, etc. Via Internet to customers on a self-serve basic [4]. In addition, students and learners can accomplish their works and assignments through cloud computing applications that enable them to share their duties with professors and students if required. Researchers in Cloud Computing discuss many challenges related to this technique. The most important of these challenges are lack of privacy and security [16] [17].

Other researchers also explain the Cloud Computing concept and how companies and universities move toward Cloud Computing [1] [3] [10].However, few of them studied some of the factors which influence university students who use Cloud Computing applications in their study including homework assignment.

From previous studies, the researcher has identified several factors which may have influence on university students in using cloud Computer Applications.

The main objective of this paper is to examine the influence of these factors on students' performance and seeks to answer to the following questions:

- Can Jordanian students work with Cloud Computing Applications?
- Do Cloud Computing Applications motivate students to use them in their study?
- What are the main problems facing Jordanian students in using Cloud Computing Applications?

The researcher has also defined a set of hypothesis, developed a questionnaire, and formed an experimental group of students to work on Cloud Computing Applications. The findings from using SPSS analysis were discussed and interpreted the correlations between different factors and behavioral intention to use Cloud Computing Applications.

II. PREVIOUS STUDIES

Cloud Computing is one of the modern terminology which has recently received an increasing attention by researchers in both theory and applications.

It was defined in several ways through its features and services that can be offered. According to U.S. National Institute of Standards and Technology (NIST), "Cloud computing is a model for enabling convenient on-demand network access to a shared pool of configurable computing resources (e.g., Networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction "[15]. Hashemi and Seyyed explained that "Cloud computing is making the pool of virtual computing resources with a focus on large scale computing resources that are connected to the network and which allows customers to be shared dynamic hardware, software resources and data, and according to their actual usage, paying costs" [9].

Based on the literature studies, the researcher defines Cloud Computing as follows: Cloud Computing is a model of services that allowed users to access and use different virtual resources in an easy way and without the need to recognize how to maintain and manage those resources.

The characteristics of Cloud Computing are discussed below:

A. On-demand access

On-demand network access is sharing a pool of computing resources by different users from different places. Users can access shared resources at anytime and anywhere [3] [2] [22]. Karim discussed that "a wide range of end users have access to the applications and data served by the cloud" [10]. Cloud Computing resources grow on-demand which led to that various applications uses cloud technology now a day [18]. Further, its being illustrated that "Using this feature when needed the customer can easily and automatically access computing facilities like server, net, storage from any provider as soon as possible." [8]. Access to Cloud Computing may take different ways depending on the user needs.

According to the National Institute of Standards and Technology's (NIST), there are three ways to access a Cloud Computing: software as a service, platform as a service, and infrastructure as a service [15]. The demand on Cloud Computing increases due to the increased accessibility of the internet and its expansion using the digital devices [6].

In addition, other researchers discussed that using service models (Infrastructure, Platform and Software) are more flexible for education [1]. They encouraged to use Cloud Computing in education due to its features and the benefits it may result in such as; reduce cost of technology, enhance communications and give better service delivery. According to Adeoye, 2015 "Cloud computing is the better ICT utilization mechanism for educational institutions teaching, learning and a service delivery requirement, for it enables wise and strategic use of technology that significantly reduces the cost"[2].

B. Ubiquitous network access

Location-independent resource pooling: The resources needed by different customers can be supplied by providers which contain storage and memory frequency Internet and virtual systems as a pool feature [8].

C. Data storage

Storing information and data through Cloud Computing does not require large storage spaces on the user's laptop and smart devices. "Cloud Storage delivers virtualized storage on demand over a network based on a request for a given quality of service (QoS)" [19]. On demand storage could be delivered by cloud storage on a network that based on a given quality service. Further, Abu El-Ala stated that Cloud Computing "includes all the services related to the infrastructure of the cloud such as physical resources (as storage devices, school servers, and national communication network, etc.)" [2].

Users can reduce their costs by using the Cloud Computing which does not require an internal power to store information [19].

D. Development and Maintenance

One of the Cloud's Computing's benefits is that people and institutions are not responsible for updating and maintaining the applications and services. This means that they can save time and money for updating and maintenance needs. "User is not responsible for where the services or applications are located or how it maintained or updated") [3].

III. UTAUT MODEL

Unified Theory of Acceptance and Use of Technology Model (UTAUT) is a model of acceptance, formulated by Venkatesh and others, integrates theories and models to measure user's intention and usage of technology [21]. The dimensions of this model as discussed by different authors [20], [5] are as follows:

- Performance Expectancy: mark the individual expectancy on how much a system will help in improving job performance.
- Effort Expectancy: Less effort and ease of use of the technologies.
- Social Factors: How others believe about using technology.
- Facilitating conditions: Infrastructure required for supporting technologies to facilitate tasks.

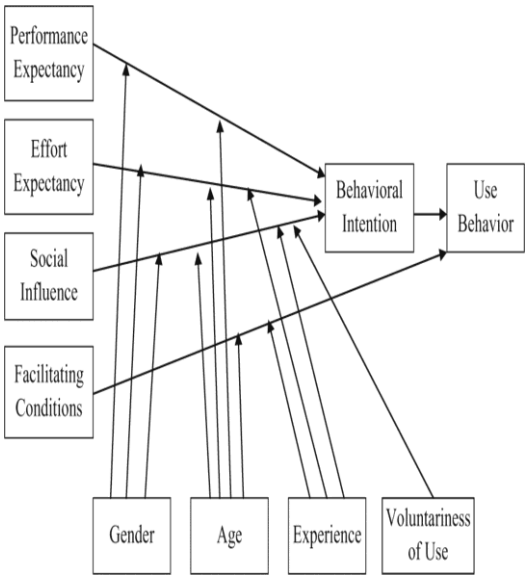


Fig. 1. Unified Theory of Acceptance and Use of Technology [21]

IV. ARCS MODEL

The ARCS model is based on a synthesis of motivational concepts and its characteristics are considered into four categories: attention (A), relevance (R), confidence (C), and satisfaction (S)" [12]. It is further explained by other researchers such as [13] [14]. Researchers have summarized the dimensions of ARCS model as follows:

- Attention: Emphasis on attracting the attention of learners in several ways, such as variety of videos, lectures and reading.
- Relevance: The importance of using instruction, command and other materials related to what users are familiar with and what they need.
- Confidence: Learners should be given a reward, a chance to succeed and increase the confidence they have.
- Satisfaction

The learner must get practical exercises and tests to increase their satisfaction with learning materials.

In addition, Keller stated that ARCS model "provides guidance for analyzing the motivational characteristics of a group of learners and designing motivational strategies based on this analysis"[11].

V. STUDY PROBLEM

Most of the researchers in Cloud Computing system are interested in examining the importance of those applications and their success in maintaining data security, availability, electronic libraries, and online education. Due to early studies, it was found that there is a lack of researches on examining the use of Cloud Computing applications by university students in studying and performing their tasks.

Different dimensions of the model were used in measuring motivation towards acceptance of those applications.

VI. RESEARCH MODEL

The research model showed in figure 2 measures the acceptance of Cloud Computing Applications and adapted from the work by [21] [12] [23].

This model discusses the effect of the certain factors which are (Performance Expectancy, Effort Expectancy, Attitude toward Using Technology, Social Influence, Self-Efficacy, Attention, Relevance and Anxiety) on the Behavioral Intension of Cloud Computing Applications. The purpose of this model is to examine the impact of the considered factors related to acceptance and motivation on Behavioral Intention to use Cloud Computing Application by university students.

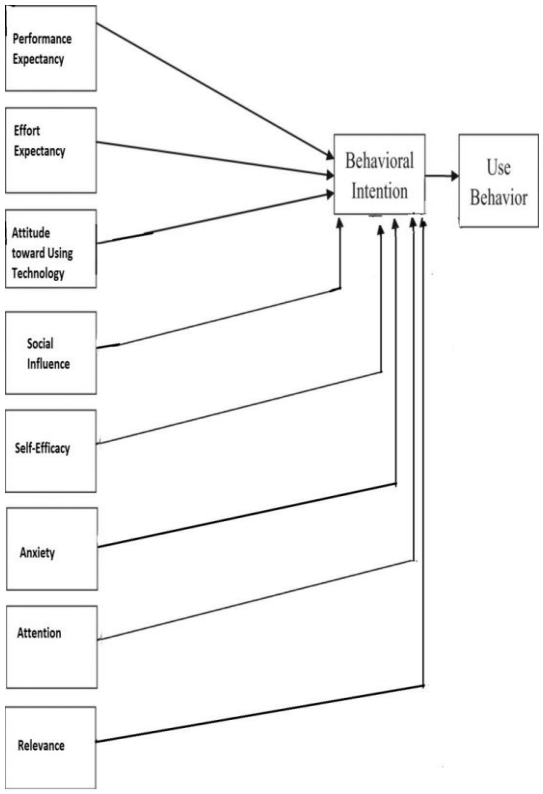


Fig. 2. Research model

VII. RESEARCH HYPOTHESIS

H0. There is no correlation between Performance Expectancy and Behavioral Intention in Cloud Computing Applications

H1. There is no correlation between Effort Expectancy and Behavioral Intention in Cloud Computing Applications

H2. There is no correlation between Attitude toward Using Technology and Behavioral Intention in Cloud Computing Applications

H3. There is no correlation between Social Influence and Behavioral Intention in Cloud Computing Applications

H4. There is no correlation between Self-Efficacy and Behavioral Intention in Cloud Computing Applications

H5. There is no correlation between Anxiety and Behavioral Intention in Cloud Computing Applications

H6. There is no correlation between Attention and Behavioral Intention in Cloud Computing Applications

H7. There is no correlation between Relevance and Behavioral Intention in Cloud Computing Applications

VIII. METHODOLOGY

A. Data Collection

The researcher collected data using the developed survey. The student participated in this study are 110. They are undergraduate and postgraduate students from different Jordanian universities. This study measures technology acceptance by adopting questions from [21][12] [23].

B. Procedure

The study uses a mix approach of questionnaire and interview as follows:

- Investigating the dimensions of both the acceptance and motivation models.
- Distributing the questionnaire online to participant.
- Interviewing students to measure their acceptance and satisfaction.
- Analyzing the data using SPSS package.

Different Cloud Computing Applications are examined by this study.

C. Experimental Group

The goals of this experiment are to demonstrate the importance of cloud computing in student's homework assignments and to examine the factors affecting students' performance through observation and survey results.

For this purpose, a group of students has created and the analysis of an electronic survey using Form in Google Docs applications has been conducted.

At the first instance some of the students were reluctant to use the Cloud Computing Applications because they were lacking of expertise. With the efforts of the researcher in highlighting the advantages and the benefits of using Cloud Computing Applications and providing some training using Google Docs Applications, the group of the participants was highly motivated to use this technology.

Thus, the participant group has conducted their assignment in an efficient way and achieved better results compared with their colleagues who have not used Cloud Computing Technology.

The experimental group got benefits from using Google Docs Application as follows:

- Creating large number of surveys for free
- Sharing knowledge, ideas and thoughts with large number of participants through a web browser.
- Distributing the surveys globally via URL or an e-mail.
- Collecting responses automatically in MS Excel for easy analysis, charts and functions.
- Helping students in choosing different question types through Google Form.

IX. RESULTS AND DISCUSSIONS

A. Descriptive Sample

The sample of the study consisted of 110 undergraduate and postgraduate students. The findings show that 64.5 % of the respondents are of age 18 to 24, 26.4% of the respondents is of age 25 to 31, 5.5% of the respondents are of age from 31 to 37 and 3.6% of the respondents are of age 45 or greater. These results are illustrated in figure 3.

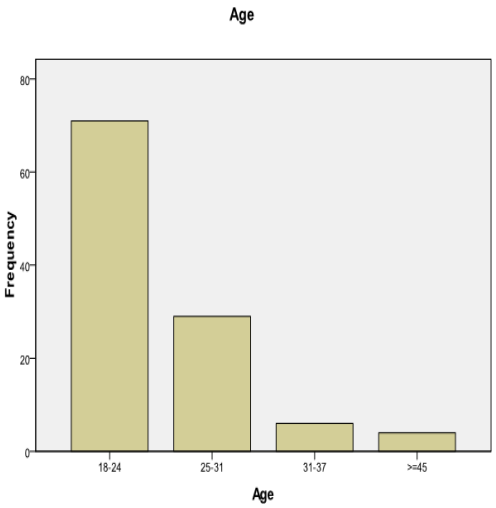


Fig. 3. Distribution of the participants' age

The results in figure 4 show those participants with computer experience which ranging from excellent (27.3%) to week (1.8) based on Likert scale.

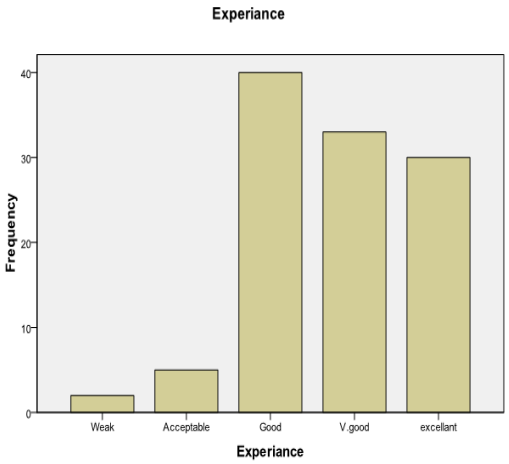


Fig. 4. Distribution of participants' computer experience

B. Statistic

The findings from the data analysis show the following:

H0. There is a significant relation between Performance Expectancy and Behavioral Intention in Cloud Computing Applications (equal =0.594 significant at level 0.01).

H1. There is a significant relation between Effort Expectancy and Behavioral Intention in Cloud Computing Applications (equal =0. 455 significant at level 0.01).

H2. There is a significant relation between Attitude toward Using Technology and Behavioral Intention in Cloud Computing Applications (equal =0.589 significant at level 0.01).

H3. There is a significant relation between Social Influence and Behavioral Intention in Cloud Computing Applications (equal =0.504 significant at level 0.01).

H4. There is a significant relation between Self-Efficacy and Behavioral Intention in Cloud Computing Applications (equal =.478 significant at level 0.01)

H5. There is a negative relation between Anxiety and Behavioral Intention in Cloud Computing Applications (equal =-0.121 significant at level 0.01).

H6. There is a significant relation between Attention and Behavioral Intention in Cloud Computing Applications (equal =0.642 significant at level 0.01).

H7. There is a significant relation between Relevance and Behavioral Intention in Cloud Computing Applications (equal =0.556 significant at level 0.01).

The results show that the students are able to use Cloud Computing Applications with different levels of knowledge through the questionnaire and observations. It also show that there is a significant relation between Performance Expectancy and Behavioral Intention in Cloud Computing Applications (equal =0. 594). However, it has been found that there is a weak correlation (equal to 0.454) between usefulness and Behavioral Intention of Clouds Computing regarding to the Performance Expectancy.

In addition, the results also show Performance Expectancy, Effort Expectancy, Attitude toward Using Technology, Social Influence, Self-Efficacy, Attention and Relevance have different levels of significant correlation with Behavioral Intention in Cloud Computing Applications with a negative correlation between Anxiety and Behavioral Intention in Cloud Computing Applications. This negative correlation may be due to the students' hesitation of using Cloud Computing Applications. Their anxiety might be caused by the fear of losing information as a result of choosing a wrong option.

Further, the results show that the experimental group has the desire to use Cloud Computing Applications. However, the survey results also show that they have some anxieties toward using Cloud Computing Applications as a result of lack of sufficient knowledge on the benefits of the cloud technology.

The results indicating that there is a motivation toward using a Cloud Computing Applications among students as explained by the significant relation between Attention and Behavioral Intention in Cloud Computing Applications.

One of the benefits of the experimental group in this study is highlighting the need for considering more factors that may have impacts on Behavioral Intention in Cloud Computing Applications. Such factors are those that may affect the use of a Cloud Computing Application in free time, planning to use it in future, and using this technology as often as possible.

X. CONCLUSIONS

This study finds out that Jordanian university students have the desire, motivation and ability to use Cloud Computing Applications in conducting their study work including homework assignment. However, instructors have to work toward encouraging and educating students on the importance of Cloud Computing Applications and their roles in getting the required information and facilitate the mechanics of completing homework. This would alleviate anxiety due to lack of sufficient knowledge on the benefits of Cloud Computing Applications.

It also concluded that students using Cloud Computing Applications are performing better than those who have not used this technology. The analysis of the results indicates that there is a positive correlation between almost all the considered factors in this study and Behavioral Intention in Cloud Computing Applications.

XI. FUTURE WORK

This study should be explored to include more students from different countries to have comprehensive results. The study may be expanded by further consideration of the relation between knowledge acceptance and other factors.

Based on the results of this study it is proposed to establish a virtual Cloud Acceptance and Usability Center (CAUC). The purpose of set center is to contains all the research outputs in the different areas of Cloud Computing from different countries to enhance cloud products, make it more usable, and

obtain comprehensive results and comparative studies among users of Cloud Computing Applications globally.

REFERENCES

[1] Abid, Muhammad Haris, Tasleem Mustafa and Muhammad Shakeel Faridi , "Cloud Computing: A general user's perceptions and security issues at Universities of Faisalabad Pakistan, " JCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, 2012.

[2] Abu El-Ala, N. S., W. A. Awad and H. M. El-Bakry," Cloud Computing for Solving E-Learning Problems," IJACSA, International Journal of Advanced Computer Science and Applications, vol. 3, No. 12, 2012 , www.ijacsa.thesai.org

[3] Adeoye, Blessing F.,"Utilization of Cloud Computing in education," Journal of global research in computer science, Vol. 6, No. 4, April 2015.

[4] Al Mourad, Mohamed Basel and Mohammed Hussain,"The Impact of Cloud Computing on ITIL Service Strategy Processes," International Journal of Computer and Communication Engineering, vol. 3, No. 5, September 2014.

[5] Attuquayefio, Samuel NiiBoi and Ghana Hillar Addo, "Using the UTAUT model to analyze students' ICT adoption," International Journal of Education and Development using Information and Communication Technology (IJEDICT), vol. 10, Issue 3, pp. 75-86, 2014.

[6] Chung, Hyunji, Jungheum Park, Sangjin Lee, and Cheulhoon Kang,"Digital forensic investigation of cloud storage services," Digital Investigation, vol. 9, Issue 2, PP. 81–95, 2012.

[7] Gorelik, Eugene, "Cloud Computing Models," Working Paper CISL# 2013-01, Massachusetts Institute of Technology, Master Thesis, 2013.

[8] Hashemi, Sajjad,"Cloud Computing Technology for E-Government Architecture," International Journal in Foundations of Computer Science & Technology (IJFCST), vol. 3, No.6, November 2013.

[9] Hashemi, Sajjad and Seyyed Yasser Hashemi,"Cloud Computing for E-Learning with More Emphasis on Security Issues," World Academy of Science, Engineering and Technology International Journal of Computer, Control, Quantum and Information Engineering, vol.7, No.9, 2013.

[10] Karim, Faten, Robert Goodwin," Using Cloud Computing in E-learning Systems," International Journal of Advanced Research in Computer Science & Technology (IJARCST), vol. 1, issue 1, 2013

[11] Keller, J. M., Motivational systems. In H. D. Stolovitch, & E. J. Keeps (Eds.), Handbook of human performance technology, 2nd Edition. San Francisco: Jossey-Bass Inc., Publisher, 1999.

[12] Keller, John," John .How to integrate learner motivation planning into lesson planning: The ARCS model approach," Paper presented at VII Semanario, Santiago, Cuba, 2000.

[13] Malik, Sangeeta," Effectiveness of Arcs Model of Motivational Design to Overcome Completion Rate of Students in Distance Educations," Turkish Online Journal of Distance Education, vol. 15, no.2: Article 14, 2014.

[14] Marshall, James and Matthew Wilson,"Motivating e-Learners: Application of the ARCS Model to e-Learning for San Diego Zoo Global's Animal Care Professionals," The Journal of Applied Instructional Design, Vol. 3, Issue 2, October 2013, www.jaidpub.org

[15] Mell, P., T. Grance, "The NIST Definition of Cloud Computing," Gaithersburg, National Institute of Standards and Technology, 2011.

[16] Padhy, Rabi Prasad, Manas Ranjan Patra and Suresh Chandra Satapathy,"Cloud Computing: Security Issues and Research Challenges," IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS) vol. 1, no. 2, December 2011.

[17] Piplode, Rajesh and Umesh Kumar Singh,"Overview and Study of Security Issues & Challenges in Cloud Computing," International Journal of Advanced Research in Computer Science and Software Engineering Research Paper, vol. 2, issue 9, 2012.

[18] Rizzardini, Rocael Hernez, Linares, Byron, Mikroyannidis, Alexander and Schmitz, Hans-Christian, "Cloud services within a ROLE-enabled Personal Learning Environment," 1st International Workshop on Cloud Education Environments (WCLOUD 2012), Antigua, Guatemala, November 2012.

[19] Stephen, Okeke,"The Study of the Application of Data Encryption Techniques in Cloud Storage to Ensure Stored Data Integrity and Availability," International Journal of Scientific and Research Publications, vol. 4, issue 10, October 2014.

[20] Thomas, Troy Devon, Lenandlar Singh and Kemuel Gaffar, "The utility of the UTAUT model in explaining mobile learning adoption in higher education in Guyana," International Journal of Education and Development using Information and Communication Technology (IJEDICT), vol. 9, issue 3, pp. 71-85, 2013.

[21] Venkatesh, Viswanath, Michael G. Morris, Gordon B. Davis and Fred D. Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, pp. 425-478, 2003.

[22] Vouk, Mladen A.," Cloud Computing – Issues, Research and Implementations," Journal of Computing and Information Technology - CIT 16, 4, pp. 235–246, 2008.

[23] Yuvaraj, Mayank,"Examining Libraries' behavioral intention to use cloud computing applications in Indian center universities," Annuals for library and information Studies, vol. 60, pp. 260-268, 2013.

The Impact of Black-Hole Attack on ZRP Protocol

CHAHIDI Badr
Mathematics and Computer Science Dept, LAVETE Laboratory Faculty of Sciences and Technical Settat, Morocco

EZZATI Abdellah
Mathematics and Computer Science Dept, LAVETE Laboratory Faculty of Sciences and Technical Settat, Morocco

***Abstract*—lack of infrastructure in ad hoc networks makes their deployment easier. Each node in an ad hoc network can route data using a routing protocol, which decreases the level of security. Ad hoc networks are exposed to several attacks such as the blackhole attack. In this article, a study has been made on the impact of the attack on the hybrid routing protocol ZRP (Zone Routing Protocol). In this attack a malicious node is placed between two or more nodes in order to drop data. The trick of the attack is simple, the malicious node declares to have the most reliable way to the destination so that the wife destination chooses this path.**

In this study, NS2 is used to assess the impact of the attack on ZRP. Two metrics measure, namely the packet delivered ratio and end to end delay.

Keywords—ZRP; Blackhole; security; Routing

I. INTRODUCTION

Wireless sensor networks are composed of a set of independent nodes capable of communicating with each other via radio waves. Communications can be direct or through other nodes called relay allowing others outside to communicate. Each node acts as a terminal and as a routing point so that each node can send packets or receive packets or re-route packets if they belong to another node.

Putting a number of radio range nodes causes the appearance of a rapidly deployed network and adapts to a number of situations where the infrastructure mode is too expensive, too long or sometimes impossible.

Ad hoc mode differs from the infrastructure mode where the nodes communicate via an access point, which can be connected to a fixed network. This type of network (Ad-hoc) which is characterized by a lack of infrastructure is used in various fields such as industrial fields for monitoring the pressure flow or others such as the military for surveillance of the battlefield or in the civil field during disasters by rescue services.

So we are dealing with ad hoc networks that use specific routing protocols where the big problem is security, because that they are designed to run in an environment of trust. Arguably the MANET is susceptible to attacks, whether active or passive.

To secure an ad hoc network, you must consider the following attributes: availability, confidentiality, integrity and authentication. Most of the research has been done with the aim of reducing energy consumption without taking into account different attacks such as the attack Black Hole.

In this section, the security requirements are presented as well as principles of routing, and the impact of the attack on the Black Hole ZRP protocol. The simulation is performed on NS-2 and the simulation results are analyzed on various parameters such as the rate of delivered packages and the time from start to finish.

In this article, a detailed explanation of the new routing protocol where it has implemented the attack black hole. The simulation was made under NS2, in the objective of studying the impact of the attack on the networks Manet. Metric two were measured to know the rate of lost packets and the end-to-end delay. As expected a decrease in performance was noted mainly in the case where the number of nodes sources is high.

Our paper is organized as follows: the second part describes the principle of routing in ad hoc networks. In the third part there is a classification of attacks. The fourth part gives more information on security in ad hoc networks and the implementation of the attack in the Protocol ZRP. The simulation of the attack and the discussion of the results are shown in Part 5. We conclude the section in Part 6.

II. ROUTING IN AD HOC NETWORKS

The routing protocols in different categories, and this according to the itinerary discovery method, according to the information exchange method or how the nodes share the job of routing them.

A. Routing classification

Given their specific characteristics (absence of fixed infrastructure, limited source of energy and ability to calculate non-secure communication links), ad hoc networks CANNOT use the WIRED NETWORK ROUTING PROTOCOL. New protocols were born with the aim to meet their needs.

These protocols can be divided into three categories according to the update method of the routing table. The first so-called proactive where each node maintains its updated routing table via a regular exchange with its neighbors. OLSR [1] (Optimized link state routing) is one of the most popular routing protocols for this category.

The second category is called REACTIVE; each node performs a demand routing. When a node wants to communicate with another, it sends the route request requests to all nodes, and expects the recipient's response, a response that contains the path to take. Among the reactive protocols it there's the AODV (Ad-hoc On-demand Distance Vector).

The last category includes the proactive and reactive, it is called Hybrid. Each node wants to send data verified the

presence of the destination within the zone using the reagent. Out of the proactive area is used to derive the road. ZRP [3] is a hybrid routing protocols known for this category.

Each category has different strengths and weak. The proactive routing consumption of bandwidth due to the regular exchange of packets for the regular updating of the routing table. As against the problem of reactive protocols is latency, due to the discovery route to each request.

B. Routing Data

To understand the attacks in Ad hoc networks can be said that each node wants to send a message checks for the destination in its routing table. If it does not exist, it starts the route discovery process is broadcast on the network a route request message type. When an intermediate node receives this packet, and it is not also the recipient and the destination is not present on the table it in turn generates a road type of packet request containing its identifier.

In the event that the route to the destination is present in the routing table, a route reply message type is returned to the source indicating the way. Figure 1 shows the route discovery process.

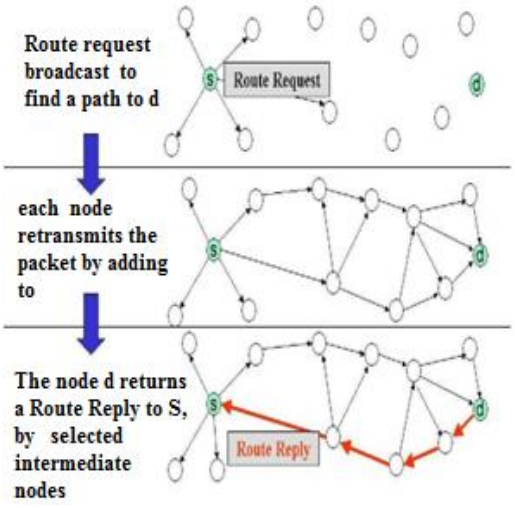


Fig. 1. route discovery process

When receiving the request reply packet from the source node, an update is made to its routing table to find out list of intermediate nodes to the destination and the associated cost. The cost is to choose between two routes to the same destination.

C. ZRP

The Zone Routing Protocol or ZRP [3], combines the advantages of both proactive and reactive approaches in a hybrid plan, taking advantage of proactive discovery in the local vicinity of a node, and using a reagent protocol for communication between the zones.

ZRP is proposed with the aim of reducing checks messages for proactive protocols and latency for reactive protocols. It is suitable for networks with a wide range and diverse patterns of mobility. For each node a routing area is defined separately. In the routing area, routes are available immediately, but outside the zone ZRP uses the route discovery process.

ZRP in each routing area comprises nodes that are a distance of max n jumps of reference node. There are two types of nodes for a routing area in ZRP [10]:

- Peripheral nodes
- Interior nodes

The nodes whose distance from a central node is less than the radius of an area are internal nodes while the node in the distance is exactly equal to the radius ρ are peripheral nodes. In Fig. 2, peripheral nodes E, F, G, K, M and Interior nodes B, C, D, H, I, J. The node is outside the node routing area A.

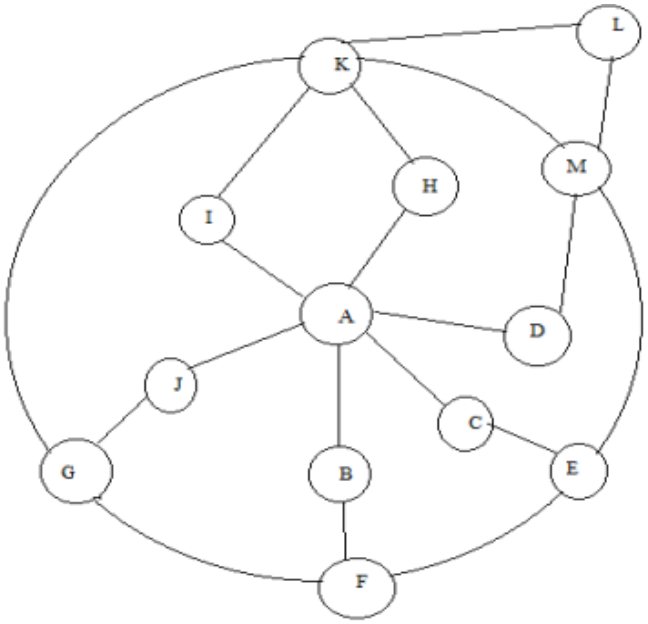


Fig. 2. Node Routing Area A with a radius = 2 jump

The source node sends a route request to the device nodes of its zone.la Route request contains the source address, destination address and a unique sequence number. Each device node checks if the destination is in its local area. If the destination is not a member of this local area, the boundary node adds its own address to the route request packet and sends it to its own device nodes.

If the destination is a member of the local area, it sends a response on the reverse path to the source. The source node uses the path recorded in the response packet to send data packets to the destination.

By adjusting the node transmission power, the number of nodes in a routing area can be controlled. Lowering the power reduces the number of nodes whose direct reach and vice versa. [10] ZRP uses both proactive and reactive routing strategies. In a routing area, the proactive strategy is used, while the reagent is used between the zones. ZRP refers to intra-zone Proactive Routing Protocol in local routing (IARP). The reactive routing is called inter-zoneRouting Protocol [12]. Its architecture is shown in Fig 3. IARP maintains nodes routing information existing in the node a routing area. The

discovery and maintenance of road is offered by IERP. If the topology of the local area is known, IERP can be used to reduce traffic.

Instead of broadcasting a package, ZRP uses the concept of broadcasting. [10] The broadcasting service is provided by the broadcasting Resolution Protocol (BRP).

BRP [11] uses an extended routing map provided by IARP, to build broadcast trees through which request packets are directed.

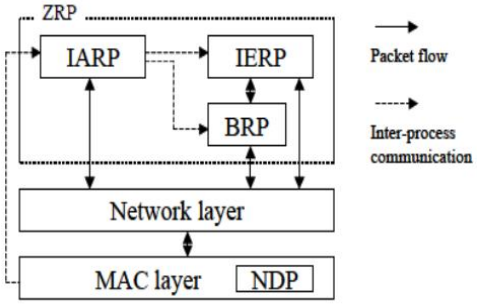


Fig. 3. Architecture ZRP [11]

III. Classification of Attacks

Before Routing protocols are exposed to various attacks that impact that this differs. Some attacks can cause the shutdown of a node by consuming their energies. Other attacks lead to a connectivity outage which influence on the packet rate issued and the time from start to finish.

Attacks in Ad hoc networks can be classified according to several criteria, such as the intelligence of the attack, its objective, the location of the attacker node, the impact of the attack on the network, etc...

- Impact of the attack: an attack can have a passive impact is to say that there's a network traffic analysis, surveillance of communications without modification of data or network operation and also without no injection of information in the network, all this in order to use this information in other attacks, such as the collection of passwords. It can be inferred that the main objective of such an attack is to know and understand how the nodes communicate with each other, and how they come together in the network. This attack is known as the "sniffing attack". [6] Another type of impact, called active, is a result of active attacks. This type of attack requires an injection of information in the network, or interacts with other nodes. Among active attacks include the attack "sleep deprivation" [7], which is to work the target in order to exhaust its battery.
- The objective of the attack: the target of the attack to a direct relationship with the type of striker. There are two types of attacker: the rational and the irrational. The first type of striker prepares his attack in order to take a direct or indirect benefit of the results of the attack. However, the objective of the second type of attack is to disrupt the proper functioning of the network. These attacks can be distinguished attack "jamming" [5]
- The intelligence of the attack: This type of attack is based on one or more layers of the OSI model. There are several types of attacks that are either of the attacks based on network layer attacks that exploit the failure of routing algorithms. The attack black hole (black Hole) is an attack that offers a shorter wrong path [8] it is based on the network layer.
- Location of the attack: the location of the attack is a very important parameter. An attack can be launched depending on the target location in the network. For example, a node that has a strategic location that provides network connectivity can be a target for an attacker seeking to isolate the network is to switch it off.

A. The BlackHole attack

An ad hoc network is susceptible to many security attacks. The blackhole is among the most known attacks. It is defined simple but effective, an attack that is based on the insertion of a malicious node having the capacity to take the identity of valid nodes on an ad hoc network since there is no physical barrier. This insertion leads to disturbances in the network and that due to the participation of all the nodes in the routing.

During this attack a malicious node exploits the vulnerability and claims to have the most reliable path to the destination. The source node takes one consideration that path is sending data to the malicious node which leads to loss of data. The main aim of such attack is to drape the packages, and to break the communication between nodes is diverting traffic to a non-existent node. . Fig 4 describes a blackhole type attack.

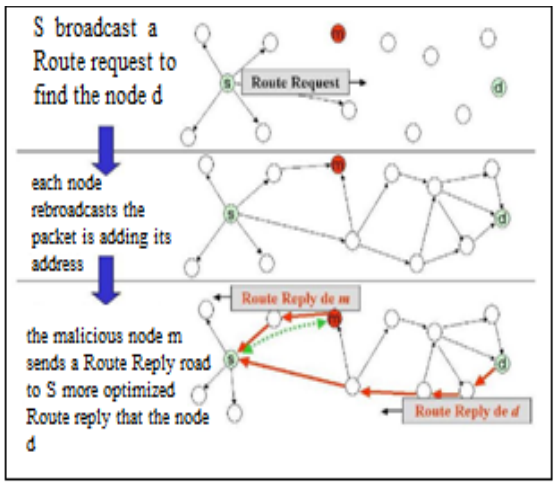


Fig. 4. Attack blackhole

When a source node wants to send data to a destination, it launches the route discovery mechanism is sending a RREQ message type. When receiving such a message by a malicious node, it responds immediately by sending a fake RREP post where he mentions he has the correct path to the destination requested with high sequence number. After receiving such a message by the source, it stops the process of discovery and ignores other RREP messages and begins sending packets to the malicious node. In turn it absorbs all the packets from

other nodes and thus the source node is attacked and its data are lost.

B. The Wormhole attack

This attack is based on two strikers who are interconnected via a link known as the tunnel. The first node in a striker this side of the network, receives packets from a legal node, the encapsulated then transmit using the tunnel to the second malicious node located in the other side of the network. The striker said node having the shortest route to the destination with the objective that it becomes the relay node. Fig 5 shows an example of a Wormhole attack [4], where two malicious nodes A and B that communicate through a tunnel which can be wired or wireless types. In this figure the nodes 3 and 7 respectively represent the source and destination. When the source wants to send given to the destination that is to say the node to node 3 and 7 in the absence of malicious nodes will be the path taken with a number 3-2-6-5-7 jump equal to 3.

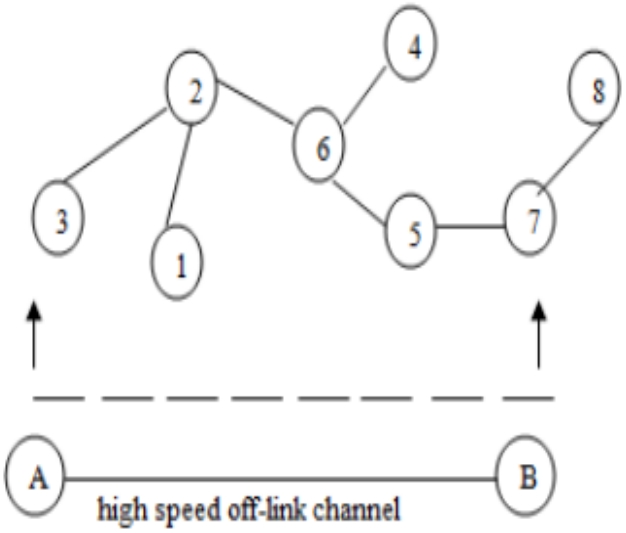


Fig. 5. Example Wormhole attack

In the case of presence of a Wormhole attack, the two nodes A and B will be activated where the transmissions take place between 3 and 7 via both malicious nodes A and B is using the Wormhole tunnel.

C. The attack RUSHING

In a type of attack Rushing [9] the malicious node responds as quickly as possible on RREQ type messages with the aim that the road through either retained him. If the proposed path is chosen, it will be to absorb all or part of the packets passing through it. Due to the high transmission speed, packets sent by the attacker will reach the destination first, pushing the source accepted her packages and throw the others. This way the attacker can easily access the communication between the transmitter and receiver.

D. Location disclosure

In the case of location disclosure attack, the malicious node role for collecting information on the location of the nodes, the set of paths and nodes involved and also other information useful on the network.

IV. Security in Ad Hoc Networks

The safety requirements for Ad hoc networks are almost identical for the wired or wireless networks with infrastructure. The security services are based on three concepts: authentication, confidentiality, data integrity and non-repudiation of users.

A. Authentication

The first concept is that authentication controls the identification of a node or entity in the network. Authentication ensures control of access to network resources. With the lack of authentication, malicious nodes can easily assume the identity of another with the aim to attack or take the privileges assigned to that node.

B. Confidentiality

Confidentiality ensures protection of information against threats that may lead to the disclosure of information. Confidentiality ensures private communication between nodes; is based on encryption. Encryption that can be applied to different levels of protocol layers. Encryption algorithms require encryption keys before sending it to the destination. However at the destination one must have the decryption key to decrypt the message.

C. Integrity

Integrity ensures protection against the traffic without prior authorization modification during transmission. Arguably, it is made to secure the system against threats that can cause change in the configuration of the system or data. This concept can be applied in a indirect way with protocols that confidentiality or authentication.

D. Nonrepudiation

Non-repudiation is made to ensure the identity of the sender and receiver. The non-repudiation of the issuer proves that the data was sent, and the non-repudiation of the receiver verifies and confirms receipt. This concept is reached on using the technology of the digital certificate.

V. Simulation of Blackhole Attack on ZRP

A. Simulation environment

In this part a study has been made on the impact of blackhole attack on the ZRP hybrid routing protocol, the NS 2.33are chose for simulation. The attacker is known in advance and simulation parameters are shown in Table 1.

Two performances are evaluated in order to infer the influence of the attack on the ZRP protocol namely the packet rate issued and the time from start to finish.

The mobility scenario is one generated using the random way point method, a method that generates a scenario in a random manner ie speed and nail mobility.

To implement the attack on NS 2 changes are made at the source code of the ZRP protocol in order to generate the new clone ZRP Protocol integrating the attack. This new protocol will be used by the attacker node while other nodes use the standard protocol ZRP.

TABLE I. SIMULATION PARAMETER

Parameter	Value
Nbr. Sources node	5, 20, 30
mobility	Absent
routing protocol	ZRP
Simulation time	200s
Packet size	512
Traffic	CBR
Network size	1000 X 1000
Total of node	50

B. Scenarios Simulations

To assess the impact of the attack on the black hole routing protocol ZRP different scenarios have been proposed:

- 1st scenario: In this first scenario simulation, all nodes using ZRP as the communication protocol are fixed, including the attacker node.
- 2nd scenario: In the second scenario, the fixed mobility is kept for all nodes; and increasing the number of node addressing two.
- 3rd scenario: In this third scenario simulation all nodes using the ZRP routing protocol for communication are mobile except attacking node.
- 4th scenario: The simulation in the fourth and final scenario simulation is the same as the third, it is made with mobile nodes except instead of an attacker node using two nodes.

This after a discussion of the results obtained in the simulation of the attack on the black hole ZRP protocol, checking the two parameters ie the rate of packets delivered and the time from start to finish. The results are as graphs and four scenarios are used to test the performance.

1) packet rate issued

The results obtained from the simulation of the attack black hole on hybrid routing protocol ZRP we see the influence of the attack.

Fig 6 illustrates the variation of the lost packet rate based on the number of source nodes, and also in different scenarios. Based on the results we see that the attack Black Hole has an impact on the ZRP protocol considered especially in cases where the number of source nodes is high. Also the rate of delivered packets decreases from the fixed case we note that in the case of mobility, which is logical as mobility increases the rate of lost packets according to the results previously obtained.

The reduction of packages delivered in the mobile case rate is not 100% on the attack, but also the mobility that has a significant impact on this metric.

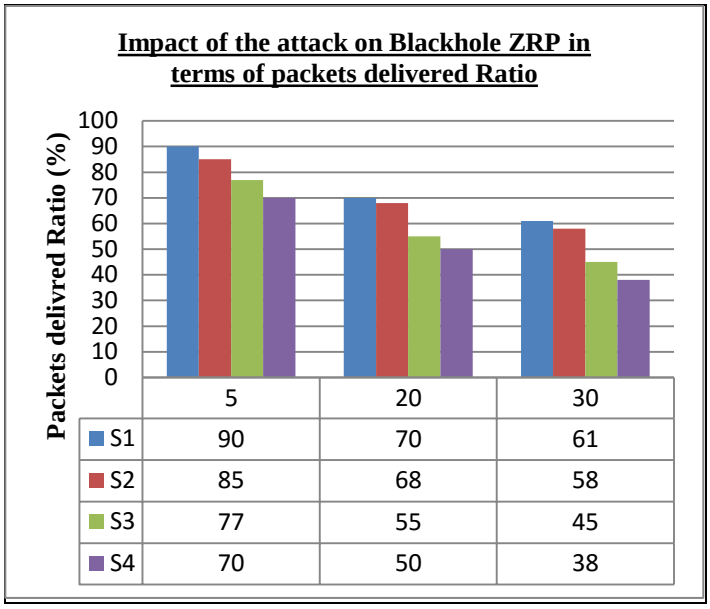


Fig. 6. Variation packets delivered ratio

2) End to End delay

Fig 7 shows the time from start to finish in different scenarios depending knew many nodes sources. From the results obtained it can be inferred that the attack has an effect on this metric especially in the case where the number of source nodes is as high in the presence of mobility.

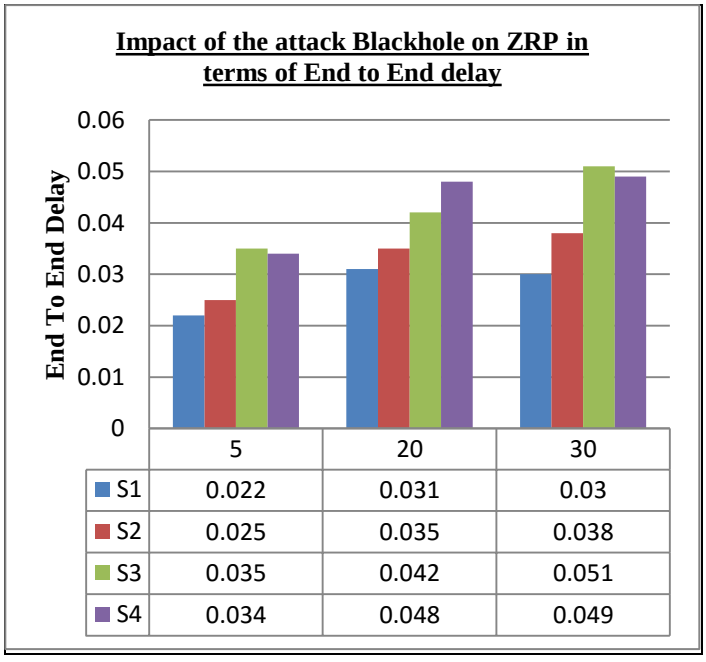


Fig. 7. Variation End to End Delay

VI. CONCLUSION

Ad-hoc networks are characterized by the absence of infrastructure, also by devices with limited capabilities in terms of calculated and energy. The lack of infrastructure is considered a strong point for this type of facility since the implementation network in an environment with minimal cost.

Each node in its network can simultaneously be a capture unit as a routing device, all this makes them vulnerable to a Manet set of security attacks; attacks that can be active or passive and influence on the confidentiality, integrity and availability of data.

These attacks found the attack Black Hole (Black Hole); a powerful attack that influence on Ad hoc networks. This attack can cause a complete network failure is absorbing the traffic as it can isolate part.

In this study we investigated the impact of the attack black hole on hybrid ZRP protocol, for we have created a clone of the protocol where we implemented the attack, the new protocol will be called by the attacker in order drape traffic.

According to the results we see that the attack has an impact on the protocol is in the fixed or mobile network case. As the rate of packets delivered decreases with increasing the number of source nodes; one can also deduce that the high number of packets lost in the case of mobility is not at 100% of the attack but also because of the mobility of the network.

For the second metric (time from start to finish), he was also influenced by the attack and in the same time by mobility, which makes sense from the results found previously.

To conclude, in such an attack traffic is diverted to a specific station or the malicious node influence on the whole of the network which induces to the injury of the MANET. The detection of such a nodes is difficult in this type of network.

REFERENCES

[1] Institute of Electrical and Electronics Engineers. IEEE Std 802.15.1-2005, Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifi-cations for Wireless Personal Area Networks (WPANs), 14 June 2005. URLhttp://standards.ieee.org/getieee802/download/802.15.1-2005.pdf

[2] Institute of Electrical and Electronics Engineers. IEEE Std 802.15.4-2006, Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifi-cations for Low-Rate Wireless Personal Area Networks (WPANs), 8 Septem-ber 2006.

[3] IEEE Std. 802.11a, "High-Speed Physical Layer in the 5 GHz Band," 2000.

[4] S. Upadhyay and B. K. Chaurasia, "Impact of wormhole attacks on MANETs", International Journal of computer science & Emerging Technologies (E-ISSN: 2044-6004) vol. 2, no. 1, (2011) February.

[5] Y. Z. T. W. W. Xu, W. Trappe, 2005. The feasibility of launching and detecting jamming attacks in wireless networks. Dans les actes de ACM international symposium on Mobile ad hoc Networking and Computing, 46–57.

[6] K. K. Z. Trabelsi, H. Rahmani et M. Frikha, 2004. Malicious sniffing systems detection platform. Dans les actes de International Symposium on Applications and the Internet, 201–207.

[7] V. N. P. M. M. K. M. Pirretti, S. Zhu et R. Brooks, 2005. The Sleep Deprivation Attack in Sensor Networks : Analysis and Methods of Defense. International Journal of Distributed Sensor Networks 2-3, 267–287.

[8] S. P. M. Al-Shurman, S. M. Yoo, 2004. Black hole attack in mobile Ad Hoc networks. Dans les actes de 42nd annual Southeast regional conference, 96–97.

[9] Y.C. Hu, A. Perrig, and D.B. Johnson. Rushing attacks and defense in wireless ad hoc network routing protocols. In Proc. of the 2nd ACM workshop on Wireless security, page 40. ACM, 2003.

[10] G. Barrenetxea, F. Ingelrest, G. Schaefer, and M. Vetterli, "SensorScope: Out-of-the-Box Environmental Monitoring," in ACM/IEEE IPSN , 2008 Baydere, S.,Safkan, Y., and Durmaz, O. 2005.

[11] Lifetime Analysis of Reliable Wireless Sensor Networks. IEICE Transactions on Communications E88-B, 6, 2465-2472

[12] L. Selavo, A. Wood, Q. Cao, T. Sookoor, H. Liu, A. Srinivasan, Y. Wu, W. Kang, J. Stankovic, D. Young, and J. Porter, "LUSTER: Wireless Sensor Network for Environmental Research," in ACM SenSys , 2007.

Design of Modulator and Demodulator for a 863-870 MHz BFSK Tranceiver

A.Neifar, G. Bouzid, and M. Masmoudi
Micro Electro Thermal Systems (METS) Research Group, Tunisia
University of Sfax, National Engineers School of Sfax

***Abstract*—This paper presents the design of low power modulator and demodulator circuits dedicated to a BFSK transceiver, operating in the 863- 870 MHZ ISM band. The two circuits were designed using ams 0.35µm technology with 3V dc voltage supply. Simulation results of the new Direct Digital Frequency Synthesizer in the modulation have shown good performances of the designed system as the Spurious Free Dynamic Range SFDR reached -88 dBc while the circuit consumes only 47.7 µW @ 43.3MHz. The demodulator has also presented a good BER of 10^{-3} @10.9 EbtoN0 and a sensitivity of about -115 dBm.**

Keywords—ISM band; FHSS; FSK modulator; BFSK demodulator; wireless sensor network

I. Introduction

The low-power market has experienced explosive growth over the past ten years by the presence of new wireless command and control technologies. This expansion is provided by the technology of wireless sensor networks of low range, which has many applications including home automation, industrial and commercial automation, peripherals for personal computers but also medical survey and health care monitoring. Actually, the requirement for vital sign monitoring has significantly increased as population aging is rapidly progressing in many industrialized countries. This grow-up is accompanied by an even more dramatic increase in the number of old people suffering from chronic diseases and disabilities as specified in [1].

Thus, several standards have been studying the implementation of wireless sensor network, such as the ultra wideband UWB (IEEE 802.15.3) standard [2], Bluetooth (IEEE 802.15.1), but mostly the standard Zigbee (IEEE 802.15.4) that is dedicated to the wireless networks of the family WPAN LR (Low Rate Wireless Personal Area Network) [3]. Therefore we propose to design a wireless sensor that will be integrated in a wireless sensor networks used for vital sign monitoring using the Zigbee protocol. The RF transceiver will operate in the 863-870 MHz ISM band, as it is available only in Europe, so presents a good field for testing new concepts in order to develop low power transceiver for short range and low data-rate applications.

As the direct conversion transceiver has shown interesting specifications like low power consumption and low manufacturing costs [4], this architecture is used for the wireless sensor.

In this paper, the design of two blocks in the transceiver is presented. A new method for FSK (Frequency Shift Keying) modulation is presented using binary scheme modulator and transistor level of a zero crossing demodulator is then realized using 0.35µm technology. The paper is organized as follows: section II describes the FSK transceiver architecture, section III presents the design and implementation of the digital modulator while section IV details the design of the digital demodulator. Finally simulation results are presented in section V and a conclusion and perspectives are given in section VI.

II. Tranceiver Architecture

The Frequency Shift Keying (BFSK) transceiver is presented in Fig.1. The BFSK modulator uses a Frequency Hopping Spread Spectrum (FHSS) technique and a Direct Digital Frequency Synthesizer (DDFS), which allows the generation of BFSK signal using hopping frequencies. Hence the digital data is synthesized in quadrature outputs signals in base band and up converted to the center frequency of the ISM band using a single-sideband up conversion mixer controlled by a local oscillator that selects either the lower or the upper sideband for the instantaneous carrier frequency.

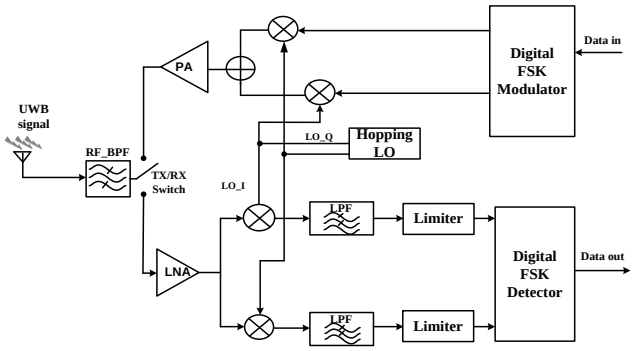


Fig. 1. Transmitter architecture

The transmitted data is carried by symbol tones at an offset of ±20 kHz from the carrier frequency, so a maximum data rate of 20 kb/s is achieved by sending two symbols per hop. Received by the antenna, the signals concourse a bandpass filter that selects the ISM band, then a low noise amplifier (LNA) which provides enough power for the signals to spread with a minimum noise.

As the ISM Band is composed of several channels, the operating one will be selected using a low pass filter followed by a limiting amplifier to convert the received signal into binary level [5], instead of using a linear and complicated

automatic gain control, as in most FM receivers. Finally the correct transmitted data will be at the output of a digital FSK Detector.

III. Design of the Digital FSK Modulator

A. Modulator Specifications

In order to select the appropriate Data rate, it is necessary to estimate the overall average power consumption of a transmitter node in the sensor networks. A thorough study in [6] has given an approximated formula for the power dissipated by the transmitter as a function of the data rate. The result is in Fig.2 and show that a data rate of 20 kb/s is selected taking into account the requirements of lower power, a BFSK tone frequency to avoid the impact of DC offset and flicker noise caused by direct conversion architecture and bandwidth efficiency.

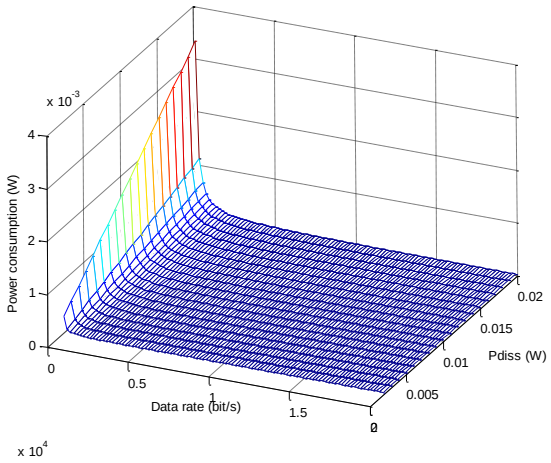


Fig. 2. Transmitter power consumption as a function of data rate and node power dissipation

As the ETSI (European Telecommunications Standards Institute) regulations require a minimum separation of 25 KHz bandwidth between two adjacent channels, and in order to follow the requirement of FHSS technology, the ISM band 863-870 MHz was divided into 58 channels with a transmission bandwidth of 80 KHz for the each transmitted signals and a separation of 40 KHz between adjacent channels. The power spectrum of a BFSK signal is reported in Fig. 3.

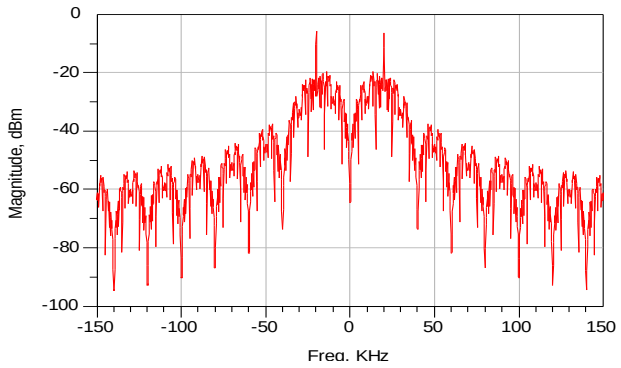


Fig. 3. Power spectrum of a BFSK signal (bit rate=20Kbps)

B. Modulator Implementation

The proposed architecture of the modulator circuit is presented in Fig.4. It is composed of a DDFS, a PN code generator and a multiplexer. The role of the DDFS is to produce digital samples from baseband sinusoidal waveforms by addressing a sine ROM (Read Only Memory) at a frequency set by a 20-bit control sequence. The PN code however, generates a random code corresponding to the hopping pattern and among which, the multiplexer will select a single code [7].

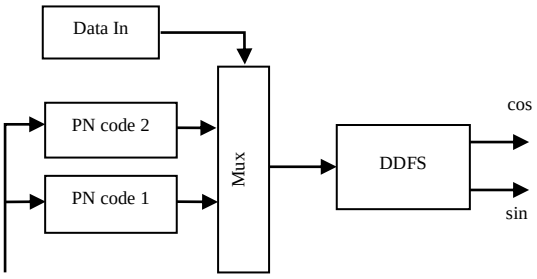


Fig. 4. Block diagram of the modulator

A minimum clock speed Fclk of 43.4 MHz was used in this modulator system, thus a smallest frequency resolution of about 41.29 Hz is obtained since the frequency control word is 20 bits fixing the frequency control resolution Fr to 41.29 Hz as:

$$F_r = \frac{F_{clk}}{2^N} \quad (1)$$

Where F_{clk} is the sampling frequency of the DDFS/DAC and N is the number of frequency control bits.

Fig.5 shows a typical architecture of a DDFS system, it is mainly composed of a phase accumulator, a sine/cosine generator and a ROM.

At each edge of the clock the PN code generates a word binary N that is used to increment the phase accumulator in the DDFS, and every N binary word will be carried in a ROM memory, as shown by the VHDL test simulator in Fig.6.

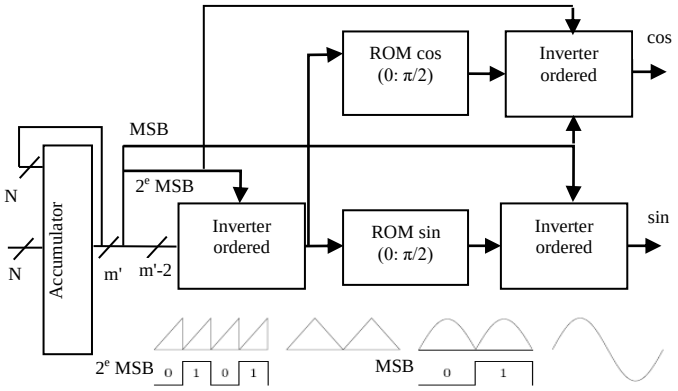


Fig. 5. Architecture of DDFS system

The length of the internal word also ensures a spurious tone of at least -72.6 dBc from the fundament frequency freeing therefore the imperfections in the DDFS [8].

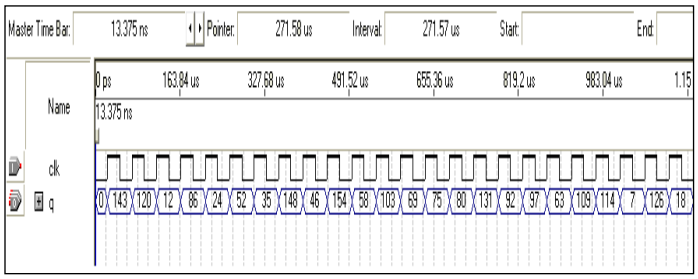


Fig. 6. PN code Simulation results

The bloc diagram of the phase accumulator is depicted in Fig.7. Generally this circuit is pipelined as m stages of L bits each as: m×L=N [7].

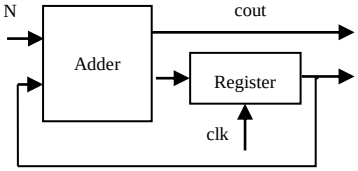


Fig. 7. Block diagram of the phase accumulator

IV. Design of the Digital FSK Demodulator

The digital demodulation is realized using a limiter that converts the received analog signals into binary levels and a zero crossing BFSK detector.

A. Design of the limiting amplifier

The limiting amplifier has a cascaded architecture. Indeed, the number of cells used, the amplification gain, the bandwidth and the consumption of each cell determine the overall performance of the circuit [9].

In order to determine the number of stages that may be used for this application, simulations of the total gain GT and bandwidth BT variations were carried (Fig.8) considering a number Z of identical cells for the whole limiter [10] as:

$$\text{Gain of each cell:} \quad G_C = G_T(^1/_Z) - 1 \tag{2}$$

$$\text{Bandwidth of each cell:} \quad B_C = \frac{1}{\sqrt{2^{1/Z}-1}} \tag{3}$$

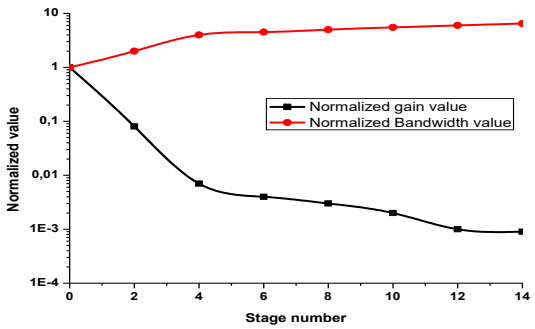


Fig. 8. Variation of gain and bandwidth based on the number of cells

Also the total power consumption of the circuit depends on the number of stages used as:

$$P_T = Z \times (G_C \times B_C)^2 \tag{4}$$

Thus, the number of limiting amplifiers has to be reduced so to meet the requirements of low consumption for the intended medical application. A compromise between the different specifications has led to the use of 7 floors of limiting amplifiers which a cell is shown in fig.9.

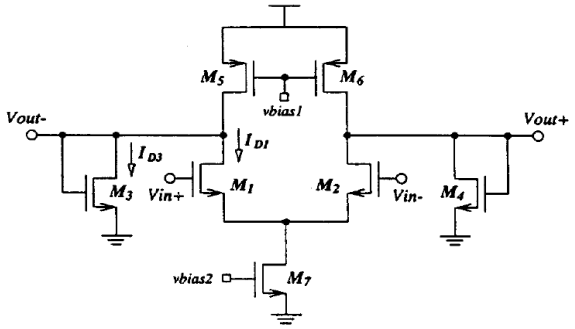


Fig. 9. Proposed limiting amplifier

The circuit is composed of a conventional simple source coupled pair [11] with a load diode biased with an independent current.

B. Design of the BFSK demodulator

The designed demodulator is shown in Fig.10. It is composed of four differentiators circuits which detect the zero crossing of I and Q signals at the output of the limiter, two OR gates and one NOR gate, a shape keeping circuit and a low pass filter.

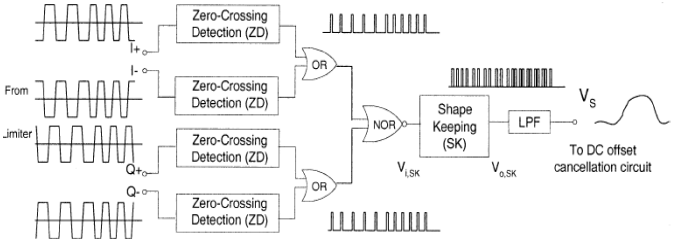


Fig. 10. Architecture of the proposed BFSK demodulator

The architecture of one ZD circuit is given in Fig.11. For every zero crossing of the input signal, the circuit generates a pulse whose width depends of the circuit component values as:

$$\tau = (R.C)Ln(2) \tag{5}$$

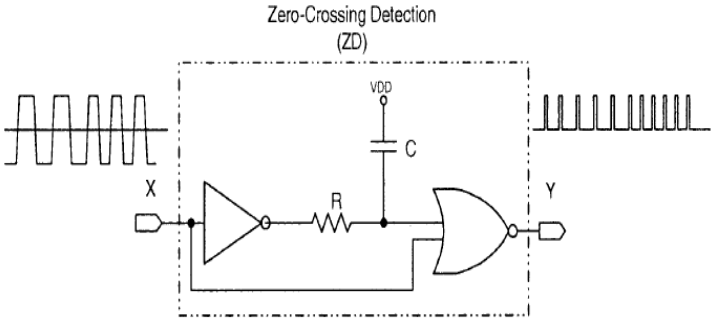


Fig. 11. Zero-crossing Detection

As the output generated pulses don't have the same width as shown at the output in Fig.11, after the collection with the logic OR and NOR gates, a shape keeping circuit (Fig.12) is utilized to fix the width of every input pulse whose value depends on the parameters of the circuit components. Indeed the width is mainly determined by the ratio of the R1-R3 voltage divider and the values of the capacitor C2 and the resistor R5 at the drain of the output transistor.

Finally a digital low pass filter is applied to the generated pulses so to filter errors and to calculate the mean value of the signals which allow determining whether it is a 0 or 1 transmitted Data. The cut-off frequency of the filter is equal to 80 KHz which is exactly the width of a single channel in the ISM band 863-870 MHz. The filter response is shown in Fig.13.

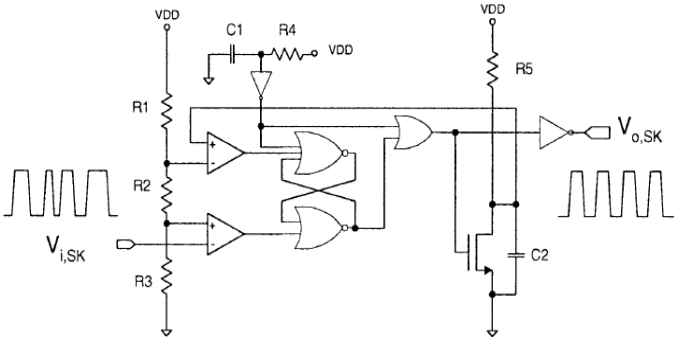


Fig. 12. The Shape keeping circuit

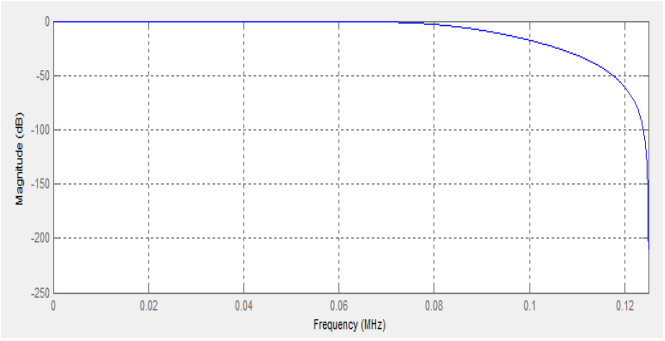


Fig. 13. Digital filter response

V. Simulation Result

The modulator simulation is first achieved with Modelsim in VHDL language, then the code was synthetized using Quartus and a chip of the whole cicuit was realized using Cadence. Fig.14 shows the simulation of VHDL maximum code of the modulator. At each edge of the clock (Clk), the PN code generates a word binary that increments the phase accumulator of the DDFS, and for each increment in this DDFS (clk2), the modulator outputs sample of sin and cos signals. Moreover, The multiplexer switches between the two PN code at each edge of the signal (s).

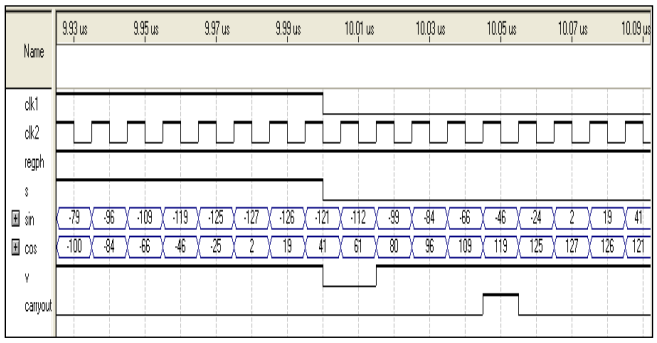


Fig. 14. Maximum code vhdl simulations

Fig.15 shows the output sine and cosine signals of the modulator. Indeed, a quadrature phase is observed between the two signals as if the sine is at its maximum, the cosine takes negative values and if it is at zero, the cosine takes its maximum.

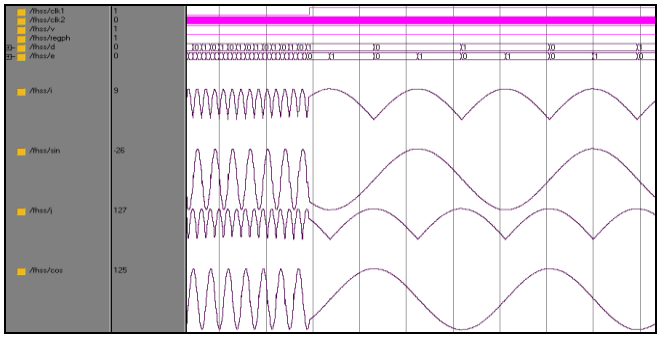


Fig. 15. Modulator Block simulation results

The chip implementation of the modulator circuit is shown in Fig.16. It was realized using ams 0.35µm CMOS standard cell library. The chip design was divided into thirteen subsystems and simulations results have shown that the average power consumed by the whole circuit is about 47.7 µW at Fclk=43.4 MHz and the Spurious-Free Dynamic Range is about -88 dBc which complies with the specifications of the ISM band.

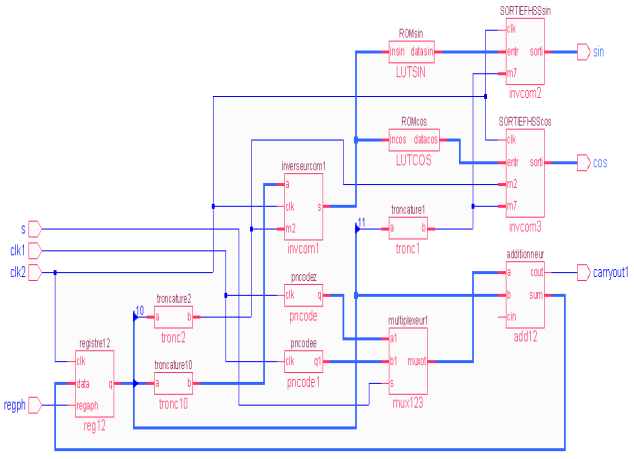


Fig. 16. Chip design

The BFSK signals at the output of the modulator were applied to the designed demodulator using an Additive White Gaussian Noise (AWGN) channel. Thus, the diagram of Fig.16 was obtained including the input data of the modulator (in blue), the output of the filter at the end of the demodulator (in red) and the data out (in pink) after integration and dump of the signals.

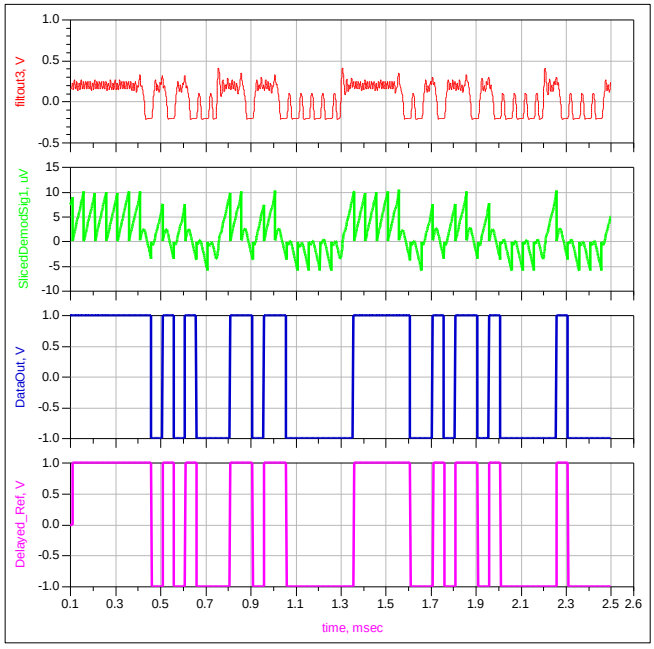


Fig. 17. Output signals of the demodulator

Finally a bit error rate simulation BER was achieved as a function of EbtoN0 expressing the signal to noise ratio performance. The result is in Fig.17 and shows that for a bit error rate of 10^{-3}, only 10.9 dB of the EbtoN0 is needed for the circuit which deeply satisfies the requirement of the application and the FSK modulation. The Demodulator circuit presents a sensitivity of about -115 dBm.

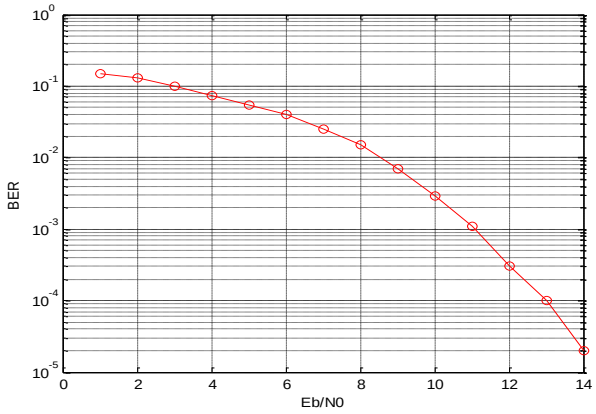


Fig. 18. BER plot versus EbtoNo(dB)

Performance comparison between the proposed modulator and demodulator circuits and other reported works in the literature are presented in table I and table II respectively.

TABLE I. PERFORMANCE COMPARISON OF THIS WORK (MODULATOR) AND REPORTED WORKS IN LITERATURE

	[12]	[13]	[14]	This work
Output (bits)	12	12	10	10
Fclk (MHz)	480	150	500	43.4
SFDR (dBc)	-80	-84	-70	-88
Power dissipation (µW/MHz)	72	500	34.4	47.7
Supply voltage(V)	2.5	1.8	1.8	3

TABLE II. PERFORMANCE COMPARISON OF THIS WORK (DEMODULATOR) AND REPORTED WORKS IN LITERATURE

	[15]	[16]	This wok
EbtoN0 (dB)	10.5	16	10.9
Sensitivity (dBm)	-114.5	-109	-115

VI. CONCLUSION

In this paper, a novel design and implementation of FHSS-FSK modulator and demodulator for a 863-870 MHz receiver was presented. First, we have presented a new method for designing the modulator exploiting the symmetry of trigonometric functions (sine and cosine) in order to reduce the spurious tones of the Direct Digital Frequency Synthetizer. Thus, a new architecture of the DDFS with small lookup table for the sine and cosine functions and pipelined phase accumulators was put into test. Simulation results showed that the designed modulator was able to generate BFSK signals with frequency hopping while consuming only 47.7 µV at 43.4 MHz.

A BFSK demodulator circuit was also designed using ams 0.35µm technology, the circuit was tested with input signals coming from the designed modulator bloc. Simulation results showed that the circuit presents 10^{-3} of bit error rate for an input signal to noise ratio of 10.9 DB while the sensitivity reached -115 dBm. Therefore, the designed circuits are suitable for FSK modulated applications as the health monitoring systems.

As a perspective of this work, we can achieve the design of the whole transceiver using the UWB technology and BPSK modulator as in [17], and compare the performances of the two designed works for a better use in this medical application.

REFERENCES

[1] Reijula, J. Using well-being technology in monitoring elderly people-a new service concept. Ph.D. Thesis, Helsinki Aalto University, Espoo, Finland, 2010.

[2] Imen Barraj · Hatem Trabelsi · Wenceslas Rahajandraibe ·Mohamed Masmoudi, "Modular baseband pulse generator for impulse-radio ultra-wideband transmitte", Electronics Letters, sep 2015.

[3] T. k. Nguyen, N. J. Oh, V. H. Le, and S.G. Lee, Member IEEE "A Low Power CMOS Direct Conversion Receiver With 3dB NF and 30KHz Flicker-Noise Corner for 915-MHz Band IEEE 802.15.4 ZigBee Standard," IEEE Trans. on Microwave Theory and Techniques 2006. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] Aarno Pärssinen, "Direct Conversion Receivers in Wide-band Systems", Springer Science business media, 2006.

[5] H. Trabelsi, Gh. Bouzid, F. Derbel and MohamedMasmoudi, "A863–870 MHz spread-spectrum FSK transceiver design forwireless sensor"

2008International Conference on Design &Technology of Integrated Systems in Nanoscale Era.

[6] H. Trabelsi, Gh. Bouzid, Y. Jaballi, L. Bouzid, F. Derbel and Mohamed Masmoudi "System level design of a low power, short range FHSS transceiver for Wireless sensor," 4eme International Conf. JTEA 06, May 2006, pp120–124.

[7] Gh. Bouzid, H.Trabelsi and Mohamed MAsmoudi, "FHSS-FSK Modulator design for wireless sensor transmitter", Trends in Applied Science Research 3 (5) 344:356, 2008.

[8] B. D. Yang, J. H. Choi, S. H. Han, L. S. Kim and H. K. Yu, "An 800-MHz Low-Power Direct Digital Frequency Synthesizer with an on-Chip D/A Converter," IEEE J. Solid-State Circuits, vol. 39, no. 5, pp. 761–774, May 2004.

[9] He Rui, Xu Jianfei , Yan Na; Ž , Sun Jie, Bian Liqian , and Min Hao, "Design and analysis of 20 Gb/s inductorless limiting amplifier in 65 nm CMOS technology", Journal of Semiconductors, Vol. 35, No. 10, 2014.

[10] Po-Chiun Huang, Yi-Huei Chen, and Chorng-Kuang Wang, "A 2-V 10.7-MHz CMOS Limiting Amplifier/RSSI" IEEE Journal of solid-state circuit, vol.35, No. 10,october 2000.

[11] Wenjun Sheng, , Bo Xia, , Ahmed E. Emira, , ChunyuXin, Ari Yakov Valero-Lopez, Sung Tae Moon, and Edgar Sanchez-Sinencio, "A 3-V, 0.35 μm CMOS.

[12] W.Akram, Jr. E.E Swartzlander, "Direct Digital Frequency Synthesis Using Piece-wise Polynomial Approximation,".in IEEE Int. Signals, Systems and Computers Conf. Paper, Vol 2, Nov. 2003, pp.2237-2241.BluetoothReceiver IC" IEEE 2003.

[13] H. Jafari, A. Ayatollahi and S. Mirzakuchaki, "A Low Power, High SFDR, ROM-Less Direct Digital Frequency Synthesizer," in IEEE Int. Microelectronics Conf. Paper, Vol, Dec. 2005, pp. 50–54.

[14] A. G. M. Strollo, D. De Caro and N. Petra, "A 630 MHz, 76 mW Direct Digital Frequency Synthesizer Using Enhanced ROM Compression Technique," IEEE J. Solid-State Circuits, Vol. 42, No. 2, pp.350–360, Feb. 2007.

[15] H.M. Eissa, Khaled Sharaf and H. Ragaie, "ARCTAN differentiated digital demodulator for FM/FSK digital receivers" The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002, Volume: 2.

[16] Hatem Trabelsi, Ghazi Bouzid, Faouzi Derbel and Mohamed Masmoudi, "Direct conversion transceiver design in the 863-870-Mhz band Application: Wireless sensor network", Proceedings of the 7th WSEAS International Conference on Circuits, Systems, Electronics, Control and Signal ¨rocessinf (CSECS'08), 2008, pp:69-73.

[17] Imen Barraj · Amel Neifar · Mohamed Masmoudi, "Three to Five Gigahertz UWB Transmitter for Vital Sign Monitoring Systems", BioNanoScience, Jun 2016 ·

Reducing the Calculations of Quality-Aware Web Services Composition Based on Parallel Skyline Service

Maryam Moradi
Department of Computer Engineering,
Yazd Branch, Islamic Azad University,
Yazd, Iran

Sima Emadi*
Department of Computer Engineering,
Yazd Branch, Islamic Azad University,
Yazd, Iran

***Abstract*—The perfect composition of atomic services to provide users with services through applying qualitative parameters is very important. As expected, web services with similar features lead to competition among the service providers. The key challenge to find an appropriate web service for composition occurs when multiple aspects of quality (such as response time, cost, etc.) are considered in the optimal composition of services. Skyline service provides the best service with consideration of qualitative parameters by using superior analysis. In this study, Skyline algorithm is used to find a set of best possible services compositions while taking into account qualitative parameters. The parallelism technique in this study, had significant impact on reducing response time and increasing the speed of service composition as well as reduction in composition calculations. The results of the analysis and evaluation of the proposed method shows optimum runtime and the best composition.**

Keywords*—service composition; parallel Skyline service; the dominant relationship; service quality

I. Introduction

In recent years, the application of web-based systems in institutions and government agencies is increasing. Introduction of web services is an effective approach in business structures to provide the required capabilities of service providers for services composition. Selection of the appropriate service on user's request is based on the service quality of services. Several different methods have been suggested to solve the problem of web services composition based on qualitative characteristics. These methods can be divided into two types of exact methods and approximate methods according to [2]. The first type is known as non-innovative methods which selects the best design from all available designs by examining and calculating the candidate's routes and thus provide a more precise answer. In the second type or innovative methods, contrary to the first type, an ideal design that is close to the best and most accurate answer will be chosen.

Due to the importance of optimal composition of web services in recent years, a lot of works have been done in the field of each method. By studying various types of innovative algorithm, one can concluded that many problems still exist to solve in web services composition based on qualitative characteristics. For example, each of these methods usually have local optimality problem alone; or in basic genetic algorithm, crossover and mutation operation acts randomly and without any guidance, which leads to degeneration of the method. Therefore, efforts to improve efficiency such as using combined methods, operators like revolution operator or adding functions to improve were performed. These techniques are provided for better speed, faster convergence, and higher efficiency in large spaces. Based on the mentioned studies, there is no specific benchmark tool for evaluating the algorithm. Although some researchers used different simulation environments or different data to compare them with each other, the results show that different methods have different disadvantages and they do not have any specific standard. Skyline algorithm method and parallelism technique are used in this proposed method in order to provide the best composition with regard to the shortest response time in high scalability.

The paper is as follows; the second section expresses the concept of Skyline service. In the third section, related works in the field of service composition by Skyline service will be discussed. The proposed method is described in the fourth section. And the fifth section analysis and evaluation and the results of the proposed method are presented and in the final section, conclusions and recommendations for the future studies are presented.

II. Skyline Service Algorithm

Skyline service includes a set of services that is not dominated by other services. In this regard, the concepts required are as follows:

A. Dominance Service and Skyline Service

In service composition, services that are better in all aspects of service quality compared to other services are dominance service.

For example, assume web service composition of travel assistance: This composition, for the user, provides travel assistance package. Examples of needed services include: travel planner, maps and weather forecasting. Weather conditions during travel are an important factor for weather service. A user enjoys a number of choices for each of these services. As noted, here limitations with similar capability arises which are discussed later. In Table 1, five maps

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

providers and four travel planner service providers are presented [2].

TABLE I. FIVE MAPS SERVICE PROVIDERS AND FOUR TRAVEL PLANNER SERVICE PROVIDERS [2]

Sid	Operation	Latency	Fee	Reputation
Map Providers				
A	Geo code	0.5	0.8	2
	Get map	1	0	2
B	Geo code	0.7	0.3	3
	Get map	2	0.5	3
C	Geo code	0.5	0.2	2
	Get map	1.5	0.9	2
D	Geo code	0.3	0.7	2
	Get map	1	0.4	2
E	Geo code	0.6	0.7	3
	Get map	0.8	0.5	3
Trip Planner Providers				
T	Search Trip	2	0	2
	Get Trip	2	0.8	2
G	Search Trip	1	0	3
	Get Trip	2	1	1
H	Search Trip	2	1	2
	Get Trip	3	1	2
I	Search Trip	3	1	3
	Get Trip	2	0	2

It should be noted that sometimes to request a web service, it requires to call a number of services. For example, to request maps provider, it requires two functions of geo code and get map. In addition, it is possible that dependency constraints exist between these two functions (like get map which is dependent to geo code). Therefore, these functions can be arranged in a sequence of dependency constraints; the production process of Skyline service composition by quality values of any services in Table 1 are described. According to Table 1, for calculation of quality for each service, records of the service must be combined. The results are shown in Table 2. Calculation of qualitative parameters is according to the equation (1):

Service name: (∑ Operation, ∑ Latency, ∑ Fee, ∑ Reputation) [2] *(1)*

TABLE II. SUMS OF EACH QUALITATIVE PARAMETER [2]

Sid	Latency	Fee	Reputation
Map Providers			
A	1.5	0.8	2
B	2.7	0.8	3
C	2	1.1	2
D	1.3	1.1	2
E	1.4	1.2	3
Trip Planner Providers			
F	4	0.8	2
G	3	1	2
H	5	2	2
I	5	1	3

By using Dominance Analysis, the best services from each group are chosen. As in each group (Map Provider: A, B, C, D, E) and (Trip Planner Provider: F, G, H, I), the lowest numerical value is selected as dominance or superior service. In the first group, A and D and in second group F and G, are selected as the dominance services. To understand better, the dominance service selection process for map provider is described. As noted, the lowest value is selected as the best services to combine that have been selected at the first stage of existing services.

TABLE III. SELECTED NUMERICAL VALUES AT THE FIRST STAGE [2]

Sid	Latency	Fee	Reputation
Map Providers			
A	1.5	0.8	2
D	1.3	1.1	2
E	1.4	1.2	3

Then, according to the second column of qualitative parameters in Table 3, the best service is selected as the dominance service. Service selection process carried out in the second group is likewise.

TABLE IV. DOMINANCE SERVICES IN THE SECOND GROUP [2]

Sid	Latency	Fee	Reputation
Map Providers			
A	1.5	0.8	2
D	1.3	1.1	2

Selected services are the best ones among available services that ultimately will be combined. All possible compositions are shown in Table 5.

TABLE V. SKYLINE IMPLEMENTATION OF SERVICE COMPOSITION [2]

C-Sky	Latency	Fee	Reputation
C-Sky (A,F)	5.5	1.6	4
C-Sky (A,G)	4.5	1.8	4
C-Sky (D,F)	5.3	1.9	4
C-Sky (D,G)	4.3	2.1	4

B. Double Progressive Algorithm

The origin of this algorithm is progressive counting principle. In this principle. After making the root node which is the parent node, latter nodes are built upon it. The rule to create each node is that selected services available in composition are different only in one service with its child nodes. For example, the root node in Figure 1 is a1b1c1 its child nodes are (a1b2c1) (a1b1c2) (a2b1c1). The number of services composition produced of Skyline are calculated in form of |A|*|B|*|C|=27.

In this algorithm, all services of Skyline are stored in the memory. General space of services composition is proliferated based on the principle of progressive counting and using the lattice.

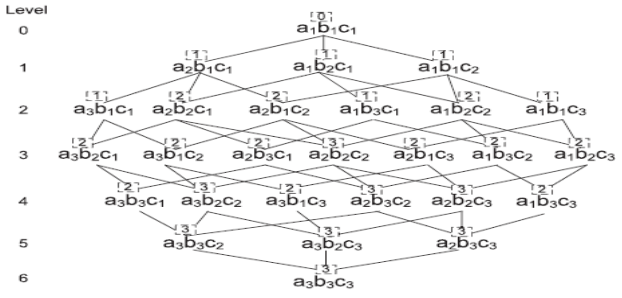


Fig. 1. Expansion lattice [3]

In Double Progressive, the algorithm sorts services composition in ascending order within an expansion lattice; as well as showing services composition of Skyline progressively [3].

It is noteworthy that in expansion lattice, a service composition that Ni (No. of composition) has higher ranking than its ancestor does not mean that Ni is dominated by its ancestor; Therefore, expansion lattice only determines the counting order between among and proves that Ni is counted after its ancestor. But for nodes that do not have parent-child relationship, such as: (a2, b1, c1) at level 1 and (a1, b3, c1) at level 2, a correct order must be ensured, since it is possible (a1, b3, c1) has less score than (a2, b1, c1); Hence, it should be considered earlier. To achieve the progressive counting, expansion lattice (T) with a heap (H) is used. Expansion lattice ensures that the parent node is considered before the child node, which is very desirable; because the score of a parent node cannot be greater than the score of child nodes. On the other hand, the heap determines the order of nodes without parent-child. The process of making compositions from the first level to the heap (H) begins with Composition Service Execution Plan (CSEP) 1. Each stage of the counting will have two sub-phases: 1. Extraction: a service composition with the lowest Ni rank, is extracted from the heap and compared with the existing Skyline. Finally, Ni will be entered into Skyline if it is not dominated and omitted [3].

C. Parent table

In progressive algorithm a node alone can be produced several times from the other parent nodes, where proliferation problem occurs because a node can have m parents, where m is number of Skylines' service, and similar child nodes can be produced when each of ancestors develop. As shown in Figure 1, number of every node shows its parent node number (index). For example, the nodes (a2, b2, c2) are placed three times in the heap, since it has three parents, and when each of them develop, (a2, b2, c2) will be produced and will be put into H. The nodes proliferation problem is associated with many computational problems, because many nodes are analyzed several times. This node can be used more than once in Skyline, causing a false Skyline [4].

The parent table is a suitable solution to solve the problem of nodes proliferation with minimal computations. Instead of considering all the ancestors, the parent table only stores the data of number of parents for a node. A basic rule is that a single node can only be placed in the heap, when all its parents are analyzed, and since the highest number of parents for a node is m, the parent table can at least be given a higher bit than number of parents in a node. This means that maximum of one bit more than the specified number of services be calculated. According to the expansion lattice, the following characteristics are followed:

Assuming the index of the S-Skyline starts with 1, the number of parents of each node is written above it.

The parent table, stores the number of parents in each node. In each comparison with another node, a figure of the number of parents will be decreased and the table will be updated with new values. Finally, each node reaches to zero, will be placed into heap. The operation ensures that all child nodes are placed into heap before parent node [3].

One of the key features of the parent table is that a table should be stored in the heap which have the lowest possible size without a single node and the ancestors be symbiotic in the heap. Heap size is a determinant factor in the overall performance of the algorithm. It should be noted that other methods to avoid the proliferation of nodes can be created.

D. Button-Up Algorithm

DPA algorithm decides with two major functions:

(F1)Heap functions: verify that the child node entered into the heap before the parent node.

(F2)Skyline comparisons: choosing the best services through existing services and taking into account the dominant relationship.

BUA algorithm is a bottom to top computational framework and presents a linear comparison strategy for better performance and scalability.

This algorithm carries out optimization and cost calculations with positive traits inherited from DPA. The algorithm's strategy is to use linear compositions while doing comparisons to select the best composition. A linear composition is to compare the results of two nodes with the next node, and finally achieving the best possible composition [4]. Figure 2 displays schema of a linear composition.

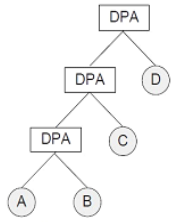


Fig. 2. linear composition [3]

III. Related Studies

Today, service composition is discussed as a key challenge in software systems. Using the appropriate method to combine services, provides favorable results than other existing approaches. It should be noted that by study and research in the service-oriented fields and service composition, new and optimal methods in this case are presented.

In this section, different methods of service composition are studied and presented and the best method is chosen among the existing methods.

Wu et al. used Dynamic Skyline Composition Algorithm for combining web services based on service quality, where the compositions of web-based services are performed dynamically. With the emergence of new services, old services will be removed and quality of service also changes. The advantages of this approach is to identify and select the best web services through a set of services based on service quality, and also using linear compositions to reduce the number of selected web services from the set of available services. The disadvantage of this study was lack of evaluating the algorithm

on a composition of web services in real and unreal data sets [5].

Benouaret et al. presented fuzzy dominance method for calculating Skyline service based on quality of service, in which key challenges including increasing number of web services and qualitative aspects have been considered. The users can enter their desired quality of service. The advantages of this new concept are α-dominance of Skyline service for composition of web-based services in high dimension that calculation in its composition has decreased. This method is also able to select the best services through quality services is undesirable. The disadvantages of this method, lack of awareness of sufficient information for the selected feature among Web services [6].

Benouaret et al. used fuzzy logic and Top-k algorithm to services' preferences and their compositions based on user requests. In this method, user preference is modeled based on fuzzy method, and the RDF is used to determine the relationship between the web services. The advantages of this study are improvements on diversity of web services compositions while maintaining service composition with the highest score. The disadvantages of this method is limits in fuzzy method in composition of services to match the user requests through using set of comparative methods [7].

In [8] ranking and clustering of web services is provided by using criteria of dominance relationship. In this research, retrieval, selection and composition of web services are done according to their increase and matching on the basis of criterion matching and using Skyline method among different qualitative parameters. The disadvantage of this method is lack of optimization at the time of service composition. In addition, reduced weight of service quality, results in decreased accuracy of retrieval services and thereby causes loss of important information. The advantage of this method is ranking suitable services on request which defines dominance relationships between services. The clustering provides prevailing interface between matched parameters. The proposed algorithms have been effective on the user's actual requests on true and false data.

Yu et al. presented the calculated Skyline service for invalid qualitative data, in which service composition in which the performances of services compositions are tested according to fluctuations in the dynamic environment. In this study, qualities of real services are provided by inherently invalid services. The optimization approach of available services may not change the quality of services over time. Therefore, multiple and sometimes conflicting qualitative criteria for the selection of each service requires using the weight characteristics. The disadvantage of this method is false understanding of weight characteristics which leads to loss of services desired by users. The advantages of this method are creating a new concept called p-dominant Skyline service that is suitable for weight characteristics and quantitative parameters. In addition, use of R-tree structure due to pruning dual structure, is effective in optimal composition web services in Skyline algorithm [9].

Alrifai et al. used Skyline service to reduce the number of candidate services composition. And by using K-means clustering algorithm, they have produced clusters of trees. This algorithm receives inputs of Skyline and returns the structure of the binary tree and defines the root. Indeed it can be said that it creates dominance relationship between the web services. And puts them in order based on the characteristics of services quality as these services belong to Skyline. The advantage of this method is removing invalid services and limitations of complex calculations and long processing time as they result in decreased efficiency [10].

IV. The Proposed Approach

As mentioned in the previous section, Skyline was introduced as an appropriate method by having advantages including high scalability, reduced process time of combining and providing the best composition. In previous studies, each methods providing the best composition of web services had problems. Parallel Skyline service algorithm, by using parallelism, increases the speed of combining huge volume of services. The steps in proposed method are as follows:

- First the user enters his request in form of input service and enters the desired output and service quality. Requested service format is as follows:

Request: (Input/s, Output/s, QoS)

In this step, preprocessing is performed on incoming request. The pre-processing is operations such as search and retrieval of required service from data base and designing an expansion graph in parallel.

Construction of expansion graph in sequential order is described in the second part. In this paper construction of each graphs and compounds in each level are parallel unlike [11]. In other words, compounds of each level due to being independent, can be done simultaneously, so the desired graph is generated faster. And by increasing the number of services involved in the composition, computing time and graph construction will have no significant increase. As Figure 3 shows, in every line of the graph, the possible composition are created in the context of parent-child rules, and the final graph is created again by using parent-child.

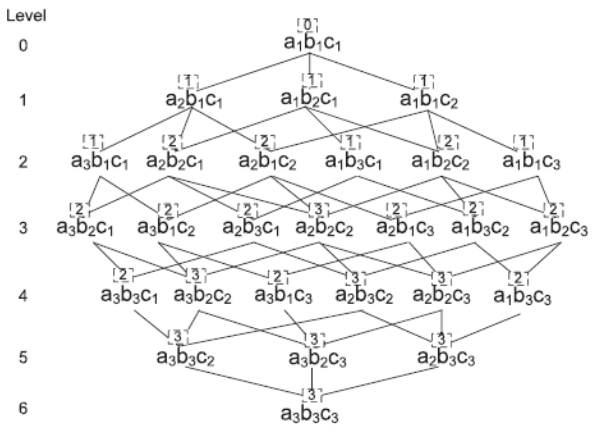


Fig. 3. Expansion Graph [3]

A. *Selecting the best service composition in each expansion graph by using a parallel linear combination*

According to Figure 4, parallel linear combination means comparing obtained combination of any comparison at all levels. For example, as shown in Figure 4, first two service composition of A, B through DPA algorithm are compared, and then the result is compared with a higher level of service composition which is C [3].

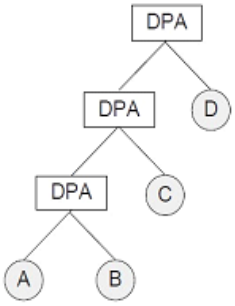


Fig. 4. linear combination [3]

After the parallel linear combination in each expansion graph, the best combination of the graph will be determined, which are defined as the Skyline Services.

B. *Selecting the best combination from existing Skyline services*

After determining any suitable combination of the two graphs, the results of each graph are compared with each other. Finally, the best combination will be presented to the user.

V. Simulation Results of the Proposed Model

For analysis and evaluation of parallel Skyline service, first implementation of proposed approach provided by Skyline algorithms were evaluated. And then the desired results with the same data on the algorithms provided by Skyline, Top-K algorithm and the proposed parallel Skyline algorithm were evaluated.

Here, the results of the proposed algorithm for the problem of finding the best composition in web services are described.

A. *Runtime*

In analyzing the results, it is observed that the addition of web services of data set, increases the runtime.

Due to the fact that the proposed model, the best service composition has been provided through parallel Skyline, thus runtime in equal number of web services declined compared to the Top-K algorithm and improvements occurred in the algorithm's runtime. Then in Table 6 runtime charts for Top-K and Parallel Skyline and Skyline algorithms are specified.

TABLE VI. Runtime (in Milliseconds) in Implementation of Parallel-Skyline and Top-K and Skyline Methods with Equal Number of Web Services

Number of Services	Skyline	Top-K	Parallel-Skyline
10000	3.4	2.12	1.35
20000	15.3	6.12	2.17
30000	37.5	13.27	4.41
40000	81.35	22.65	8.65

In Fig. 5 these three methods in low scale and equal number of services are compared and shown.

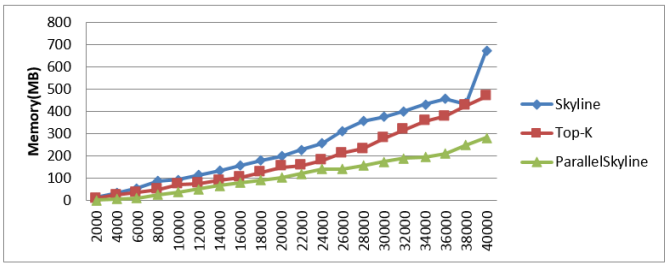


Fig. 5. comparison of runtime in low scale

These three methods are compared with each other in high scale and higher number of services.

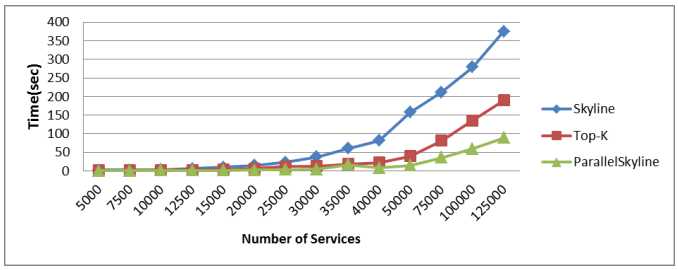


Fig. 6. comparison of runtime in high scale

As can be seen, Skyline method due to using parallel technique, has better composition speed than Top-K method.

B. *Memory Usage*

Another issue which can be analyzed in results is memory usage by proposed algorithm of Parallel-BUA.

Whereas Parallel-BUA method candidate services for composition has decreased; therefore, memory usage is less than Top-K and Skyline algorithms, and improvement occurs in memory usage. In Table 7 charts for the algorithms of Top-K and Parallel Skyline and Skyline are specified.

TABLE VII. Comparison of Memory Usage in Top-K and Parallel Skyline and Skyline Methods with Equal Web Services in Low Scale (MB)

Number of Services	Skyline	Top-K	Parallel Skyline
10000	95.45	73.32	37
20000	178.95	150	101
30000	375	278.5	175
40000	675	469	282

In Fig. 7 these three methods in low scale and equal number of services are compared and shown.

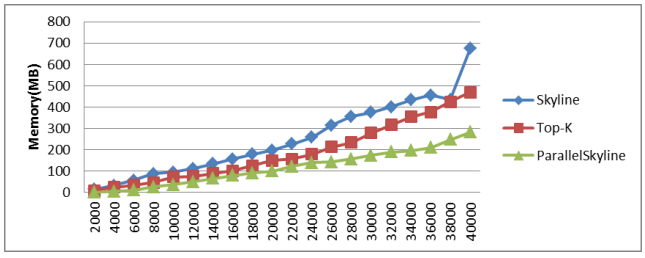


Fig. 7. Comparison of memory usage in Top-K and Parallel Skyline and Skyline methods in low scale

These three methods are compared with each other in high scale and higher number of services.

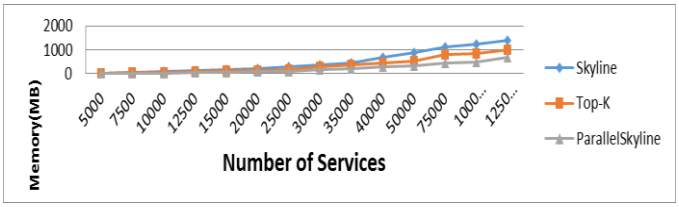


Fig. 8. Comparison of memory usage in high scale

As you can see, Skyline method due to use of parallelism techniques and the use of more powerful processor, use less memory compared to the Top-K method.

In addition, in parallel Skyline method compared to Skyline service, number of services and sub-services is selected dynamically and it is done according to user requests. Then, by a Top-K filter is applied on user requests, and user creates services based on the quality and through Parallel Skyline the best composition is shown to the user. This method consumes less memory than the BUA and Top-K.

VI. Conclusion and Future Works

The right composition of atomic services to provide services to users of is a challenge in web services. In this regard, choosing suitable atomic service required applying the right qualitative parameters is very important. Many methods have been suggested to solve the problem web services composition in regard to quality characteristics. As expected, web services with similar features lead to competition among the service providers. The key challenge to find an appropriate web service for composition occurs when multiple aspects of quality (such as response time, cost, etc.) are considered in the optimal composition of services. In this study various methods of service composition are discussed and compared. It was shown that previous methods had limitations in terms of computation and complexity. By using parallel Skyline service, selection speed and diversity has increased in a large scale space.

Suggested future works:

- Use parallel Skyline service to choose reliable services in social networks
- Use parallel Skyline service to predict quality of fuzzy services to solve the problem of inconsistency in fuzzy services.
- Use parallel Skyline service to solve inconsistent services

References

[1] Nuri S. and Tafakor F. (2013). Selecting optimal location services in service composition by using location service quality, mapping and geographic information Journal, p. 1-12

[2] Wang, L., Shen, J. And Yong, J. (2012). A Survey on Bio-inspired Algorithms for Web Service Composition", Proceedings of the IEEE 16th International Conference on Computer Supported Cooperative Work in Design, pp. 569-574.

[3] Yu, Q. And Bouguettaya, A. (2013). Efficient Service Skyline Computation for Composite Service Selection, IEEE Transactions on Knowledge & Data Engineering,vol.25, pp. 776-789.

[4] Yu, Q. And Bouguettaya, A. (2010). Foundations for Efficient Web Service Selection", Springer Dordrecht Heidelberg London New York.

[5] Wu, J., Chen, L. And Liang, T. (2014). Selecting Dynamic Skyline Services for QoS-based Service Composition, Applied Mathematics & Information Sciences an International Journal, vol. 8, PP. 2579-2588.

[6] Benouaret, K., Benslimane, D. And HadjAli, A. (2011). On the Use of Fuzzy Dominance for Computing Service Skyline Based on QoS, IEEE Conference Web Services, PP. 540-547.

[7] Benouaret, K., Benslimane, D., Hadjali, A. And Barhamgi, M. (2011). Top-k Web Service Compositions Using Fuzzy Dominance Relationship, International Conference on Services Computing, PP. 144-151.

[8] Skoutas, D., Sacharidis, D., Simitsis, A. and Sellis, T. (2010). Ranking and Clustering Web Services Using Multicriteria Dominance Relationships, IEEE Tranactin on Services Computing,Vol.3, pp.163-177.

[9] Yu, Q. and Bouguettaya, A. (2010). Computing Service Skyline from Uncertain QoWs", IEEE Tranaction on Services Computing, vol. 3, pp. 16-29.

[10] Alrifai, M., Skoutas, D. and Risse, T. (2010). Selecting Skyline Services for Qos-Based Web Service Composition, International World Wide Web Conference Committee, PP. 11-20.

[11] Chen, L., Cui, B. and Lu, H. (2011). Constrained Skyline Query Processing against Distributed Data Sites, IEEE Transactions on Knowledge and Data Engineering ,vol. 23,PP. 204-217.

A New Strategy to Optimize the Load Migration Process in Cloud Environment

Hamid Mirvaziri
Assistant professor of computer engineering,
Shahid Bahonar University of Kerman
Kerman, Iran

ZhilaTajrobekar
Student of Computer Engineering, Islamic Azad University of Kerman
Kerman, Iran

***Abstract*—Cloud computing is a model of internet-based service that provides easy access to a set of changeable computational sources through internet for users based on their demand. Load balancing in cloud have to manage service provider resources appropriately. Load balancing in cloud computing is the process of load distribution between distributed computational nodes for optimal use of resources and have to decrease latency in order to prevent a situation in which some nodes overloaded and some others under-loaded or be in the idle mode. Load migration is a potential solution for most of critical conditions such as load imbalance. However, many load migration methods are only based on one purpose. Practically, considering just one objective for migration can be in contrary to the other objectives and may lose optimal solution to work in existing situation. Therefore, having a strategy to make load migration process purposeful is essential in cloud environment. The main idea of this research is to reduce cost and increase efficiency in order to be compatible with cloud different conditions. In the recommended method, it is tried to improve load migration process using several different criteria simultaneously and apply some changes in previous methods. The simulated annealing algorithm is employed to implement the recommended strategy in the present research. Obtained result show desired performance and efficiency in general. This algorithm is highly flexible by which several important criteria can be calculated simultaneously.**

Keywords—cloud computing; load balancing; migration; virtual machines; simulated annealing

I. INTRODUCTION

One of the modern developments in the internet is introduced by Cloud Computing (CC) technology. This technology becomes quickly popular due to its properties in which every kind of facilities offers to the users in the form of a service [13]. The CC is an internet-based service model as it provides easy access to a set of changeable computing resources through internet for users based on their demands. In such mode, users try to access based on their needs regardless of where the service is located or how it is delivered. Various types of computing services try to offer such services to the users. Some of these computing systems are cluster computing, grid computing and recently CC. There are some services provided by CC architecture based on IT customers' needs [4]. Naturally, any new changes and concepts in IT environment has its own specific problems and complexities; using CC is not an exception and put many challenges in front of experts in this field, such as load balancing, security, reliability, ownership, backing up data, data portability and supporting different platforms. Considering the importance of migration in load balancing of CC, it is going to improve this process in this paper [2]. This article organized as follows: In section two previous works will be reviewed. In section three our proposed method is stated. This method is evaluated in section four and finally there is a conclusion in the last section.

II. LITERATURE REVIEW

A. Cloud Computing

CC platform is a completely automatic and service providers let the users buy, remote creation, dynamic scalability and system management [9]. Operational and capital costs can be reduced by CC [5]. In addition, CC systems are elastic; the amount of resources available to a server has capable of increasing or decreasing. However, it covers all what really a cloud server is [9].

The CC is a new internet-based service becomes common for users to provide different services. Many different and wide sources can be used instead of local or remote servers in CC services. There is not any standard definition for CC. Generally, it is including of a set of known server, offering services and resources to different and demanding clients. Distributed computers provide those demand-based services. The main advantage of CC is quick reduction of hardware costs and computation and capacity increasing [5].

B. Details of Cloud system and its properties

A Cloud system is composed of three main parts: clients, data center and distributed servers. Each part has its own specific role and purpose defined in the following.

- **Clients:** end users interact with clients to manage cloud-related data.
- **Data center:** a data center can created to save data and applications.
- **Distributed servers**: parts of cloud hosting different programs all over the internet while a cloud user thinks that the programs run in his machine [5].

C. Cloud implementation models

There are four models for cloud implementation as follows [14].

- **Private Cloud:** It is an infrastructure of cloud computing only working for an organization and

managed by the same organization or a third party organization.

- **Public Cloud:** It is also known as external cloud and describes CC as its main and traditional meaning. The services prepared dynamically through internet in the form of small units by a third party distributer who lends the resources in share to the users.
- **Community Cloud:** also known as group cloud in which cloud infrastructure shared and supported by several organizations with common goals in security mission or considerations. This cloud can be managed by the same or a third party organization. Since the costs divided between fewer users than in public clouds, this choice is more expensive than public cloud but has more privacy, security and compatibility with policies.
- **Hybrid Cloud:** a cloud in which its infrastructure composed of two or more types of clouds (private, community, public).

D. Load balancing based on migration of virtual machines and efficiency

- Load balancing between physical hosts is necessary for cloud environment in order to improve efficiency of data centers by increasing in throughput and decreasing in system latency. There are many physical hosts in a data centers therefore, based on migration of virtual machines between physical hosts, load balancing plays an important role to provide stable and highly efficient services. It is possible to have a situation in which physical hosts loaded excessively, that is, the number of virtual machines being run in physical hosts is more than the average. Therefore, running services are incapable of guarantee the needs. The efficiency of data centers servicing can be improved by migration of virtual machines from heavy loaded hosts to the light ones [16].
- Two-stage load balancing algorithms OLB+LBMM: A two-stage scheduling algorithm recommended to mix scheduling algorithms of Opportunistic Load Balancing (OLB) and Load Balance Min-Min (LBMM) for using better execution and maintaining load balancing of the system. The OLB scheduling algorithm keep any node in idle mode to reach the load balancing objective; the LBMM scheduling algorithm applied for reducing running time of any task on the node. Therefore, it decreases the total running time. This algorithm used in three-level CC networks in which efficiency and utilization criteria are considered. Such hybrid algorithm helps in effective use of resources, increases efficiency and offer better results than honey bee algorithm, random sampling and active clustering [4].
- Min-Min algorithm: It is started by a set of unassigned tasks. First of all, the minimum ending time of each task is found. Then, the least minimum is selected among these minimum times which is the minimum time between all tasks exist on each resource. After that, the task scheduled on the related machine within the minimum time. Now, the running time for all other tasks updated on the machine by adding up the assigned task running time to other tasks running time and the assigned task removed from the task list, assigned to the machine. This process repeated till all tasks assigned to the resources. However, there is a problem in this method which can be resulted in starvation [4]. In this algorithm, utilization of resources, overload, throughput, latency and efficiency are considered from load balancing criteria [13].
- A Lock-free multiprocessing solution for load balancing: this solution recommended because it is refused to use shared memory in comparison to other multiprocessing solutions of load balancing and keeps user session by locking. The memory this method obtained by applying Linux core and helps the improvement of general efficiency of load balancer in a multi core environment through running several load balancing processes in one load balancer [11].
- Ant Colony Optimization: A model presented in which unique ants act as very common insects. They have very limited memory and show individual behavior in order to have a big random performance. The ants are working together to find food sources and make use of food source for transferring food toward the colony simultaneously [10].
- Load Balancing Mechanism based on Ant colony and Complex Network Theory (ACCLB): It is presented in a federation of open CC with the purpose of overcoming complexity and problems of dynamic load balancing which use scale-free and small-world properties of complex networks to reach better load balancing. This method improved many aspects of the related ant colony algorithms and recommended to achieve load balancing in distributed systems. Moreover, it overcomes heterogeneity, compatible with dynamic environment, well in resisting against faults and has appropriate scalability since it helps the improvement of system efficiency [4]. Several studies showed that dynamic changes of criteria to calculate the possibility function for an ant in order to select a neighbor node, should have high efficiency for ant colony optimization algorithm, however such issues are not considered in this case [15]. In this algorithm in which complex network theory is used, utilization of resources, scalability and efficiency is considered from load balancing criteria [4].
- Join-Idle-Queue Algorithm: It recommends a load balancing algorithm for dynamic scalability of web services. This algorithm provides LB in large scale by distributed distributers. First, balance the load of idle processors in distributors is happened for any idle processor to access any distributor and then assign tasks to the processors in order to reduce the queue length of each processor. Removing load balancing task from vital route of demands processing, this algorithm reduces system load effectively, no connectional overload occurred at the time of tasks entrance and the

real latency not increased. The environment this algorithm used in is cloud data centers in which latency and overload considered as efficiency criteria. This algorithm is capable of running relatively optimal when applied for web services. However, it cannot be used for web services of nowadays dynamic contents due to scalability and reliability [4].

E. Simulated Annealing Algorithm

- It is an algorithm for optimization issues inspired from nature. This algorithm introduced by works of Kirkpatrick and Cerny et al. in 1983 and 1985 respectively. It is used for solid substance reaches to the mode in which put arranged well with minimum energy. In this method, put the substance in high temperature and then it decline gradually in order to put the substance in a mode in which it is arranged and has minimum energy. In this algorithm, each point of searching space (s) regarded as a mode of substance (searching space for our algorithm is all possible modes for migration of load between virtual machines) and the E(s) function which is the energy function (fitness function) should be minimized. The purpose is to transfer a substance mode (problem answer) from starting point (initial population) toward optimal mode (optimal answer). Evolutionary algorithms such as this one begin with an initial solution which is the starting point for moving toward optimal solution produced randomly. This means, a simple solution can be regarded initially but it is not necessarily the optimal one and it is just produced for using the algorithm and moving toward more optimal solution [6].

III. Recommended Strategy

Suppose that *m* is a host in which *N* virtual machines exist. *n* tasks will be under service. In servicing, it is possible to some of virtual machines overloaded and therefore their energy consumption and latency increases while the efficiency decreases. In such cases, some methods required for migration of services from overloaded virtual machines to those without overloading; on the other hand, load migration occurs. The purpose of this research is to offer a strategy that solves the problem of load migration on virtual machines considering different criteria such as energy consumption, efficiency etc. the recommended strategy is as follows:

In this strategy, at first each criteria of load migration problem are modeled and relationship between them are expressed. Then, annealing algorithm and modeling the problem in form of a population, a population with optimal fitness obtained. The optimal population will determine the order of load migration on virtual machines.

A. Criteria of the strategy

The following criteria considered in this research and it is tried to optimize all of them simultaneously.

1) load volume

Since it is possible for a virtual machine to be overloaded during task running in different aspects such as processor, memory or network, a criterion set for load volume of virtual machine, this criterion is a combination of processor, memory and network loads as follows:

$$loadvolume = \frac{1}{1-CPUutilization} \cdot \frac{1}{1-MEMutilization} \cdot \frac{1}{1-NETutilization} \quad (1)$$

In equation (1), $CPUutilization$ ،$MEMutilization$ و $NETutilization$show the efficiency rate of CPU, memory and network [17].

2) Energy consumption

Excessive increase of CC networks result in increasing of energy consumption in data centers intensively which is a vital problem and global concern for industry and society. The following equation used to calculate the criterion of energy consumption [3].

$$EnergyConsumption = Static\ energy + Dynamic\ energy \quad (2)$$

In equation (2), *Static energy* is the static rate of energy consumption and *Dynamic energy* is the dynamic rate of energy consumed by the service on the virtual machine which is obtained by the following equation:

$$Dynamic\ energy = \alpha . time . Speed^2 \quad (3)$$

$$Speed = \frac{f_{\text{cpu}}}{f_{\text{maxcpu}}} \quad (4)$$

In equation (3), α is relative coefficient, *time* is the execution time of application and *Speed* is the processor speed. In equation (4), f_{cpu} is normal frequency and f_{maxcpu} is maximum frequency of the CPU [3].

3) Resource utilization

Resource utilization depends on considering balance in using resources. The following equation used to calculate it; by simplifying the equation into CPU, memory and network we have:

$$Resource\ Utility = (Mem\ Utility - CPU\ Utility) + (Mem\ Utility - Net\ Utility) \quad (5)$$

In equation (5), $CPU\ utility$ ،$Mem\ utility$ and $Net\ utility$are efficiency rate of CPU, memory and network respectively [18].

4) Migration cost

Migration cost of different tasks may be different significantly regarding various settings of virtual machine and feature of the task volume. Most of the load balancing methods is highly efficient but unfortunately load migration cost, is ignored during designing the method which resulted in overloading and recommended methods become useless for cloud environment. Therefore, this criterion considered in the present research. The following equation used for calculating migration cost [1, 14].

$$Migration\ Cost = a . MigTraffic + b . MigTime + c . MigEnergy + d . MigDowntime \quad (6)$$

In equation (6), a, b, c, d are weights of the cost criteria which their sum should be equal to one. $MigTraffic$ is total network traffic of migration process, $MigEnergy$ is total energy consumed by migration process, $MigTime$ is the migration time and $MigDowntime$ is idle time emerged by migration process.

$$MigTime = \frac{V_{mem}}{M_{TR}} \quad (7)$$

$$MigEnergy = E_{sour} + E_{dest} + E_{net} \quad (8)$$

$$E_{sour} = V_{mem} \cdot P_{sour} \quad (9)$$

Simply suppose that in offline migration, $MigTraffic$ is the very V_{mem} which is equal to migrated memory in virtual machines in equation (4-7), M_{TR} is the memory transfer rate. In equation (8), E_{sour} ،E_{dest} ،E_{net} are amount of extra energy consumed by the source, destination and network interfaces. In equation (9), P_{sour} is the amount of energy needed for transferring one megabit to the resource node.

5) Fault Tolerance

In the present research it is tried to solve the load migration problem without violating fault tolerance required for services. Expressing this criterion, it is supposed that if amount of fault tolerance level of each service is k_i then the following equation guarantee fault tolerance of each service not to be violated.

$$\sum_{j=1}^{m} placement[i][j] - \sum_{j=1}^{k_i} placement[i][j] \geq N_i\,,\ i = 1, \dots, n \quad (10)$$

Where N_i is base number of virtual machines used by *i'th* service. Moreover, in $placement[i][j] = a$ which denotes the number of virtual machines put on *j*'th host by *i*'th service [16].

6) Efficiency

This criterion calculated by both hops time and waiting time parameters. Hop time is the time spent for transferring load from overloaded virtual machines to the ones without overloaded. Waiting time is the time in that virtual machines is preparing to receive services 12].

7) Implementation

The above mentioned criteria is combined linearly and considered as fitness function. Weight of each criterion determined regarding the importance of each one has. In addition, upper and lower limit of load can be considered for each virtual machine by which if the existing load in a virtual machine is more than the upper limit, the load must transfer to the machines with loads less than the lower limit. Efficiency rate of CPU in all virtual machines calculated by equation as follows (4-11) [4].

$$VMutility = \frac{totalRequestedMips}{totalMips\ for\ that\ VM} \quad (11)$$

$$HostBw = \sum current\ allocated\ bandwidth\ for\ VMs\ for\ host \quad (12)$$

$$HostRam = \sum current\ allocated\ Ram\ for\ VMs\ for\ host \quad (13)$$

$$Sum = \sum Uvm \quad (14)$$

Equation (12) and equation (13) expresses total bandwidth and total memory assigned to virtual machines on each host respectively. Equation (14) delineated the efficiency rate of all virtual machines. The virtual machine which its CPU efficiency is more than the upper limit of load means excessively loaded and needs migration of load to another virtual machine with lower load limit. Following equations is used to calculate upper limit of the load [18].

$$temp = Sum + \left(\frac{HostBw}{\sum Bw\ for\ all\ hosts}\right) + \left(\frac{HostRam}{\sum Ram\ for\ all\ hosts}\right) \quad (15)$$

$$T_{Upper} = 1 - \Big(\big((Puu * temp) + Sum\big) - \big((Pul * temp) + Sum\big)\Big) \quad (16)$$

In equation (15), the $temp$ variation is the sum of considered criteria. In equation (16), T_{Upper} is the upper limit of the load. Spare rate of CPU with Puu as upper probability and Pul as lower probability maintained for each host. Moreover, P_l is lower limit for spare capacity of the CPU. Spare capacity of CPU means that number of running services in CPU is lower than its capacity. The virtual machine that its CPU efficiency is less than the lower limit of the load becomes under-loaded and the overloaded machines are appropriate for migration of load to it. If CPU efficiency is less than 30%, the lower limit of load is always 0.3 [18].

$$if\ CPU\ utilization\ is < 30\% => T_{Lower} = 0.3 \quad (17)$$

$$if\ CPU\ utilization\ is \geq 30\% => T_{Lower} = 1 - \big((Pl * temp) + Sum\big) \quad (18)$$

Where T_{Lower} is the lower limit of the load [18].

Fitness function calculated through equation (18) in which there is $c1 + c2 + c3 + c4 + c5 + c6 = 1$ where $c1, c2, c3, c4, c5, c6$ are the weights of fitness calculation criteria, each one considered $\frac{1}{6}$ by default. However, as mentioned previously, weight of each criterion can be determined regarding the importance of each one.

a) $fitness =$
$c1 * loadvolume + c2 * PowerConsumption + c3 * Resource\ Utility + c4 * Migration\ Cost + c5 * fault\ tolerance + c6 * performance.$

IV. Evaluation of Proposed Algorithm and Several Studied Algorithm in the Related Works Chapter

Our intended strategy consist of six criteria "load volume, energy consumption, resource utilization, migration cost, fault tolerance and efficiency" to calculate fitness and obtained by equation (18). Proposed strategy will be compared with several algorithms of related works and equation (18) will be evaluated for them which show in the form of a chart. Fitness function of the recommended strategy considered for all these algorithms: OLB + LBMM, Min-Min, Max-Min, Ant Colony and Artificial Bee Colony and the comparison is demonstrated a table. As shown in the table, just the recommended strategy is based on SA algorithm which is including all criteria and can present an appropriate strategy for making load migration process purposeful in cloud environment. The following chart is the comparison between the recommended strategy and other algorithms.

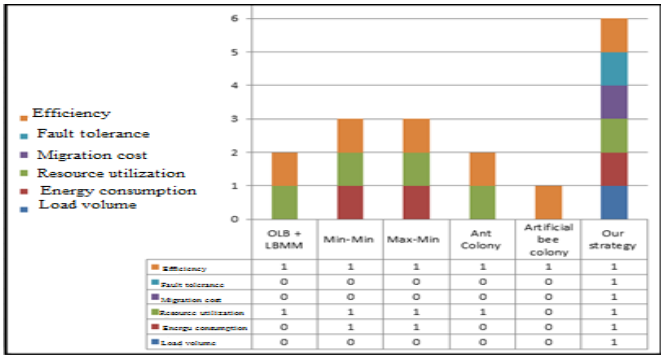


Fig. 1. comparing the recommended strategy with other algorithms based on the number of important criteria for making migration process purposeful in form of bar chart

1) Operation of recommended strategy (pseudo code)

At first, the services is assigned to the virtual machines in different hosts randomly and then the upper and lower limit of the load calculated in each virtual machine on each host and if there is at least one overloaded virtual machine, the load migration strategy activated and the load migration is done. In addition to finding a load-less combination, the algorithm calculates the above-mentioned determined criteria using fitness function produces fitness of each solutions of the population. Our purpose is to return a combination without overload and a better fitness. Then, the n number of better solutions (with higher fitness) is kept and the new population is produced by means of Crossover. After that, the amount of fitness of each solution in the new population is calculated and again the n number of better solution is kept; this process continued until the end condition of the algorithm occurred. Pseudo code for the recommended strategy is as follows:

Load Migration Algorithm

1: **Calculate** upper and lower bound load per every virtual machine for every host.
2: **if** there is at least one virtual machine with over load **then**
3: **Initialize** initial population by generating a random migration load from any virtual machine with over load to any virtual machine with minimum load.
4: **repeat**
5: by Crossover generate new population by migration load from any virtual machine with over load to any virtual machine with minimum load
6: Mutate new population
7: calculate fitness function for every member of new population
8: **until** termination conditions meet.
9: **end if**

a) Temperature and fitness function in the recommended strategy based on SA algorithm

As previously explained in SA algorithm, temperature is also one of the main parameters of this algorithm. Therefore, in this section, testing temperature in different values and calculating fitness function based on the related temperature shows that the more temperature decreased the more appropriate fitness value is obtained. Thus, the major task now is to determine the temperature situation. This main part is temperature reduction schedule which start at (1000˚C), and finish at (0-1˚C) and it is a linear function ($T_{k+1} = T_k * a$) [19, 20]. According to the calculation of temperature reduction rate, temperature variations chart and its effect on optimization is based on the points selected from the above table shown in the following figure.

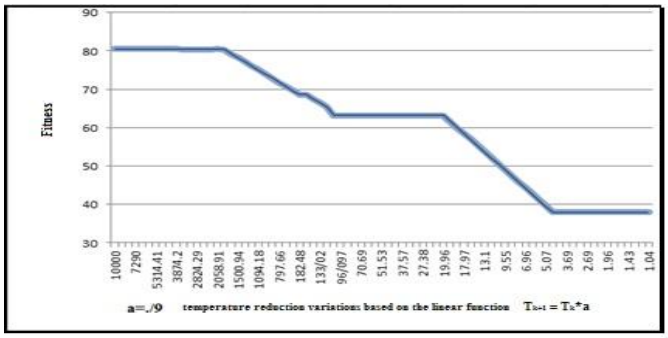


Fig. 2. Fitness and temperature reduction variations based on the linear function

2) Implementation and the strategy analysis

The algorithm simulated in MATLAB software by 5 hosts, 3 virtual machines and 60 services. A population consist of 10 solution is considered in the simulation. The recommended strategy algorithm starts up by random assignment of services to virtual machines in different hosts. Now, two modes may occur in this state for the algorithm:

1st mode: There is not any overloaded machine in the hosts after random assignment as shown in figure 3 in which each square denotes a host and it is clear, there are 5 hosts including 3 virtual machines, each one has different services.

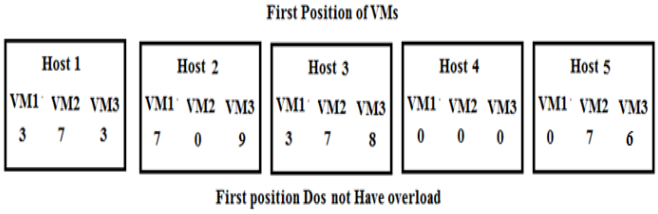


Fig. 3. primary random assignment without overload

In this mode, although there is not any overload, an appropriate load balancing algorithm guarantees,using of that equal amount of all existing resources at any moment. Therefore, since the primary random assignment without overload may not use all virtual machines for services assignment, the algorithm run in this mode and returned a population with appropriate order including 10 solution shown in figure 4 in which number of chromosomes is more than one row that is, there is a population includes 10 chromosomes each one has to return a fitness value. Therefore, there are 10 solutions and each solution is related to fitness of each population. Among these 10 answers, one with optimal fitness function can be selected as order of assigning the services to the virtual machines. In fact, selection process of optimal

fitness is that the first and the least values of fitness selected as the optimal solutions.

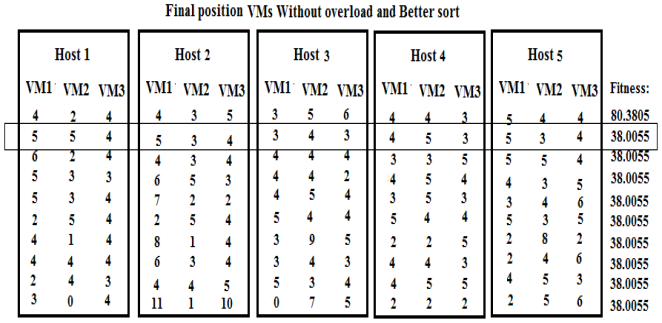


Fig. 4. population without overload with better order

2nd mode: it is assessment of the pattern for overloaded virtual machines. That is, one or more of the virtual machines in the hosts include overload after running of the algorithm and production of primary chromosome.

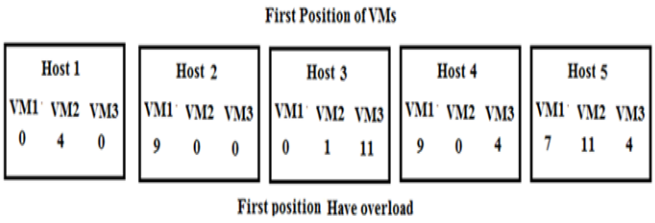


Fig. 5. primary random assignment including overload

The above figure shows that the algorithm performs until finding a situation without overload in this mode. Thesprocess is continued until finding a chromosome in which there is not any overloaded virtual machine. After finding a mode without overload for all virtual machines the algorithm ended and the output, is containing primary chromosome without overload obtained from the algorithm. The number of algorithm repetitions to find such chromosome and the output is shown in figure 6.

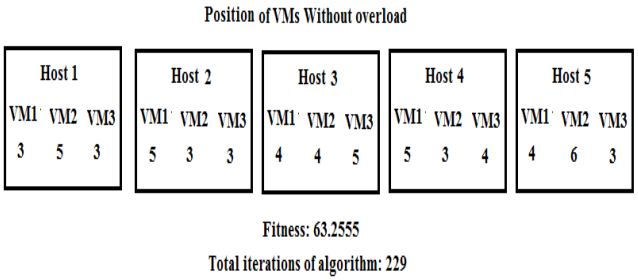


Fig. 6. population without overload with better order

The algorithm returns the first chromosome without overload as the solution. This solution may not have an appropriate fitness function therefore the process continues again and returns other chromosome without overload along with their fitness function (figure 7). The same as previous, a chromosome with optimal fitness function can be selected as the order of assigning services to the virtual machines.

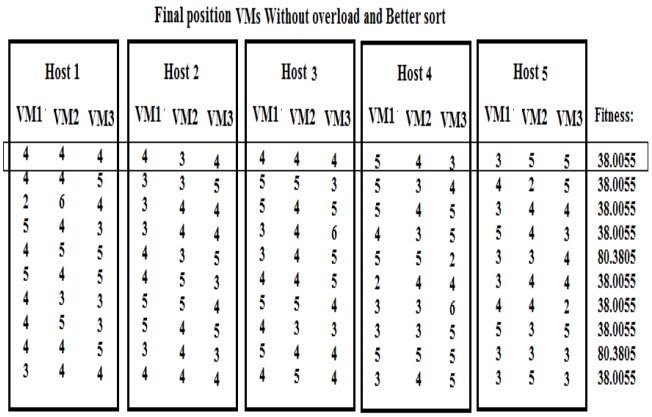


Fig. 7. population without overload with better fitness funtion and order

Moreover, the maximum number of genetic algorithm to be run is set to 1000 repetitions in order to prevent endless running. Generally, for determining whether virtual machines is overloaded or not, first CPU efficiency is calculated for each machine and the upper and lower limits of load is calculated for each host beside. The lower limit of load is calculated by the equation mentioned in previous parts if the CPU efficiency is more than 30% (the lower limit of load determined as 0.3 for each host in this research). Now, we have efficiency and the lower and upper limits of virtual machine. Therefore, the virtual machine with CPU efficiency more than upper limit of load is overloaded and needs migration of load to the other virtual machine with lower limit of load while the virtual machine with CPU efficiency less than lower limit of load is under-loaded and appropriate for migration of load from overloaded virtual machines.

V. CONCLUSION

In this paper, it is tried to study the differences resulted from applying different criteria in load balancing process in cloud environment. As mentioned before, load balancing process in cloud environment is very important which has high effective in applying cloud services. In the present research, a strategy is introduced in order to optimize migration process in load balancing. Considering studies done in this research, a criterion as the purpose of migration in load balancing process can be in contrary to other purposes, and the optimal solution for dealing with current situation may be lost and such algorithm neither used in all cloud conditions nor lead to the best results. Considering this result, using an algorithm capable of considering several load-balancing criteria in load migration process and optimize them simultaneously may overcome such defect to somehow. Thus, after studying load-balancing challenge, the present research uses simulated annealing method, which is an algorithm for optimization issues and inspired from nature, for utilizing the criteria simultaneously.

After evaluating of the recommended method, the method could prove its efficiency. In addition, the recommended method is highly flexible and developed in a way that several load-balancing criteria used easily in this method, the number of criteria can be increase or decrease. The most important innovation of this research is consideration of several criteria at the same time as the purpose of migration and using simulated

annealing for this reason. This makes the using of recommended algorithm possible in cloud environments with different conditions. The obtained results shows, the importance of using load-balancing process in cloud environment. According to previous studies, an introduced algorithm for load balancing does not have always the best result and it is depending on various factors but making migration purposeful can improve load-balancing process significantly. It is recommend to use dynamic programming algorithms in the future works to improve migration process in load balancing algorithms.

REFERENCES

[1] Aikebaier A, Enokido T, Takizawa M. "Trustworthy Group Making Algorithm in Distributed Systems".Human-centric computing and information sciences 1, No. 1, pp. 1-15, 2011.

[2] Begum S, Prashanth C.S.R., "Review of Load Balancing in Cloud Computing". IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No. 2, pp. 343-352, 2013.

[3] Beloglazov A, Buyya R., "Energy Efficient Resource Management in virtualized Cloud Data Centers". 10th IEEE/ACM Int Conf. Cluster, Cloud and GridComputing , pp. 826-831, 2010.

[4] Kansal N.J, Chana I., "Cloud Load Balancing Techniques : A Step Towards Green Computing".IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 1, pp. 238-246, 2012.

[5] Padhy R. P, Rao G. P., "Load Balancing in cloud computing Systems". Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering Bachelor Thesis, Orissa, India, pp. 1-46. 2011.

[6] Kirkpatrick S, Gelatt C.D, Vecchi M.P., "Optimization by Simulated Annealing". Science, New Series, Vol. 220, No. 4598, May 13, pp. 671-680, 1983.

[7] Kokilavani T, George Amalarethinam D.I., "Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing". International Journal of Computer Applications, Vol. 20, No. 2, pp. 42-48, 2011.

[8] Beloglazov A, Buyya R., "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers", Concurrency and Computation: Practice & Experience, Volume 24 Issue 13, pp. 1397-1420, 2012.

[9] Membrey P, Hows D, Plugge E., "Load Balancing in the Cloud", Chapter 13, pp.211-224, 2012.

[10] Nishant K, Sharma P, Krishna V, Rastogi N, Rastogi R., "Load Balancing of Nodes in Cloud Using Ant Colony Optimization", 14th International Conference on Modelling and Simulation, pp.3-8, 2012.

[11] Pathak K.K, Yadav P.S, Tiwari R, Gupta T.K., "A Modified Approach for Load Balancing in Cloud Computing Using Extended Honey Bee Algorithm", IJRREST: International Journal of Research Review in Engineering Science and Technology, Vol. 1, Issue 3, pp. 1-8, 2010.

[12] Rashmi K.S, Suma V, Vaidehi M., "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud". Special Issue of International Journal of Computer Applications on Advanced Computing and Communication Technologies for HPC Applications (ACCTHPCA), pp. 31–35, June 2012.

[13] Sran N, Kaur N., "Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing". International Journal of Engineering Science Invention, Vol. 2, Issue 1, pp. 60-63, 2013.

[14] Wang S.C,Yan K.Q, Liao W.P, Wang S.S., "Towards a Load Balancing in a Three-level Cloud Computing Network". Chaoyang University of Technology Taiwan, R.O.C. 978-1-4244-5540-9/10/$26.00 ©2010 IEEE, pp. 1-6, 2010.

[15] Zhang S,Yan H,Chen X., "Research on Key Technologies of Cloud Computing". International Conference on Medical Physics and Biomedical Engineering, Hebei Province, China, pp. 1791–1797, 2012.

[16] Yao L, Wu G, Ren J, Zhu Y , Li V., "Guaranteeing Fault-Tolerant Load Requirement Balancing Scheme" Published by Oxford University Press on behalf of The British Computer Society, pp. 1-8. 2013.

[17] Wood T, Shenoy P, Venkataramani A, Yousif M., "Black-box and Gray-box Strategies for Virtual Machine Migration". 4th USENIX Symposium on Networked Systems Design, Implementation, Cambridge, April 11–13, pp. 229–242, 2007.

[18] Xu J, Fortes J.A., "Multi-objective virtual machine placement in virtualized data center environments". In Green Computing and Communications (GreenCom), IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom), pp. 179-188, 2010.

[19] Jonathan R, Wolfgan K, and Jurgen W., “Temperature Measurement and Equilibrium Dynamics of Simulated Annealing Placements”;IEEE Transactions On Computer-Aided Design.VOL. 9. NO. 3, March, 1990.

[20] Sakamoto Sh, Tetsuya Oda, Elis K, Makoto I, Leonard B and FatosXh. (2013).” Performance Analysis of WMNs Using Simulated Annealing Algorithm for Different Temperature Values”; Seventh International Conference on Complex, Intelligent, and Software Intensive Systems.

Investigating the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author

Emad AlSukhni
Computer Information Systems Department
Yarmouk University
Irbid, Jordan

Qasem Alequr
Computer Information Systems Department
Yarmouk University
Irbid, Jordan

***Abstract*—Twitter is one of the most popular social network sites on the Internet to share opinions and knowledge extensively. Many advertisers use these Tweets to collect some features and attributes of Tweeters to target specific groups of highly engaged people. Gender detection is a sub-field of sentiment analysis for extracting and predicting the gender of a Tweet author. In this paper, we aim to investigate the gender of Tweet authors using different classification mining techniques on Arabic language, such as Naïve Bayes (NB), Support vector machine (SVM), Naïve Bayes Multinomial (NBM), J48 decision tree, KNN. The results show that the NBM, SVM, and J48 classifiers can achieve accuracy above to 98%, by adding names of Tweet author as a feature. The results also show that the preprocessing approach has negative effect on the accuracy of gender detection. In nutshell, this study shows that the ability of using machine learning classifiers in detecting the gender of Arabic Tweet author.**

***Keywords*—Social Networking; Data Mining; Sentiment Analysis; Sentiment Classification; Gender Detection; Twitter**

I. INTRODUCTION

Nowadays, the existence of many social websites such as Twitter, Facebook, Myspace and blogs that make the internet a large repository of different type of data. These media allow different type of users from different cultures and languages to communicate and share their opinions, and experience with others.

These opinions represent many kind of information (political, sport, technology, etc.) that come from different sources. Such a large repository of data and information sparked the attention of researchers and companies to take advantage of this data for various purposes. Sentiment analysis or opinion mining is a field aims to extract or predict the polarity of people opinions in specific areas. This is considered as a challenging task for sentiment analysis.

Gender detection is a sub-field of sentiment analysis for extracting and predicting the gender of a Tweet author. Most researchers studied gender detection for Tweet writers in different language such as English, European and other languages. However, in Arabic language there is a few researchers studied gender detection. In this study, we focused on Arabic opinions Twitter. Some of these studies have been investigated only gender aspect as a core attribute which can be a good indicator of the author of Tweet as in [1, 2]. Other studies investigated not only gender but also other attributes such as age for example in [3, 4].

Twitter website is our interest of study. We analyzed Tweets, which are small texts that consist of maximum 140 characters each). The Tweets are classified based on their writers' gender into two classes male and female. Twitter is considered as one of the largest social media website widespread in the world that has a huge number of users and a large amount of data in different languages from different places. Many researchers have studied the users Tweets for many purposes such as extracting political opinions, spam detection, etc.

The importance of knowing the gender of the Tweet author may help governments to make their policies and help companies in handling commercial issues. Thus, many social websites collect some information about their users when register such as age, gender, location, and others.

The main purpose of this research is to detect the gender of the writer of an Arabic Tweet by classifying them into two classes (male or female). This problem can be considered as a binary text classification (TC) problem. In this study, we used five classifiers KNN, NB, NBM, SVM, and J48 decision tree to test their ability in predicting the gender of the Tweet author.

Research Questions

In this study we are trying to answer the following questions:

Q1: Are data mining techniques able to identify the gender of an Arabic Tweet author with a significant accuracy?

Q2: What are the best classifier(s) to predict the gender of a Tweet author?

Q3: What is the effect of preprocessing techniques on classification accuracy in gender detection domain?

Q4: What is the effect of adding an author name as a feature on classifiers accuracy in gender detection domain?

Q5: What is the effect of adding the number of words and average word length in the Tweet as features on classifiers accuracy in gender detection domain?

The reset of the paper is organized as following: Section II reviews the previous works. Section III presents the methodology. Section IV discusses the experimental results. Finally, section V presents conclusions and future work.

II. Literature Review

Many researchers have studied gender detection of the writer of Twitter website and other social media users in different languages. However, a few of them have investigated Arabic language Tweets. In this section we list the most important of these studies with their results.

A. Gender Detection research on Multi-language

Rao et al. in [3] studied many author attributes such as age, gender, regional, and political orientation to classify Twitter users based on each attribute. They investigated the use of SVM algorithm over a set of features to classify user attributes (e.g. age, gender, regional). They built a large dataset manually and also used crawling Tweets. Their task was to detect a gender of Tweet author whether male or female based on the content of the Tweet. Their goal was to show if the language has an impact on detecting attributes of the author based on his/her Tweets. They used three classification models, first sociolinguistic-feature model; which is based on finding a lot of keywords. They also studied the writing styles effects in Tweet author gender and age detection. They extracted a list of words to be used in SVM classifier. The second model they used the N-gram feature with SVM classifier. The results of detecting gender of author showed that the SVM classifier slightly outperformed than sociolinguistic-feature and also n-gram with 72.33%.

Burger et al. in [5] used statistical models to detect the gender of unknown users from different places with different language. They used a huge dataset of Tweets from Twitter website labeled with male and female. The experiments on this dataset were conducted to show the accuracy of these models. They used WEKA tool to apply machine learning algorithms such as SVM, NB, and balanced Winno2. The result of the first experiment showed that the NB accuracy is 67.0%, and balanced Winnow2 accuracy is 74.0%, and SVM accuracy is 71.8%.

Liu and Ruths in [6] studied the relationship between the first name and detecting the gender of users who write English Tweets, and how this can improve the accuracy of gender detection. They collected a dataset from Twitter website randomly. Then, they have introduced idea of knowing and labeling the gender of each Tweet throughout profile picture. To ensure the accuracy of labeling they used the Amazon mechanical truck, which approved that the accuracy gives a good indicator of labeling. The core classifier they used was SVM; they applied some of the features as methods to be used in SVM such as top keywords, key-top stems, key-top n-gram, that all differentiate between the two genders. The results of these methods achieved high accuracy with SVM as 87.1%.

Marquart et al. in [7] investigated how to increase the predictive way of detecting users with both age and gender attributes from different social media such as Twitter, blogs, reviews, and others based on English and Spanish languages. They have used three features; content-based feature related to frequency of words, and a stylistic feature related to readability and spelling issues. In the evaluation step, they used SVM as the core classifier and used two approaches first label-power set which transforms multi-label problem into single label problem. They also used chain classifiers; which determine the dependency between two classes, and determine which class is good predictor to the other. They showed that gender is a good feature to use in predicting age.

Modak and Mondal in [8] studied gender classification using machine learning techniques; such as Naïve Bayes, maximum entropy and decision tree. In their study, they focused on the name of a user rather than on content-based of texts to classify it into male and female written. They collected different names from the web and form a labeled corpus. In their study they tested the three classifiers. The results showed that maximum entropy has achieved the highest accuracy in comparison with other classifiers.

Deitrick et al. in [1] studied gender identification of Tweets author for the English language using simple stream-based neural network. They collected a huge amount of data from Twitter website and then divided it into three different feature groups, 1-gram, 2-gram, and other features. After that they split the dataset into two files; one file containing the training set used in modified balanced Winnow. While the other file, containing testing dataset, used to evaluate the balanced Winnow. The algorithm has achieved 82% accuracy using entire set of features and 97.89% precision.

Mikros in [9] investigated the authorship of attribution and author gender detection or author profiling using Greek blogs. Blogs were chosen because people can write their opinion on the blogs. He collected the corpus from different blogs. They focused on two features of text content; first classical stylometric features, which depends on the vocabulary "richness", word length, and word frequencies. As for the second type of features, they used modern features which depends on character bigram, word gram. The classifier they used is SVM which is suitable for binary classification problem. The results of their experiment showed that accuracy of gender identification achieved 82.6%

Koppel et al. [10] studied automatically detect the gender of formal document authorship. They focused in their research on classification based on the writing style. The research tested two assumptions based on some previous research. First they assumed that no difference in writing between man and woman in formal texts. But in the second assumption, there is a difference between the two genders where it can be used to classify text of unknown authors. They built a dataset named BNC (Britch national corpus), and applied machine learning algorithms. Finally, they proved that male differs in writing pronouns and some types of noun modifiers in comparison with females.

Sap et al. [11] derived predictive lexica for age and gender using regression and classification models from words based on social media websites such as Facebook and Twitter. They collected a dataset mainly from Facebook in English language. The lexica has achieved 91.9% accuracy in gender detection.

Volkova et al. in [12] introduced an analysis of important difference between male and female in subjective language in Twitter website using three languages English, Russian and Spanish. They studied how the gender of Tweets play an important role in the sentiment classification. They developed two corpuses one for gender detection while the second for sentiment analysis. In their research, they showed that included author gender as a feature can significantly improve subjectivity and polarity classification with all tested languages.

Ugheoke in [13] studied gender detection for Tweet author. He focused on Twitter website because of its popularity in the world. Some of features that helped to be as indicators of Tweet author gender such as user profile, behavior of Tweets user which is related to number of Tweets per day and the number of replies, the linguistic style, and the social network were used. he relies on the name of user profile that checked by US census data (American names) for manual labeling the dataset. They divided the texts into separated words then also used a stemmer to reduce the number of keywords. The experiments show that, SVM classifier has achieved 86.8% accuracy with no name inference, and 95.3% accuracy with associated author name.

B. Gender Detection on Arabic Language

Few researchers have studied gender detection for the Arabic language, here we show these studies as follows.

Estival, et al. in [14] developed an application which can detect author attributes or demographic information; such as name, age, gender, level of education, from Arabic emails. They used two email corpuses for Arabic and English languages. They used a questionnaire to check and analyze the personality of the author of email such as age, gender, name and level of education. Many machine learning classifiers were used in their experiments; such as SVM, KNN, and decision trees (J48) combined chi-square and information gain. In gender detection, the result showed that SVM without feature selection technique achieve high accuracy over other classifiers.

Alsmearat et al. in [2] investigated gender identification on Arabic articles using the Bag Of Words (BOW) feature in the selection phase. The proposed technique works by estimating each word frequency in each document. They collected their dataset from Arabic news websites manually. They also collected Modern Standard Arabic for both genders. To reduce the number of words they used light stemming technique. To reduce feature selection they applied four algorithms (correlation analysis, Principle Component Analysis, correlation-based subset evaluator, Relief F) after dividing the dataset into five versions to show if there are any relationships between words (stylistic differences). In the classification phase they used many classifiers such as Naïve Bayes, KNN, and SVM, and then applied them on the five versions separately. Results showed that NBM and SVM achieved high accuracy on the first version (original version). In other versions and by applying feature selection techniques, the results showed a negative impact compared with results of the original version due to lack of information. On the other hand, they studied the impact of stemming on the dataset (Arabic light stemmer), The results showed that no significant impact on the accuracy. But when they applied stemming with best feature selection technique (sub dataset) the results showed NBM achieved good results over other classifiers.

III. RESEARCH METHODOLOGY

This section describes the research methodology that consists of 4 steps as shown in Fig. 1: collect Tweets form Tweeter, text preprocessing, gender classification, and evaluate the result.

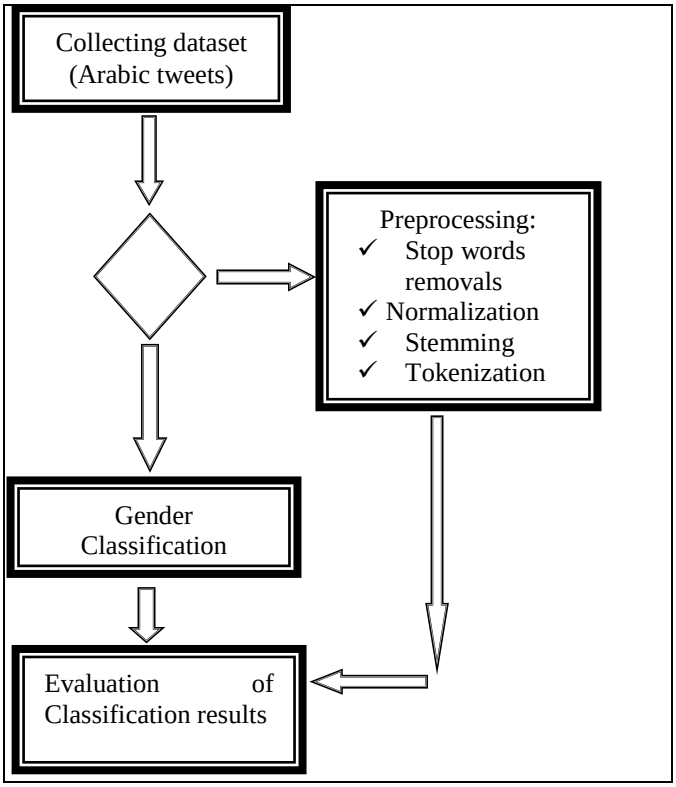


Fig. 1. Schematic overview of the methodology

In this research, we considered Twitter website as the target population to collect user Tweets from to be used in our experiments. There are many terminologies used in Twitter such as:

- Tweet: special text written by each user to represent his/her opinion about any topic.
- ReTweet: any user may republish any Tweet written from other users to be appeared in his/her profile.
- Followers: users follow individual user and see his/her Tweets.
- Hashtag (#): special symbol used to group any Tweets contain it such as (#sport) this Hashtag will group every Tweet that include this word.

In Twitter website when a user register, the user need just to enter username, email, and password to complete the registration, so no extra information could be used to determine the gender of the users. Figures 2 and 3 show two Twitter profiles one for a male user and another for a female user.

Fig. 2. Female Twitter Profile

Fig. 3. Male Twitter Profile

A. Dataset of Tweets

We collected large number of Tweets from Twitter website that exceeds 8000 Tweets in Arabic language, mainly in Jordanian dialects. We selected them from different domains for different users. The dataset consists 4017 Tweets written by males and 4017 Tweets written by females. Each tuple in our dataset consist of the following attributes:

1) The Tweets
2) Name of a Tweet author
3) Gender
4) Tweet Average length of word in
5) Number of words in the Tweet.

During the collecting phase, we considered some points to determine the gender of the Tweets' writers such as:

- Profile user name was used as a good indicator (User names are written either in Arabic or in English language) such as (‘علي’, ‘Ali’), and the profile picture to identify the correct gender (male or female).
- We used a lot of (Hashtages) to search about user profiles (‘رؤيا’,’#بس_بقول’,’#الجامعة الاردنية’, etc.) and then we visit the followers to collect other Tweets.
- We focus on Tweets, which are written by original users and excluded Tweets that were reTweet it (which is wrote by other users).
- We excluded Tweets that are written in newspaper articles or television reports.
- During building the data collection, we took approximately 50 Tweets for each author.
- Tweets have English words are excluded.
- We classify the dataset into male and female manually.

Sample of male Tweets with their authors are shown in table I.

TABLE I. SAMPLE OF MALE TWEETS

Profile User Name	Arabic Tweets
Fadi ‘فادي ‘	يعني لو درست لاب كان نجحت بس أنا “ متخلف”
Mohammed ‘محمد ‘	ثلاث ساعات نايم من مبارح لا ومش “ عارف انام كمان”
Khaled ‘خالد’	“مفكرينا لاجئين هؤن”
Omar ‘عمر’	"صباح الخير ياعرب"
Abd-Rahman ‘عبدالرحمن’	"البيت في الشتوية عبارة عن مكان ممل للغاية واحلى اشي فيه الأكل والنوم فقط لا غير"
Shaheen ‘شاهين’	"حد عنده راس للبيع. بدي راس ثاني"
Bahaa ‘بهاء’	"الدوام بكره زهق وملل ...بداية اسبوع ممتعه"
Mammon ‘مأمون’	"الجو بنعس كتير"
Tareq ‘طارق’	" هاي شكل واحدة تنتحر"
Moaied ‘مؤيد’	"الواحد بضل يتحمض للمخميس وبس ييجي الخميس بطلع مثل يوم الثلاثاء"
Salama ‘سلامه’	" طول ما الله موجود ! فالامل ابدا ما حيموت"
Nabil ‘نبيل’	" السواقة في هيك جو خرافية مش شايف متر قدامي"
Hasan ‘حسن’	" جمال العيون في النظرة مش في اللون"
Faris ‘فارس’	" نصيحة اليوم حبوا بعض وفي كل وقت من الاوقات بلابد ان تكرهو بعض"
Mosab ‘مصعب’	" خذو الحكمة من افواه المغردين"
Sameer ‘سمير’	" السناب عندي عبارة عن تغطية مباشرة لكل شوارع جدة و هي بتغرق"
Ahmed ‘احمد’	" حتى لو برد و تجمدنا و جلدنا بس برضه الشتى احلى من الصيف و ناموسه و قرفه"

Sample of female Tweets with their authors are shown in Table II.

TABLE II. SAMPLE OF FEMALE TWEETS

Profile User Name	Arabic Tweets
Aseel ‘ اسيل ‘	“اقنع امي انو انا ما باكل مقلوبه”
Hanaa ‘ هناء ‘	“ أكتر شخص مسالم وهادي بالحياة بس بحسك دايماً مكتئبة من الحياة”
Marh ‘مرح’	"مقدار السعادة انه لهلا سهرانة بدون ما اتذكر بكرة لازم اصحى على 6"
Salsabeel ‘سلسبيل’	“ ما أشنع البنت الي بتسوي حالها مهتمة ومحروقة عأشياء الرياضة خلص ماشي انت حلم كل شاب عربي بس اسكتي”
Sammer ‘سمر’	"بس عشان نكون واقعيين، الحياة بدها شخصية باردة ومكبرة عقلها"
Haneen ‘حنين’	"صباح الباص إلي راح علي"

Rand 'رند'	"نفسي اغسل أموال بهل بلد عشان افيدها بمشاريعي"
Tasneem 'تسنيم'	"زهقانة حالي"
Wasen 'وسن'	"صباحكم جميل..بكلمات أخرى .. صباح القرود"
Araam 'غرام'	"مش عارف ليش عندي شعور انه اليوم رح يكون احلى يوم"
Yara 'يارا'	" أسوأ اشي انه تحس حالك بتمشي بسرعة بتصير تضيّع بالوقت زي المحروم"
Amani 'أماني'	" لا تجبر حالك على شي مسببلك قلق ، خلي قاعدتك في الحياة الشي اللي ما بسعدني ما بلزمني"
Randa 'رندا'	" يلا حبايبي اللي مش عاجبه انفولو بسرعة"
Losi 'لوسي'	" علمتني الحياه انو ما اصدق غير اللي بشوفو بعيني او بسمعو باذني غير هيك لا لانو هالعالم صارت تعشق الهشت عشق"
Ronza 'رونزا'	" الإشي الوحيد الحلو بحياتي حالياً هو الأكل"
Rawan 'روان'	" الجاكيتات الجوخ الطويله ، فقط للرجل صاحب القامه الطويله غير هيك عبث"
Jodi 'جودي'	" عد ما رجعت من عنده و اتطلعوا ع بعض، لم كل صحابه و ما خلّى كلمة عليها و بلش يتسلى عليها"

B. Limitations and Assumptions

- The dataset represent the Jordanian dialects of Arabic language.
- Some users may use fake profile name that does not refer to the gender, such as a male user may use a female name.
- Collected profiles of famous users, newspapers, or any profile that uses the Arabic standard Arabic language are not considered.

C. The Preprocessing Phase

In this research, we study if the preprocessing stage has any impact on the quality of the results. Preprocessing stage consists of two major steps: 1) removing stop words and 2) stemming.

According to [18], using the Weka tool can make the preprocessing step by applying Saad light stemmer which performs the following things:

1- Normalized words
   - o Remove diacritics
   - o Replace آ أ ا with ا
   - o Replace ة with ه
   - o Replace ى with ي
2- Stem prefixes
   - o Remove Prefixes:كالـ, فالـ الـ, والـ.
3- Stem suffixes
   - o Remove Suffixes:ها , ان, ات, ون, ين.

D. Classifications

The dataset tuples are classified into two classes; male and female. We applied supervised machine learning classifiers to study the accuracy for each of them in detecting the author gender. Basically, classification is an approach aims to predict a class label that is unknown. Classification consists of two main stages: it builds the model from the training dataset, and then making a prediction.

In our research, we have used five data mining classifiers as listed below:

1) Key Nearest Neighbor (KNN): By using similarity and dissimilarity measures, the classifier works to estimate the distance between unlabeled documents and all documents in the training set as in [15]. For instance, if we want to classify the document x, it calculate the distance between x and documents in training set then after finding the k nearest documents to x, the classifier assign the document x to the class that have the large number of documents near of x. The Euclidean distance is used as a conventional method for measuring distance between two documents, d_1 (w_{11}, w_{12},…, w_{1n}) and d_2 (w_{21}, w_{22},…, w_{2n}):

$$E(d_1, d_2) = \sqrt{\sum_{i=0}^{n}(W_{2i} - W_{1i})^2} \quad \text{.....(1)}$$

2) Naïve Bayes (NB): Worked based on the probability theorem of conditional probability, mainly it is used for binary classification. In this classifier, the features of each document do not depend on the other features to predict the class, In the below the equation used estimate the probability of class.

$$P(Ci \mid X) = (\ P(X \mid C_i)\, P(C_i)\ ) / P(X) \quad \text{.....(2)}$$

3) Naive Bayes Multinomial (NBM) : The multinomial model of naïve Bayesian classification algorithm captures the word frequency information in document. NBM take into account the word frequency of each word as in [17].

4) Support vector machine (SVM): This algorithm works based on structural risk minimization principle from the computational learning theory. It divides the training set into two groups then try to find the hyperplane that is far from two groups as in [15]. Finding the optimal hyperplane based on the following formula:

$$F(X) = B_0 + B^T X \quad \text{..... (3).}$$

Where (B) is weight vector and (B_0) is a bias.

The closest training documents to hyperplane are called support vectors. Then the distance x and the hyperplane is estimated based on the below equation:

$$\text{Distance} = |\ B_0 + B^T X| / (\|B_0\|) \quad \text{..... (4).}$$

And then find the margins (distance) between the document (x) and the hyperplane from both sides of two groups, the margin that is represented by

$$M = 2/|B| \quad \text{..... (5).}$$

According to [15], SVM has several advantages over other techniques, such as it is robust in high dimensional spaces, any feature is important, they are robust when there is a sparsely of samples.

5) Decision tree: the decision tree classifier works by creating a classification tree, where each non-leaf node corresponds to a feature name and its children corresponds to a feature value. The Decision Tree classifier is a supervised machine learning approach that often used in a text classification domain. It requires two sets: a training set and test set. the Decision Tree Classifier creates a binary tree where the child nodes are instances of the classifier. In other words, this algorithm partitions the training set from the bottom to the top and then it picks up one attribute each time

and then the most information gain attribute is used to split the tree.

IV. Experiment and Evaluation

A. Performance Measures

Yang and Liu in [19] lists many of measurements to test the performance of classifiers such as:

1) True Positive (TP): If the instance is a positive and classified as positive.

2) False Negative (FN): If the instance is a positive and classified as negative.

3) True Negative (TN): If the instance is negative and it is classified as negative.

4) False Positive (FP): If the instance is negative but it is classified as positive.

Accuracy: It is the ability to predict categorical class labels. This is the simplest scoring measure. It calculates the proportion of the classified instances correctly:

Accuracy = (TP + TN) / (TP+TN+FP+FN) …….(6)

Sensitivity/Recall: Sensitivity is the proportion of the actual positives, which are correctly identified as positives by the classifier. It is also called true positive rate.

Sensitivity = TP / (TP +FN) …….(7)

Precision: it is a measure of the retrieved instances that are relevant.

Precision= TP/ (TP+FP) ……. (8)

B. Experiment Results and discussion

We evaluate the performance of the selected classifiers in classifying the gender of the Arabic Tweets' author. The rest of this section describes the results of experiments that have been designed and conducted to answer the research questions of this study. Most of the research use cross validation technique and splitting percentage in their classification experiments. In this study, we use cross-validation (10 Folds) and splitting percentage (66% train, 33% test). Because of the limited space of this Paper, we include all the cross-validation based results and the summery of splitting percentage based results.

Experiment 1: Evaluation classifiers without preprocessing.

In this experiment we test the performance of the selected classifiers in classify the gender of the Arabic Tweets' author without applying preprocessing step. As shown in Table III, the NBM classifier outweighs to other classifiers so as to achieve a better of accuracy of (62.49%) and recall (63%) to 5021 instances that are correctly classified manually (2532 instances for females and 2489 for males) as shown in Table III. The SVM classifier is improved slightly the precision (64%) compared with other classifiers. We notice that the NBM classifier outperforms the other classifier in correctly classified female instances than male instances.

TABLE III. Rate of Classifiers Performance without Preprocessing Step

Classifier	Accuracy	Recall	Precision
KNN	54.00%	0.393	0.557
Naïve bayes (NB)	57.39%	0.554	0.577
J48 (decision tree)	57.91%	0.488	0.597
SVM	61.63%	0.529	0.641
NBM	62.49%	0.630	0.624

Experiment 2: Evaluate the effect of preprocessing classification accuracy in gender detection domain.

In this experiment, we test the performance of the selected classifiers in classifying the gender of Arabic Tweets' author with applying the preprocessing step. As shown in Table IV and Figure 4, the NBM classifier outperforms the other classifiers. It achieved promising results with 61.27% accuracy in which the total number of correctly classified instances was 4923 (including 2507 instances of them for female and 2416 for male). Based on the precision measure SVM classifier achieved good result with 61%. We notice that NBM classifier outperforms the other classifier in correctly classified female instances than male instances. In another hand; the accuracy of KNN classifier is the lowest result with 53.43%.

TABLE IV. Classifiers Performance with Preprocessing

Classifier	Accuracy	Recall	Precision
KNN	53.43%	0.407	0.546
J48 (decision tree)	57.09%	0.494	0.584
Naïve bayes (NB)	57.55%	0.575	0.576
SVM	59.99%	0.539	0.614
NBM	61.27%	0.624	0.610

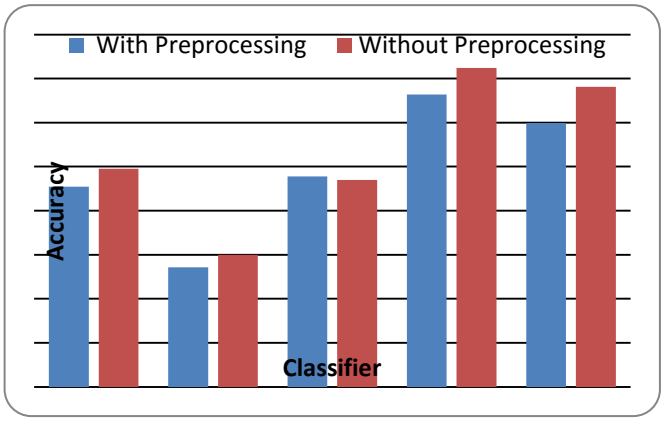


Fig. 4. Accuracy Results of Classifiers with and without Preprocessing

By comparing the accuracy of the classifiers, we conclude that the preprocessing step has a small negative effect on the accuracy of all classifiers. This results answer the third research question Q3.

Experiment 3: Evaluate the effect of adding an author name as a feature on classifiers accuracy in gender detection domain.

This experiment is designed to evaluate the effect of the author name feature on the performance of the selected classifiers which classifying the gender of Arabic Tweets' author with and without preprocessing. We add the name of the Tweet's author as a new feature in to the dataset to test the effect of this feature on the accuracy of the classifiers.

The result of this experiment shows the accuracy of detect the gender of the Tweet's author is significantly improved by adding Tweet's author name as a feature in the dataset as shown in the Figure 5. Moreover, we notice the same significant effect is achieved with and without applying preprocessing step. It is also clear that the accuracy of the top three classifiers become convergent.

TABLE V. EVALUATION OF CLASSIFIERS ACCURACY (WITH NEW TEXT FEATURES AND AUTHOR NAME ADDED)

Classifier	Accuracy (With Author Name added) without Preprocessing	Accuracy (With Author Name added) with Preprocessing	Accuracy Improvement ratio
Naïve bayes (NB)	77.29%	74.96%	25.75%
KNN	91.39%	83.67%	40.91%
NBM	98.49%	98.19%	36.55%
SVM	98.69%	98.25%	37.55%
J48 (decision tree)	98.69%	98.29%	41.32%

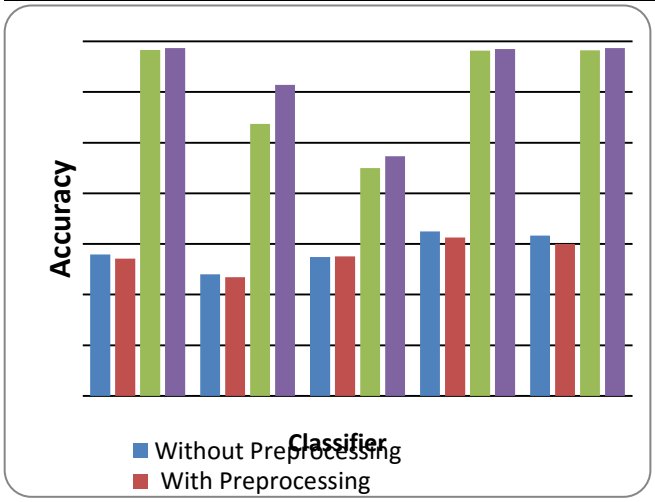


Fig. 5. Classifiers Accuracy (With Author Name added)

It is noticeable that after adding the author name feature to the dataset, the accuracy of the J48 classifier become very close to the accuracies of NBM and SVM classifiers. So, we can conclude that adding the author name feature to the dataset has the significant effect on the J48 classifier accuracy. Table V shows that the J48 and SVM classifiers outperform other classifiers with 98.69% accuracy either applying preprocessing or without applying. In this experiment the total number of correctly classified instances is 7929 by J48 (including 3994 female Tweets and 3935 male Tweets) and total number of correctly classified instances is 7929 by SVM (including 3990 female Tweets and 3939 male Tweets). We also notice that J48 and SVM classifiers have correctly classified both male and female written Tweets with the same accuracy.

According to Ugheoke T in [13], there is a relationship between the Tweets written in American English language and the name of Tweet's author that has an enhancement on accuracy of the gender detection. Thus, We can conclude from the result of this experiment that the effect of adding author names of Arabic language Tweets has the similar effect of adding author names of English language Tweets on the gender detection.

Experiment 4: Evaluate the effect of adding the number of words and average word length in the Tweet as features on classifiers accuracy in gender detection domain.

In order to get more improvement on classifiers accuracy, we added two features into our dataset that includes Tweet's author name. These two features are average word length and the number of words in the Tweet. So, this experiment is designed to evaluate the effect of the average word length and the number of words features on the accuracies of the classifiers into which classifying the gender of Arabic Tweets' author, Table VI gives such results.

TABLE VI. EVALUATION RESULTS OF CLASSIFIERS ACURACY WITH ADDING NAME OF TWEET'S AUTHOR AND THE NUMBER OF WORDS AND LENGTH OF WORD IN TWEET WITHOUT PREPROCESSING

Classifier	without preprocessing			preprocessing		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
KNN	69.60%	0.719	0.687	67.97%	0.722	0.666
Naïve bayes (NB)	73.38%	0.740	0.731	72.91%	0.733	0.727
NBM	99.06%	0.985	0.996	98.45%	0.977	0.992
SVM	99.35%	0.995	0.993	98.97%	0.987	0.992
J48 (decision tree)	99.50%	0.996	0.994	98.86%	0.989	0.988

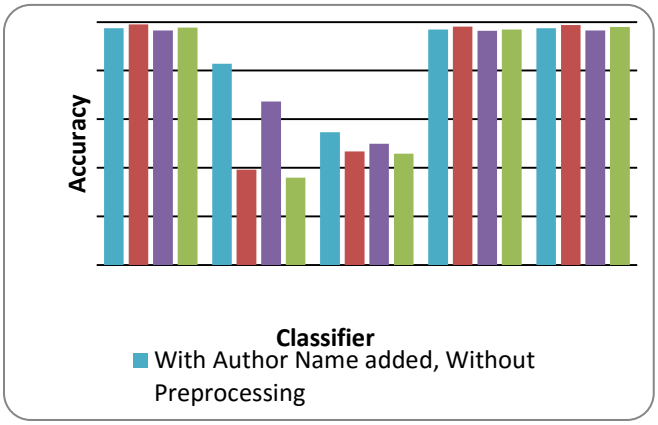


Fig. 6. Classifiers Accuracy adding the number of words and average word length in the Tweet as features

In this experiment, we applied the selected classifiers to study the impact of adding the name of the Tweet's author, the number of words and the length of word to the dataset without applying the preprocessing step. As shown in Table VI and Figure 4, the J48 classifier outperforms the other classifiers and it achieves the highest accuracy with 99.50% which the total number of 7994 correctly classified instances is 7994 (including 4001 for female and 3993 for male), with recall of

0.996%, but in precision we notice that the NBM classifier achieves slightly better results with 0.996%. Although the J48 classifier outperforms the other classifiers, the accuracies of SVM and NBM are very close to the accuracy of J48. We conclude adding the number of words and the average word length has a minor positive effect on top thee classifiers J48, NBM and SVM. KNN. On other hand, adding the number of words and the average word length has a negative effect on KNN and Naïve bayes classifiers as shown in Table 6 and Figure 6.

C. Summary of classifiers Cross-Validation based accuracy results

Table VII and Figure 7 present the summary of the effect of each studied feature on the classifiers accuracy. The results show that the accuracy of the classifiers without preprocessing are vary from 57% to 62%. The accuracy of all classifiers slightly decreased with applying preprocessing. The classifiers accuracy significantly increased by adding author names as a feature. The accuracy of the three beast classifiers slightly increased by adding two text features (number of words and average word length in the Tweet).

Form the table VII and Figure 7, we can answer all the research questions especially the first two questions; the answer of the first question is: yes, data mining techniques is able to identify the gender of an Arabic Tweet author because three classifiers got 99% accuracy. Regarding the second question answer, the three classifies (J48 (decision tree), NBM and SVM) have the best results in classification the Tweet author gender.

TABLE VII. SUMMARY OF CLASSIFIERS CROSS-VALIDATION BASED ACCURACY

Classifier	Without preprocessing	With Preprocessing	Adding Author Name	Adding New Text features
J48 (decision tree)	57.91%	57.09%	98.69%	99.50%
KNN	54.00%	53.43%	91.39%	69.60%
Naïve bayes (NB)	57.39%	57.55%	77.29%	73.38%
NBM	62.49%	61.27%	98.49%	99.06%
SVM	61.63%	59.99%	98.69%	99.35%

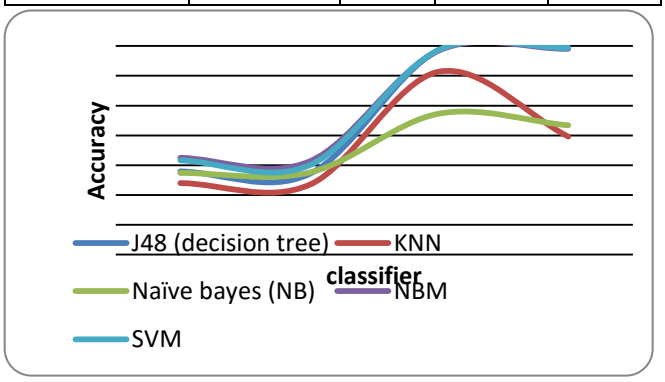


Fig. 7. Summary of classifiers Cross-Validation based Accuracy

D. Summary of classifiers Splitting Percentage based accuracy results

To give more creditability to our results, we rerun the experiments with Splitting Percentage of (66%).

In nutshell, we evaluate the performance of the selected classifiers in classifying the gender of Arabic Tweets' author based on the specific Splitting Percentage. From the Splitting Percentage based results, we can get the same above mentioned conclusions which give us more confidence in our findings.

TABLE VIII. SUMMARY OF CLASSIFIERS SPLITTING PERCENTAGE BASED ACCURACY

Classifier	without preproc-essing	with preproc-essing	Author Name added	With New Text features and Author Name added
KNN	53.69%	53.18%	77.85%	67.24%
J48 (decision tree)	57.79%	56.00%	88.17%	74.45%
Naïve bayes (NB)	57.39%	58.23%	98.64%	99.52%
NBM	60.72%	59.91%	98.79%	99.93%
SVM	60.68%	59.22%	98.79%	99.59%

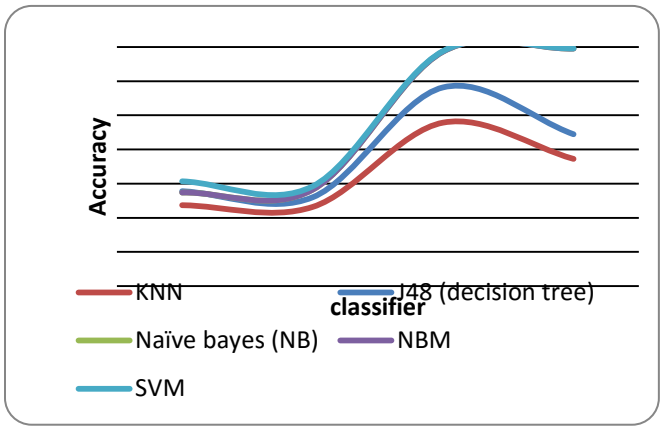


Fig. 8. Summary of classifiers Splitting Percentage based Accuracy

Table VIII and Figure 8 shows the results of many experiments. Let us start with the results of the classifiers without preprocessing on the dataset, the NBM classifier outperforms the other classifiers and achieved promising results with accuracy of 60.72% in which the total number of correctly classified instances was 1659 (including 854 instance of them for female and 805 for male). We notice that the NBM classifier outperforms the other classifier in correctly classifying female instances than male instances. Table 8 shows that KNN is the lowest classifier result with accuracy of 53.69%.

Table VIII, also shows the results of the experiment with apply preprocessing on the dataset. Notice that, the NBM classifier outperforms the other classifiers, into which the accuracy of 59.91% in which the total number of correctly classified instances 1637 (including 849 instance of them for female and 788 for male). We notice that the NBM classifier outperforms the other classifier in the correctly classified female instances than male instances. On the other hand, KNN is the lowest classifier with accuracy of 53.18%.

We test effect of adding the Tweet's author name to the dataset with and without preprocessing. As shown in Table 8 without preprocessing, the J48, SVM classifiers outperform other classifiers with 98.79% for both accuracy in which the

total number of correctly classified instance is 2699 for J48 (including 1348 for female and 1351 for male) for both classifiers. We also notice that the J48 and SVM classifiers outperform the other classifier in correctly classifying male instances than female instances. On the other hand, NB classifier got the lowest accuracy of 77.85%.

In the last experiment, we add the number of words and the average word length to each Tweet in the dataset, which already has names of Tweet's authors. The last column in Table 8 shows the results this experiment. The results show that the NBM classifier outperforms the other classifiers and it achieves better results in accuracy with 99.93%, in which the total number of correctly classified instance is 2703 (including 1333 for female and 1370 for male). On other hand, KNN classifier has the lowest accuracy which is 67.24%.

V. Conclusion and Future Work

This research aims to test the ability of many machine learning classifiers, such as J48, KNN, Naïve Bayes, NBM and SVM in detecting the gender of Arabic Tweet's writers. We collect the dataset that contains 4017 Tweets as a first step for the purpose of this study. The results show that the classifiers can be used to detect the gender of the Tweet's author. We also test the effect of preprocessing on the accuracy of the classifiers that were under testing. The results show a negative effect of preprocessing on the accuracy of all classifiers. Moreover, this study tests the effect of adding author names and word features on the accuracy of the classifiers that were under testing. The results show significant positive effect of adding the names of Tweets' author on the accuracy of all classifiers, the accuracy of J48, NBM and SVM classifiers achieved above 98%. Overall, the results of all classifiers in recall and precision measures are significantly improved. The results also show that there is a slightly positive effect result in adding the number of words and the average length of Tweet's words on the accuracy of the J48, NBM and SVM classifiers. On other hand, the results shows a significant negative effect on the accuracy of KNN and Naïve Bayes.

Overall results demonstrate that it is possible to use machine learning classifiers to detect the gender of Arabic Tweet's author. We got the same findings with both cross-validation and splitting percentage (66%) on preparing the dataset for our experiments. During the experiments, we notice that NBM, J48 and SVM classifiers achieve the best results in ability to classify female instances more than male instances. This leads to conclude that the possibility to detect female Tweets written in more accurate than male Tweets.

In future, we plan to add different dialects of other Arab countries and also collect Tweets written in standard Arabic language in our dataset. We also planning to study gender detection of Tweet's author who uses both English and Arabic languages in the same Tweet.

References

[1] Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., & Hu, W.,"Gender identification on Twitter using the modified balanced winnow". Communications and Network, 4(3), pp 189–195, 2012.

[2] Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R., "An extensive study of the Bag-of-Words approach for gender identification of Arabic articles". In proceeding of Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference, IEEE, pp. 601-608. 2014.

[3] Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M., "Classifying latent user attributes in Twitter". In Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM, pp. 37-44, 2010.

[4] Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M. F., Davalos, S., Teredesai, A., & De Cock, M., "Age and gender identification in social media". Proceedings of CLEF 2014 Evaluation Labs, 2014.

[5] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G.,"Discriminating gender on Twitter". In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguisticsl, pp. 1301-1309, 2011.

[6] Liu, W, and Ruths, D., "What's in a name? using first names as features for gender inference in Twitter". In Symposium on Analyzing Microtext. 2013.

[7] Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M. F., Davalos, S., Teredesai, A., & De Cock, M., "Age and gender identification in social media". Proceedings of CLEF 2014, 2014.

[8] Modak, S, and Mondal, A., "A Comparative study of Classifiers Performance for Gender Classification", IJIRCCE, 2(5), pp 4214-4222, 2014.

[9] Mikros, G. K.., "Authorship Attribution and Gender Identification in Greek Blogs", Methods and Applications of Quantitative Linguistics, pp. 21–32, 2012.

[10] Koppel, M., Argamon, S., & Shimoni, A. R., "Automatically categorizing written texts by author gender". Literary and Linguistic Computing, 17(4), pp. 401-412, 2009.

[11] Sap, M., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Schwartz, H. A., "Developing age and gender predictive lexica over social media". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 1146-1151, 2014.

[12] Volkova, S., Wilson, T., & Yarowsky, D., "Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media". In Proceedings of EMNLP, pp. 1815-1827, 2013.

[13] Ugheoke, T. O., "Detecting the Gender of a Tweet Sender", pp 1-60, 2014.

[14] Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B., "TAT: an author profiling tool with application to Arabic emails". In Proceedings of the Australasian Language Technology Workshop, pp. 21-30, 2007.

[15] Gharib, T. F., Habib, M. B., & Fayed, Z. T., "Arabic Text Classification Using Support Vector Machines". IJ Comput. Appl., 16(4), pp192-199, 2009.

[16] http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html.accessd-28/12/2015. Accessed in March 10th,2016.

[17] Saad, M. K.., "The impact of text preprocessing and term weighting on Arabic text classification", Doctoral dissertation, The Islamic University-Gaza, 2010.

[18] Motaz K. Saad and Wesam Ashour,"Arabic Morphological Tools for Text Mining", 6th ArchEng International Symposiums, EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, pp. 112-117, 2010.

[19] Yang, Y., and Liu, X., "A re-examination of text categorization methods". In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp 42-49, 1999.

[20] Nguyen, D. P., Trieschnigg, R. B., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., & de Jong, F. M. G., "Why gender and age prediction from Tweets is hard: Lessons from a crowdsourcing experiment". In Proceedings of COLING, 2014.

AUTHORS PROFILE

Emad Mahmoud Alsukhni obtained his PhD in from Ottawa University in Canada in (2011), he obtained his Masters' degree in Computer and Information Science from Yarmouk University in (2003), and obtained his Bachelor degree in Computer Science from Yarmouk University in (2003). Alsukhni is an assistant professor at the Faculty of Information Technology and Computer Science at Yarmouk University in Jordan. Alsukhni research interests include Computer Networks, Information Retrieval, Sentiment analysis and Opinion Mining, and Data Mining. He is the author of several publications on these topics.

Qasem Ibrahim Alequr obtained his Master degree in Computer Information Systems from Yarmouk University in Jordan in (2016), he obtained his Bachelor degree in in Computer Information Systems from Yarmouk University in (2008), and . Alequr research interests include Sentiment analysis and Opinion Mining, and Data Mining

Diagnosis of Diabetes by Applying Data Mining Classification Techniques

Comparison of Three Data Mining Algorithms

Tahani Daghistani, Riyad Alshammari
Health Informatics Department, College of Public Health and Health Informatics
King Saud Bin Abdulaziz University for Health Sciences (KSAU_HS)
King Abdullah International Medical Research Center (KAIMRC)
Minisrty of National Guard Health Affairs
Riyadh, KSA

***Abstract*—Health care data are often huge, complex and heterogeneous because it contains different variable types and missing values as well. Nowadays, knowledge from such data is a necessity. Data mining can be utilized to extract knowledge by constructing models from health care data such as diabetic patient data sets. In this research, three data mining algorithms, namely Self-Organizing Map (SOM), C4.5 and RandomForest, are applied on adult population data from Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia to predict diabetic patients using 18 risk factors. RandomForest achieved the best performance compared to other data mining classifiers.**

Keyword—Diabetes; Data mining; Self-Organizing Map; Decision tree; Classification

I. Introduction

Saudi Arabia is facing financial challenge due to the prevalence of diabetes. The Ministry of Health (MOH) in Saudi Arabia and Institute for Health Metrics and Evaluation (IHME) implemented, as collaboration, the assessment of burden based on the direct cost of diabetes from integrated health information system in 2014 [1]. Based on the established system, the current estimated cost of diabetes is 17 billion Riyals (US $4.5 billion) with expectation to increase the cost to 27 billion Riyals (US $7.2 billion) in the case that undiagnosed people are documented. Moreover, if pre-diabetes people become diabetes the cost will increase to 43 billion Riyals (USD 11.43 billion). The cost includes medications, visits, and lab tests, which also varies based on the patient's stage. The high cost of treating diabetes plus the expected growth of diabetes will put Saudi Arabia face to face with financial and health challenges in near future. Prevention, monitoring and controlling are the most effective actions to face such a health care challenge.

Data mining techniques assist health care researchers to extract knowledge from large and complex health data. With the evolution of information technology, data mining provides a valuable asset in diabetes research, which leads to improve health care delivery, increase support to decision-making and enhance disease management [2]. Data mining techniques include pattern recognitions, clustering, classification and association.

Diabetes is one of the main topics for medical research due to the longevity of the diabetes and the huge cost on the health care providers. Early detecting of diabetes ultimately reduces cost on health care providers for treating diabetic patients [3-8], but it is a challenging task. For early detecting of diabetes, researchers can take advantage of the patient's health care data to convert raw data into meaningful information and extract hidden knowledge by applying data mining such as decision tree or SOM to construct an intelligent predictive model.

SOM or Kohonen maps is a machine-learning tool that is used to analyze heterogeneous data and provides supervised or unsupervised learning model [9-11]. Hence, SOM maps high dimensional data to be more meaningful by identifying similarities. In this research article, decision trees, namely C4.5 and RandomForest, are compared with SOM to build a classification model to predict diabetic patients using retrospective data collected from hospital database systems. The data sets are extracted from the hospital information management system from the Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia. The National Guard Health hospitals provide optimum health care to their employees, dependents, other eligible patients and private patients. The data sets are collected from four hospitals in the three largest regions in Saudi Arabia in terms of populations. The hospitals are: i) King Abdulaziz Medical City (SANG) in Riyadh, Central Region; ii) King Abdulaziz Medical City in Jeddah, Western Region; iii) Imam Abdulrahman Al Faisal Hospital in Dammam, Eastern Region; and v) King Abdulaziz Hospital in Alahsa, Eastern Region. The contribution of this study is utilizing the data mining techniques to construct intelligent predictive model using real healthcare data that are extracted from hospital information systems using 18 risk factors.

The rest of the paper is organized as follows. The literature review is given in Section II. Methodology is presented in Section III. Results and discussion are given in Section IV. Finally, conclusions and future work are presented in Section V.

II. Literature Review

In the literature, SOM has been applied in health care data. Mäkinen et al. [12] used SOM algorithm to detect association

between certain risk factors and complications. They used SOM as an unsupervised method to cluster biochemical profiles. A 7 x 10 grid of hexagonal map units with Gaussian neighborhood function were used to present similarities and differences between variables. Tirunagari et al. [13] applied SOM to cluster heterogeneous diabetes data. They were able to reduce the dimensionality of the data and demonstrate the similarities between patients by placing them in groups using the U-matrix. As a result, the profiles of patients who need self care management were grouped clearly and easily were identified.

In another study, Tirunagari [14] used the SOM to recognize the behavior of self care based on survey data collected from type I diabetic patients. The visualization result improved understanding pattern of various behaviors as well as detecting patients who need to adjust their lifestyle. Zarkogianni et al. [15] proposed personalized hybrid model by combining Compartmental Models **(**CMs) and Self-Organizing Map. The model helped patients with Type I Diabetes Mellitus to predict the metabolic behavior. Luboschik et al. [5] used SOM as part of an early detecting system to predict Neuropathy complications in diabetic patients. By using the computational and visual methods of SOM, they were able to identify characteristics of diabetic Neuropathy patients.

Other data mining algorithms had been applied to classify diabetic patients. Farran et al. [16] used non-laboratory attributed to classify the diabetes by applying 4 data mining models that were logistic regression, k-nearest neighbors (k-NN), multifactor dimensionality reduction and Support Vector Machines (SVM). They achieved an accuracy of 85% for diabetic patients. Barakat et al. [17] applied SVM on data collected from a national survey in the Sultanate of Oman that investigated the prevalence of diabetes mellitus. They achieved a sensitivity of 93% and 94% for accuracy and specificity.

Moreover, Ganji et al. [18] used (FCS-ANTMINER) on public diabetes data set (Pima Indians Diabetes data set [19]). They obtained an accuracy of 84%. Huang et al. [20] employed three data mining algorithms that were Naive Bayes, IB1 and C4.5 to predict diabetes on data gathered from Ulster Community and Hospitals Trust (UCHT) between 2000 and 2004. They were able to achieve an accuracy of 98%. Furthermore, Al Jarullah A. [21] employed C4.5 data mining algorithm on Pima Indians Diabetes data set [19]. He achieved an overall accuracy of 78%.

From the literature review, data mining algorithms have been used to predict diabetes using public data or private data. However, the data sets are either small in size (less than 10,000 records) or collected from one region (mostly one hospital). In this research study, the data sets are collected from 4 large hospitals in Saudi Arabia. The model extracted from the data could assist in improving healthcare plans that are delivered for diabetic patients.

III. METHODOLOGY

To achieve the study objectives, study method consists of several phases, which are collection of data and attribute selection, data mining algorithms and evaluation criteria.

A. Data sets and Attributes Selection

In this work, the data sets are collected from Ministry of National Guard Health Affairs (NGHA) databases from the highest three populated regions in Saudi Arabia, where the databases have all patients visit information such as laboratory and medications, etc. These regions are: central region (Riyadh city), western region (Jeddah city) and eastern region (Alahsa and Dammam cities). The latest Saudi census showed that more than 66% of the country total population lives in these three regions and the largest city on these three regions are Riyadh city (The capital and the largest city in the Central region); Jeddah (the largest city in the Western region; iii) Dammam; and Alahsa (the largest two cities in the Eastern region) [22].

The data sets consist of 66,325 diabetic and non-diabetic instances. The study used data from the hospital Information System in MNGHA from the 2013 to 2015. Hospital databases are extremely exposed to inconsistent values, noisy and missing input values from the data because the data are coming from heterogeneous sources. There are several considerations that are followed and assured throughout the data extraction process by the information systems in MNGHA to insure the accuracy of the data. In addition, the data sets are gone through manual inspections to ensure the data are consistent and accurate.

All adult patients who have diabetes are included while pediatric diabetic patients are excluded. The data used for the study did not include identification information in order to not violate the patient privacy.

Detailed information about demographic variables is summarized in Table1. Furthermore, the data set divided into training and test data sets as follows:

- Data from 2013 to 2014 represents a training set that is used to construct and train the model.
- Data from 2015 represents a test set that is used to test the model and estimate the accuracy rate.

The data sets consist of a total of 18 attributes. The attributes include gender, age and region as demographic variables; patient's measurements such as BMI and blood pressure in addition to 11 various lab tests. The Data sets contain 36,811 male (55.50%) and 29,514 females (44.50%), all of them at least 14 years old and older. More than half of the total patients (64.47%) have diabetes; male diabetic patients represent 36.34% of the total diabetic patients, while female diabetic patients represent 28.13% of the total diabetic patients as shown in Table 1.

TABLE I. DISTRIBUTION OF DEMOGRAPHIC VARIABLES

Variables	*Diabetes*	*Non-diabetes*	*Total*
Region			
Central	34039 (62.87%)	20102 (37.13%)	54141 (81.63%)
Eastern	8012 (72.28%)	3073 (27.72%)	11085 (16.71%)
Western	708 (64.42%)	391 (35.58%)	1099 (1.66%)
Gender			
Male	24104 (36.34%)	12707 (19.16%)	36811 (55.50%)
Female	18655 (28.13%)	10859 (16.37%)	29514 (44.50%)
Age			
13-19	118 (0.28%)	460 (1.95%)	578 (0.87%)
20-34	1120 (2.62%)	2947 (12.51%)	4067 (6.13%)
35-44	2070 (4.84%)	2416 (10.25%)	4486 (6.76%)
45-64	16226 (37.95%)	7723 (32.77%)	23949 (36.11%)
65-84	20602 (48.18%)	8447 (35.84%)	29049 (43.80%)
>85	2623 (6.13%)	1573 (6.67%)	4196 (6.33%)
Overall total	42759 (64.47%)	(35.53%)	66325

Lab test data are described statistically and summarized in Table 2 in order to provide more understanding of lab tests data which are considered as attributes in the study.

TABLE II. STATISTIC DESCRIPTION OF LAB TEST DATA

	N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
eGFR	66325	2.00	220.00	78.3311	40.83988
MCV	66325	.00	129.80	86.9547	7.58909
MCH	66325	12.20	59.80	28.0319	2.91036
MCHC	66325	27.10	373.00	317.5521	38.98827
RDW	66325	10.00	99.00	15.2312	2.42988
Plt	66325	3.00	999.00	273.7028	125.00519
MPV	66325	.00	90.00	8.5508	1.38118
WBC	66325	.00	319.60	9.3527	5.81372
RBC	66325	1.18	8.71	4.1693	.83756
Hgb	66325	5.10	232.00	114.5618	26.71571
Hct	66325	.04	54.70	.9110	4.44215
Valid N	66325				

B. Data Mining Algorithms:

R software [23] is used to employ SOM algorithm in order to predict diabetes patients. Kohonen package in R implements SOM as unsupervised algorithm as well as supervise algorithm. The *bdk* and *xyf* are supervised functions of SOM in R. The returned output obtained from calling both functions is used for prediction in this study.

Since SOM has a number of parameters, selecting the appropriate parameters, such as type of SOM, network size and training algorithm, is important. Parameters have direct impact on the classification performance as well as computational time [9]. The values for parameters are summarized in Table 3. On the other hand, Weka [24] data mining tool is used to run C4.5 and RandomForest decision trees using the default parameters.

TABLE III. SOM PARAMETERS

Parameters	***Meaning***	***Value***
Grid	To determine the size of map	20 x 20, hexagonal
Rlen	To determine number of iterations	1000
Alpha	To determine the learning rate for start and stop	[0.05, 0.01]
Radius	To determine the initial neighborhood. The value is decreased during training linearly	[90%, 85%]

C. Evaluation Criteria:

To select the best performance data mining algorithms in predicting diabetic patients, two standard matrices have been applied, which are Recall and Precision. Recall, Eq. 1, will reflect the number of diabetic instances who are correctly classified, which we need in such system. It is calculated using:

$$\text{Recall} = TP/(TP + FN) \quad (1)$$

While Precision, Eq. 2, represents the relevant instances that are correctly classified. It is calculated using:

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

True Positive (TP) implies that diabetic patients who are classified as diabetic patients, whereas False Negative (FN) implies that diabetic patients who are classified as non-diabetic patients. On the other hand, False Positive (FP) implies that non-diabetic patients who are classified as diabetic patients. Commonly, the best learning algorithm is going to be selected based upon the performance of the classifiers in terms of high Recall, and Precision.

IV. RESULT AND DISCUSION

In Table 4, two different measurements were calculated for each algorithm for assessing how well each model and to be used to evaluate algorithm's performance compared to each other. C4.5 and RandomForest achieved Recall and Precision over 90% on the training data set while SOM (bdk and xyf) was able to achieve Recall and Precision over 79% on the training data set.

To choose the best algorithms in terms of high performance, according to the evaluation criteria, all algorithms are evaluated on an unseen data set (test data set). The algorithm/model who achieved the highest performance in terms of high Recall and Precision is considered to be the best one. It can be seen that RandomForest achieves the highest Recall and Precision on the test data set as indicated in Table 4.

TABLE IV. RESULT OF THE CLASSIFIERS

Models	***Recall***	***Precision***
SOM- bdk – Training data	0.79	0.83
SOM- bdk – Test data	0.69	0.48
SOM- xyf – Training data	0.79	0.84
SOM- xyf – Test data	0.79	0.46
C4.5 – Training data	0.965	0.92
C4.5 – Test data	0.776	0.67
RandomForest – Training data	1.0	1.0
RandomForest – Test data	0.904	0.68

The reason behind that SOM could not perform higher than decision trees due to the fact that the SOM constructs the its model from only the first SOM grid layer. The multi-layer classification capability of SOM could improve the performance. However, the multilayer capability is not available in R software [23].

In this study, SOM and decision tree techniques are applied to predict diabetic patients using 18 risk factors (attributes). The most common risk factors among the model constructs from the algorithms are as the following: i) gender; ii) age; iii) blood pressure; iv) Body Mass Index (BMI); v)

region; and vi) several lab tests such as Hematocrit (Hct), hemoglobin (Hgb), Platelet count (Plt) and Mean Platelet Volume (MPV).

The extracted knowledge from the research conducted among the samples (patient records) from MNGHA can be generalized to the wider diabetic population in Saudi Arabia since the data sets (samples) are collected from the largest populated region in Saudi Arabia where more than 66% of the total country population lives.

V. CONCLUSION

Model constructed from the data mining algorithms could help to support decision making in different fields including health care field. In this research, real health care data sets have been collected from MNGHA databases that contain 18 attributes. Furthermore, three data mining algorithms have been evaluated, namely SOM (bdk and xyf), C4.5 and RandomForest to construct data mining models to predict diabetic patients using real health care data sets.

The results show that the constructed data mining model could assist health care providers to make better clinical decisions in identifying diabetic patients. Additionally, the model could be further developed for patient protection. In the future, the results can be utilized to create a control plan for diabetes because diabetic patients are normally not identified till a later stage of the disease or the development of complications.

ACKNOWLEDGMENT

This study was funded by the King Abdullah International Medical Research Center (KAIMRC), National Guard, Health Affairs, Riyadh, Saudi Arabia with research grant No. SP15/064.

REFERENCES

[1] Mokdad AH, Tuffaha M, Hanlon M, El Bcheraoui C, Daoud F, et al. (2015) Cost of Diabetes in the Kingdom of Saudi Arabia, 2014. J Diabetes Metab 6: 575

[2] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.

[3] Li, R., Zhang, P., Barker, L. E., Chowdhury, F. M., & Zhang, X. (2010). Cost-effectiveness of interventions to prevent and control diabetes mellitus: a systematic review. Diabetes care, 33(8), 1872-1894.

[4] Lin, J. H., & Haug, P. J. (2006). Data preparation framework for preprocessing clinical data in data mining. In AMIA Annual Symposium Proceedings (Vol. 2006, p. 489). American Medical Informatics Association.

[5] Luboschik, M., Röhlig, M., Kundt, G., Stachs, O., Peschel, S., Zhivov, A., ... & Schumann, H. (2014). Supporting an Early Detection of Diabetic Neuropathy by Visual Analytics.

[6] Nuwangi, S. M., Oruthotaarachchi, C. R., Tilakaratna, J. M. P. P., & Caldera, H. A. (2010, December). Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes. In Information Technology for Real World Problems (VCON), IEEE 2010 Second Vaagdevi International Conference on (pp. 3-8).

[7] Wang, K. J., Adrian, A. M., Chen, K. H., & Wang, K. M. (2015). An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. Journal of biomedical informatics, 54, 220-229.

[8] Shivakumar, B. L., & Alby, S. (2014, March). A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes. In Intelligent Computing Applications (ICICA), IEEE 2014 International Conference on (pp. 167-173).

[9] Ballabio, D., Vasighi, M., & Filzmoser, P. (2013). Effects of supervised Self Organising Maps parameters on classification performance. Analytica chimica acta, 765, 45-53.

[10] Wehrens, R., & Buydens, L. M. (2007). Self-and super-organizing maps in R: the Kohonen package. Journal of Statistical Software, 21(5), 1-19.

[11] Wijayasekara, D., & Manic, M. (2012, June). Visual, linguistic data mining using Self-Organizing Maps. In Neural Networks (IJCNN), The 2012 International Joint Conference on (pp. 1-8). IEEE.

[12] Mäkinen, V. P., Forsblom, C., Thorn, L. M., Wadén, J., Gordin, D., Heikkilä, O., ... & Groop, P. H. (2008). Metabolic phenotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes. Diabetes, 57(9), 2480-2487.

[13] Tirunagari, S., Poh, N., Aliabadi, K., Windridge, D., & Cooke, D. (2014, December). Patient level analytics using self-organising maps: A case study on Type-1 Diabetes self-care survey responses. In Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on (pp. 304-309). IEEE.

[14] Tirunagari, S., Poh, N., Hu, G., & Windridge, D. (2015). Identifying Similar Patients Using Self-Organising Maps: A Case Study on Type-1 Diabetes Self-care Survey Responses. arXiv preprint arXiv:1503.06316

[15] Zarkogianni, K., Litsa, E., Vazeou, A., & Nikita, K. S. (2013, November). Personalized glucose-insulin metabolism model based on self-organizing maps for patients with Type 1 Diabetes Mellitus. In Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on (pp. 1-4). IEEE.

[16] B. Farran, A. M. Channanath, K. Behbehani, T. A. Thanaraj, Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwait– a cohort study., BMJ Open 3 (5). 2012

[17] N. Barakat, A. Bradley, M. Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, Information Technology in Biomedicine, IEEE Transactions on 14 (4) (2010) 1114–1120.

[18] M. F. Ganji, M. S. Abadeh, A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis, Expert Systems with Applications 38 (12) (2011) 14650 – 14659

[19] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, Johns Hopkins APL Technical Digest 10 (1988) 262–266.

[20] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients'data, Artificial Intelligence in Medicine 41 (3) (2015) 251–262.

[21] A. Al Jarullah, Decision tree discovery for the diagnosis of type ii diabetes, in: Innovations in Information Technology (IIT), 2011 International Conference on, 2011, pp. 303–307.

[22] Central Department of Statistics & Information (CDSI), Statistical yearbook, http://www.cdsi.gov.sa/ar/1805/, June, 2016

[23] R, (n.d.), r-project.org, Retrieved 15 November 2015, from https://cran.r-project.org/bin/windows/base/

[24] Machine Learning Group at the University of Waikato. (2015). *Weka 3: Data Mining Software in Java.* Retrieved December 23, 2015 from http://www.cs.waikato.ac.nz/ml/weka/

Finding Non Dominant Electrodes Placed in Electroencephalography (EEG) for Eye State Classification using Rule Mining

Mridu Sahu[1], N.K.Nagwani[1], ShrishVerma[2]
[1]Department of Computer Science and Engineering, NIT Raipur, C.G., India
[2]Departmet of Electronics and Telecommunication, NIT Raipur, C.G., India

***Abstract*—Electroencephalography is a measure of brain activity by wave analysis; it consist number of electrodes. Finding most non-dominant electrode positions in Eye state classification is important task for classification. The proposed work is identifying which electrodes are less responsible for classification. This is a feature selection step required for optimal EEG channel selection. Feature selection is a mechanism for subset selection of input features, in this work input features are EEG Electrodes. Most Non Dominant (MND), gives irrelevant input electrodes in eye state classification and thus it, reduces computation cost. MND set creation completed using different stages. Stages includes, first extreme value removal from electroencephalogram (EEG) corpus for data cleaning purpose. Then next step is attribute selection, this is a preprocessing step because it is completed before classification step. MND set gives electrodes they are less responsible for classification and if any EEG electrode corpus wants to remove feature present in this set, then time and space required to build the classification model is (20%) less than as compare to all electrodes for the same, and accuracy of classification not very much affected. The proposed article uses different attribute evaluation algorithm with Ranker Search Method.**

Keywords—Electroencephalography (EEG); Most Non Dominant (MND); Ranker algorithm; classification; EEG

I. INTRODUCTION

The MND feature subset selection is a part of corpus preprocessing, and it is useful for classification model building as a supervised learning .Classification is one of the task performed by data mining tools and applicable in different area of biomedical electrical devices such as EEG, ECG(Electrocardiograms), EMG(Electromyography), EOG(Electrooculography), Actigraph devices etc. These devices are popular devices for recognizing of different types of disease like Sleep Apnea diagnosis[1] using ECG, driving drowsing using EEG[2],EEG and electromyography (EMG) enable communication for people with severe disabilities [20],muscles activity using EOG[3],and military operation using EEG[21] etc. These are the motivational points for proposed work because the article finds those positions electrode they are less responsible for classification then the removal of those electrodes minimize the size of devices. The present work is performed with EEG electrode data having 16 electrodes and 14892 instances [4,5]. This uses the instance based classifier (K*), because based on statistic of data and nature of data spread over the corpus found it is best among other classifier the result of this present in literature [6, 7], [28], [33], [38]. Method selects either one electrode, two electrode or three electrodes based on how much search space the corpus wants to reduce. Its outcome generated from different attribute selection search with attribute evaluation techniques [8], [37]. Here it is 11 different combination of search with evaluation techniques. Then generating rules using Apriori algorithm [9], it gives frequent electrodes which are placed in ranked as a last four sequences, it also depends how many last feature ranked matrix the corpus wants to create. Here it is 11*4 , where 11 are a Row value and 4 is a column value. Ranker Search with different attribute evaluation algorithms shown in Figure [1].Rankers Algorithm is an algorithm useful for ranking of attributes by their individual evaluation [10]. Here three attribute evaluation methods are defined.

1) **Info Gain Attribute evaluation**: Evaluate the worth of an attribute by measuring the gain ratio with respect to the class.

2) **Classifier Attribute Evaluation**: Evaluate the work of n attribute by using a user specified classifier.

3) **OneR Attribute Evaluation**: Evaluate the work of an attribute by using the oneR classifier.

II. ASSOCIATION RULE MINING

Association Rule Mining is used here for obtaining frequent set they are correlated with each other using support and confidence parameters [11- 13].

Support is define as how frequently a specific item set occur in the data base (the percentage of transactions that contain all of the items in the item set, here the set of items are electrodes present in corpus and the transaction is the different method used for evaluation).

Confidence is the probability that items in RHS (Right Hand Side) will occur given that the items in LHS (left hand side) occurs. It Computed as

Confidence (LHS) =>Support (LHS U RHS)/ Support (LHS) Electrode1 => Electrode2 [0.588, 0.88]

If Electrode1 is selected in MND set, then Electrode2 also selected in MND set if it will satisfies minimum support and minimum confidence value. Left hand side electrode as Antecedent and Right hand side electrode [RHS] as consequent

frequency.

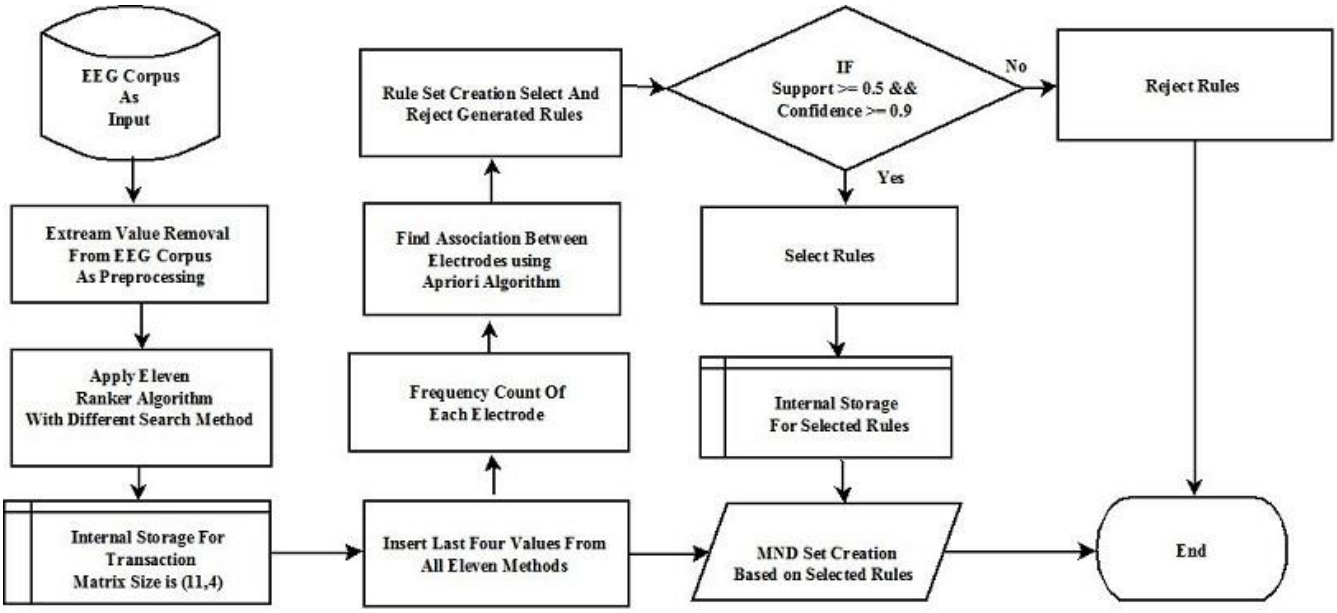


Fig. 1. Flowchart of Proposed Work

III. ELECTROENCEPHALOGRAPHY (EEG)

EEG is useful for measuring brain activity. During the test very little electricity is passed between the electrodes and skin. EEG usually takes 30-60 minutes. The technician will put a sticky gel adhesive on 16 to 25 electrodes at various spots on our scalp [14]. There are various spatial resolution of EEG systems like 10/20, 10/10, 10/5 systems as relative had surface based positioning system. The international 10/20 system a standard system for electrode positioning with 21 electrodes extended to higher density electrode such as 10/10 and 10/5 systems allowing more than 300 electrode positions [15].

Here the proposed methodology is used in 10/20 system with 16 electrodes (AF3, F7, F3, AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4).

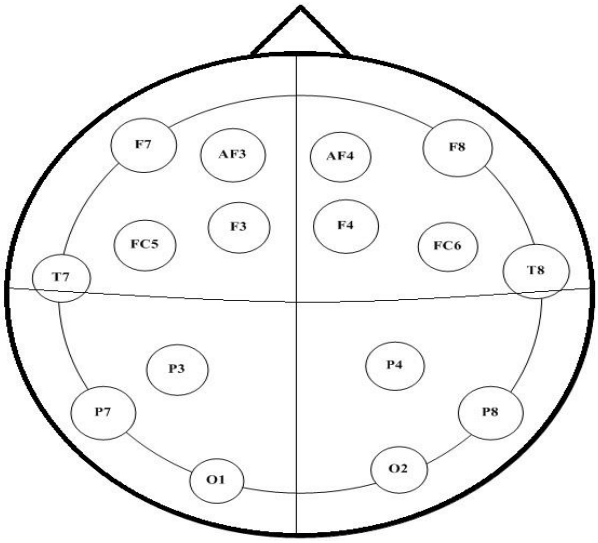


Fig. 2. EEG positioned electrodes 10/20

TABLE I. SEARCH METHOD USED WITH DIFFERENT ATTRIBUTE EVALUATORS

SEARCH METHOD + ATTRIBUTE EVALUATOR	TRANSACTION
Ranker + InfoGainAttributeEval	T1
Ranker + ChiSquaredAttributeEval	T2
Ranker + ClassifierAttributeEval	T3
Ranker + CorrelationAttributeEval	T4
Ranker + CVAttributeEval	T5
Ranker + FilteredAttributeEval	T6
Ranker + GainRatioAttributeEval	T7
Ranker + OneRAttributeEval	T8
Ranker + ReliefFAttributeEval	T9
Ranker + SignificanceAttributeEval	T10
Ranker + SymmetricalUncertAttributeEval	T11

IV. EEG CORPUS

The Corpus consists of 14980 instances with 15 features each (14 features representing the values of Electrodes and one as eye state (Boolean Variable)).Statistical Evaluation finds extreme values present in the corpus, here thirty eight instances **(186, 899, 10387, 10674, 10675, 10676, 10677, 10678, 10679, 10680, 10681, 10682, 10683, 10684, 10685, 10686, 10687, 10688, 10689, 10690, 10691, 10692, 10693, 10694, 10695, 10696, 10697, 10698, 10699, 10700, 10701, 10702, 10704, 10707, 10708, 10709, 11510 and 13180)** declared as extreme values in this, removal of it makes new corpus and it is having 14942 instances . The stored corpus as ordered to able to analyze temporal dependency 8220(55.01%) instances of the corpus corresponds to the eye open and 6722(44.99%) instances to the eye closed. EEG eye state dataset was donated by Rosler and Suendermann from Baden-Wuerttemberg Cooperative State University (DHBW), Stuttgart, Germany [4]. The output of the corpus "1" indicates

the eye-closed and "0" indicates the eye-open state.

V. Extreme Value Removal

The extreme value removal is a part of data cleaning step for data mining. The procedure for applying the extreme value theorem is to first establish that the function is continuous on the closed interval [16]. The next step is to determine the critical points in the given interval and evaluate the function at these critical points and at the end points of the interval. If the function f(x) is continuous on closed interval [a, b] then f(x) has both a maximum and a minimum on [a, b] [17]. In proposed method inter-quartile range [IQR] is used for extreme value calculations. IQR is major of variability based on dividing the dataset into quartiles [18].

VI. Feature Subset Selection

Feature Subset Selection is a task of data mining tool[25,26] ,it selects optimal feature subset for classifying the dataset but the literature shows the subset of optimal feature may or may not be optimal[19],[22-24]. The proposed work is searching Most Non Dominant features (MND) from the feature set. This performed by ranker algorithm and with different search methods. The outcome of this step is ranks of electrodes placed in scalp. Proposed work used different 11 algorithms for obtaining the ranks of electrodes (most to least dominant).

TABLE II. Transaction in Matrices with Four Last Dominant Attributes

Transaction	L4	L3	L2	L1
T1	O2	F7	FC5	F3
T2	O2	F7	FC5	F3
T3	FC6	O2	FC5	F7
T4	P7	O1	FC5	T7
T5	F7	AF4	F8	AF3
T6	O2	F7	FC5	F3
T7	F7	FC5	O2	F3
T8	FC6	O2	FC5	F7
T9	F3	F4	O2	P8
T10	P8	O2	F3	F7
T11	F7	FC5	O2	F3

VII. Classification

Classification is the task of data mining and it is a supervised learning. To classify EEG signals, various classification techniques present in literature [34-38]. The instances present in corpus for eye state recognition using EEG, these instances are classified in to two different classes, zero is for eye opened state and one is for eye closed state. The instance based classifier is a type of lazy classifier [27], and proposed method uses K* is a type of instance base classifier, after extreme value removal and attribute selection. The literature shows there are various statistical measures are used for analysis of classification outcomes generated from classification process [29-32].

VIII. Proposed Methodology for MND Set

The proposed methodology is use full for finding non-dominant feature from feature set. If "n" number of features are used for classification of eye state recognition then the space and time requirement is very high but if using less no of features obtained from proposed method then this will save time and space requirement. MND set electrodes are always a most non-dominant electrodes they are less responsible for classification accuracy. The flowchart shows in figure [1], and described steps shows, how to get MND from feature subset results generated from previous step.

S. No.	LHS		RHS	Support	Confidence	Lift
1	{}	=>	{F7}	0.8182	0.8181818	1
2	{FC6}	=>	{FC5}	0.1818	1	1.375
3	{FC6}	=>	{O2}	0.1818	1	1.375
4	{FC6}	=>	{F7}	0.1818	1	1.222222
5	{P8}	=>	{F3}	0.1818	1	1.571429
6	{P8}	=>	{O2}	0.1818	1	1.375
7	{F3}	=>	{O2}	0.5455	0.8571429	1.178571
8	{F3}	=>	{F7}	0.5455	0.8571429	1.047619
9	{FC5}	=>	{F7}	0.6364	0.875	1.069444
10	{O2}	=>	{F7}	0.6364	0.875	1.069444
11	{FC5, FC6}	=>	{O2}	0.1818	1	1.375
12	{FC6, O2}	=>	{FC5}	0.1818	1	1.375
13	{FC5, FC6}	=>	{F7}	0.1818	1	1.222222
14	{F7, FC6}	=>	{FC5}	0.1818	1	1.375
15	{FC6, O2}	=>	{F7}	0.1818	1	1.222222
16	{F7, FC6}	=>	{O2}	0.1818	1	1.375
17	{F3, P8}	=>	{O2}	0.1818	1	1.375
18	{O2, P8}	=>	{F3}	0.1818	1	1.571429
19	{F3, FC5}	=>	{O2}	0.3636	0.8	1.1
20	{F3, FC5}	=>	{F7}	0.4545	1	1.222222
21	{F3, F7}	=>	{FC5}	0.4545	0.8333333	1.145833
22	{F3, O2}	=>	{F7}	0.4545	0.8333333	1.018519
23	{F3, F7}	=>	{O2}	0.4545	0.8333333	1.145833
24	{FC5, O2}	=>	{F7}	0.5455	1	1.222222
25	{F7, FC5}	=>	{O2}	0.5455	0.8571429	1.178571
26	{F7, O2}	=>	{FC5}	0.5455	0.8571429	1.178571
27	{FC5, FC6, O2}	=>	{F7}	0.1818	1	1.222222
28	{F7, FC5, FC6}	=>	{O2}	0.1818	1	1.375
29	{F7, FC6, O2}	=>	{FC5}	0.1818	1	1.375
30	{F3, FC5, O2}	=>	{F7}	0.3636	1	1.222222
31	{F3, F7, FC5}	=>	{O2}	0.3636	0.8	1.1
32	{F3, F7, O2}	=>	{FC5}	0.3636	0.8	1.1

Fig. 3. Rule Generated from Apriori Algorithm

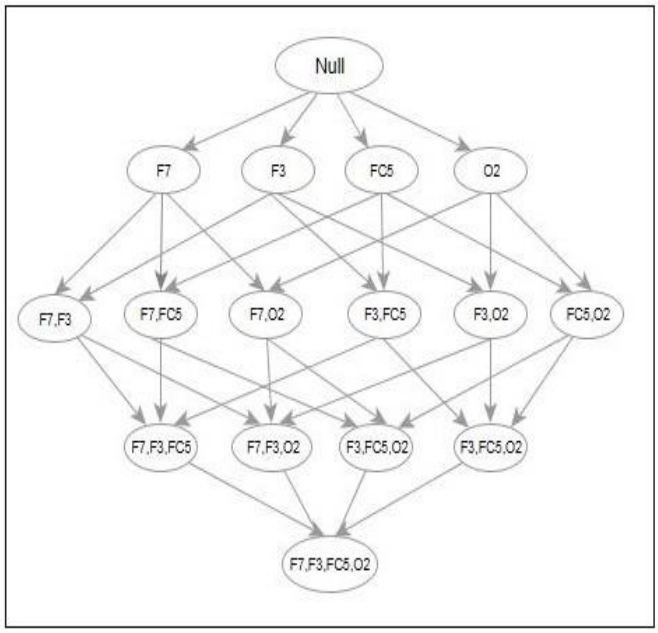


Fig. 4. Lattice occurs by Association Rule Mining

Algorithm 1: Most Non Dominant (MND) Feature Set Generation Algorithm

Input: D, I , T , C, MSupport.

//D = Set of Electrodes.
// I = Total Instances
//T = Transactions
//C = Corpus
// L= Class label {0, 1}
//MSupport=Minimum Support

Output: MND set

Segment(C); // *call this for creating a training, testing and validation set creation.*
For i=1 To 11 do
For j=1 To 4 do
T[R] [C] =LRS(C);
// *Call function for last 4 values from different feature ranker search with evaluation techniques*
//Transaction Matrix insertion for Item Set (Electrode) placed last 4 positions.

MND=Apriori(T,Msupport);
//Calling Apriori for frequent set generation for

End

Function Definition for Segment Creation from Corpus

Segment(C)
{
T= 60/100 *(C); *//Training Set Creation*
R= C-T;
Te = 50/100*(R); *//Testing Set Creation*
R= R-Te;
V= R; *//Validation Set Creation*
Return (T, Te, V)
}

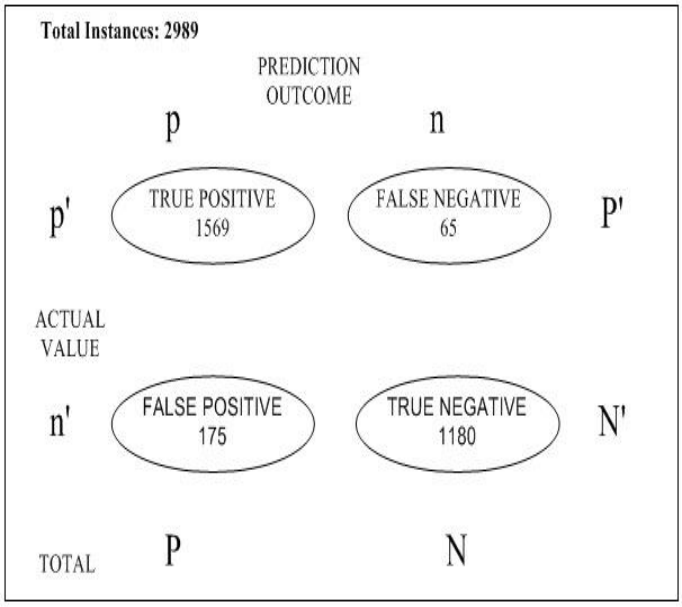


Fig. 5. Confusion Matrix on Removal of F7, FC5, O2

Function Definition for Last Ranked Set (LRS)

LRS(C)
{

For i=1 To 11 do
For j=1 To 4 do

Ran[j] =Ranker (i); *//Last 4 ranked value search and stored in array*
End //for End

Return (Ran[j]);

End
}

Function Apriori Algorithm for Frequent Set Mining

Apriori (T, mSupport)
{
//T is the database and mSupport is the minimum support
F_1= {frequent items};

For (k= 2; F_{k-1}! =∅; k++)
{
Ck= candidates generated from F_{k-1}
//Cartesian product F_{k-1} x F_{k-1} and eliminating any k-1 size item set that is not frequent

For each transaction do {
//increment the count of all candidates in Fk that are contained in T

F_k = candidates in C_k with minimum Support}

} end for inner for Return $\cup_k L_k$; }

IX. Result and Analysis

This study used Ranker Search with Attribute Evaluation technique for MND set creation shown in table[1], then for rule generated using association rule mining this task performed by using Apriori algorithm ,all the generated rules are shown in figure[3],and the lattice shown in figure [4],shows how many frequent set to be considered for rule generation , the rules which is having minimum support and confidence is highlighted in figure[3],this gives frequent items (Electrode) set ,here it is {FC5,O2,F7}. This set declared as MND set, removing of this electrodes from EEG corpus sufficiently decrease the space and time requirement to built the classification model. The accuracy towards the classification changed very less and this analysis outcome shown in table [3], figure[6]. The Confusion matrix shown in figure [5] and ROC curve shown in figure [7], evaluate the classifier performance here the classifier is Instance based classifier (K*), the classification accuracy is computed and it is mapped in table [3].

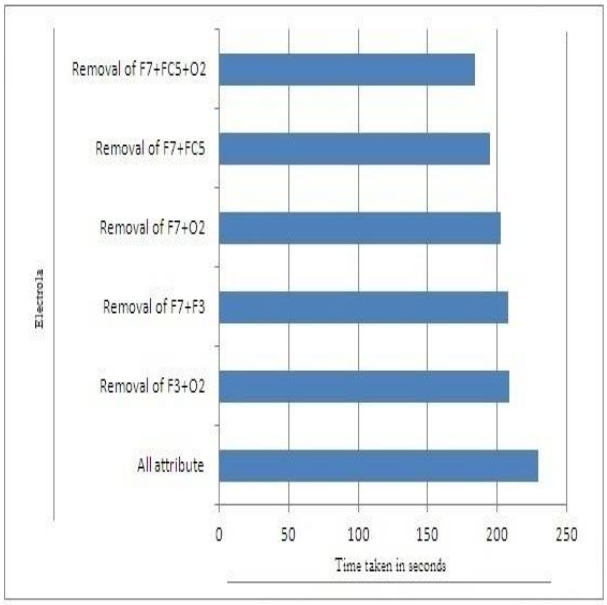


Fig. 6. Time duration with Removal of Different Attributes

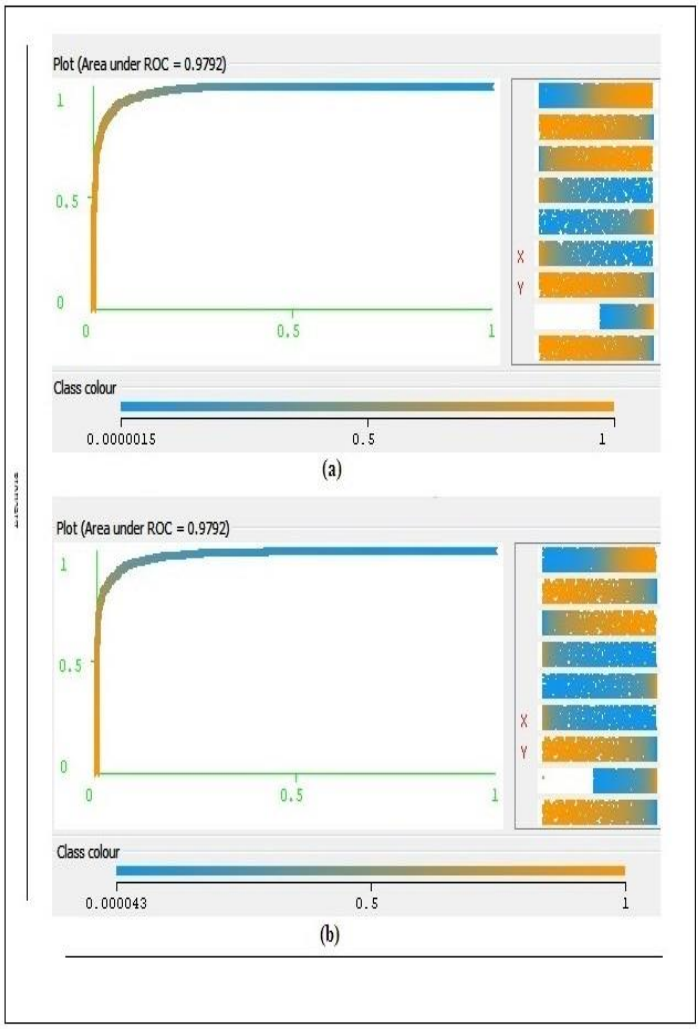


Fig. 7. ROC Curve removal on removal of F7, FC5, O2 (a) Threshold Value as 0 (b) Threshold value as 1

X. Conclusion

This is the first study to investigate the characteristics of Most Non Dominant feature from feature space they are less responsible to build the classification model, the MND set always gives concept which feature removal sufficiently reduce space and time requirement to build the classification model. This result is tested with EEG corpus to investigate eye state, either it is closed or open. Approximate 20% of time is saved by removal of these three most dominant features as compare to all attributes considered for classification.

TABLE III. RESULT ANALYSIS AFTER REMOVAL OF ATTRIBUTES FROM FEATURE SET FROM EEG DATA SET

Electrode Removal	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC	PRC	Accuracy %	Error %	Time Taken in sec.
Removal of F7+F3	**0.937**	**0.07**	**0.938**	**0.937**	**0.937**	**0.87**	**0.986**	**0.987**	**93.7103**	**6.2897**	**208.47**
Removal of F7+FC5	**0.927**	**0.08**	**0.929**	**0.927**	**0.927**	**0.85**	**0.983**	**0.984**	**92.74**	**7.26**	**194.78**
Removal of F7+O2	**0.939**	**0.066**	**0.94**	**0.939**	**0.939**	**0.88**	**0.985**	**0.986**	**93.9445**	**6.0555**	**202.76**
Removal of F3+O2	**0.947**	**0.057**	**0.948**	**0.947**	**0.947**	**0.89**	**0.989**	**0.989**	**94.714**	**5.286**	**209.24**
Removal of F7+FC5+O2	**0.92**	**0.089**	**0.921**	**0.92**	**0.919**	**0.84**	**0.979**	**0.98**	**91.9706**	**8.0294**	**184.06**

ACKNOWLEDGEMENT

This research is supported by the National Institute of Technology, Raipur and thanks to WEKA machine learning group as well as Rosler and Suendermann from Baden-Wuerttenberg cooperative state university (DHBW), Stuttgart, Germany for providing EEG corpus.

REFERENCES

[1] Jong-Min Lee, Dae-Jin Kim, In-Young Kim, Kwang-Suk Park, Sun I. Kim.: Detrended fluctuation analysis of EEG in sleep apnea using MIT/BIH polysomnography data, Computers in Biology and Medicine, Volume 32, Issue 1, Pages 37–47 (2002).

[2] M. V. M. Yeo, X. Li, K. Shen, and E. P. V.Wilder-Smith.: Can SVM is used for automatic EEG detection of drowsiness during car driving.Safety Science, vol. 47, no. 1, 115–124, (2009).

[3] Bulling A., Roggen D., and Troster G.: Wearable EOG goggles:eye-based interaction in everyday environments. ACM (2009).

[4] O. Rosler and D. Suendermann.: First step towards eye state prediction using EEG, Proceedings of the International Conference on Applied Informatics for Health and Life Sciences. Istanbul.Turkey(2013).

[5] T. Wang, S.U. Guan, K.L.Man and T.O. Ting.: EEG Eye State Identification Using Incremental Attribute Learning With Time-Series Classification. Hindawi Publishing Corporation Mathematical Problems in Engineering,vol- (2014).

[6] Mridu Sahu, N. K. Nagwani, Shrish Verma, Saransh Shirke,: An Incremental Feature Reordering (IFR) Algorithm to Classify Eye State Identification Using EEG, Volume 339, 2015, pp 803-811 edn., India: Information Systems Design and Intelligent Applications Advances in Intelligent Systems and Computing.

[7] Mridu Sahu, N.K. Nagwani, Shirish Verma, Saransh Shirke: Performance Evaluation of Different Classifier for Eye State Prediction using EEG Signal, Singapore: International Conference on Knowledge Engineering (ICKE 2015).

[8] G. Holmes, A. Donkin and I. H. Witten: Weka: A machine learning workbench, Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, Retrieved 2007-06-25, 1994.

[9] Sheng Chai, Jia Yang, and Yang Cheng: The Research of Improved Apriori Algorithm for Mining Association Rules. Proceedings of the Service Systems and Service Management, (2007).

[10] Ying Zhang , Almut Silja Hildebrand , Stephan Vogel: Distributed language modeling for N-best list re-ranking, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, July 22-23,(2006) Sydney, Australia

[11] Liu, B., Hsu, W., Ma Y.: Integrating Classification and Association Rule Mining. KDD, (1998).

[12] Rakesh Agrawal , Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the 20th International Conference on Very Large Data Bases, p.487-499, September 12-15, (1994).

[13] Sunita Sarawagi , Shiby Thomas , Rakesh Agrawal: Integrating association rule mining with relational database systems alternatives and implications, Proceedings of the international conference on Managementofdata,p.343-35,Seattle,Washington, USA(1998).

[14] Niedermeyer E. and da Silva F.L. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Lippincot Williams & Wilkins. ISBN 0-7817-5126-8 (2004).

[15] Towle, Vernon L.; Bolaños, José; Suarez, Diane; Tan, Kim; Grzeszczuk, Robert; Levin, David N.; Cakmur, Raif; Frank, Samuel A.; Spire, Jean-Paul: The spatial location of EEG electrodes: Locating the best-fitting sphere relative to cortical anatomy. Electroencephalography and Clinical Neurophysiology 86 (1): 1–6. Doi:10.1016/0013-4694(93)90061-Y. PMID 7678386 (1993).

[16] Stephen J. Robert: Extreme value statistics for novelty detection in biomedical signal processing. In Proceedings of the 1st International Conference on Advances in Medical Signal and Information Processing. 166—172 (2002).

[17] V. Glandola , A. Banerjee and V. Kumar,: Anomaly Detection: A Survey, (2007).

[18] Pour, T. Gulrez, O. AlZoubi, G. Gargiulo, and R. Calvo.: Brain-Computer Interface: Next Generation Thought Controlled Distributed Video Game Development Platform in Proc. of the CIG, Perth, Australia (2008).

[19] Pham and D. Tran.: Emotion recognition using the emotivepoc device. Lecture Notes in Computer Science, vol. 7667(2012).

[20] O. Ossmy, O. Tam, R. Puzis, L. Rokach, O. Inbar, and Y. Elovici.: MindDesktop - Computer Accessibility for Severely Handicapped. Proc. of the ICEIS. Beijing, China (2011).

[21] J. van Erp, S. Reschke, M. Groojen, and A.-M. Brouwer: Brain Performance Enhancement for Military Operators in Proc. of the HFM. Sofia. Bulgaria. (2009).

[22] Yu L., Liu H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. Mach Learn International Workshop Then Conf. 856(2003).

[23] N. Kwak and C.H. Choi.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE Trans. Pattern Analysis and Machine Intelligenc. vol. 24. no. 12,1667-1671,(2002).

[24] P. Langley.: Selection of Relevant Features in Machine Learning. Proc. AAAI Fall Symp. Relevance (1994).

[25] J. Jaeger, R. Sengupta, and W.L. Ruzzo.: Improved Gene Selection for Classification of Microarrays. Proc. Pacific Symp. Biocomputing, 53-64 (2003).

[26] R. Kohavi and G. John.: Wrapper for Feature Subset Selection. Artificial Intelligence. vol. 97. no. 1-2, 273-324 (1997).

[27] A. Webb.: "Statistical Pattern Recognition. Arnold" (1999).

[28] T. Cover and J. Thomas.: Elements of Information Theory. New York. Wiley. (1991).

[29] García, Salvador, et al.:A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing 13.10 , 959-977(2009).

[30] Sokolova, Marina, and Guy Lapalme.: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45.4, 427-437(2009).

[31] Demšar, Janez. : Statistical comparisons of classifiers over multiple data sets ,the Journal of Machine Learning Research 7 ,1-30(2006)

[32] T. Wang, S.U. Guan, K.L. Man and T.O. Ting.: EEG Eye State Identification Using Incremental Attribute Learning With Time-Series Classification. Hindawi Publishing Corporation Mathematical Problems in Engineering, vol- (2014).

[33] John G. leary, Leonad E Tigg.: K*: An Instance-based Lerner Using an Entropic Distance Measure, (1995).

[34] P.A. Est´evez, C. M. Held, C. A. Holzmann et al.: Polysomnographic pattern recognition for automated classification of sleep-waking states in infants. Medical and Biological Engineering and Computing, vol. 40, no. 1, 105–113 (2002).

[35] K. Polat and S. G¨unes: Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform .Applied Mathematics and Computation, vol. 187, no.2, 1017–1026 (2007).

[36] K. Sadatnezhad, R. Boostani, and A. Ghanizadeh.: Classification of BMD and ADHD patients using their EEG signals.Expert Systems with Applications,vol. 38, no. 3,1956–1963, (2011).

[37] N. Sulaiman, M. N. Taib, S. Lias, Z. H. Murat, S. A. M. Aris,and N. H. A. Hamid: Novel methods for stress features identification using EEG signals. International Journal of Simulation: Systems. Science and Technology. vol. 12, no. 1, 27–33 (2011).

[38] T. Nguyen, T. H. Nguyen, K. Q. D. Truong, and T. van Vo.: A mean threshold algorithm for human eye blinking detection using EEG. Proceedings of the 4th International Conference on the Development of Biomedical Engineering in Vietnam, 275–279.Ho Chi Minh City, Vietnam (2013).

[39] Thomas M. Cover, Joy A. Thomas: Elements of Information Theory Chapter 2 John Wiley & Sons, Inc. Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1 (1991).

Evaluation of Fault Tolerance in Cloud Computing using Colored Petri Nets

Mehdi Effatparvar
Sama Technical and Vocational Training College
Islamic Azad University, Ardabil Branch Ardabil, Iran

Seyedeh Solmaz Madani
Department of Computer and Informatics Engineering
Islamic Azad University, Ardabil, Iran

***Abstract*—Nowadays, the necessity of rendering reliable services to the customers in business markets is assumed as a crucial matter for the service providers, and the importance of this subject in many fields is undeniable. Design of systems with high complexity and existence of different resources in network cloud leads the service providers to intend to provide the best services to their customers. One of the important challenges for service providers is fault tolerance and reliability and different techniques and methods have been presented for solving this challenge so far.**

The method presented in this paper analyzes the fault tolerance process in interconnected network cloud in order to avoid problems and irreparable damages before implementation. In the offered method, the fault tolerance was evaluated aiding colored petri nets using Byzantine technique. Summary of results analyzed by cpntools and demonstrated reliability. It was concluded that upon increase of requests, the fault tolerance is reduced and consequently reliability is also reduced and vice versa. In other word, resources management is under impact of requested services.

Keywords—Cloud Computing; Fault Tolerance; Colored Petri Nets; Reliability

I. INTRODUCTION

Upon increasing development of information technology, a great volume of computations was created that its implementation by supercomputers was very costly and not available for all. Therefore, a new technology in the name of cloud computing was emerged. Using this technology, all systems existing in a network are assumed as a computational resource and may be used for computation. Accordingly, a great and powerful source is created using the network-connected systems resources that are able to perform complex and great operation.

The cloud computations provided the requirements for using computing resources shared by the computers in different networks that are different from each other in structure as well as geographically may be situated within different intervals.

Currently, various definitions of cloud computing have been presented. The definition used in this paper is as follows:

The cloud computing is a computational model therein lots of systems are connected to each other as private or public networks to provide dynamic and scalable infrastructure for the applied programs, data storage and files. Upon emergence of this technology, the computations cost, applied programs hosting, content storage, and delivery of services was reduced considerably. The idea of cloud computing principally is formed based on "reuse of technological capabilities" [1].

One of the important problems in cloud networks is management of time and resources, because as the viewpoint of different users, one of the most important criteria in selection of resources is implementation time. Whatever the implementation time is shorter and has lower cost, will be more appropriate as the viewpoint of users.

Cloud computing customers don't need to pay any cost for management, commissioning, maintenance and increasing the scale of their service for traffic control. They only ought to invest their cost and time for web development and pay easily the needed resources they want to provide for the web.

The cloud computing has frequent advantages, but also there are a few reasons for caution; risks such as losing services in case of occurring any problem or failure of cloud computing provider services or closure of their business. The legal problems are created when personal information has been stored in international level and security concerns are also created when the users have lost the control on their data protection. Hence, unilateral services are provided to the users for compensating the probable loss at the time of occurrence of a disaster. Fault tolerance is one of major concerns in applied programs implementation. In order to implement the applied programs correctly and reduce the effect of error on them, at first the error ought to be predicted, later managed and controlled. The fault tolerance methods are offered for prediction of these errors and performance of a proper action before fault.

The fault tolerance is the ability of a system for continuing the fulfillment of supposed tasks even in case of error. In fact, in this paper, byzantine fault tolerance technique in interconnected network cloud was used to avoid the irreparable loss, and colored petri nets were used to show the reliability of this technique.

In continue, the paper structure is as follows: in second chapter (2), the fundamental concepts related to cloud computing and fault tolerance, colored petri nets and reliability related to the subject are explained. In third chapter (3), a useful selection of previous studies and researches and similar to the subject of paper are explained briefly and usefully, and in fourth chapter (4), the proposed method is described. Finally, to show the accuracy of proposed method, in fifth chapter (5), a case study is provided for evaluation and modeling of proposed method.

II. Fundamental Concepts

Currently, different definitions of concepts are provided based on their application and performance. In this paper, the concepts were defined based on the application and type of usage. The cloud computing should not to be mistaken with network computations, because in network computations, the options and information are in general provided only on the servers of a specific company, whilst the cloud computations are much more bigger and include several companies and a large number of servers and equipment.

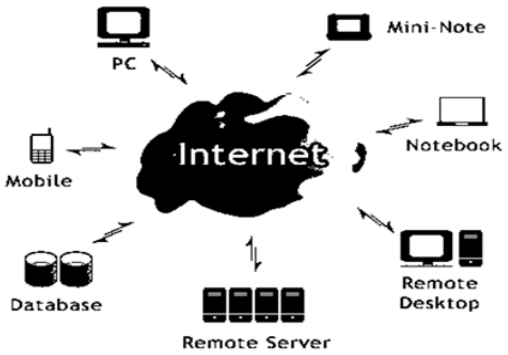


Fig. 1. Cloud computing model

A. Fault tolerance

Whereas in cloud networks one of the important topics is fault tolerance, in other word, it is necessary for the service receivers to have the reliability and security of their stored data in the highest level, therefore the service providers to retain their customers and render appropriate services ought to use specific mechanisms for applying security and high reliability factor.

The purpose of fault tolerance in cloud systems is that in case of occurrence of error, the system to have the ability of tolerating to the happened occurrence and can continue its process. In this type of systems, definitions of error, fault and failure are presented to associate the difference between them in the mind of reader. Therefore:

- Failure: whenever a system doesn't perform its expected job correctly, a failure has been occurred.
- Fault: The reason of failure occurrence is existence of an error in the system.
- Error: The reason of fault is existence of an error in the system.

B. Colored petri net

These nets present a graphic and clear exhibition of system together with a mathematical approach and can show the communication patterns, control patterns and information processes. These nets provide a framework for analysis, validation and evaluation of performance, the basis of petri nets has been formed based on the graph and informally it is a two-part directional graph consisted of two elements of time and transition. These nets are status-based and not event-based and it makes the explicit status modeling of each case possible. Petri nets provide models of structural and behavioral aspects of a discrete event system. Moreover, provide a framework for analysis, performance validation and evaluation, and reliability [2].

Colored petri nets figure 2, provide exacter models of complex asynchronous processing systems. In these nets, contrary to the petri nets, the tokens are distinguishable from each other, because each one of tokens has traits in the name of color. This type of nets provides exacter and detailer modeling from complex asynchronous processes. The tokens may be different from each other, so that a property called color is added to each token. The arcs may include mathematical phrases consisting of combination of color sets and variables related to them. Guard is a Boolean expression that is attributed to a transition and creates conditions for activation of input arc. In colored petri nets, each one of places, arcs and transitions depending on their color have their own guard [2].

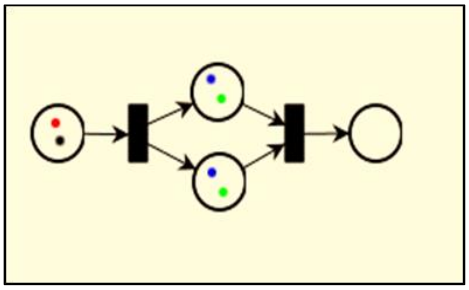


Fig. 2. A model of colored petri net

C. Reliability

However, nowadays the cloud nodes have been considered by the people so much, nonetheless one of the major challenges that cloud computing face it, is procedure of data protection and applying security for users processes. The security that is provided in the cloud environment is very important for the organizations and the people, because a few organizations suppose the transfer of important applied programs and their sensitive data to a general cloud environment as a great risk. Therefore, to reduce these concerns, a cloud provider must make this confidence that the customers can reserve their security and privacy control on the applied programs, so the cloud providers to convince their customers about the security issues ought to perform actions such as service level agreement. This agreement is a document that specifies the relationship between provider and receiver and indeed is a legal agreement between service provider and customer.

Following items are raised in the agreement about satisfying and ensuring the customer [3]:

- Identification and definition of customer needs
- Simplification of complex problems
- Reduction of grounds for conflict between users
- Persuading to discourse about encounters and disputes
- Omission of unreal expectations
- Presentation of a framework for easier perception

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

The cloud computing has not always provided continuous reliability.

III. Previous Works

Within recent years, various methods were propounded for evaluation of fault tolerance and reliability in cloud networks and these methods were studied in various researches. A few cases are analyzed in this paper.

Zh.Shan simulated and analyzed the performance in grid system in consideration of priority queues and applying priority for tasks and subtasks. This model has been made based on stochastic petri nets and analysis of its performance was provided based on properties of petri nets [4].

In another model, the modeling and simulation of tasks scheduling procedure in grind environment was analyzed and studied to reach the maximum operational power and reliability. This study is based on the queue systems and has no formal definition and ultimately the simulation was made based on the petri nets [5].

Genetic algorithm that is a hybrid evolutionary algorithm was presented for solving independent tasks scheduling problem in cloud network. The main objective in this algorithm is finding a solution therein the overall implementation time is minimum. Whereas genetic algorithm searches throughout the problem environment and is weak in local search, upon its combination to thermal simulation that is a local search algorithm, it is attempted to remove this defect and so the combination of advantages of these two algorithms were used. The chromosomes are exhibited by RFOH algorithm. Genetic algorithm includes a random population producer, elitist selection operator, repetitive combination and mutation aiding thermal simulator. Based on the fitness function, selection operator selects half of the best chromosomes from population. The combination operator is running and so the new children are generated for the next generation [6].

In a method, to evaluate the reliability of byzantine fault in a system, byzantine fault tolerance technique was used. In this method, interconnected network cloud was used. This method includes the characteristics such as an automatic job scheduling instrument that allows the job plan to be provided automatically to several heterogeneous clouds, and a massage system that creates the secure connection between the cloud and interconnected networks cloud and a fault tolerance adjudication system was presented as well. In this method, to prove the accuracy of method, two virtual machines were tested experimentally [7].

A method was presented for tasks distribution modeling and reliability calculation in cloud networks with star topology therein the tasks scheduling for reaching to the appropriate quality level as one of important and outstanding fields in cloud networks was analyzed. In this method, the reliability in grid services was examined and using the colored petri nets, a model was offered for computation. In this study, the grid environment has star topology and consequently RMS is connected with all resources in the grid. As specified, task of RMS is receiving tasks consisted of a series of subtasks from user and later distribution of the corresponding subtasks on the resources available on the grid. The general schema of job of tasks receipt and division thereof to subtasks by RMS and distribution among the resources is shown in fig. [3].

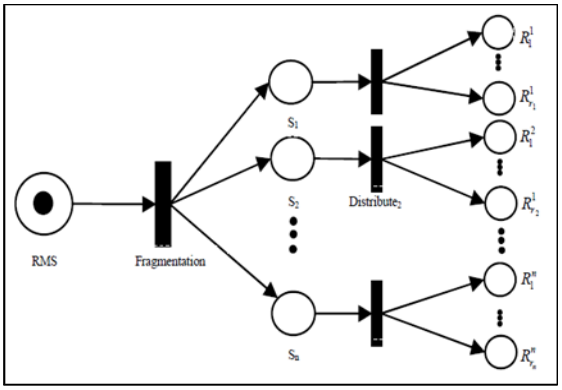


Fig. 3. Breaking the tasks and distribution of subtasks by RMS [3]

RFOH algorithm was presented for task scheduling with fault tolerance in computational grid aiding colored petri nets. This method records the history of fault event in resources in a table called fault event history table in information server of grid. Each row of FOHT table for each resource includes two columns; one column shows the history of failure event in that resource and the other one specifies the number of tasks implemented by that resource that was modeled by colored petri nets [8].

The modeling along with reliability computation methods with lower performance was studied and only a few case and specific studies were applied on the grid. In a few other methods that Y-Sh.Dai et al computed reliability in grid services, it was concluded that these studies provide the computations related to reliability and performance in grid services and only a solution for their calculation and maximization was presented and no virtual model thereof was presented [9].

In another method, the workflow procedure in grid environment was examined based on service and using UML charts [10]. In another method, workflow modeling in grid environment was discussed using simple petri nets and FhRG software. In this paper, the resulted petri network was presented, but this network was simple and allocated for the said software [11].

The hybrid HPSO algorithm in fact is combination of PSO and thermal simulation that upon using thermal simulation gets away from falling into a local optimum. To map the scheduling problem to a solution in this paper, the research environment is assumed as NxM dimensions, therein N is referred to the number of subtasks and M to the number of resources. Each particle is comprised of M pars and each part has N independent tasks that show the combination of N tasks on M machines [12].

A new scheduling algorithm was designed based on two basic Min-Min and max-Min algorithms so that the advantages of these two algorithms were used and yet the faults were covered. The criterion for selecting the proposed algorithm, the algorithm selected from two foregoing algorithms, is standard deviation of tasks completion period on resources. Min-Min

algorithm includes two stages; at the first stage, the waiting time for each subtask is specified and in the next stage, the tasks are sorted and marked based on job completion time on descending basis.

The tasks are allocated to their corresponding resources based on the priority of resources and this process is continued until all duties existing on MT are processed. Max-Min algorithm is similar to Min-Min algorithm with this difference that at the first stage, the tasks and jobs are sorted and marked based on the job completion time on ascending basis [13].

In another method, genetic algorithm was presented for solving the dependent tasks scheduling therein two important parameters of service quality including time and cost were taken into account. In this algorithm, instead of production of initial population randomly, disturbed variables were used. The combination of genetic algorithm advantages to the disturbed variables resulted in distribution of produced solutions by this algorithm throughout the research space and avoided the early convergence in the algorithm and the better solutions and products to be achieved within shorter time, and the algorithm convergence speed to be increased [14].

In a method, an algorithm was presented using queue theory for reduction of programs running cost in cloud network environment. The algorithm presented in this method is system-oriented and in addition to considering the performance and productivity factors focused on cost parameter that is raised more in business cloud environments [15].

IV. PROPOSED METHOD

The proposed method is formed based on byzantine fault tolerance. The interconnected clouds, in other word multiple clouds that each one has a specific policy and management and also is administrated differently were studied and analyzed using this technique. Colored petri nets were used to show the fault tolerance in cloud network aiding this technique. In addition, to avoid extra costs before implementation and execution phase, the respective method was simulated by colored petri nets aiding cpntools to evaluate the reliability in the proposed method.

Byzantine fault tolerance is not used for single cloud networks considering reducing the network reliability [7].

In continue, to introduce the method, at first byzantine tolerance and thereafter reliability computation procedure in the proposed method was analyzed and ultimately upon evaluation of fault tolerance and computation of reliability aiding colored petri nets and presentation of an executive model of proposed method in cpntools, obtained results were computed and exhibited.

A. *Byzantine fault tolerance technique in interconnected cloud computing*

Whereas cloud networks are assumed as one of main principles of computational systems, reliability in these systems is one of essential concerns, therefore the error potential in this type of systems is due to high simplicity. Hence, different techniques were developed for fault tolerance in these systems and in this paper byzantine technique was used.

In fact, byzantine fault is appeared when therein the server may provide any response optionally simultaneous to emergence of problems such as enemy attacks, user fault and software fault. A solution for avoiding this problem is use of byzantine fault tolerance (BFT). Byzantine fault tolerance figure 4, is the name of a problem that is occurred in distributed systems that is occurred due to creation of intentional failures.

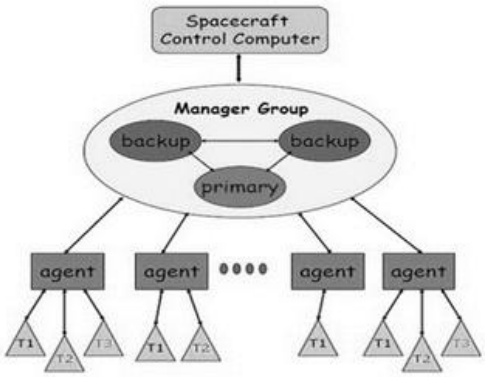


Fig. 4. Byzantine fault tolerance

FT-FC (Fault Tolerant – Fault Control) is one of fault tolerance frameworks in interconnected network cloud that that has various features such as having an automatic job scheduling tool that provides the possibility of showing job plan automatically to several heterogeneous clouds and a massage-based system that makes the safe communication between the cloud and FT-FC and also includes a fault tolerance adjudication system. Figure 5, exhibits the system created based on FT-FC framework. The modules shown in the figure are described in continue.

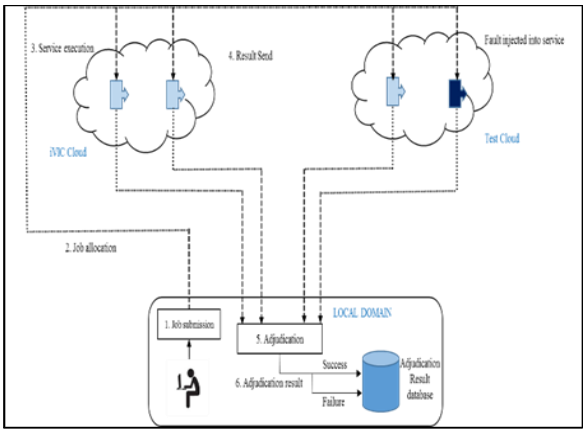


Fig. 5. System created based on FT-FC framework

In figure 5, it is shown for two clouds titled Test, iVIC with FT-FC framework. The purpose is expression of clouds independent from each other. The used modules respectively are as follows:

- Job submission: Division of jobs automatically among distinct clouds.
- Job allocation: The jobs are allocated to N-Copy services.

- Service execution: Each cloud has a backup that is switched thereto in case of occurring any error.
- Result send: Sending the results to adjudication node by services.
- Adjudication: Decision making of adjudication node in time management and accuracy of the result of executive process.
- Adjudication result: Specifying the accuracy and inaccuracy of obtained results.

The obtained results are stored in local database. In consideration of the foregoing, at the next stage, the reliability evaluation in this method should be assessed and in continue its procedure is described.

B. Reliability computation procedure

In this study, to compute the reliability, a failure rate is labeled to each one of nodes. However, all of these rates are designed dynamically in the model so that to be closer to the reality. To assess the reliability, equation (1) is used [16].

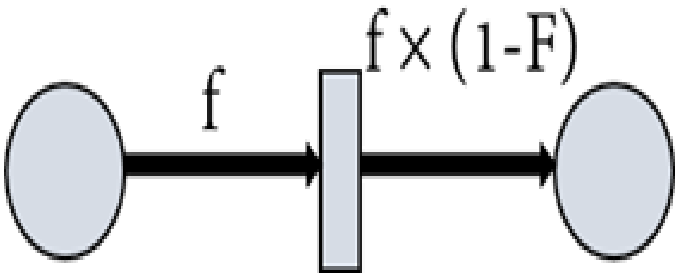


Fig. 6. equation (1) used to assess the reliability

In above equation, f denotes reliability and F the error rate or failure factor. After that each time in case of error occurrence, the reliability was computed, is averaged and updated with the new event of failure and new reliability.

In other word, in each cloud, an error factor or fault may be occurred due to byzantine error reasons. Equation (1) is used for evaluation of fault tolerance and a backup is used in each cloud for upraising the fault tolerance threshold limit of byzantine framework.

C. Creation of executable model

In this paper, colored petri nets and CPN Tools were used for creation of executable model. Reliability is evaluated in FT-FC framework using byzantine fault tolerance upon labeling the failure rate in each cloud. In continue, figure 6, shows the executable model of a distinct cloud network.

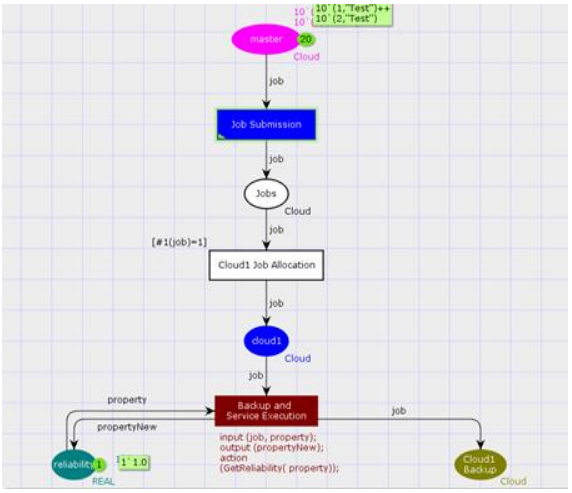


Fig. 7. Executable model of a distinct cloud network

V. CASE STUDY

In this paper, to show the applicability of method and accuracy of offered method, an example of cloud network with five distinct clouds were analyzed due to high complexity. Metric analysis of reliability is provided using the proposed method and simulation of executable model.

In order to evaluate the accuracy of performed simulation, Master provides five commands for sending to the clouds that each one of these commands were specified in the name of Job Properties. To show the dynamism in job, each command is repeated 10 times and 30 times; in other word, the possibility of data exchange between Master and clouds was shown.

figure 8, is a general schema of presented model that was simulated based on the explained assumptions and obtained results are analyzed in continue.

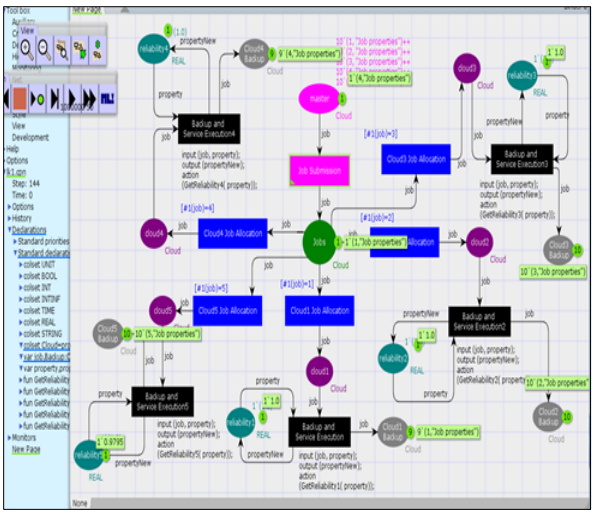


Fig. 8. General schema of presented model

In table 1, summary of results of execution of number of different jobs on the presented method is shown by byzantine fault tolerance technique. The failure rate in model is dynamic.

TABLE I. SUMMARY OF RESULTS OF PROPOSED MODEL

Id	Job Submission	Fault rate	Reliability	Average Reliability
cloud1	10	0.21	0.895	0.954
cloud2	10	0.11	0.945	
cloud3	10	0.13	1	
cloud4	10	0.09	1	
cloud5	10	0.14	0.93	
cloud1	30	0.21	0.368474398	0.777050282
cloud2	30	0.11	0.8945	
cloud3	30	0.13	0.874225	
cloud4	30	0.09	1	
cloud5	30	0.14	0.74805201	

According to the results obtained from simulation of studied method in this paper, the effect of increase and decrease of number of jobs on fault tolerance value and finally total reliability of system is observable.

In other word, the resources management considering the increase of requests will intensify the faults and the fault tolerance is declined. As a result, reliability of system is reduced. figure 9, shows the results of reliability evaluation with 10 and 30 tasks for each job in the simulated model.

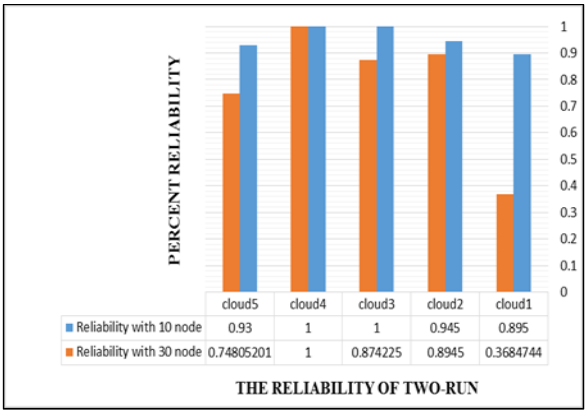


Fig. 9. Reliability evaluation by failure rate for 10 and 30 requests for each job

Byzantine fault tolerance n interconnected clouds that are stronger in management and have powerful infrastructure and are managed by the great cloud producers such as Amazon and Google is higher than byzantine fault tolerance with optional resources clouds that include lots of users and combined their computational resources.

In optional resources, due to dynamism of environment, reliability of a system is a critical issue and a byzantine fault tolerance system was presented for solving this problem. If number of our resources is 3F+1, F tolerates byzantine error. Commonly, infrastructure of clouds with optional resources is cheaper and more dynamic than greater clouds, but has lower power and reliability. Moreover, communication links between modules is not reliable [17].

In table 2, results obtained from execution of number of different works on byzantine fault tolerance method in clouds with optional resources is shown.

TABLE II. RESULTS OBTAINED FROM BYZANTINE FAULT TOLERANCE IN CLOUDS WITH OPTIONAL RESOURCES

Id	Job Submission	Fault rate	Reliability	Average Reliability
cloud1	10	0.21	0.7465553	0.85702715
cloud2	10	0.11	0.797493651	
cloud3	10	0.13	0.935	
cloud4	10	0.09	1	
cloud5	10	0.14	0.8060868	
cloud1	30	0.21	0.361424654	0.637855983
cloud2	30	0.11	0.64561702	
cloud3	30	0.13	0.576647074	
cloud4	30	0.09	1	
cloud5	30	0.14	0.605591166	

In figure 10, a comparison between proposed model and byzantine fault tolerance method is provided in clouds with optional resources.

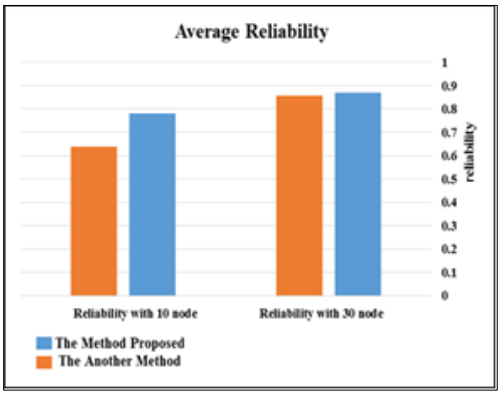


Fig. 10. Comparison between proposed model and byzantine fault tolerance method in clouds with optional resources

VI. CONCLUSION

The method presented in this paper analyzed byzantine fault tolerance in interconnected network clouds and colored petri nets were used to evaluate the reliability aiding this method that has strong math support. According to the results obtained from simulation, the fault tolerance using offered method is appropriate in comparison to the fault tolerance method in clouds with optional resources.

One of advantages of proposed method is combination of byzantine technique and colored petri nets that reliability was evaluated and analyzed using provided modeling. It was concluded that upon increase of requests, the fault tolerance is reduced and consequently reliability is also reduced and vice versa. In other word, resources management is under impact of requested services. One of limitations of proposed method for the subsequent activities is that in case the fault gets beyond the threshold limit and cloud tolerance, how reliability may change or which technique can be used to answer this question if the best option is tolerance or fault removal.

REFERENCES

[1] Asgharpour, Mohammadjavad, Multi-criteria Decision Making, 2011, 10th Ed., Tehran, Press Institute of University of Tehran;

[2] Kristensen L.M., Wells L. Lensen K., "Coloured Petri Nets and CPN Tools for modeling and validation of concurrent systems," International Journal on Software Tools for Technolog Transfer (STTT), no. Springer Berlin / Heidelberg, pp. 213-254, 2007.

[3] Entezari Maleki, Reza; Abdollahi Azgami, Mohammad, Subtasks Distribution Modeling and Computation of Reliability in Grid Services with Star Topology using Colored Petri Nets, March 2009, 14th Annual Conference of Iranian Computer Association;

[4] Shan Zh., Lin Ch., Ren F., Wei Y., "Modeling and Performance Analysis of a Multiserver Multiqueue System on the Grid", Proceedings of the The Ninth IEEE Workshop on Future Trends of Distributed Computing Systems (FTDCS'03), 2003.

[5] Caron E., Garonne V., Tsaregorodtsev A.,"Defination,modeling and simulation of a grid computing scheduling system for high throughput computing", Future Generation Computer Systems 23, pp. 968-976, Elsevier,2007.

[6] Cruz-Chavez, M. , Rodriguez-Leon, A. , Avila-Melgar, E. , Juarez-Perez, F. , Cruz-Rosales, M. and Rivera-Lopez, R. ″Genetic-Annealing Algorithm in Grid Envirinment for Scheduling Problems″, Security-Enriched Urban Computing and Smart Grid Communications in Computer and Information Science, Springer, Vol. 78, pp. 1-9.2010.

[7] Garraghan.p and Townend.P and J. Xu, "Byzantine FaultTolerancein Federated Cloud Computing" Proceedings of The 6th IEEE International Symposium on Service Oriented System Engineering, pp. 280-285, 2011.

[8] Khanli, L., Etminan Far, M., Ghaffari, A., Reliable Job Scheduler using RFOH in Grid Computing, Journal of Emerging Trends in Computing and Information Sciences, Vol. 1, No. 1, pp. 43- 47, 2010.

[9] Dai Y.S., Levitin G., "Optimal Resource Allocation for Maximizing Performance and Reliability in TreeStructured Grid Services", IEEE Transaction on Reliability, Vol. 56, No.3, September 2007.

[10] BenDaly Hlaoui Y., Jemni BenAyed L.; "Toward an UML-based Composition of Grid Services Workflows",Research Unit of Technologies of Information and Communication, Tunisia, ACM, AUPC'08, July 2008.

[11] Neubauer F., Hoheisel A., Geiler J., "Workflow-based Grid applications", Future Generation Computer Systems 22, pp. 6-15, Elsevier, 2006.

[12] Chen, R., Shiau, D. and Lo, SH., Combined Discrete Particle Swarm Optimization and Simulated Annealing for Grid Computing Scheduling Problem, Lecture Notes in Computer Science, Springer, Vol. 57, pp. 242 – 251, 2009.

[13] Etminani, Kobra, Design and Implementation of a Scheduling Algorithm in Grid based on two Min-Min and Max-Min Algorithms and Evaluation of its Performance, 2007, Thesis of Ferdowsi University of Mashhad.

[14] Gharooni fard, G., Moein darbari, F., Deldari, H., Morvaridi, A., Scheduling of scientific workflows using a chaos- genetic algorithm, Procedia Computer Science, Elsevier, Vol. 1, No.1, pp. 1445- 1454, 2010.

[15] Afzal, A., McGough, A.S., Darlington, J., "Capacity planning and scheduling in Grid computing environment", Journal of Future Generation Computer Systems 24 , pp.404-414, 2008.

[16] Farjaminejad.F and Harounabadi.A , "Modeling and Evaluation of Performance and Reliabilityof Component-based Software Systems using Formal"International Journal of Computer Applications Technology and ResearchVolume 3– Issue 1, 73 - 78, 2014.

[17] Zhang .Y and Z.Zheng and M. Lyu, "A Byzantine Fault Tolerance Framework for Voluntary-Resource Cloud Computing" IEEE 4th International Conference on Cloud Computing, pp. 444-451, 2011.

Reputation Management System for Fostering Trust in Collaborative and Cohesive Disaster Management

Sabeen Javed, Hammad Afzal, Fahim Arif
National University of Sciences and Technology
Islamabad, Pakistan

Awais Majeed
Bahria University
Islamabad, Pakistan

***Abstract*—The best management of a disaster requires knowledge, skills and other resources not only for relief and rehabilitation but also for recovery and mitigation of its effects. These multifaceted goals cannot be achieved by a single organization and require collaborative efforts in an agile manner. Blind trust cannot be applied while selecting collaborators/team members/partners therefore good reputation of a collaborator is mandatory. Currently, various Information and Communication Technology based artifacts, for collaborative disaster management, have been developed; however, they do not employ trust and reputation as their key factor. In this paper, a framework of reputation based trust management system is proposed for the support of disaster management. The key features of framework are Meta model, Reputation Indicator Matrix and Computational algorithm, deployed using Service Oriented Architecture. To evaluate the efficacy of the artifact, a prototype is implemented. Furthermore, an industrial survey is carried out to get the feedback on the proposed framework. The results support that the proposed reputation management system provides significant support in collaborative disaster management by assisting in agile and smart decision making in all phases of disaster management cycle.**

Keywords*—reputation; trust; reputation management; disaster management; collaborators; collaborative management

I. INTRODUCTION

Trust is such an investment in a society that returns strong and peaceful society whereas reputation is an asset that makes social interactions smooth and comfortable. The importance of trust and management in society, particularly in a disaster and emergency situation, cannot be ignored as managing such situations demands the collective role of a society. The stakeholders involved in such situations focus on the cohesive and collaborative efforts. These collaborative efforts demand sharing of skills, knowledge and other financial resources. It is also evident that collaborators involved in disaster management have different cultural, social and organizational background which may obstruct the efforts of collaboration.

Sharing of resources and skills in other domains like e-government, e-commerce and e-business, is made possible through Virtual Organization (VO), a form of collaborative networks. The focal point of VO is doing tasks and projects together by making collaboration among a number of entities, and ensuring agility [1]. It is a successful experiment for accomplishing the required goals by making a temporary alliance of partners [2] and the reason of its success is the opportunity of cohesive and collaborative efforts, it provides. Many Information and Communication Technology (ICT) based solutions for collaboration have been developed. These solutions have not only facilitated the VO form of collaboration in e-commerce and e-government but also in social networks. Moreover, to facilitate the collaborative and cohesive efforts for disaster management, several ICT-based solutions like SAHANA [3] and Oasis [4]are developed. These systems were more focused on considering the agile collaboration of VO, however, reputation and trust were not much focused in these systems. Since collaborative efforts engage precious resources, therefore, the role of trust and reputation becomes vital. Neither an organization nor a government can put a blind trust in anyone when there are issues of saving human lives and utilization of precious resources. Therefore, it is apparent that reputation based trust is such a building block of disaster management, without which the goal of efficient disaster management cannot be achieved in its true essence.

It is recognized that in ICT, considerable importance to the reputation of collaborators for disaster management is not given and the solutions presented in other domains cannot be directly applied to disaster management, keeping in view the dynamic context of trust.In this research, a conceptual framework for disaster management is proposed that lays emphasis on reputation and trust. The framework includes indicators/factors, having impact on the reputation of collaborators. These indicators/factors are extracted and deduced from the existing extensive literature available on disaster management. We have also gathered the requirements from stakeholders which are working in this domain. An algorithm to calculate the reputation score of collaborators/partners is proposed. Moreover, software architecture for reputation based trust management system is also proposed that helps in the exchange of required information among the heterogeneous systems of disaster management organizations/agencies. Using this heterogeneous platform, disaster management agencies will be able to get support in more collaborative and efficient. We have conducted a field survey involving a number of large and effective organizations involved in disaster management. The results of survey, on one hand, demonstrate the efficacy of our proposed framework.

The structure of rest of the paper is as follows: Section 2 provides a detailed literature review comprising some of the basic definitions such as Trust and Reputation Management, followed by related work in the field including Feedback Aggregation Models and existing ICT based Disaster

Management Systems. Proposed Framework is presented in Section 3, followed by implementation details. Finally, evaluation and results of survey are presented followed by conclusions.

II. LITEARTURE REVIEW

This section reviews the state of the art in the domain of trust, reputation, reputation management system, disaster management and virtual organization. Moreover, the applicability of VO to disaster management is discussed.

A. Trust

Trust is an uncertain situation, faced by an individual, in which beneficial or harmful result is reliant on the other person's behavior. Deutch defines trust as the situation when an individual faces an ambiguous path in which the outcome is either beneficial or harmful depending on the behavior of another person[5].The nature of trust is identified as dynamic in context[6]. Considering this characteristic, several trust management systems(applicable to different domains) have been developed. These systems are either reputation-based or policies-based. For reputation-based systems, trust is evaluated on the history of performance or interaction whereas for policy-based systems, trust assessment is based on credentials for issuance of access[6]. Furthermore, trust has three properties; identification, qualification and consistency. *Identification* includes traditional encryption and authorization; *Qualification* checks and analyze whether the subject entity has the required criteria; and *Consistency*, the most difficult one, is checked through formal certification or feedback [7, 8]. In another study[9], Gallivan classified trust into following five types.

1) Knowledge based trust is attributed with previous performance history of the subject entity.

2) Swift trust builds quickly within temporary teams. It is recommended when temporary teams have to achieve critical goals in short time.

3) Characteristic based trust is based on the qualities of the subject entity.

*4) Institutional based trust*develops with the help of guarantors.

5) Justice based trust illustrates procedural justice that is ensured through fair procedure.

B. Reputation and Reputation Management Systems

Reputation is the evaluation of trustworthiness of an entity that it can perform a task. A reputation system manages the reputation of entities [10]. Reputation is calculated based on the feedback of associative entities and assists in developing trust among the community/team members. Due to this, reputation based trust is given vital importance in virtual communities like social networks and e-commerce. Several reputation based trust management systems have been developed in different domains that facilitate the relevant authorities in decision making in different contexts.Amazon[11], epinions and eBayuse web based reputation systems. For reputation management, multi-criteria assessment approach is proposed by [2].

C. Feedback Aggregation Models

Feedback aggregation models are an integral part of reputation systems. These models help in aggregating and compiling feedback score obtained from different internal and external sources to produce a cumulative reputation score. Table 1 shows different types of feedback aggregation models with examples as described in[10].

TABLE I. TYPES OF FEEDBACK AGGREGATION MODELS

Models/ Network Type	Reputation Calculated through	Examples
Sum and Mean	Summation and then normalization through mean	Amazon, eBay and epinions
Flow Network	Transitive iterations	Google's PageRank
Markov Chain	Weighted Directed Graph	PowerTrust
Fuzzy Logic	Describe linguistically	REGRET
Bayesian	Probability Distributed Function	-

D. Importance of Trust in Collaborative Disaster Management

Collaborators have different cultural and organizational background, therefore, trust and reputation become important for effective collaboration. This understanding persists in the disaster management stakeholders and is identified in [12]. The author presents trust as a key element in quickly formed teams for DM. Considering the quick formation of emergency respondent teams, their study identified that swift trust is a key to achieve strong collaboration in quickly formed temporary teams.

Another study [13], carried out after 9/11 attack, identified that Trust is catalyst for effective collaboration in disaster management. Their work was carried out to reveal the causes behind the lack of unified command during the incident. The author suggested that multi-disciplinary training and education, through exercises, drills and other means, can improve the disaster management. The author also recommended that the mechanism for comprehensive and timely information sharing among the respondent agencies must be in place. Besides this, the formation of joint operation teams also improves the inter-agency collaboration. The author suggested that these are the ways for developing an understanding of the respondent agencies' culture, improve information sharing and enhanced integrated response.

In another research[14], the importance of inter-agency cooperation, during a disaster and emergency situation, was evaluated. It was explored that working together for managing disasters in an influential way greatly depends on the trust among the team members. It was also recognized that training through drills and other means also enhances the skills and knowledge of the respondents. This research also suggested that the use of ICT can improve the information sharing process. Realizing the importance of trust in DM, an information repository framework for emergency response information system was proposed in[15]. Risk engine is also incorporated in this framework for evaluating any kind of risks involved for decision maker. This is a generalized model and ignores the dynamic contextual nature of trust.

E. DM Collaboration through ICT-based Solutions

An Emergency Management Accreditation Program (EMAP) identifies thirty categories for emergency response operations. Twelve categories require collaborative and cohesive efforts. These categories are *mass sheltering, human services, public health and medical services, debris management, population protection, restoration of transport system, fatality management, fire protection, donation management, resource management, public works& engineering* and *damage assessment*.To handle the tedious tasks of disaster management, disaster management cycle is divided into four phases[16]:

1) Preparation and Planning: contingency plans are made and actions are defined for any expected disaster in this phase.

2) Response: this phase depicts all the operations that are carried out immediately after a disaster comes.

3) Recovery: it includes all the actions that are carried out to recover the infrastructure and the routine life.

4) Mitigation: actions are taken to alleviate and lessen the effects of a disaster.

By understanding the importance of cohesive and collaborative efforts in DM, various ICT-based solutions are developed for efficient disaster management. SAHANA and OASIS are two examples of it.Open Advance System for disaster and emergency management (OASIS)[16] is a co-funded project by European Commission. The purpose of this project is to facilitate the communication, collaboration and decision making in a disaster and emergency situation. Three hierarchical (strategic, operational and tactical) levels are defined for OASIS along with various other modules like IT Framework, Operational Tools and Tactical Situational Object (TSO)[17]. Moreover, OASIS also resolves the interoperability and security issues that can be faced by legacy IT systems. SAHANA is an open source disaster management system. It has various modules like Organization Registry, Missing Persons Registry, Request Management System, Volunteer Management, Shelter Registry, Situation Awareness and Inventory Management[3]. SAHANA has been deployed in various disaster situations and has proved to be a good facilitator in managing disasters.

F. VO and its Applicability to DM

VO is a temporary alliance of companies that can quickly share their core competencies to exploit the market opportunities[2]. This coalition forms another organization where company's boundaries are smudged. The advantages over the conventional way are the agility, the disperse-ability of collaborators, the integration of resources and the digitization of whole process [1]. For reputation building and managing, the approach of multi-criteria assessment is used in[2]. The lifecycle of VO consists of four phases:

1) Identification: goals and objectives are identified.

2) Formation: collaborators are evaluated and selected.

3) Operational: to meet the goals and objectives, operations are carried out by the selected collaborators.

4) Dissolution: when goals are met, temporary alliance is dissolved.

The stakeholders, involved in disaster management, recognize and admit that neither a government nor an organization can handle large-scale disasters on its own; therefore it is the responsibility of each individual and each organization to play its role collectively. Since VO also focus on the collaboration among a number of organizations/ collaborators and doing tasks together by introducing agility in the process; therefore, this concept seems applicable to DM.

III. PROPOSED FRAMEWORK

The proposed framework focuses on the reputation based trust management of collaborators involved in disaster management operations. This framework has three main characteristics: (i) it is applicable to all phases of disaster management cycle. (ii) it is applicable to the operational categories (defined in EMAP) that need collaboration. (iii) it is applicable to three hierarchical levels which are defined in the OASIS i.e. reputation can be managed at these three levels. The framework consists of four components:

1) Meta model for Reputation Indicators (RI)

2) Reputation Indicator Matrix

3) Computational Algorithm for Reputation Calculation

4) Service Oriented Architecture (SOA) for information extraction and exchange.

Further details of these components are described below:

A. Meta Model for Reputation Indicators

The first component of our framework provides the Meta model for reputation indicators. This Meta model consists of indicator/factors/criteria having impact on the reputation based trust of the collaborators. These factors are deduced from framework, protocols, standards and other published work of disaster management. The deduced factors are related to the operational categories defined in EMAP. Our meta-model divides these factors into two categories based on the four phases of disaster management cycle. The first category is called 'Pre-disaster' whereas other is called 'During & Post/After disaster'. Both categories consist of nine factors each. Our proposed Meta model is shown in the Fig. 1.

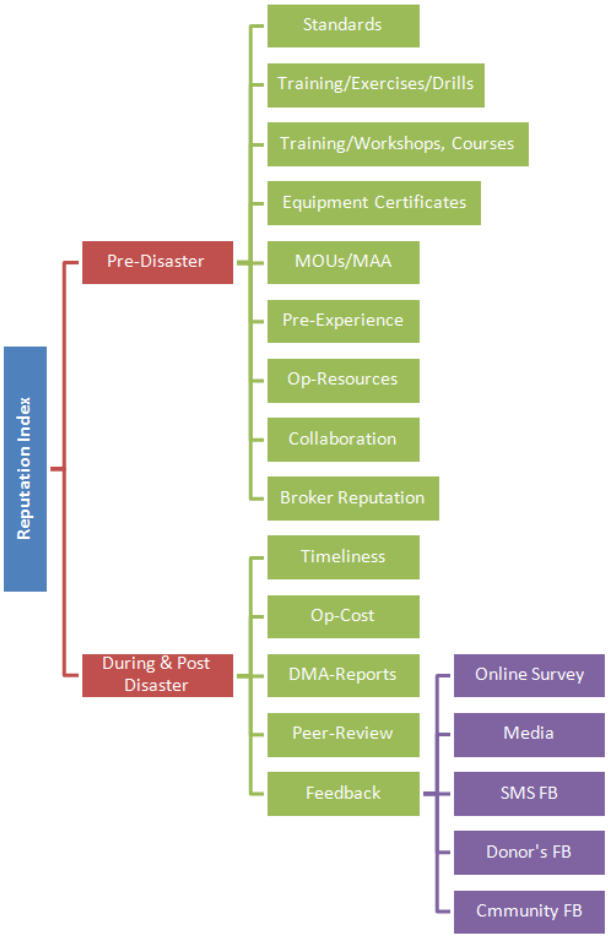


Fig. 1. Proposed Meta Model for Reputation Indicators

Meta model for reputation indicators consists of factors/indicators having impact on the reputation of collaborators. These indicators are extracted from the literature of disaster management and briefly discussed here.

Standards: Various standards (like Humanitarian Accountability Partnership and SPHERE) for managing disasters are developed. These standards are defined and developed by different disaster management stakeholders and ensure quality of humanitarian services. The organizations that follow such standards are considered trustworthy as it shows their level of preparedness.

Training: Training/educating of collaborators is considered extremely important for disaster management. The training can be either in the form of *Courses/Workshops* or in the form of *Drills/Exercises*. To make awareness among the collaborators, workshops and training courses for disaster management are important. Similarly, *Drills and Exercises* are meant to train the people/organizations for an emergency and disaster situation. The other purpose for this factor is to give them an opportunity to do coordinative efforts so that trusting relation can be built among themselves. Organizations like United Nations International Strategy for Disaster Reduction (UNISDR), National Disaster Management Authority (NDMA) of Pakistan and Federal Emergency Management Agency (FEMA) of United States regularly organize such exercises and drills. Moreover, these organizations also conduct workshops and offer courses to make awareness about disasters.

Collaborations: Collaborators, that have participated in their related operational category in a past disaster and performed better, are also trustworthy. Organizations having local or international collaborations with other disaster management organizations are also considered to be trustworthy.

Previous Experience: A collaborator's previous experience, in their related operational category, is also an important factor to evaluate their competency.

Broker Reputation is another concept that can assist in evaluating the reputation of collaborators. This concept is applied in ICT solutions developed for other domains like e-commerce and social networks. Broker reputation management system gives the details about the reputation of the collaborator under consideration. It is helpful to consider this reputation.

Operational Resources depict the competency of the collaborators. The quality and quantity of these resources play important role in a disaster situation. This factor is given considerable importance in the framework of NDMA and Hyogo. Memorandum of Understandings (MoUs) and Mutual Aid Agreements (MAAs) form another group of criterion that is regarded as significant in disaster management.

The second category of Meta Model helps in evaluating the reputation of the partner after they have performed on ground. In other words, this category considers the performance history of the collaborators. For this, *operational costs* and *timeliness* are key factors as efficient usage of resources and agile response are mandatory in a disaster situation. Besides this, Disaster Management Agency's report about the collaborator's performance is also important. *Peer's review*about its partner is also significant in establishing the reputation. *Feedback* from other stakeholders can also be an important tool for evaluating the reputation of the collaborators. These stakeholders include affected population, media, donors and different communities. The tasks and projects which are funded by various donor organizations should be monitored. To make it transparent, donors' feedback is important to establish the good reputation of the disaster management agency and its collaborators. Peers can also review their colleagues as they have worked on ground with them. *Media reports* are also important. Feedback by various communities is also required to evaluate the reputation of collaborators. In this regards, professionals, students and other volunteers can better evaluate the collaborators. Similarly, *online survey* and *SMS feedback* from affected population can be the other means to get feedback.

B. Reputation Indicator Matrix

The second component of the framework encompasses the aspects of VO. This component is called the Reputation Indicator matrix. It is identified that in e-commerce form of VO; financial, organizational, operational and third party recommendations are important perspectives that need to be considered while forming VO[2]. Therefore, to establish a

relationship of deduced Reputation Indicatorswith these perspectives, a Reputation Indicator matrix is developed as shown in Table 2. This matrix identifies all the perspectives that are covered by reputation indicators. These perspectives are ***F****inancial*, ***O****rganizational*, ***Op****erational*, ***T****hird party*, ***E****xternal* and ***C****ompetency of collaborators*,collectively named as **FOOTEC**. *Financial* perspective covers the financial soundness. *Organizational* perspective covers the management system and control of an organization. Functional reliability is covered in *Operational* perspective. Recommendations about the collaborators are included in *Third Party*. Factors that are external to the organization are covered in *External perspective* whereas *Competency* covers the skills of a collaborator in a particular area of its expertise.

TABLE II. REPUTATION INDICATOR MATRIX

Reputation Indicators		Type					
		F	O	Op	T	E	C
Before Disaster	Standards					Y	Y
	Exercises, Drills			Y			Y
	Workshop, Courses, Certificates					Y	Y
	Equipment Certification					Y	
	Mutual Aid Agreement		Y	Y			
	Previous Experience/ Existing Reputation in the related field		Y				Y
	Operational Resources	Y					Y
	Any sort of collaboration with government/non-governmental institutions at local level		Y	Y			
	Broker Reputation		Y		Y		
During and After Disaster	Timeliness			Y			Y
	Operational Cost			Y			Y
	Feedback		Y		Y	Y	
	Peer's Review		Y	Y			Y
	DMA Evaluation Reports	Y	Y	Y			Y

The RI Matrix facilitates in categorization of the RIs and helps in identifying the relationship of these categories with VO. The computational algorithm helps in quantifying the reputation score hence make the decision making easy and efficient for the relevant authorities. SOA approach for information extraction and exchange fulfills the constraints of integration and interoperability. This issue is further elaborated in Section IV.

C. Computational Model for Reputation Calculation

For each collaborator, reputation score of the identified factors is calculated using a computational algorithm that is based on sum and mean model. Sum and mean model is one of the state of the art models for reputation calculation. The algorithm for reputation calculation is shown in Fig 2.

D. SOA for Information Extraction

As the information related to these indicators is held by different systems and stakeholders, information exchange can be difficult. Therefore for integrating heterogeneous systems and multiple sources of information, a Service Oriented Architecture (SOA) is proposed for developing such a reputation management system. The proposed SOA model is shown in the Fig 3. It consists of following three components

1. Get the disaster type as different disaster types' demands different capacities, resources and skills.
2. Get/check the weight assign to the factors.
3. Get the degree of satisfaction/reputation score of all pre-disaster factors to anticipate the reputation.
4. Get the sum and mean 'δ' for each factor's degree of satisfaction/reputation score.
5. Now apply the equation 3.1 to aggregate the value of each factor depending on the weight assign to the criterion.

$$\rho(f_n) = \omega_n \delta \quad \rightarrow (3.1)$$

Where $\rho(f_n)$ indicates the aggregated reputation value of a factor f_n, ω_n represents the weight assign to that factor while δ represents degree of satisfaction of a factor by a particular partner.

6. Check the reputation of the subject entity in the broker database (if it exists).
7. Now calculate the anticipated reputation by summing up the calculated values of the factors.
8. Task is assigned to each partner and a unique id is assigned to this task.
9. When a partner has actually performed on the ground then get the reputation scores of post disaster factors for each assigned task.
10. Repeat the step 2-5 for post disaster factors.
11. Now add the anticipated and actual reputation scores.
12. Normalize the final reputation score like using z-normal distribution.

Fig. 2. Algorithm for Reputation Calculation

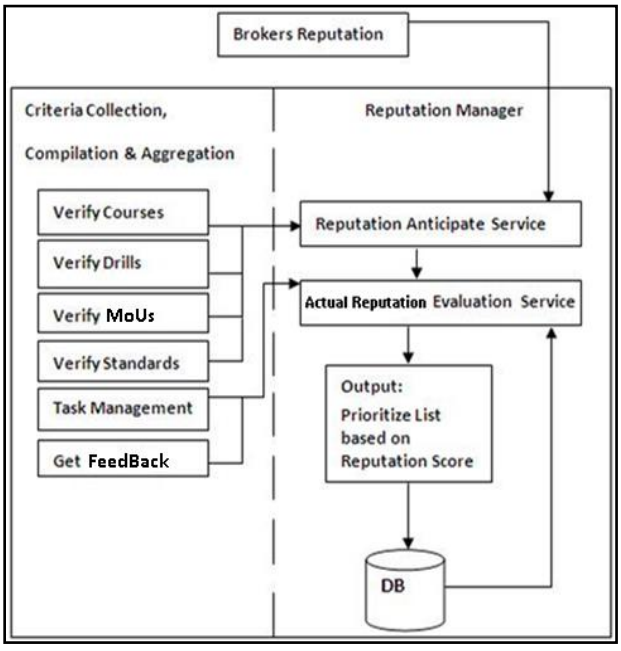


Fig. 3. SOA for Information Extraction and Reputation Calculation

1) Criteria Collection, Compilation and Aggregation: This component collects the feedback score from internal and external sources.The scores are then compiled and aggregated

to get a cumulative reputation score for different factors. Some web servicesfor the extraction of feedback score are also defined in this component.

2) Reputation Manager: This componentgets the cumulative reputation score of different factors.Mathematical model 1 mentioned in Fig 2 is then applied based on different weights assigned to these factors. This results in a prioritized list of partners based on their reputation score.

3) Broker Reputation: It checks the broker's database for reputation score of the partners

IV. Prototype Implementation

For the proof of concept, a prototype is implemented to evaluate the quality and effectiveness of the system in a disaster and emergency situation. Ahypothetical scenario is consideredto simulate a flash-flood situation. Flash flood is a type of flood that develops in few minutes. A web based application is developed to demonstrate the functionalities of our Reputation Management System (RMS). The web application comprises of a number of web services. These services are as follows:

1) Manage Partners
2) Manage Courses/Workshops
3) Manage Drills/Exercises
4) Manage Standards
5) Manage Operation Resources
6) Manage Equipment Certificates
7) Reputation Manager
8) Manage Tasks

With the help of aforementioned hypothetical scenario to handle disaster of flash floods, the usage of RMS is described in three phases while keeping in view the disaster management cycle. First is called *Pre-disaster*, consisting of preparation and planning; second is called *During-disaster* encompassing the response phase while recovery and mitigation is merged into third phase named as *Post-Disaster*.

A. Pre Disaster

A Disaster Management Agency (DMA) plans and prepares itself for the upcoming rainy season. From experience, the DMA has learnt that extensive rainfall in rainy season causes the flash flood in the low-lying areas which results in severe damages to the affected areas. It is also considered that the DMA was not able to handle such incidents alone. Considering these constraints, the DMA identifies collaborators/partners for relief and rehabilitation purposes. This process of identifying the potential threats of flash flood and based on this, the identification of collaborators is recognized as the *identification phase* of VO. The administration of the DMA decides to use the RMS for managing and tracking the reputation of these collaborators/partners. For this, the administrator uses the '*add a partner*' web service

Manage a Partner Service

The administrator can add any identified collaborator by entering the following data in the RMS for that collaborator.

- Partner ID:A unique numeric number given to collaborator.
- Partner Name: Name of collaborator
- Competency ID: Competency to be selected from a predefined list; e.g. Shelter Management.

The web interface developed for our application is given in Appendix in Fig A-1.

Manage Courses/Workshops Service

After assignment of an identity in the RMS, the collaborators use their applications to invoke their relevant web services of the RMS. It is the responsibility of the collaborators to add their data which is related to pre-disaster factors. This includes all the information of courses/workshops, drills/exercises, operational resources and all other pre-disaster factors which the collaborators possess. The interface for insertion of courses/workshops, developed for a collaborator, is shown in the Appendixin Fig A-2. The data required to be entered by collaborator is as follows:

- Course/Workshop ID
- Course/Workshop Name
- Conducting Organization

Similarly, a collaborator can enter data of its other competencies including Drills/Exercises, Standards, Operational Resources and Equipment Certificates. After insertion of pre-disaster data by the collaborators, the administrator of the DMA can assign score to each factor. The web interface is shown in Fig A-3.

B. During Disaster

During flash floods, thousands of people are evacuated through boats since roads and bridges are destroyed. Displaced people immediately need shelter, food supply and medical services. The DMA initially needs collaborators for the following tasks and missions as shown in Table 3. In this chaotic situation, the evaluation and assessment of potential collaborators is a challenge therefore the DMA uses the RMS for the selection of collaborator. The administrator invokes the Reputation Manager service for getting the prioritize list. While doing this, the DMA is in the '*formation*' phase of VO for disaster management. It will be in the '*operational*' phase of VO when tasks are assigned to the partners.

TABLE III. Summary of Tasks and Missions

Missions/ Tasks	Details
Required shelters	30, 000 tents
Water and food supply	For 50, 000 people
Health and medical	For 70, 000 people

Reputation Manager Service

The administrator selects the required competency, assigns weight to pre-disaster factors and then gets a prioritize list of competent partners along with their reputation score in the selected competency. The reputation score is normalized on the scale of 1 to 10. The interface for the reputation manager service is shown in Fig A-4 in Appendix A.

Manage Tasks Service

Based on the prioritize list, the administrator selects the collaborator and assigns the tasks by invoking 'assigning a task' service. For the purpose of assigning a task, the administrator only enters the relevant information in the RMS to track the record in future. The interface, for this purpose, is shown in the Fig A-5 in Appendix A.

C. Post Disaster

Recovery and mitigation phases start after the response phase. In these phases, the RMS collects, aggregates and compiles the task-based feedback about different collaborators. When the task is finished, team is dissolved that shows the *dissolution* phase of VO. Firstly, the administrator of DMA evaluates the collaborators' performance based reputation by considering timeliness and operational cost. For this, the administrator assigns task-wise score to each collaborator. The interface for this purpose is shown in the FigA-6 in Appendix-A.

Peers also assess their partners as they have worked on ground with their partners/collaborators. Peers can invoke *peers feedback service*. The feedback is given for the task in which they have coordinated and collaborated. The interface for this purpose is shown in Fig A-7 in Appendix-A.

V. Results and Survey

An online survey was conducted for getting feedback on the proposed framework. Thirty national and international organizations involved in disaster management were contacted. Positive response was received from six organizations who agreed to participate in the survey. These organizations include Pakistan's national organizations as well as international organizations having offices in Pakistan. These organizations include United Nations Development Program (UNDP), NDMA, Engineering Directorate, Earthquake Rehabilitation and Reconstruction Authority (ERRA), Focus and Concern. The survey consisted of a questionnaire form having one open-ended and nine closed-ended questions. The closed-ended question used different scale for getting the feedback about the extracted factors and the overall framework. The questions, their results and analysis of each question are shown graphically in this section. Different scales are used which are mentioned with each question.

Q1. How significant is the reputation of a partner/member organization for various disaster management activities (pre- and post disaster operations)?

Fig 4 illustrates that 100% participants considered the reputation of a partner as 'extremely important'. This undisputed response from the participants substantiates the idea of the current research.

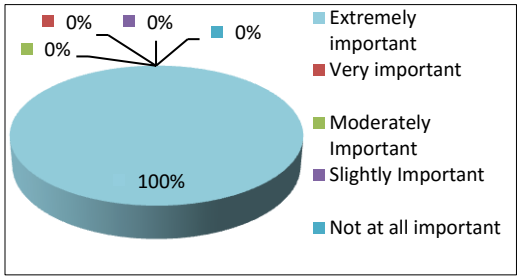


Fig. 4. Significance of Collaborators' Reputation

Q2. How often do you check or assess the reputation of your potential partners before doing any sort of collaboration with them?

Fig 5 reveals that 83% participants assess the reputation of their potential partners every time whereas 17% assess it very often. This response supports our background study in which it was identified that stakeholders involved in disaster management realize the value of reputation of the collaborators.

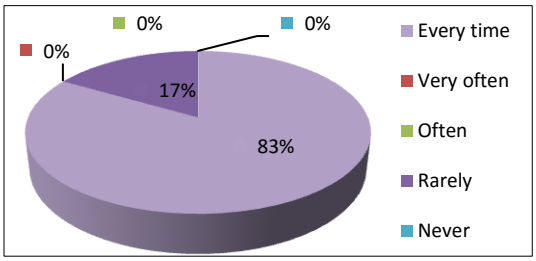


Fig. 5. Evaluation of Collaborators' Reputation

Q3. Does your organization follow any established policy/mechanism for checking/evaluating potential members/partners before doing any kind of collaboration for disaster operations?

It is exposed from the participants of the survey that their organizations have a policy/mechanism for checking/evaluating potential partners as shown in Fig 6. 83% participants informed that they have well-established mechanism for evaluating the partners whereas 17% respondents said they have a mechanism but not well-established. However, the participants were unwilling to expose their mechanisms due to restrictions of the rules and regulations of their organizations.

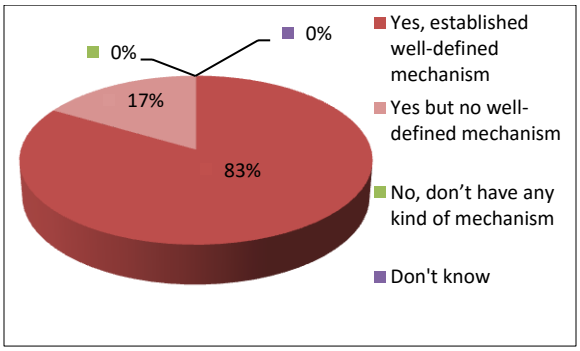


Fig. 6. Presence of Reputation Management Mechanism

Q4. In your opinion, managing the reputation and trust at three hierarchical levels (i.e. strategic, operational and tactical) is:

Figure 7 shows that the respondents evaluated the idea of managing the reputation, at three hierarchical level, as important. 50% considered it as extremely important, 33% as very important while 17% considered it as moderately important.

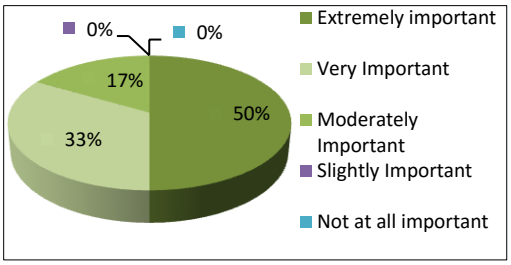


Fig. 7. Reputation Management at Three Hierarchical Levels

Q5. Classification of reputation factors into 'Pre- Disaster' and 'During & Post Disaster' operations is:

Fig 8 shows that 50% respondents evaluated our idea of dividing the extracted factors/indicators into two categories as 'absolutely appropriate' whereas 50% evaluated it as 'appropriate'. In short, all supported this idea of classification.

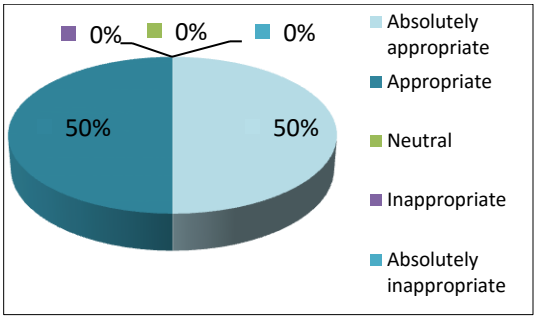


Fig. 8. Importance of Reputation Indicators' Classification

Q6. Please rate the following factors based on their helpfulness in reputation assessment. 1 represents 'extremely unhelpful', 2 represent 'moderately unhelpful', 3 represents neither 'unhelpful' nor 'helpful', 4 represents 'moderately helpful' and 5 represents 'extremely helpful'

The purpose of this question was to evaluate the relevance of extracted factors while evaluating the reputation of the collaborators/partners. The results are shown in Table 4 and 5 with respect to two groups defined in meta model.

Q7. How much weight (based on importance), would you like to give these factors to calculate overall trust/reputation of a member/partner organization?

To evaluate the helpfulness of each factor, three groups are defined based on the results of question 6 and the response of this question. Every group has its own criteria for the insertion in this group. The criteria are shown in Table 4.

TABLE IV. CRITERIA FOR PRIORITIZE GROUPS

Group	Criteria
High-priority	got more than 80% votes as 'extremely helpful' does not get any vote of neither a) 'neither helpful nor helpfulness' nor b) below this scale
Medium-priority	votes of either a) 'extremely helpful' or b) 'moderately helpful'
Low-priority	a vote of 'neither helpful nor unhelpful' is casted but no vote below this scale
Drop-out	get more than 50% of either a) 'neither helpful nor unhelpful' or b) below

TABLE V. CATEGORIZATION AND RANKING OF FACTORS IN PRIORITY GROUPS

Group	Rank	Factor's Name	Category	Average Weight
High-priority group	1st	Broker Reputation	Pre-disaster	5
		Previous Experience	Pre-disaster	5
		Community Feedback	Post-disaster	5
	2nd	Timeliness	Post-disaster	4.8
	3rd	Standards	Pre-disaster	4.7
Medium-priority group	1st	Exercises/Drills	Pre-disaster	4.5
		Operational Resources	Pre-disaster	4.5
		Courses/Workshops	Pre-disaster	4.5
	2nd	Local Collaboration	Pre-disaster	4.2
		Peer's Review	Post-disaster	42
Low-priority group	1st	Donor's Feedback	Post-disaster	4.5
		Operational Cost	Post-disaster	4.6
	2nd	Surveys	Post-disaster	4.3
		Equipment Certificates	Pre-disaster	4.3
	3rd	MoUs/MAAs	Pre-disaster	4.2
	4th	DMA Report	Post-disaster	4
	5th	Media	Post-disaster	3.7
Drop-out	--	SMS Feedback	Post-disaster	-

Based on these criteria, Table 5 shows the ranking of each factor along with the average weight assigned by the respondents to each factor.

Q8. Would you like to suggest any new factor which can help in assessing the reputation of a partner? Please also describe the reason.

The respondents' suggestions along with the remarks on the suggestions are shown in Table 6.

TABLE VI. SUGGESTIONS FROM RESPONDENTS

Suggestions	Remarks
Staff competency and aptitude	A detailed discussion, with the respondent, was carried out regarding this suggestion. He agreed that
No	The respondent was satisfied with the deduced factors.
Partners style of data management and reporting is very much important	This suggestion is point towards standardized data management and reporting which has been covered in the pre-disaster factor i.e. standards
The factors identified are all-encompassing. However, one might	This suggestion points towards the standards a collaborator is

add information management systems, reporting mechanisms and quality control / monitoring and evaluation capacity as one additional factor	following.
Community awareness and involvement in all kind of DRR and DRM activities	Community involvement is being done through feedback and awareness through courses/workshop and drills/exercises.

Q9. Do you think a combination of Information Technology and people's activities in disaster response operations can play an important role in managing reputation and trust of involved organizations?

50% participants, in response to this question voted for 'extremely helpful', 33% supported it as 'very important' whereas 17% evaluated it as 'moderately important' as shown in Fig 9. Such a response from the participants depicts that the introduction of ICT-based solution for reputation based trust management will be helpful for the stakeholders involved in disaster management.

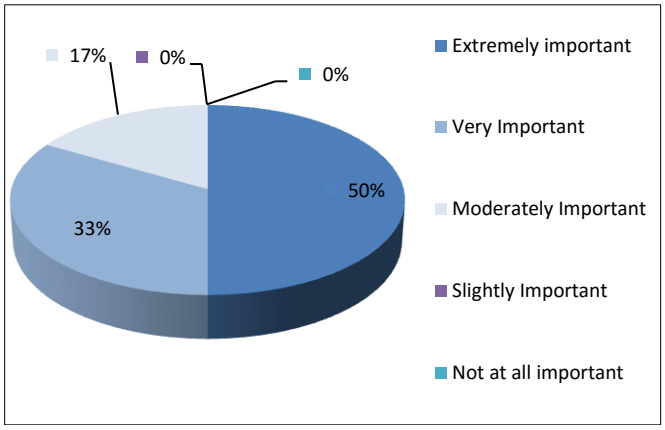


Fig. 9. Importance of Information System for Reputation Management

Q10. Do you agree that the integration of existing Information Systems of participating organizations and/or development of Online Integrated Information Services for disaster response organizations can be helpful in trust/reputation management?

All participants were agreed for introducing the IT standards in the integration of reputation management system with the legacy system of disaster management organizations as shown in Fig 10.

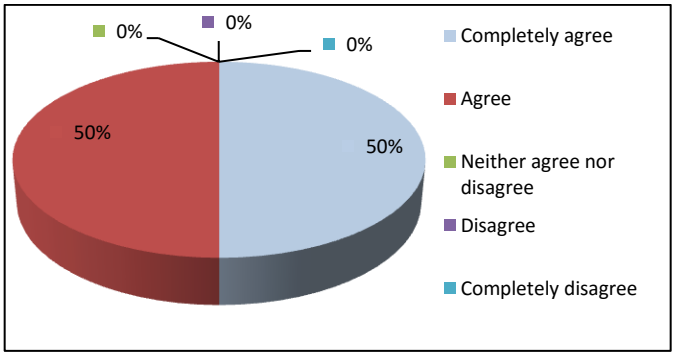


Fig. 10. Need of Online Integrated System

VI. Conclusion

This research presents a substantial base for reputation management of collaborators involved in disaster management. Keeping in view, the state of the art approach for reputation management in different domains, multi-criteria reputation assessment and an aggregation model for reputation calculation are proposed. The multi-criteria of the proposed framework illustrate the priorities of disaster management's stakeholders. These are the characteristics that are expected to be present in the collaborators hence help in establishing the reputation of the collaborators. The aggregation model helps in quantifying the reputation score in such a manner that decision making becomes easy and efficient for the relevant authorities.

The applicability of the proposed framework is evaluated through prototype implementation. It is evident from the prototype that information exchange among heterogeneous systems is not problematic hence multi-criteria monitoring is made possible for the stakeholders of disaster management. Besides this, the survey results show substantial support from the participant organizations.

Collaborators often fail in performing their duties while managing disasters. It is concluded that the reputation management system acts as a silent observer of their reputation and helps the authorities in smart decision making for disaster management. This research contribution can be refined by introducing risk management aspect in the framework hence the basis for reasoning about the involved risk (while selecting collaborators) can be presented. Moreover, information extraction (about reputation of collaborators) from media reports and social networks can also be searched semantically which can help in being more transparent.

References

[1] F. Ye and G. Li, "Study on virtual organization product development based on total lifecycle agile manufacturing in the networked collaborative environment," in Technology and Innovation Conference, 2006. ITIC 2006. International, 2006, pp. 556-561.

[2] T. J. Winkler, J. Haller, H. Gimpel, and C. Weinhardt, "Trust Indicator Modeling for a Reputation Service in Virtual Organizations," in ECIS, 2007, pp. 1584-1595.

[3] M. Careem, C. De Silva, R. De Silva, L. Raschid, and S. Weerawarana, "Sahana: Overview of a disaster management system," in Information and Automation, 2006. ICIA 2006. International Conference on, 2006, pp. 361-366.

[4] O. Konsortium, "OASIS project Executive Summary," ed, 2005.

[5] M. Deutsch, "Cooperation and trust: Some theoretical notes," 1962.

[6] S. Hall and W. McQuay, "Review of trust research from an interdisciplinary perspective-psychology, sociology, economics, and cyberspace," in Aerospace and Electronics Conference (NAECON), Proceedings of the IEEE 2010 National, 2010, pp. 18-25.

[7] K.-J. Lin, H. Lu, T. Yu, and C.-e. Tai, "A reputation and trust management broker framework for web applications," in e-Technology, e-Commerce and e-Service, 2005. EEE'05. Proceedings. The 2005 IEEE International Conference on, 2005, pp. 262-269.

[8] J. Lee and K.-J. Lin, "Context-aware distributed reputation management system," in e-Business Engineering, 2008. ICEBE'08. IEEE International Conference on, 2008, pp. 61-68.

[9] M. J. Gallivan, "Striking a balance between trust and control in a virtual organization: a content analysis of open source software case studies," Information Systems Journal, vol. 11, pp. 277-304, 2001.

[10] O. Hasan, "Privacy preserving reputation systems for decentralized environments," Thse de doctorat en informatique, INSA de Lyon (Sep. 2010), 2010.

[11] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," Internet Computing, IEEE, vol. 7, pp. 76-80, 2003.

[12] R. Zolin, "Swift trust in hastily formed networks," DTIC Document2002.

[13] T. L. Currao, "A new role for emergency management: fostering trust to enhance collaboration in complex adaptive emergency response systems," DTIC Document2009.

[14] Østensvig, "Interagency cooperation in disaster management: partnership, information and communications technology and committed individuals in Jamaica," Masters thesis at the Norwegian University Of Life Sciences, Ås, Norway. At http://www. islandvulnerability. org/caribbean. html# jamaica, 2006.

[15] Eryilmaz, M. Cochran, and S. Kasemvilas, "Establishing trust management in an open source collaborative information repository: An emergency response information system case study," in System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on, 2009, pp. 1-10.

[16] M. Couturier and E. Wilkinson, "Open advanced system for improved crisis management (OASIS)," in Proceedings of the 2nd International Information Systems for Crisis Response and Management (ISCRAM) Conference, Brussels, Belgium, 2005.

[17] Henriques and D. Rego, "OASIS Tactical Situation Object: a route to interoperability," in Proceedings of the 26th annual ACM International Conference on Design of Communication, 2008, pp. 269-270.

APPENDIX

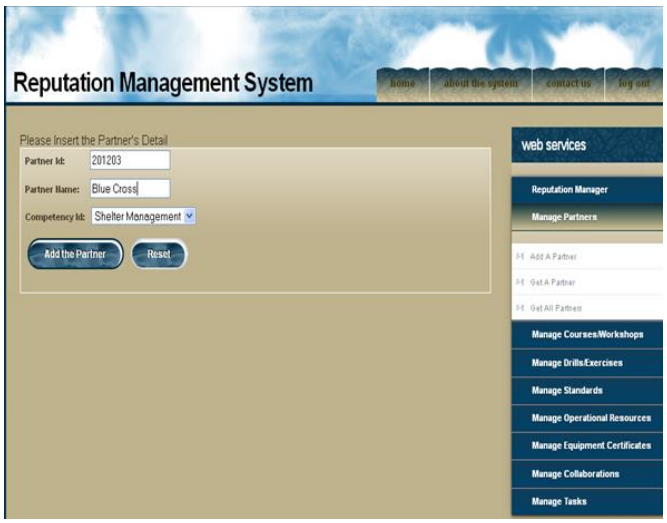


Fig. A-1. A web interface showing 'Add a Partner' service

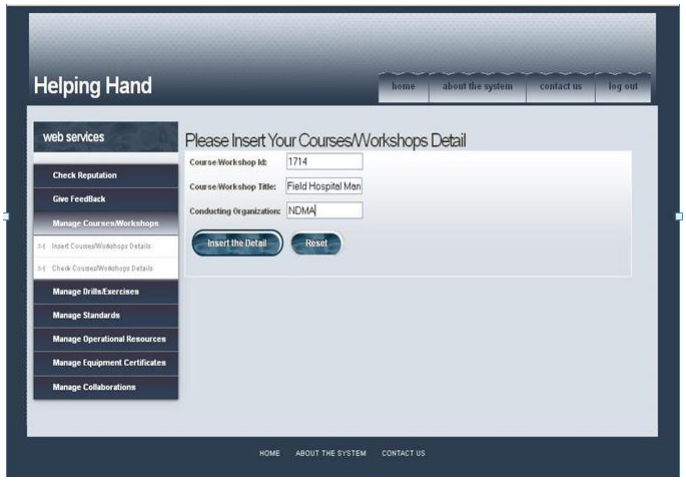


Fig. A-2. A web interface for collaborators where they can Insert Courses/Workshop Data

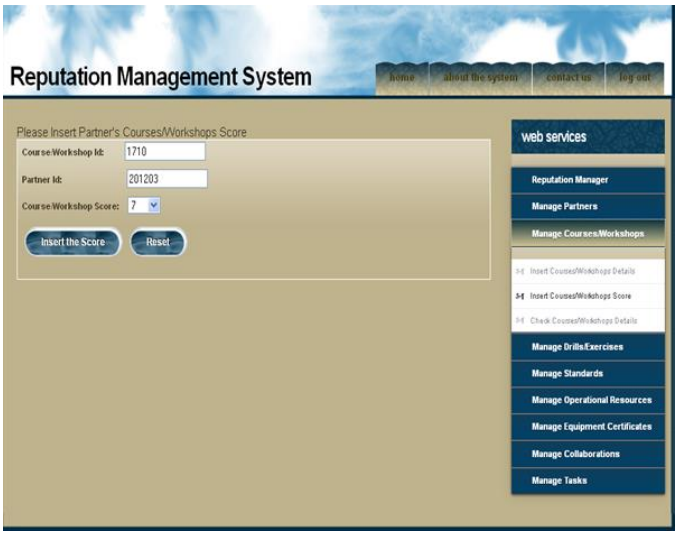


Fig. A-3. Disaster Management Agency can assign Reputation Score to Collaborator's Course/Workshop

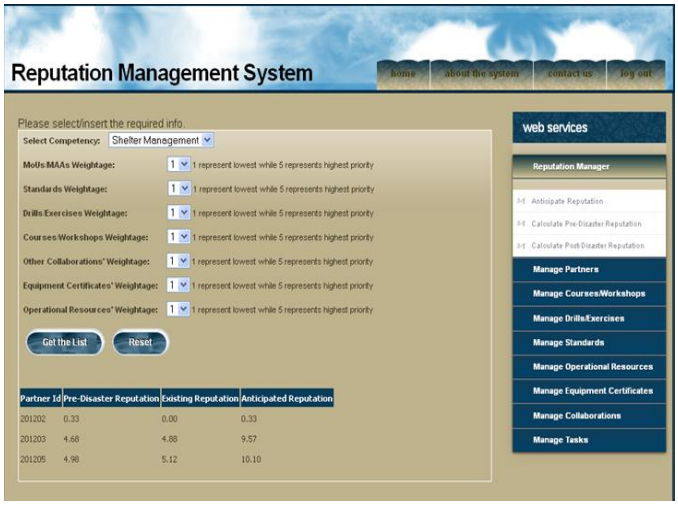


Fig. A-4. Administrator can enter weights for different factors. Based on these values, a Prioritize List of partners is generated based on Reputation Score

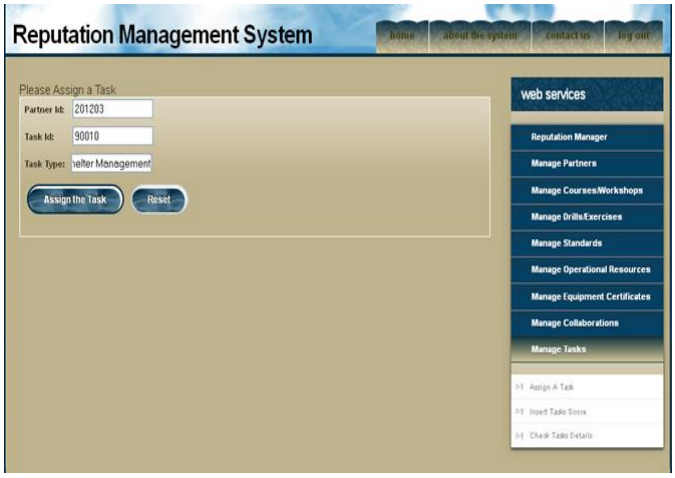


Fig. A-5. A web interface of 'Assign Task to a Collaborator' service

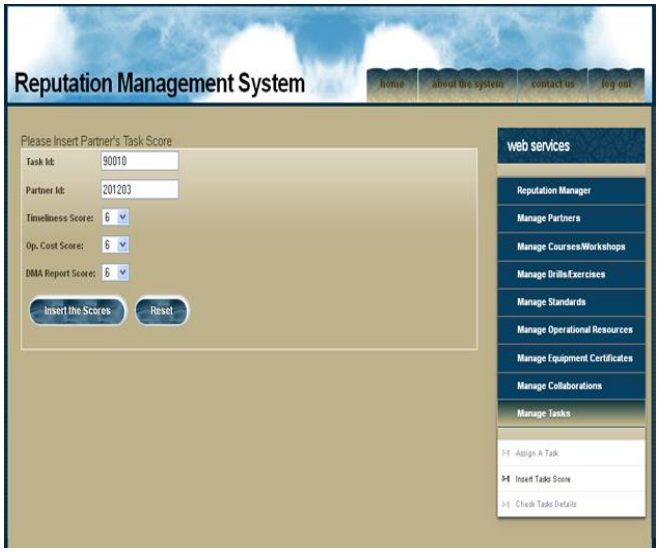


Fig. A-6. A web interface for admin to enter 'Task Score' to different collaborators in Post-Disaster phase

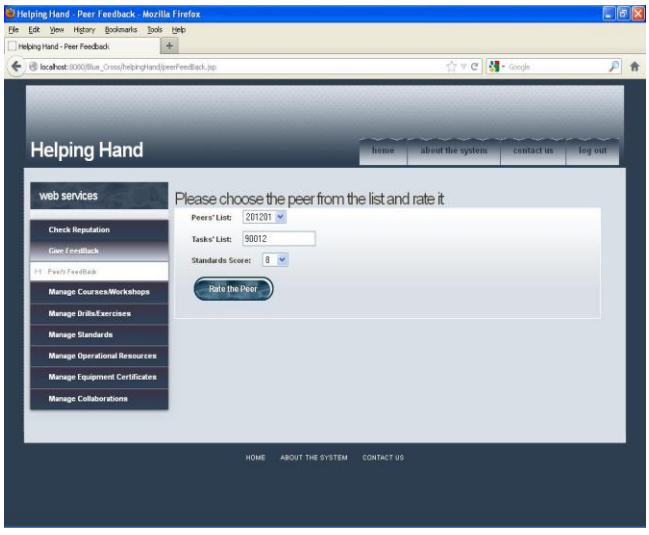


Fig. A-7. Peer's Feedback

Indirect Substitution Method in Combinable Services by Eliminating Incompatible Services

Forough Hematian Chahardah Cheriki
Department of Computer Engineering
Yazd Branch, Islamic Azad University
Yazd, Iran

Sima Emadi*
Department of Computer Engineering
Yazd Branch, Islamic Azad University
Yazd, Iran

***Abstract*—Service-oriented architecture is a style in information systems architecture with the aim of achieving loose coupling in communication between software components and services. Service, here means software implementation, is a well-defined business function that can be used and be called in various processes or software. An organization can choose and composite the Web services that fulfill its intended quality of service. As the number of available Web services increases, choosing the best services to composite is challenging and is the most important problem of service composition. In addition, due to the utilization of systems in dynamic environments, service characteristics and users' needs are constantly faced with changes which lead to deterioration of service, unavailability and quality loss of services. One of the ways to deal with this challenge is substitution of a Web service with another service, which is done at the runtime and dynamically. Substitution is both direct and indirect. Though there are many related works in the field of direct substitution, still no work is done for explaining substitution based on the indirect method, and works were conducted only on direct substitution. In this method, there are many problems such as the incompatibility of important services in composition. To solve the problems in this method and other challenges in this paper, considering a subset of inputs and outputs, qualitative parameters and service composition, simultaneous and dynamic service composition and use of the fitness function of genetic algorithm to compare the compositions are done. In addition, in substitution, a table which contains the best possible substitutes with dynamic updates through multi-threading techniques is provided. The results obtained by the analysis and evaluation of the proposed method, indicates the establishment of compatibility between the services, and finding the best possible substitute to reduce substitution time.**

Keywords*—component; indirect substitution; SLA; service composition; quality of service

I. INTRODUCTION

Through the discovery and development of Web services, an organization can choose and composite the Web services that satisfy its business needs and service quality. At the same as the number of available Web services increases, choosing the best services becomes more challenging for the given practice. Service quality plays an important role in the selection and composition of services. During the composition of services in a workflow or application to complete the process until the final result, services should be compatible in regard to the given the inputs, outputs and functions. The main problem of service composition is that the values of service quality may change from initial estimates during the implementation. The service may be unavailable or unreliable or may not offer other suitable solution. Thus, services should be evaluated dynamically to complete the program. Changes in the values of service quality may lead to failure of the expected adoption of the program for maintaining certain cases such as costs and response time. Two main issues of which adverse events require -the need for re-selection of service and services reprogramming include: extra time for selection process and lack of service composition compatibility with other service quality constraints [1]. In this paper, a method for composition of Web services that perform re-selection and prevent the deviation from the constraints of service quality, after reprogramming through the definition and evaluation of a potential substitution is proposed. In the proposed method re-selection and substitution are done only when it is needed. To help re-selection and substitution, services will be filtered based on their importance, because only the services that provide optimal solutions are considered in planning processes [1].

Substitution means the alternation of one component instead of another component; so that during the movement, the same component output be produced and meet the same needs with the replaced output component. Substitution means compatibility of Web service with the client requests and better setting than other competing Web services [1]. Substitution is both direct and indirect, till now there are many related works in the field of direct substitution. The main problem and error method in a direct substitution is that the compatibility between services in substitution process is ignored due to the lack of regulation of services in the substitution process. In addition, one cannot benefit the substitution services composition in this method and there is no possibility of automatic substitution. To solve this problem and to create compatibility between the services in substitution, indirect substitution method is used in regard to SLA violations. Every service includes its specific quality of service (QoS). In substitution process, one is trying to prevent the failure of qualitative constraints; However by maximizing the qualitative characteristics which should be augmented such as reliability and by minimizing the qualitative characteristics that should be reduced such as cost, maintaining the upper and lower limits to the maximum extent of assumed constraints and maintaining compatibility between available services, substitutability becomes possible. Finding compatible Web services to be replaced with a Web service after being selected is essential in each of the following events [2]:

- Web service should be eliminated during the runtime.

* Corresponding author

- Better service need to be available through a new Web service.
- A new version of the selected Web service should be available.

Substitution of Web services is possible only by fulfilling the following two conditions [2]:

- Web services should have the same functionality.
- They should be able to obtain their respective interfaces. The services must be compatible with each other for substitution.

The second section of this paper describes the concepts and terminology related to substitution and service composition. In the third section, the proposed method is described. In the fourth section, results of evaluation are discussed. Then in the fifth section a review of the earlier works is presented. Finally, the result of the research is provided in the sixth section.

II. Concepts and Definitions

A. direct and indirect substitutions

The direct substitutes are feasible because they can replace the failing service without further adjustment to the plan. Logically, it is very easy to compute the number of direct substitutes. [1].

In a direct substitution, the violated service can be replaced regardless of a subset of the inputs and outputs with exactly equal the input and output of a service, and without creating a service composition to achieve equal functionality.

Indirect substitutes are feasible service instances that can replace the failing service after adjustments with other service replacements are made to the resulting plan. [1].

B. Service quality traits

Quality of service is a subset of non-functional traits, which is different from various perspectives [2, 3, and 4].

In general, a non-functional trait divides into types of quantitative and qualitative traits [5]. A qualitative trait of the proposed method is defined here:

1) Response Time

The time it takes in which a service performs its task. This time is long as from the time of user's request to the time in which the answer is received, which is obtained in accordance with existing equations in Table 1 [5].

2) Service Reliability

If the service continue to work correctly and consistently in the specified period and under certain conditions. Despite the availability, reliability is defined in terms of time interval rather than time instant, and is obtained by the equations in Table 1.

3) Availability

The content and timeliness that service is immediately available for the operation and performance of its functions. Availability is related to system failure. The availability is obtained by existing equations in Table 1 [5, 6].

4) Cost

The cost for using Web services is obtained by the equations in table 1 [7].

TABLE I. Qualitative Parameters of Service and Their Calculations

Qualitative parameters	formula
Response time	Response time = Execution time+ Network time Response time= process time+ Transfer Time+ Latency Time
Reliability	Reliability=1- Probability of Failure
Availability	Availability=Uptime / (Up time +Down time) Availability =MTBF[1] / (MTBF+MTTR[2])
Cost	Total Cost = Service execution Cost+ (Network transportation/Transaction) Cost

5) SLA rules

It is a legal document format based on XML language, which consists of the three parties of the contract, guarantee terms and service terms; Figure (1) [8].

Fig. 1. SLA format [8]

Agreement contexts consist of general information such as parties and life cycle of the agreement. This information includes the address and profile of service producer, consumer, etc. The service terms consists of two parts: service reference and service properties. In service reference, an availability URL to the services is determined, and in service properties, information on quality parameters and indices is determined. Service terms are a key element of the SLA. The Guarantee terms consists of the quality objectives and financial agreements [8].

Main SLA requirements include [8]:

- SLA format should be a clear definition of a service, so that the consumer should understand the service functions
- Provide a level of service efficiency.
- Methods of monitoring service parameters and regulatory reporting format must be defined.
- Penalties when services are not met.

[1] Mean Time Between Failure
[2] Mean Time to Repair

C. static service compositions

Static service composition is created at design time and software system architecture. Adopted components will be selected, connected, and finally compiled and deployed. This case is suitable if serviced components rarely be changed or do not change in general.

D. Dynamic service composition

Service environment is a dynamic and very flexible environment. The new service will be available every day, and number of service providers will be growing. Ideally service processes must be able to accommodate to environmental changes and customer requirements with minimal user intervention.

E. Composite service

Composite service is created by composition of multiple services. Existing services in composite service might be implemented in different locations and in various fields. But they should interact with one another to achieve a goal. Service composition is referred to process of services development ranging from conventional services and compositing services.

III. Describing the Proposed Method

In [1], an approach is proposed in which due to the use of qualitative parameters in genetic algorithms and the techniques based on the total weight in the objective function of the algorithm, the potential diversion of restrictions during the program implementation has decreased, that leads to finding the best possible solution in accordance with qualitative traits of user's request. However, the method has limitations and the problems. This method is directly focused on replacing static service composition that limits the turnover in the composition and lack of focus on service composition which leads to service incompatibility. Thus, service incompatibility and composition limitations are among the most important problems in this approach. Another limitation of this method is lack of automatic service substitution, and as a result time consuming substitution process.

In this paper, indirect substitution is used to solve problems and mentioned limits as well as substitution optimization. In the proposed method substitution process dynamically improved by enjoying the [1] algorithm. The ability to build composite services, incompatibility problems of substitution and composition have been solved. In addition, in the proposed method, a solution for automatic substitution and reduction of process time is provided.

The proposed method consists of three main steps.

1) Preprocessing

- Selecting from the database.
- Receiving the requested qualitative parameters
- Receiving the requested weight of qualitative parameters
- Receiving the incoming and outgoing requests

2) Service composition

- Logical composition step
- Creating qualitative model
- Physical composition step
- services filtering
- Composition algorithm step
- Creating a service composition
- Creating the composition
- Finding the best composition
- Genetic Algorithm
- fitness function of the proposed model Reprogramming

3) Reprogramming

- Service Substitution
- Updating the substitution table

A. pre-processing phase

Pre-processing is all operations that must be performed before service composition, so that composition process be done according to the requests and qualitative parameters of user's requests along with maintaining the limits and SLA rules.

- The first step: selection of database repository

In this step, according to user requests, demand-services are called from UDDI database. These services include all similar requested services or services that are similar in the input or output.

- The Second Step: Receiving the requested qualitative parameters

In this step, a value must also be considered for each qualitative parameters of availability, response time, reliability, cost and substitution. These parameters' input values range between zero to one, and is determined by the user's request.

- The third step: Receiving the requested weight of qualitative parameters

Considering the fact that in this study, composition operations are carried out on the basis of the users' requested weight of qualitative parameters, in this step the user enters the requested weight of each qualitative parameter.

- The fourth step: Receiving the incoming and outgoing requests

In this step, the user request his desired input and output with respect to the functionality of services.

In [1] service composition is performed statically and through the genetic algorithm. It involves different stages and steps. To improve and solve the concerning problems in the proposed model, the process is modified and composition is done dynamically.

B. Service composition phase

1) Logical composition step

- Creating qualitative model

2) Physical composition step

- Services filtering

3) Composition algorithm step

- Creating a service composition and compatibility
- Fitness function

1) The first step: Logical composition

The first step to obtain the optimal composition of Web services is creating a suitable model to describe qualitative characteristics. This model must be agreed by the client and service provider. The qualitative model can be circular, parallel, serial or probable. To calculate the qualitative model, based on the type of limits and qualitative model, the pattern of aggregation functions in Table (2) can be used. Sequential method is used in this proposed method.

TABLE II. AGGREGATION FUNCTIONS FOR CALCULATION OF SERVICE QUALITATIVE PARAMETERS [1]

Attribute	Dimension Type	Constraint Type	Aggregation function			
			Sequential Invocation	Probabilistic Invocation	Structured Cycles	Parallel Invocation
Cost	Decreasing	Upper	$\sum_{i=1}^{n} cost$	$\sum_{i=1}^{M} Pi * Cost(s)$	K*cost(s)	$\sum_{i=1}^{P} cost$
Response Time	Decreasing	Upper	$\sum_{i=1}^{n} RTim$	$\sum_{i=1}^{M} Pi * RTime(si)$	K*Time(s)	MAX(s0..sp)
Availability	Increasing	Lower	$\prod_{i=1}^{N} Ava$	$\prod_{i=1}^{M} Pi * Avail(si)$	Avail(s)k	$\prod_{i=1}^{P} Ava$
Reliability	Increasing	Lower	$\prod_{I=1}^{N} Rel($	$\prod_{i=1}^{M} Pi * Rel(si)$	Rel(s)k	$\prod_{I=1}^{P} Rel($

In this step of the proposed model, unlike [1], in order to improve the composition method, the values of input and output are received from the user in the pre-processing. Then, based on these values, only those services which include the requested input and output or a subset of users' request would be called. In fact, at this stage of the proposed model, filtering operation is performed on the service call on the basis of functionality. And in any composition, searchable input and output are specified by user which leads to the method's dynamic trait. The called services will be elected as a candidate. This stage which leads to selection of services is called logical composition. This set of services in form of a set of workflows as candidate services move to the next step which is the physical composition. For example, suppose the user requested service S with input and output of (A, B.C, D). As a result, services with a subset of the input and output such as S1 (A, B.J, K) and S2 (A, H.C, D) and S3 (A.F) etc. will be called.

2) The second step: physical composition

At this stage, workflows are filtered based on user's requested qualitative parameters. Then to perform service composition they will be entered to the composition algorithm. In the given example, called services will be filtered qualitative characteristics entered by the user. In this case, it is assumed that the user requests the values of qualitative parameters in Table (3).

TABLE III. EXAMPLE OF REQUESTING USER'S QUALITATIVE PARAMETERS

availability	0.28
reliability	0.28
Response time	0.91
cost	0.91
substitution	0.11

However, if called service S1 has qualitative characteristics less than requested characteristics. The service will not be considered as a candidate for this composition. In fact, called services are filtered based on qualitative parameters.

C. Filtering substitutable services

Unlike [1], in the proposed model, a filtering operation based on qualitative parameters and weights is done dynamically before calculating the quality of services on the basis of CIFs in Table 1. As a result, the quality of service is calculated only for services that include qualitative parameters and weights; therefore additional and unnecessary calculations will be avoided. This filter is rarely applied in linear form due to its complexity. The qualitative model for every workflow is calculated by existing equations in Table (1), and thus quality of service for each workflow can be obtained.

After the pre-processing filter and reduction of candidate services we enter into algorithm phase. As mentioned in the previous section, unlike [1], composition operation is done dynamically based on the received input and output in the pre-processing phase as well as during the construction of composite service. After creating different service plans, the fitness function of genetic algorithm is used to compare and select the best composition with the highest fitness. For example, suppose you have a service consists of two inputs and outputs. In this method, based on the input and output received in preprocessing phase, three lists will be created and called services will be entered according to the inputs and output as shown in Table (4).

TABLE IV. AVAILABLE LISTS IN THE PROPOSED METHOD

Services in which their inputs are a subset of user's requested inputs	Input	LISTØ
Service in which their outputs are a subset of user's requested outputs	Out put	LIST1
Proper composition occurs through inputs and outputs and a subset	Creating service composition	LIST2

Various scenarios intended for creating List2 and service composition are as follows:

- The first scenario

According to equation (1), if the inputs of selected service in List2, are the subsets of the outputs in listØ service and if its outputs are a subset of inputs in list1 service, the condition of service composition, which is the accessibility to user's requested functionality, will be established and the compatibility will not be violated. Thus, the service in list2 will be considered for composition with services in listØ and list1 and it will be removed from list2. As a result composition process will be successfully performed, and the resulting service plan will be displayed in the output. The first scenario is given in example (1).

- The second scenario

As mentioned in Equation 2, if one of the conditions in the first scenario is violated; for example if inputs of the selected service are equal to ListØ outputs but its outputs is not a subset of List1, it results to incompatibility. Thus selected service in list2, becomes the basis among the services based on input and output to create service composition on the next survey. In other words, a service in which its outputs are a subset of the inputs in list2 service and its inputs are a subset of outputs in the mentioned service, will be sought for, and service composition will be formed. This cycles continues till the service composition condition is established, and service composition that includes a subset of the input and output according to user's request be made. Thus in the second scenario, creating a service composition with compatibility is done. For a better understanding, the proposed algorithm is expressed in example (1).

The user requested service S with the inputs of A, B and outputs of C, D. According to the proposed solution, three list is created. In the listØ S1 and in list1 service S2 are called. As the example shows, the two first rows, compatibility and service composition without creating a service composition is established like the first scenario. In the third row, S9 violated the terms of compatibility, therefore, among the set of candidate services, service S10 in which its inputs are a subset of the outputs in service S9 and its outputs are a subset of the inputs in service S8, are called. This cycles continues till compatibility condition is fulfilled. In this example, through the service composition of S9 and S10, composition problems are solved and compatibility can be achieved. Therefore, by implementation of S8, S10, S9, and S7 compatible composition is created through composite service.

Equation (1) of the first scenario:

$$\begin{cases} \text{if input list2 } \hat{c} \text{ output listØ} \\ \& \\ \text{if output list2 } \hat{c}\text{binput list1} \end{cases} \Rightarrow$$

Composite Services in list Ø &list2 & list 3 Equation (2) of the second scenario:

$$\begin{cases} \text{if input list2 } \hat{c} \text{ output listØ } \cap \text{output list 2 } \not\subset \text{input list1} \\ \& \\ \text{if input list 2 } \not\in \text{out put list Ø} \end{cases}$$

$\Rightarrow$ Research in data Service input & output ĉ list2 then composite Services listØ & list1&list2

TABLE V. SERVICE COMPOSITION ALGORITHM

	User's request	**List ∞**	**List $_1$**	**List $_2$**	**Output**
	S(A,B.C, D)	S_1(A,B.J, K)	S_2(A,H.C, D)	S_3(J,K.A,H)	S_1,S_3,S_2
First scenario	S(A,B.C, D)	S_4(B.C)	S_5(H.D)	S_6(B.C,H. T)	$S_4,S_6.S_5$
Second scenario	S(A,B.C, D)	S_7(B.F)	S_8(C.D)	S_9(F.K) S_{10}(K.C)	S_7,S_9,S_{10},S_8

This algorithm is carried out dynamically and constantly by using multi-threaded technique. After composition and creating different service plans, the fitness function of genetic algorithm is used to compare and select the best composition with the highest fitness. In addition, by entering the compositions into fitness function of the algorithm, SLA rules and violations will be investigated; as the qualitative parameters will be determined by entered weights by the user and calculation of qualitative model according to the table (2), and in case of violations, they will be outdated.

D. Fitness function

In the proposed model, after composition process, fitness function according to equation (3) is used to compare the compositions and presenting the best composition.

Given the importance of qualitative parameters in this method, a fitness function using the total weight which transfers multi objective problem to a single objective problem is used. As previously mentioned, the weights are selected based on the user's preferences and needs. Just as shown in equation (3), the composition from the previous steps is called wj; and w1, w2, w 3, w4 and w5 weights are provided by the user. In accordance with the calculation contract of qualitative model in table (1), the availability of all the services in the Wj composition are multiplied. Thus, the compositions' rate of availability will be obtained. Equation (4) shows the calculation of the qualitative parameters rating for each composition. Reliability and substitution composition are calculated in the same way. As Table 1 shows calculation contract for qualitative model, addition is used for rating the cost parameters and response time of composition. This means that cost of services composition are added together and the result is the rating of Wj composition cost. Similarly, the calculated response time will be entered to fitness function of the. In this function, as the reduction of cost parameters and response time are superior standard of composition, these parameters are placed in the denominator of fraction. Finally, the composition with best fitness will be selected and displayed in the output.

Equation (3) of fitness function [1]:

$$\text{Fitness} = w_{1*\text{Availability (wj)}} + w_{2*\text{Reliability (wj)}} + w_{3*\text{replacebility (wj)}}$$

$$W_4*\text{Cost (wj)} + w_{5*}\text{Response Time(wj)}$$

Equation (4) rating qualitative parameters of composition

$$\begin{cases} wsi \;\; i = 1{:}\,n \;\; \prod_{i=1}^{n} wsi \quad Availability \\ wsi \;\; i = 1{:}\,n \;\; \prod_{i=1}^{n} wsi \quad Relability \\ wsi \;\; i = 1{:}\,n \;\; \prod_{i=1}^{n} wsi \quad Replasebility \end{cases}$$

n= Number of Service

$$\begin{cases} wsi \quad i = 1{:}\,n \quad \sum_{i=1}^{n} wsi \quad Cost \\ wsi \;\; i = 1{:}\,n \quad \sum_{i=1}^{n} wsi \quad Respanse\ Time \end{cases}$$

E. Substitution phase

This phase involves the following operations.

- reprogramming
- Updating the substitution table

Substitution operation is performed in reprogramming step. If the service is faced with failure due to SLA violations, or if the user requests a service substitution, service substitution operation is performed. In [1], by any failure or service violation, reprogramming will be considered and if it is reasonable, reprogramming phase for substitution begins. Therefore, in case of the need for substitution, algorithm for each service for each violated service will be performed, which increases the substitution time.

First, three questions in relation to the reprogramming arises: when do we need to perform reprogramming? Where to begin reprogramming? How to perform reprogramming? To answer the first question, events which require reprogramming have been studied. Lack of service availability, breakdown in proper service response within a time period as well change of qualitative QoS before the implementation and due to the election of paths in substitution or changes in the number loops, the real qualitative values of the program are different from the estimated values.

In case of failure or services unavailability, the algorithm must be re-run to find service substitution, and seek to have service optimization. There are two reasons for this optimization. First, it can find better qualitative values. Second, it can produce a more acceptable substitution program [1].

In the proposed method to reduce the need of re-programming, the algorithm investigates the substitutable service before and after the implementation of the program. In this case, if the obtained service be replaced in the previous implementation, there would be no need for additional re-programming and extra process time. On the other hand, in case of a limited response time, if the time limit be diverted by the re-optimization, then the algorithm will use the best substitute in the previous implementation. Normally, when there is substitute, in most cases, substitution is suitable.

In the proposed model, according to the dynamically of service composition method, with any changes in qualitative parameters the best possible composition will be calculated during the composition. Therefore, in the first phase of re-programming, with the very changes in the qualitative parameter which is considered as SLA violations here, or a user's requests and failure, the program automatically runs the algorithm in the proposed model described in proposed composition model. The results can include various services and service plans along with the fitness number. Each are stored in a table. These results are dynamically updated in a specified interval. And if there is a need for reprograming and if the numeric is less that fitness value of Service plan in the table substitution will be done automatically with minimal time without running genetic algorithm and in regard to regulations of SLA.

The substitution process in the proposed model is done indirectly. Unlike [1] which is focused only on services with equal input and output and uses direct substitution, in this method, service composition in which its inputs are a subset of inputs and outputs of user's requests will be created. As a result, the possibility of substitution is available, leading to compatibility maintenance and enhancing the range of substitute services. In addition, in the event of service failures due to compatibility maintenance and if necessary, service plan can be replaced; while in [1] there is only the possibility of replacing the service.

The proposed method enjoys a higher speed compared to [1]. And composition time is significantly reduced. To achieve this all processes are running in the background. As mentioned in composition method phase, multi-threading techniques is used in the proposed algorithm. To implement service composition and construction of the composite service, a thread and to update the substitution table another parallel thread is considered.

IV. The Results of Evaluation

Services data sets are produced by the program and to test the method, different numbers of services are used. Since the in proposed method, dynamic and indirect service composition and service substitution is done. Memory consumption is higher than static service composition method and direct substitution; which is considered normal. In addition, due to the dynamic composition, time taken to find the best composition in the algorithm increases, but this increase is seen only in the first survey. The proposed method will be evaluated from four aspects of composition time, re-programming time, memory usage and rate of failure.

A. service composition time

In this study, an increase in the number of services in the proposed algorithm due to the dynamic composition, composition time increase in providing a suitable initial composition; while in a direct composition algorithm, composition time decreases due to static composition. The results are shown in Table 6 and Fig. 2.

TABLE VI. THE RESULTS OF COMPOSITION TIME EVALUATION

Count	Static composition Time1	Dynamic composition Time2
5000	0. 76	0.007
50000	0.77	0.45
100000	0.78	1.54
150000	0.78	3.68
200000	0.79	6.96
250000	0.76	10.43
300000	0.8	15.49

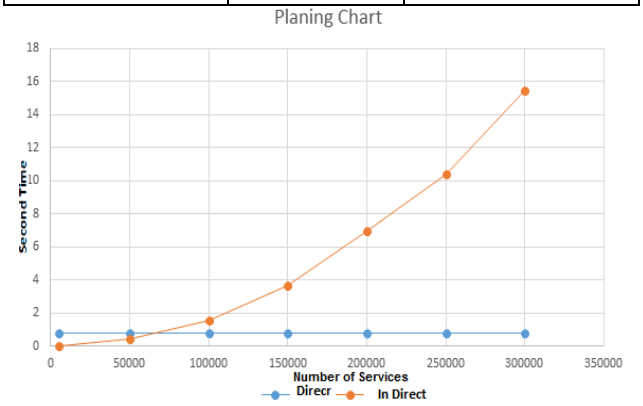


Fig. 2. The results of composition time evaluation

The results of composition time evaluation

B. memory usage

In the proposed model, the number of services is increased due to the dynamic approach. As a result, the amount of memory usage is increased compared to the direct substitution. The results of the evaluation in accordance with the memory usage are shown in Table 7 and Fig 3

TABLE VII. THE RESULTS OF MEMORY USAGE EVALUATION

Count	Static composition Memory1	Dynamic composition Memory 2
5000	34.6	34.4
50000	50.1	56.6
100000	66.8	79.2
150000	68.3	89.7
200000	78.2	108.6
250000	89.6	123.2
300000	101.1	141.9

Fig. 3. The results of Memory Usage evaluation

C. evaluation of reprogramming time (service substitution)

In the proposed model, due to use of updating technique, implementing the best solution in parallel in the background, time of reprogramming will be close to zero, and can be ignored. Also, due to using a linked list in the calculations instead of arrays, the time is reduced compared to direct substitution algorithm. The results are shown in Table 8 Fig. 4.

TABLE VIII. EVALUATION OF REPROGRAMMING TIME (SERVICE SUBSTITUTION)

Count	Direct Replacement Time1:	Indirect Replacement: Time2
5000	0.005	0.00021
50000	0.0051	0.00026
100000	0.0052	0.00042
150000	0.0061	0.00046
200000	0.0063	0.00051
250000	0.0072	0.00053
300000	0.0074	0.00059

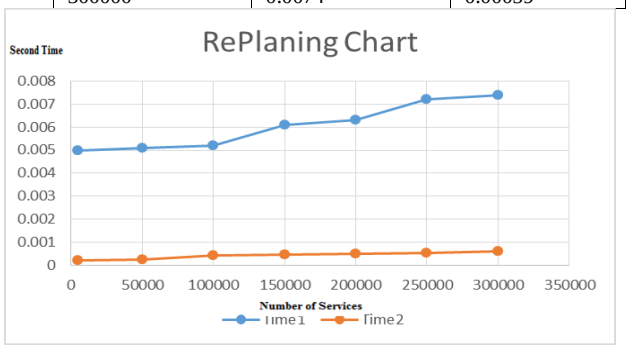


Fig. 4. Evaluation of reprogramming time (service substitution)

D. evaluation of failure rate

In the proposed method, failure rate is close to zero due to using updating technique in the table of the best solution. Since in re-programming, we always have the best answer in the table and there is no need for calculation. Although direct algorithm in [1], since computing must be performed in the reprogramming, to find best composition for the user, it is likely that there is no possibility of reprogramming and the algorithm may fail. In testing cases, the proposed algorithm had no failure. However, in cases where users demand high quality service and low cost and low response time, the

proposed algorithm is likely to fail. In case the user requests according to the table (8), the proposed algorithm is likely to fail after one or two times of reprogramming.

TABLE IX. EVALUATION OF FAILURE RATE

availability	0.9
reliability	0.9
Response time	0.3
cost	0.3
substitution	0.11

However, due to the fact that in simulation, a service which becomes unavailable, will be outdated from the algorithm, but this rarely happens in real space. In the direct method [1], despite adopting substitution, the failure rate is high since the user's requested service was not along with so high qualitative weights users.

V. RELATED WORKS IN THE FIELD OF SERVICE SUBSTITUTION

In this section, different methods of Web services substitution are introduced.

Helal AL et al. used the concept of substitution for service composition. The way it works is as follows: by selecting a qualitative model of services and identifying the cumulative function and calculating qualitative model of each service, a set of service examples is considered as a candidate and by measuring the substitutability of each candidate service by using the filtering technique and nearest neighbor algorithm in this technique, candidates set is improved, and then by using genetic algorithm and its fitness function through total weight, services substitution and reprogramming will be done. The advantage is that one can make the best choice for service composition. In this study direct substitution is focused, i.e. the service will be replaced without any adjustments. This method is useless in case of incompatibility between the services. Thus, the defects of this study is lack of focus on indirect substitution and compatibility [1].

Yu et al. have used an approach based on graph theory. In general, the main idea behind this theory is to show a service as a node. Links indicate the relationship between the services. The costs are qualitative characteristics (cost and delay). The advantages of this method are optimum runtime and memory usage, respectively. One of the weakness in this methods is the lack of scalability [9].

Sheng et al. used a method based on backward theory for service composition. The main idea of this approach is that services are selected step by step. To choose a service at each step, the selection algorithm moves one step backward and checks the selected services to ensure the best service is selected. If the selected service is approved, it will be called. The advantages of this method is fault tolerance. Thus by deterioration of a service, another efficient service will be replaced. The disadvantages of this method is increased processing time [10].

Zhang et al. presented a heuristic algorithm based on taboo search for dynamic service substitution. They used the graph of candidate service; and by using simulation, they evaluated the efficacy and performance. Simulation results show that the proposed algorithm is very good in the substitution in large-scale space. The advantages of this method is it ensures service availability and uninterrupted process and its weaknesses is focusing only on substitution algorithm [11].

Wu et al. presented a cluster-based approach for service substitution. Concepts of logical service, real service, and the cluster service and the relationship between these services had been studied. The proposed method consists of two steps: Finding the expired real service dependent to the logical service and choosing a real service from the service cluster for substitution of violated service.

In this method, compatible services are put in a cluster and can be replaced by another. The advantage of this method is increased speed and reliability of service composition. However, if none of the cluster services are available, user's requests remain unanswered, which is as a drawback in the method [12].

Li et al. presented Web Service Composition based on QoS with Chaos Particle Swarm Optimization. In this study, based on desired qualitative parameters, services are selected, then selected services are entered into the algorithm and finally provides the best composition. Increased speed of service selection and service compositions is one of the advantages of this method. The drawback of this method is lack of attention to parallelism and inconsistent data [13].

Alrafai et al. provided an approach for using Skyline service for Web service composition based on QoS; in which integration of Web-based service composition were evaluated dynamically and without defect. The advantages of this method is division as Skyline calculations can be provided in parallel in groups without changing the final result. This is done by using Pad Skyline algorithmic framework for parallel processing of Skyline request in divided groups. The optimization technique within the group and multi-dimensional filtering for each group is performed. In particular, Skyline local points along with the request as the filtered points to help identifying services in the areas of poor quality on any site are sent through Skyline service. Another advantage of this method is reduction in the response time to user requests and increased speed of Web service composition. The method is affordable and effective for specified service composition. The drawback of this method is lack of investigation in limited and a structured environment. [14].

Lu et al. have provided Web Service Composition Based on Integrated Substitution and Adaptation. They showed that substitution and adaptation complementary are and believe that the integration of adaptation and substitution provides the design of highest flexibility and performance over time for substitution and running the service. In this study, web service composition is based on adaptation and using substitution. They studied substitution at two static and dynamic levels and proposed a dynamic substitution approach. The advantages of this research is service composition without passing through the adaptation and by considering the substitution and automatic service composition with increased workflow functionality for the composition, as well as more flexibility in the composition. The drawbacks may be a lack of focus on

timing constraints in the system and describing similarities based on non-functional parameters [15].

Kuang et al. are focused on the challenge of substitution through behavioral analysis of services. They achieved this by Security Operation Center approach and a formal definition of behavior in Web services. In this study, a formal definition for service composition by means of complex behavior π and calculus formulas as well as conception of behavioral substitution of services through simulations by using formulas and tools has been evaluated. Simulations showed that behavioral substitution of services can be improved based on behavioral analysis of services through using formulas and mathematical calculations. The drawbacks of this research is lack of focus on providing the tool for automated substitution and lack of compatibility with different conditions dynamically [16].

VI. CONCLUSION

In the proposed method, incompatibility problems in composition and substitution of services are figured out by considering qualitative parameters, requested inputs and outputs of user and SLA rules by techniques of creating linear list, fitness function table by total weight and update table of the best alternatives with parallel multi-threading approach. The algorithm efficiency in reprogramming and substitution of the service significantly increased. In this method, as a process is active in algorithm's background and substitution is done dynamically and at the same time it needs to run in parallel environments. In this model, due to dynamic composition and creating composite services based on a subset of the input and output as well as an updated table with the best alternatives, the memory usage increases to some extent compared to direct substitution and static composition. Therefore, studying and improving the memory usage may be a future research.

REFERENCES

[1] H. Al-Helal and R. Gamble, ''Introducing Replaceability into Web Service Composition'', IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 7, NO. 2, PP.198-209 ,2014.

[2] V. Andrikopoulos, S. Benbernou, and M. Papazoglou, "On the Evolution of Services", IEEE Transactions on Software Engineering, Vol.33, No. 3, PP. 609-628, 2012.

[3] R. Iordache and F. Moldoveanu," QoS-Aware Web Service Semantic Selection Based on Preferences", 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, vol. 69, pp. 1152–1161 ,2013.

[4] K. Saeedi, L. Zhao, P. R. Falcone Sampaio," Extending BPMN for Supporting CustomerFacing Service Quality Requirements", IEEE International Conference on Web Services (ICWS), PP. 616 - 623 ,2010

[5] DZG.Garcia, MBF.de Toledo,"Achieving autonomic web service integration:a quality of service policy based approach", Journal of International Transactions on Systems science and applications, vol.3, pp. 41-63, 2010.

[6] S. Bosse, M. Splieth,M. Turowski," Multi-Objective Optimization of IT Service Availability and Costs", Magdeburg Research and Competence Cluster for Very Large Business Applications, Faculty of Computer Science, Otto von Guericke University Magdeburg Germany,vol.147,pp.142-155,2016.

[7] A.Eleyan, L.Zhao, "Extending WSDL and UDDI with Quality Service Selection Criteria", In: Proceedings of the 3rd International Symposium on Web Services,pp.1-10, 2010.

[8] M. Alhamad, T. Dillon, E. Chang, "Conceptual SLA Framework for Cloud Computing", 4th IEEE International Conference on Digital Ecosystems and Technologies,pp. 606 – 610,2010.

[9] H.Q.Yu, S.Reiff-Marganiec,"Web Service Composition Methods:A Survey",Information Sciences,Vol280,PP.218-238,2014.

[10] Q.Z.Sheng,X.Qiao,A.V.Vasilakos,"Selection of QoS Support on Artificial Immune Network Classifier for Dynamic Web Service Composition",International Conference on Computational Intelligence and Security,PP.643-646,2014.

[11] C .Zhang ,H .Chen and J .Du , "A Tabu Search Approach for Dynamic Service Substitition in SOA Applications " , Services Computing Conference (APSCC), 2011 IEEE Asia-Pacific ,INSPEC Accession , PP. 284 – 289,2011 .

[12] L .Wu , Y .Zhang and Z .Di, " A Service-cluster Based Approach to Service Substitution of Web Service Composition " , IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD), , PP. 564 - 568 ,2012.

[13] W.Li,H.Yanxiang, "Web ServiceComposition based on Qos with Chaos Particle Swarm Optimization", 6th International Conference on wireless Communications Networking and Mobile Computing,PP.1-4,2010.

[14] J. Wu, L. Chen, T. Liang," Selecting Dynamic Skyline Services for QoS-based Service Composition", Applied Mathematics & Information ciences An International Journal.,vol. 8,PP. 2579-2588, 2014.

[15] L.Chen,R.Chow,"Web Service Composition Based On Integrated Substitution and Adaptation", IEEE International Conference on Information Reuse and Integration,pp. 34 – 39,2008.

[16] L.Kuang, Y.Xia, SH.Deng ,J.Wu,"Analyzing Behavioral Substitution Of Web Services Based On π-Calculus", IEEE International Conference on Web Services International College Wales Swansea, pp. 441 – 448,2010

Optimum Access Analysis of Collaborative Spectrum Sensing in Cognitive Radio Network using MRC

Risala Tasin Khan
Institute of Information Technology
Jahangirnagar University
Savar, Bangladesh

Md. Imdadul Islam
Computer Science and Engineering
Jahangirnagar University
Savar, Bangladesh

Shakila Zaman
Institute of Information Technology
Jahangirnagar University
Savar, Bangladesh

M. R. Amin
Electronics and Communication Engineering
East West University
Dhaka, Bangladesh

***Abstract*—The performance of cognitive radio network mainly depends on the finest sensing of the presence or absence of Primary User (PU). The throughput of a Secondary User (SU) can be reduced because of the false detection of PU which causes an SU from its transmission opportunity. The factorization of the probability of correct decision is a really hard job when the special false alarm is incorporated into it. Previous works focus on collaborative sensing on the normal environment. In this paper, we have proposed a collaborative sensing method in Cognitive radio network for optimal access of PU licensed band by SU. It is shown performance analysis of energy detection through different cognitive users and conducts a clear comparison between local and collaborative sensing.In this paper, the maximal ratio combining diversity technique with energy detection has been employed to reduce the false alarm probability in the collaborative environment. The simulation result showssignificant reduction of the probability of misdetection with increasing in the number of collaborative users.We also analyze that MRC scheme exhibits the best detection performance in collaborative environment.**

Keywords*—Fusion center; Local energy detection; Maximum Ratio Combining; Spectrum Sensing;Receiver Operating Characteristics

I. INTRODUCTION

Day by bay spectrum demand is increasing fast with the rapid growth of a new high data rate and wireless devices. Since frequency allocation is fixed and the users do not use the spectrum all time, so it introduces significant underutilization of the available frequency. Cognitive Radio (CR) becomes as a solution to this scarcity by providing more utilization of the spectrum resources which is capable to fulfilling the demand of to be available anyplace, anytime, when needed [1]. CR is an adaptive and smart system that can automatically detect the hidden spectrum hole and provide unused licensed spectrum to the cognitive user by dynamic spectrum sharing. In Cognitive Radio Network (CRN), Primary User (PU) denotes as authorized user who uses the licensed frequency and has higher priority to access the specified spectrum and Secondary User (SU) denotes as unlicensed user who is responsible for sensing the movement of PU's and use the spectrum when PU is in stationary mode.

One of the most significant components of CR is reliable spectrum sensing. Spectrum sensing is the process of discovering unused spectrum which is allocated to PU and make awareness about the existence of PU. Due to shadowing and multipath fading, this is a great challenge to execute spectrum sensing in the hidden terminal and know the status of an instantaneous spectrum. The performance of a good CRN exclusively depends on how accurately the SU can sense the existence or nonexistenceof a PU. Various traditional techniques have been used to implement the spectrum sensing such as matched filter, cyclostationary detection, energy-based detection algorithm which are discussed in [2] and [3]. Energy detection is the most popular and general sensing method due to its less implementation complexity and superiorvelocity which is also known as semi-blind detection method [4]. In this method, correctly threshold value selection is more significant to measure the performance. Since the one or more sensing parameters are unknown in the dynamically changing environment, recent studies [5]–[7]have focused on improving the performance of the detection method. In this paper, we have used noise level estimation to choose significant threshold value to meet constant false alarm rate. For quick and reliable spectrum detection and reduce false alarm, Collaborative Spectrum Sensing (CSS) has been introduced [8]–[11]The purpose of collaborative spectrum sensing induces new design and optimization challenges, such as transmission delay, security risk and energy consumption [12]–[15]. In CSS, secondary users send their local sensing information to the Fusion Center (FC). Then FC fuses the received signal information to decide about absence or presence of PU [16]. In [17], different data fusion scheme is used to optimize the detection performance in FC. At FC different diversity scheme can be implemented to make the final decision such as Maximum Ratio Combining (MRC), Selection Combining (SC) and Square-Law Combining (SLC) [18]. This paper includes energy detection in CSS with Maximum Ratio Combining (MRC) scheme. Maximum ratio combining diversity technique is used to analysis best possible

access in collaborative environment. Under MRC scheme, multiple cognitive users received the sensing result and send to the FC using data fusion, where the data from multiple cognitive radios are combined by MRC linear combiner. Then an energy detection is used to dealings the MRC combiner output. In [17], Collaborative spectrum sensing is used to verify the efficiency of detecting spectrum holes by different combing scheme.

The goal of this paper is the analysis of the optimum energy detection based on the different parameter in local and collaborative spectrum sensing with MRC using traditional energy detection.The rest of this paper is structured as follows. Preliminary models of local and collaborative energy detection are discussed in Section II. The comprehensive system architecture of collaborative sensing with MRC is also discussed in section II. In section III, simulation parameter and results are given to analysis the optimum spectrum sensing in CRN. Finally, we conclude this paper in section IV.

II. System Model

A. Local energy detection for CR users

In the case of PU's information is unknown in the Cognitive radio network, most popular PU's detection method is Energy detection. By following fig 1 for known time interval a bandpass filter is used to select collected frequency and bandwidth for energy detection method. SNR is estimated by using channel SNR estimator. Then energy of received signal is measured by magnitude squaring device with an integrator. Measured test statistics is compared with a predefined threshold τ to produce information about the existence of the Primary user. Spectrum sensing using energy detection is formulated with two hypothesis test also known as binary signal detection. If there is no primary user then hypothesis result produce H_0, otherwise it produces H_1 that indicatethe presence of Primary user.

Mathematically the two hypothesis-testing can be formulated as,

$$x_k[t] = \begin{cases} w_k[t]; & H_0 \\ \alpha_k e^{j\theta_k} s[t] + w_k[t]; & H_1 \end{cases}$$

Where $x_k[t]$ denoted as received signal at k^{th}Secondary user with k = 0, 1, 2, 3.....*Nr* which is *independently* and *identically distributed* [19], $\alpha_k e^{j\theta_k}$is the channel gain between PU and SU, $s[t]$ is the PU's transmitted signal that follows the Gaussian random process with zero mean and variance $\sigma_s^2$and $w_k[t]$ denotes the white noise.

To determine the efficient performance of energy sensing method the test estimation is defined as,

$$T(x) = \frac{1}{M}\sum_{k=0}^{M} x_k[t]^2 \quad ; \qquad (1)$$

Where M is the size of observation vector and $T(x)$ is text statistics.

Two important detection probability parameters are used to measure the performance of any detection method that is P_D and P_{FA}. P_D indicates the probability of detecting the existence of the signal of PU on the required frequency when it is actually present. Since PU is surrounded by various interference, P_D should be as large as possible to make the correct decision. This detection condition can be written as,

$$P_D = \Pr(Signalisdetected|\, H_1) = \Pr(T(x) > \tau|H_1)$$

$P_{FA}$is the probability of choosing H_1 but H_0 is true, that means CR's decides that primary user is detected as on mood butactually there is no primary user. To utilize more transmission opportunities of unutilized spectrumP_{FA} should be as small as possible. That can be expressed as,

$$P_{FA} = \Pr(Signalisdetected|\, H_0) = \Pr(T(x) > \tau|H_0)$$

Another important parameter is used to determine the performance of Cognitive user that is the probability of misdetection P_{MD}. This condition occurs when a cognitive user (SU) choose H_0 but H_1 is true. In this case performance of SU detection will decrease and it can be formulated as,

$$P_{MD} = \Pr(Signalisnotdetected|\, 1)$$

The existence decision of a PU can be estimated by comparing the decision matrix$T(x)$ and a fixed predefine threshold τ. It is very important to identify required threshold value that may change based upon environment condition that is related to the distance between PU and SU.

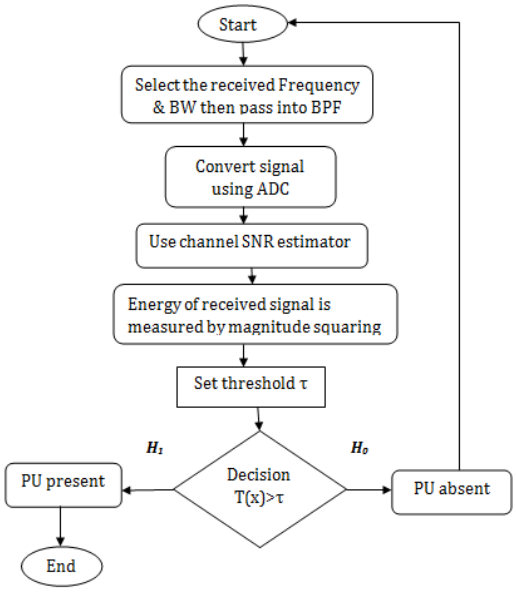


Fig. 1. Local Energy Detection for Cognitive user

Threshold value can be estimated by given equation that is based upon signal noise, detected signal energy, the size of observation sample and noise variance. Since the distance between Cognitive user and primary user changes, it is difficult to estimate signal energy. So threshold value is chosen to meet constant false alarm rate using noise level estimation.

That is calculated by given equation [20],

$$\tau = \sqrt{\frac{\sigma_s^2}{M} Q^{-1}(P_{FA})} \quad (2)$$

where,
τ = threshold,
σ_s^2= noise Variance,
M = size of observation vector or sample.

To estimate probability of correct detection we can choose the required threshold value from equation (2) and the probability of detection can be formulated as [21],

$$P_D = Q\left(\left(\frac{\tau}{\sigma_s^2} - \gamma - 1\right)\sqrt{\frac{\delta_t f_s}{2\gamma+1}}.Q^{-1}(P_{FA})\right) \quad (3)$$

where,
f_s = sampling frequency,
$Q(.)$ = complementary distribution function of standard Gaussian,
δ_t= sensing time or duration.

Then P_{FA} is given by,

$$P_{FA} = Q\left(\left(\frac{\tau}{\sigma_s^2} - 1\right)\sqrt{\delta_t f_s}\right) \quad (4)$$

The probability of misdetection is complement of P_D which can be formulated as,

$$P_{MD} = 1 - P_D \quad (5)$$

From [21] for a target false alarm P_{FA}, misdetection probability is given by,

$$P_{MD} = 1 - P_D$$
$$= 1 - Q\left(\frac{1}{\sqrt{2\gamma+1}}\left(Q^{-1}(P_{FA}) - \sqrt{\delta_t f_s \gamma}\right)\right) \quad (6)$$

Therefore, P_{FA} is related to targeted detection probability which is formulated as [21],

$$P_{FA} = Q\left(\sqrt{2\gamma+1}\left(Q^{-1}(1-P_{MD}) - \sqrt{\delta_t f_s \gamma}\right)\right) \quad (7)$$

B. Collborative Spectrum sensing Including MRC scheme

a) Formulation: To decreases false alarm probability, misdetection and also mitigate the hidden problem in the CRN, collaborative spectrum sensing has been introduced. In this case, multiple Cognitive users sense the spectrum band collaboratively. According to fig 2, data fusion method is used in all cognitive radio users to sense their spectrum independently but they do not make any decision to get the opportunity to transmit. All individual node transmit their sensing information to a central Fusion Centre (FC) using local sensing method. Then FC makes the final decision whether the SU transmit or not using the information of PU present or absent.

If we consider a CRN with n number of user where n = 1, 2, 3.....Nr, the probability of collaborative detection is formulated as [22],

$$Q_D = 1 - (1 - P_D)^n \quad (8)$$

Then collaborative false alarm probability is given by,

$$Q_{FA} = 1 - (1 - P_{FA})^n \quad (9)$$

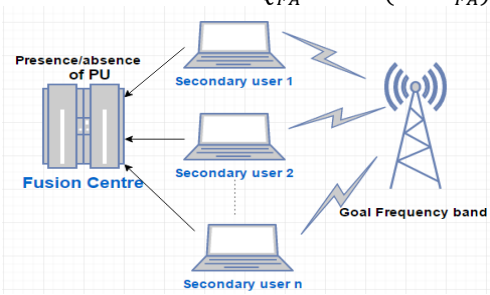


Fig. 2. State diagram of collaborative spectrumsensing

In FC, various combining techniques are used to combine the collected sensing information that comes from n number of independent cognitive users. In this paper, we have considereda case where FC make a decision whether the PU is absent or present using Maximum Ratio Combining scheme. According to fig 3, multiple cognitive radio users directly forward their sensing decision to the FC where the collected information is combined by an MRC scheme using linear combiner.

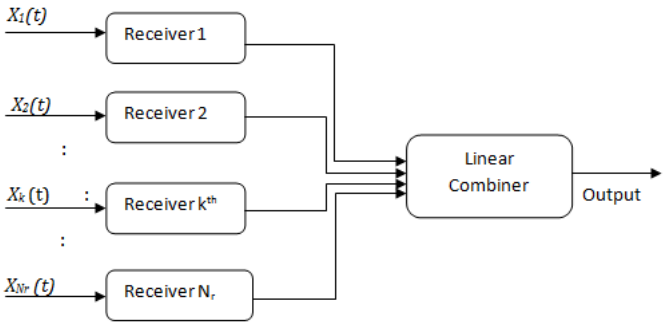


Fig. 3. Block diagram of Maximum ratio combining using N_r receiving antenna

Using Complex envelop of receive signal for k^{th}branches in MRC technique the corresponding received linear combined signal is calculated as [23],

$$\tilde{y}[t] = \sum_{k=1}^{N_r} \xi_k \tilde{x}_k[t]$$
$$= \sum_{k=1}^{N_r} \xi_k[\alpha_k e^{j\theta_k} s[t] + \tilde{w}_k[t]] \quad (10)$$

where,
$\alpha_k e^{j\theta_k}$= complex factor or channel gain in fading,
ξ_k= complex weighted factor for each channel that characterize the linear combiner,
$\sum_{k=1}^{N_r} \alpha_k e^{j\theta_k} s[t]$ = complex envelop of output signal,
$\sum_{k=1}^{N_r} \tilde{w}_k[t]$= complex envelop of output noise.

In this case, two hypothesisis mathematically formulated as, $\tilde{x}_k[t] = \begin{cases} \sum_{k=1}^{N_r} \tilde{w}_k[t]; & H_0 \\ \sum_{k=1}^{N_r} \alpha_k e^{j\theta_k} s[t] + \tilde{w}_k[t]; & H_1 \end{cases}$

Therefore, the MRC technique produces an instantaneous output SNR that is denoted as γ_{mrc} which maximizes the detectionprobability in collaborative spectrum sensing manner. Thatproduce by summarizing all individual users instantaneous SNR using linear combiner and given by [23],

$$\gamma_{mrc} = \sum_{k=1}^{N_r} \gamma_k \quad (11)$$

and,

$$\gamma_k = \frac{E_s}{N_0}\alpha_k^2 \quad (12)$$

where,
γ_k= instantaneous SNR for the individual k^{th}receiver where k = 1, 2, 3.....Nr,
E_s= symbol energy,
N_0 = one-sided noise spectral density.

Then the targeted collaborative detection and false alarm probability under MRC combiner scheme can be given by, $P_{MD} = 1 - (1 - Q\left(\frac{1}{\sqrt{2\gamma_{mrc}+1}}\left(Q^{-1}(P_{FA}) - \sqrt{\delta_t f_s \gamma_{mrc}}\right)\right))^n$ (13)

$P_{FA} = 1 - (1 - Q(\sqrt{2\gamma_{mrc}+1}(Q^{-1}(1-P_{MD}) - \sqrt{\delta_t f_s \gamma_{mrc}}))^n$ (14)

b) System Architecture:

In fig 4, a flowchart is proposed that shows the energy detection steps with MRC scheme in collaborative sensing manner. In this technique, each SU sends their local information to FC and FC calculate the energy by linear combiner and also calculate weighted and fading complex factor then produce γ_{mrc}using equation(11).Finally, calculate test statistics and compare with predefined threshold τ, then make a decision about the existence of PU. MRC system can improve the performance and the bandwidth efficiency in CRN. In this paper, a new algorithm, maximum ratio combining with energy detection is proposed for collaborative sensing environment where itshows step by step procedure to improve the performance over frequency sensing channel.

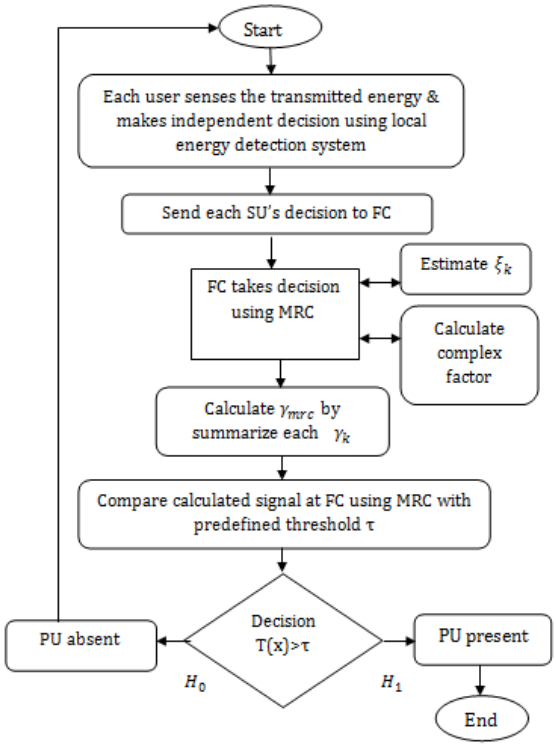


Fig. 4. Collaborative Energy detection using MRC

Algorithm 1 Steps to calculate Collaborative Spectrum Sensing efficiency with MRC

Step 1: Each CR's collect signal independently and pass through Bandpass Filter (BPF).
Step 2: Channel SNR (γ) estimation.
Step 3: Set τ from equation (2) that consider constant false alarm using noise level estimation.
Step 4: Estimate σ_s^2.
Step 5: Select thenumber of users (n) as thecollaborative sample.
Step 6: Each CR's Report the sensing information (H_1 or H_0) to FC.
Step 7: FC takesthe final opportunistic decision using MRC.
Step 8: Calculate weighted factor (ξ_k) for each SU's.
Step 9: Calculate Complex factor$\alpha_k e^{j\theta_k}$.
Step 10: Calculate γ_{mrc}using equation (11).
Step 11: Compute test statistics $T(x)$.
Step 12: Compare $T(x)$ with τ and return any one of two hypothesis decision.
Step 13: ComputeP_D, $P_{FA}$and P_{MD}using required parameters.
Step 14: Calculate $Q_D$and $Q_{FA}$for estimate Collaborative sensing efficiency.

III. Simulation And Result

A. Simulation parameters

Table I include the simulation parameters followed by the above scheme. To calculate local detection probability, false alarm probability and collaborative detection with MRC following parameters are considered.

TABLE I. Parameters for Calculating PD, PMA and PFAfor Local and Collaborative Environment

Parameter	Description	Value
τ	Threshold	.002-.04
γ	Instantaneous SNR	2-15 db
δ_t	Sensing time	50-150 ms
f_s	Sampling frequency	10-400 Hz
σ_s	Variance	.001
n	Number of user	1-30
M	Size of sampling vector	20-60
P_{FA}	False alarm probability	.001-.5

B. Simulation Results

To identify the tradeoff between the probability of detection and probability of false alarm, Receiver Operating Characteristics(ROC) analysis has been used. This section provides simulation and analytical result to verify and compare the ROC curves in different scenarios. All figures show that theoretical results are closely meet with simulation result. Therefore, maximum confidence level is achieved. At first, we show the performance and tradeoff between probability of detection and false alarm for non-cooperative sensing environment which is very important to compare the sensing efficiency in a collaborative manner. For a perfect reporting channel, the basic requirement is to identify the threshold. In this paper, the threshold selection is carried out by considering present conditions of noise level using constant false alarm rate method from equation (2).

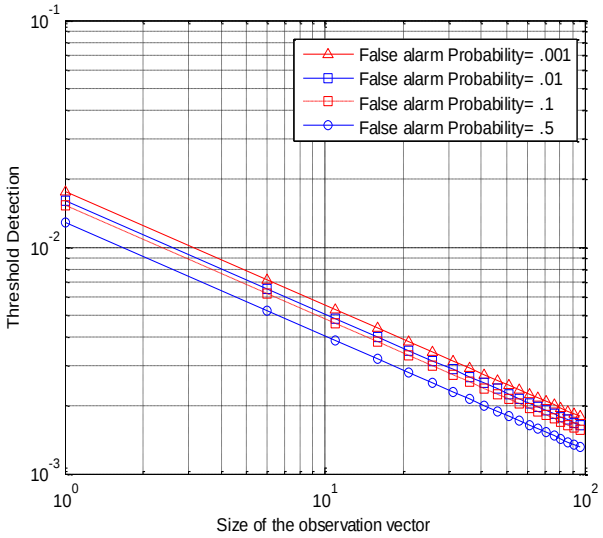


Fig. 5. Complementary ROC curves of threshold detection over size of observation vector for constant false alarm rate

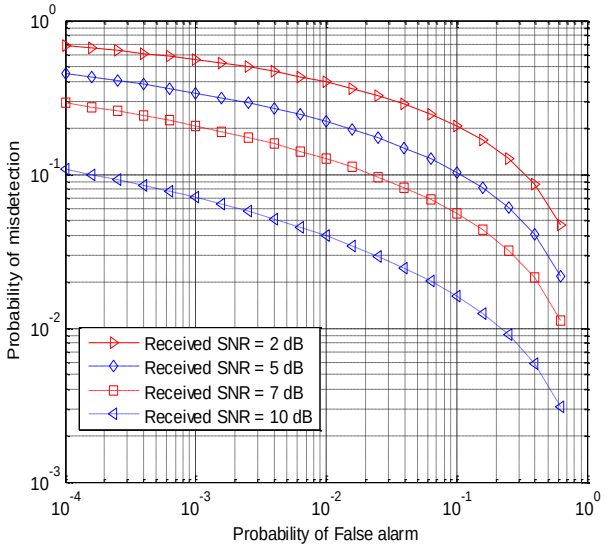


Fig. 6. Complementary ROC curves of energy detection over AWGN

Fig 5 shows ROC curves to identify perfect threshold value for different constant false alarm rate. Fig 6 shows the effect of probability of false alarm on misdetection where each user sees different SNR. We observe from this figure that for large SNR the probability of misdetection over false alarm will decrease. It also shows that energy detection works better for higher SNR. Fig 7 shows the performance of an energy detector for fixed false alarm rate which varies with frequency. From Fig 7, it is observed that when the P_{FA} decreases, the P_{MD}also gradually decreases. For low false alarm rate probability of misdetection remain low. Fig 8 demonstrates that for a large spectrum sensing period the performance of energy detection increases with less approximate error.

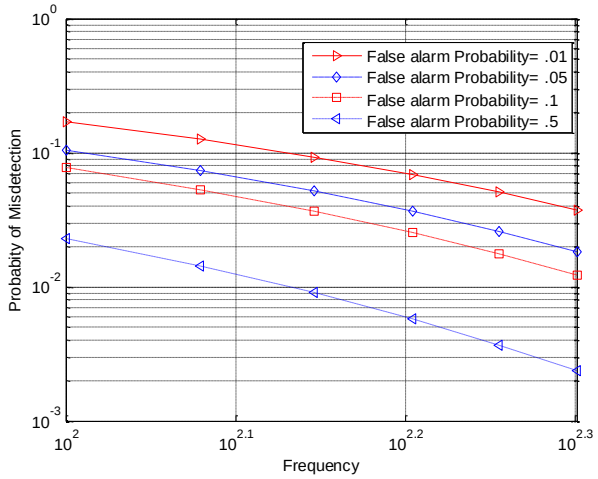


Fig. 7. Variation of the probability if misdetection against frequency for fixed false alarm rate

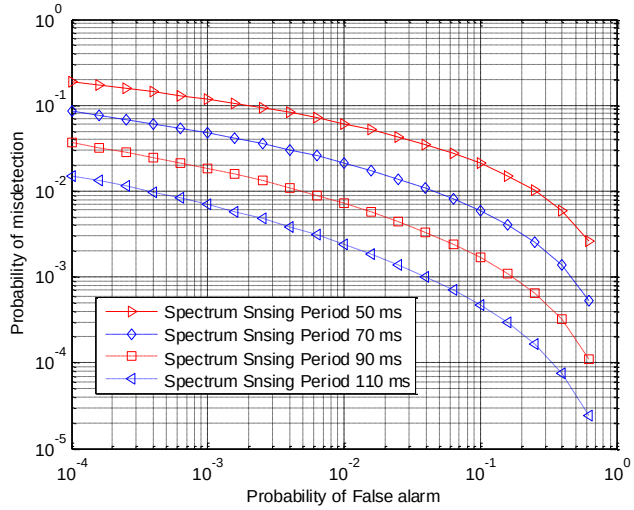


Fig. 8. ROC curve for the probability of misdetection VS probability of false alarm for various sensing period

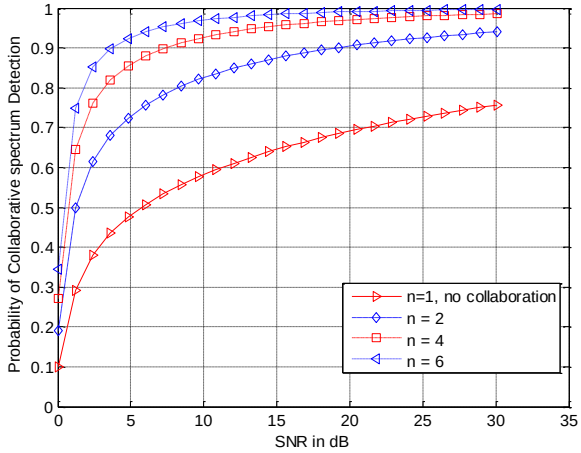


Fig. 9. Variation of probability of detection against SNR for different number of receiving antenna

Now we discuss about the performance of collaborative spectrum sensing cases. In these following simulations, it has been observed that collaborative spectrum sensing works better than non-collaborative environment. Figs 9-11 manifest that the probabilities of collaborative detection will decreases when either the number of collaborative user *n* or SNR decreases. Large number ofcollaborative user produces better performance with less misdetection in FC. Fig 10 shows the effect of number of collaborative users on the probability of collaborative detection. It has been observed for the figure that the collaborative detection rises with the sensing time. Fig 12 shows the ROC curve of comparison between normal collaborative (without diversity scheme) and collaborative detection using MRC diversity. We observe from this figure that MRC scheme exhibits the best detection performance with energy detection in collaborative environment though it requires channel state information withcomplex factor discussed in section II.

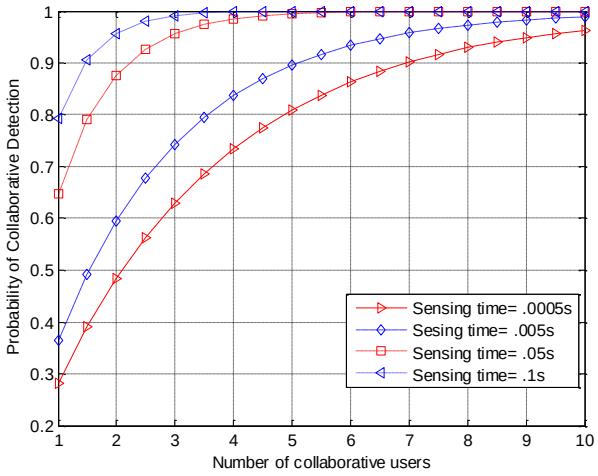


Fig. 10. ROC curve for energy detection VS. number of users for different sensing period in collaborative manner

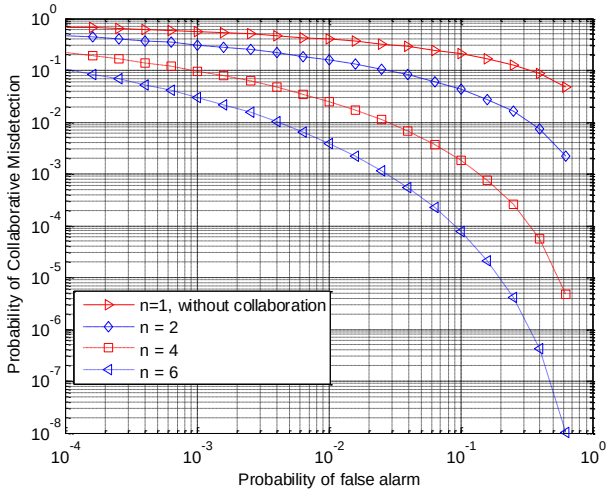


Fig. 11. Complementary ROC curves for collaborative energy detection against false alarm for different number of receiving antenna

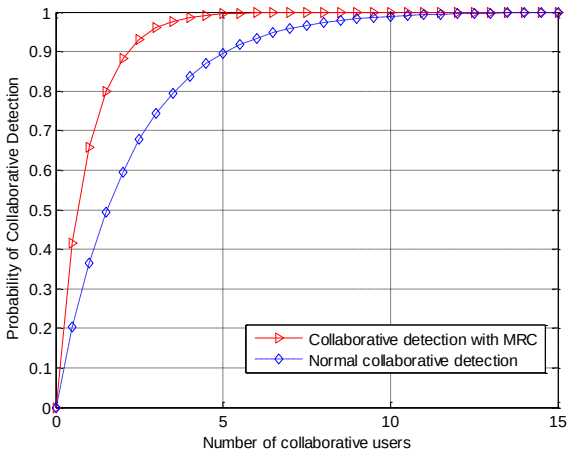


Fig. 12. Complementary ROC curve of collaborative detection with MRC

IV. Conclusion

In this paper spectrum sensing concepts are re-evaluated with collaborative sensing using MRC by considering different dimensions of the spectrum space. We have used energydetection method to sensing unused spectrum. In this method, probability of correct decision completely depends upon appropriate value of threshold. In our paper, an adaptive thresholdalgorithm is used to select suitable threshold value in vigorously changing environment. We have focused on optimum spectrum sensing in cognitive radio network based on differentrequired parameters. This paper proposes a new architecture in collaborative spectrum sensing with MRC diversity. This System provides efficient spectrum sensing and consequently leads to enhanced CR performance. From the simulation results, it is observed that there is significant reduction of the probability of misdetection with increasing in the number of collaborative user.

References

[1] S haykin, "Cognitive radio: Brain-empowered wireless communication", IEEE Journal Selected Areas in Communications, vol.23, no.2, pp.201-202, Feb.2005.

[2] J. G. Proakis, "Digital communications" ed: McGraw-Hill, New York, 1995.

[3] Aamir Zeb Shaikh, Dr. Talat Altaf, "Collaborative spectrum sensing under suburban environments", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.7, 2013.

[4] Y. Zeng, et al., "A review on spectrum sensing for cognitive radio: challenges and solutions," EURASIP Journal on Advances in Signal Processing, vol. 2010, pp. 2, 2010.

[5] Daniela Mercedes Martnez Plataa, ngel Gabriel Andrade Retiga, "Evaluation of energy detection for spectrum sensing based on dynamicselection of detection threshold", International Meeting of Electrical Engineering Research, ENIINVIE-2012.

[6] Deepak R. Joshi , Dimitrie C. Popescu and Octavia A. Dobre, "Adaptive spectrum sensing with noise variance estimation for dynamic cognitive radio systems", IEEE, 2010.

[7] Adeel Ahmed, Yim Fun Yu,James M.Nora , " Noise Variance estimationfor spectrum sensing in cognitive radio" AASRI Conference on circuitsand signal processing, 2014.

[8] Y. Liang, L. Lai, and J. Halloran, "Distributed algorithm for collaborative detection in cognitive radio networks", Communication Control and Computing, pp. 394 -399, 2009.

[9] H. Li, H. Dai, and C. Li, ”Collaborative quickest spectrum sensing via random broadcast in cognitive radio systems”, IEEE Global Telecommunications Conference, pp. 1-6, 2009.

[10] H. Teng-Cheng, W. Tsang-Yi, and H. Y.-W. Peter, ”Collaborative change detection for efficient spectrum sensing in cognitive radio networks”, IEEE Vehicular Technology Conference, pp. 1-5, 2010.

[11] H. Li, M. Junfei, X. Fangmin, L. ShuRong, and Z. Zheng., ”Optimization of collaborative spectrum sensing for cognitive radio”, NetworkingSensing and Control, pp. 1730-1733, 2008.

[12] Yue Wang, Justin P. Coon, Angela Doufexi, ”Energy-efficient spectrum sensing and access for cognitive radio networks. Vehicular Technology”, IEEE Transactions on, Vol.61(2), pp. 906-912, 2012.

[13] Mesodiakaki A., Adelantado F., Alonso L., and Verikoukis C., ”Energy effi- ciency analysis of secondary networks in cognitive radio systems” In Communications (ICC), IEEE International Conference on, (pp. 4115-4119). IEEE, June 2013.

[14] Adelantado F., and Verikoukis C., ”Detection of malicious users in cognitive radio ad hoc networks: A non-parametric statistical approach” Ad Hoc Networks, Vol.11(8), pp.2367-2380, 2013.

[15] Wang, W., Wu, K., Luo, H., Yu, G., and Zhang, Z., ”Sensing error aware delay-optimal channel allocation scheme for cognitive radio networks”, Telecommunication Systems, Vol. 52(4), pp. 1895-1904, 2013.

[16] D.cabric, S.Mishra, R.Brodersen , ”Implementation issues in spectrum sensing for cognitive radios” in: Proc. Of Asilomar Conf. on Signals, System and Computers, vol.1, pp.772-777, 2004.

[17] D. Teguig , B. Scheers and V. Le Nir, ”Data fusion schemes for cooperative spectrum sensing in cognitive radio networks”, Communications and Information Systems Conference (MCC), Military, IEEE 8-9 , pp:1 7 Print ISBN: 978-1-4673-1422-0, Oct.2012.

[18] F. F. Digham, M. S. Alouini, and M. K. Simon, ”On the energy detection of unknown signals over fading channels”, IEEE Trans. Commun., vol. 55, no. 1, pp. 2124, Jan. 2007.

[19] E. Visotsky, et al., ”On collaborative detection of TV transmissions in support of dynamic spectrum sharing”, in New Frontiers in Dynamic Spectrum Access Networks, First IEEE International Symposium on, pp. 338-345,2005.

[20] Deep Raman and N.P.Sing , ”An Algorithm for spectrum sensing in Cognitive Radio under noise uncertainty”, International journal of Future Generation communication and Networking, Vol.7, No. 3, pp. 61-68, 2014.

[21] Ahmed El Shafie, ”Optimal Spectrum Access for Cognitive Radios”, Wireless Intelligent Networks Center (WINC), Nile University, Giza, Egypt. 24, arXiv:1208.4508v5 [cs.IT], 24 Mar 2013.

[22] Xiaoge Huang, Ning Han, Guanbo Zheng, Sunghwan Sohn, Jaemoung Kim, ”Weighted-Collaborative Spectrum Sensing in Cognitive Radio” ,Communications and Networking in China, CHINACOM Vol.07. Second International Conference on, 2007.

[23] Simon O. Haykin, Michael Moher, ”Modern Wireless Communications”, ISBN-10: 0130224723.

A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA

Amit Gupta
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

Ali Syed
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

Azeem Mohammad
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

Malka N. Halgamuge
School of Computing and Mathematics
Charles Sturt University
Melbourne, Victoria

***Abstract*—In the last five years, crime and accidents rates have increased in many cities of America. The advancement of new technologies can also lead to criminal misuse. In order to reduce incidents, there is a need to understand and examine emerging patterns of criminal activities. This paper analyzed crime and accident datasets from Denver City, USA during 2011 to 2015 consisting of 372,392 instances of crime. The dataset is analyzed by using a number of Classification Algorithms. The aim of this study is to highlight trends of incidents that will in return help security agencies and police department to discover precautionary measures from prediction rates. The classification of algorithms used in this study is to assess trends and patterns that are assessed by BayesNet, NaiveBayes, J48, JRip, OneR and Decision Table. The output that has been used in this study, are correct classification, incorrect classification, True Positive Rate (TP), False Positive Rate (FP), Precision (P), Recall (R) and F-measure (F). These outputs are captured by using two different test methods: k-fold cross-validation and percentage split. Outputs are then compared to understand the classifier performances. Our analysis illustrates that JRip has classified the highest number of correct classifications by 73.71% followed by decision table with 73.66% of correct predictions, whereas OneR produced the least number of correct predictions with 64.95%. NaiveBayes took the least time of 0.57 sec to build the model and perform classification when compared to all the classifiers. The classifier stands out producing better results among all the classification methods. This study would be helpful for security agencies and police department to discover data patterns and analyze trending criminal activity from prediction rates.**

Keywords*—Data Mining; Classification; Big Data; Crime and Accident

I. INTRODUCTION

Technologies provide companies new ways to gather talents of innovators working outside corporate margins. Corporate companies create real prosperity when they combine technology with new ways of doing business and storing data at a standard. There is a need to store data as the Computer technology and the use of Internet has heightened the use of social media such as Facebook and Twitter. The increase in social media urges the need for collecting, storing and processing data for company's development. Analyzing this big data is a challenging process, and therefore the need for certain tools and techniques that are significant in sorting huge amounts of data becomes extremely important. Data Mining is one of the disciplines that is used to convert raw data into meaningful information and knowledge [1]. Data mining searches and analyses large quantities of data automatically by discovering, learning and knowing hidden patterns, trends, and structures [2] and it answers questions that cannot be addressed through simple query and reporting techniques [3]. Data Mining is broadly classified into two categories [4], Predictive Data Mining: that deals with the use of few attributes from a dataset and foretells the future value, or it could also be said that the developing model of the system as per given data. On the other hand, Descriptive Data Mining: finds patterns that describe the data, in other words, presenting new information based on the available dataset trends available.

With the use of new tools and techniques, the offenses and accidents are tracked, monitored and reduced; but at the same time, people are getting more knowledgeable about different crimes and ways to perform them with information available online at their fingertips. The use of technology such as surveillance cameras, speed detection devices, fire and burglary alarms, has helped various monitoring and tracking easier than ever. The types of software that are used today, stores huge amount of data that is collected every day [5]. A particular data set related to crimes and accidents from Denver city, USA has been obtained, and data mining techniques are applied to analyze and find information. The criminal activities and accidents show that there is an increase in death rates in the USA [6]. The major cause of road accidents is drink driving, over speed, carelessness, and the violation of traffic rules [5]. Assessing the cause of crimes is extremely important as it makes taking precarious measures easier.

Education or informing police depends on these assessments. Additionally, the cause of these accidents is only preventable if they are tracked and evaluated to inform police in taking measures for minimizing it and bringing awareness to public. This paper is organized as follows. In Section II, we introduce the dataset and attributes in it, and how the data was collected and pre-processed. It also lists and explains the selected classification algorithms. Section III outlines the results obtained by using two different test methods and also the dataset is analyzed on different criteria's giving us insight on trends and patterns of incidents that have occurred in the due course. Section V concludes the paper.

II. Materials and Methods

This paper has used the predictive method of data mining where the particular attribute value is predicted based on other related attributes. A few classification algorithms: BayesNet, NaiveBayes, OneR, J48, Decision Table and JRip are used in this paper to predict the outcomes of collected statistical data.

A. Data Collection

Data is collected from statistical websites: US City open data census and official government site of Denver city from the year 2011 to 2015, and this data is based on the National Incident-Based Reporting System (NIBRS) where the data is updated every day. This dataset excludes crimes related to child abuse and sexual assault as per legal restrictions law. This Dataset contains 15 attributes and 372,392 instances.

TABLE I. Attribute Description for Classification

Attribute Name	Description
Incident-ID	Unique identification number for a particular incident.
Offense-ID	Unique identification number related to particular Offense.
Offense-Code	Code associated to each offense type
Offense-TypeID	Different types of offenses
Offence-CategoryID	Offenses grouped / assigned into categories.
First-Occurrence-Date	Date incident first occurred on.
Last-Occurrence-Date	Date incident last occurred on.
Reported-Date	Date on which the incident was reported.
Incident –Address	Address of the location where an incident happened.
GeoX	Geographical location
GeoY	Geographical location
District-ID	Name of the district where an incident took place.
Precinct-ID	Precinct name where an incident occurred.
Neighbourhood-ID	Nearby location to the incident
Incident Type	Type of incident (crime/accident)

B. Data Pre-processing

The raw data obtained does not give any information in the form it appears. The raw data stored could contain errors due to multiple reasons like, missing data, inconsistencies that arise due to merging data, incorrect data entry procedures, and so on [7]. Deriving meaningful information from the raw data requires preprocessing of data that converts real-time data into computer readable format. The phases involved in data processing are as shown in Fig. 1.

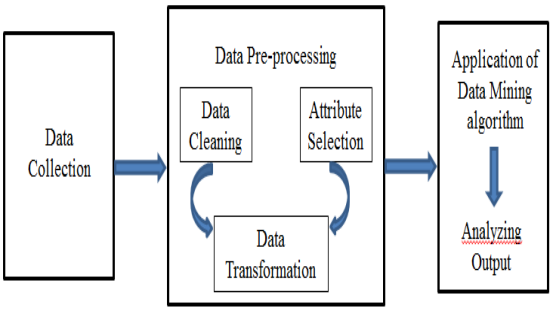


Fig. 1. Data processing of crime and accident dataset obtained for Denver City the USA

The preprocessing is an important phase in data mining. This stage involves the attribute selection, data cleaning, and data transformation [8]. This process starts off with data collection, then the required features or attributes have been selected from the raw data, ready for analysis. Then Data cleaning was performed by eliminating the errors and missing values, with the correction of syntaxes, for example, the address attributes. Finally, the data is prepared and transformed into a suitable and readable format for the data-mining tool to generate.

C. Classification Algorithms

A number of classifications and algorithms are available, and few of them have been selected and used. Below table presents the method used and gives a brief description of the approach and how it is matched with the classifier. The classifiers that are selected are Bayesian, decision trees, and rules based which are outlined in Table 2.

TABLE II. Classification Methods Used in this Study and Description of the Methods

Classifier	Description
NaiveBayes	This supervised learning algorithm is a probabilistic classifier and uses statistical method for each classification.
J48	J48 is an algorithm that generates decision tree using C4.5 algorithms an extension of ID3 algorithm and is used for classification.
JRip	It implements a propositional rule learner called as “Repeated Incremental Pruning to Produce Error Reduction (RIPPER)” and uses sequential covering algorithms for creating ordered rule lists. The algorithm goes through 4 stages: Growing a rule, Pruning, Optimization and Selection [9].
BayesNet	Bayes Net model represents probabilistic relationships among a set of random variables graphically. It models the quantitative strength of the connections between variables, allowing probabilistic beliefs about them to be updated automatically as new information that becomes available. It is a directed acyclic graph (DAG) G that encodes a joint probability distribution, where the nodes of graph represent random variable and arc represent correlation between variables [10].
OneR	A simple classification that produces one rule for each predictor in the data and then the rule with smallest total error is selected [11].

Decision Table	Builds a simple decision table majority classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation.

D. Data Analysis

This study deals with applying the stated classification algorithms in Table 2, to the crime and accident dataset obtained from Denver city, and compared the outputs/results of the classification methods. The analysis is performed based on varied outputs attained from identified number of correct instances and less execution time taken to build the model. The evaluation also helps to gain insights onto which incidents are high in number overall, during a given period of time, and how the trends have been for the last five years.

The software used for this analysis and application of algorithms is Weka (Waikato Environment for Knowledge Analysis, version 3.7). This software allows people to compare different machines to learn algorithms on datasets [11] that contain a collection of visualization tools and algorithms. It is useful for predictive modeling and analyzing data, along with graphical user interfaces for easy access to this functionality [12].

III. Results and Discussions

Results obtained this study are based on different test options: k-fold cross-validation and percentage split criteria.

A. Prediction: k-fold validation

This study has used K-fold cross validation (k=10) method. This method runs the test 10 times, and the first 9 times is used for training, and the final fold is for testing [3] [13], and we have also used the percentage split approach for comparing the outputs and performance of used algorithms. Performances and outputs of each classifier method obtained are compared and presented in Table 3.

TABLE III. Classifiers Accuracy on the Dataset Based on 10-Fold Cross Validation Test Mode

Classification Method	Correctly Classified Incidents	Incorrectly Classified Incidents
NaiveBayes	66.80%	33.19%
Bayes net	68.74%	31.25%
J48	73.54%	26.45%
OneR	64.95%	35.04%
Decision Table	73.66%	26.34%
JRip	73.71%	26.28%

JRip classifier has identified a number of incidents correctly with 73.71%, followed by Decision Table having correct classification rate of 73.66% compared to other classifiers and OneR has determined least correct instances with 64.95%.

TABLE IV. Classifier Execution Time and Root Mean Square Error on the Dataset Based on 10-Fold Cross Validation Test Mode

Classification Method	Time to Build the Model (Seconds)	Root Mean Squared Error
NaiveBayes	0.57	0.460
Bayes net	4.34	0.461
J48	0.87	0.440
OneR	0.81	0.592
Decision Table	18.6	0.435
JRip	21.27	0.440

Execution time is higher for JRip with 21.27 sec and Decision Table with 18.6 sec, while NaiveBayes time to build the model was the least with 0.57 sec, with J48 and OneR time for a model build is 0.87 sec and 0.81 sec, respectively.

There are different performances and measures that are calculated based on the confusion matrix produced by the algorithms. Fig. 2 portrays the model of confusion matrix also known as contingency table. In this matrix, each row exhibits the actual class and column exhibits the predicted class [11].

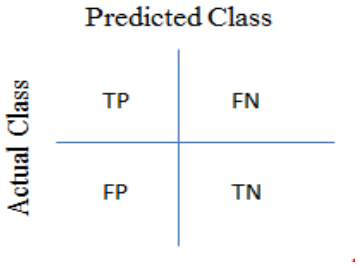


Fig. 2. Confusion Matrix representation

TP (True Positive) and TN (True negatives) are instances correctly classified as a given class and FP (False Positive) and FN (False Negative) are the instances falsely classified as a given class. Other measures are: Precision - % of selected items that are correct and are calculated as Precision (P) = TP / (TP+FP) and Recall - % of correct items that are selected and the calculation for it is Recall (R) = TP / (TP+FN) [14]. With the help of Precision and Recall is calculated F-Measure (F) - the Harmonic mean of precision and recall, calculated as F=2*R*P/(R+P).

TABLE V. Performance Measures Calculated Based on Confusion Matrix Using 10-Fold Cross Validation

Classifier	TP Rate	FP Rate	Precision (P)	Recall (R)	F-Measure (F)
NaiveBayes	66.80%	53.30%	66.50%	66.80%	66.60%
Bayes net	68.70%	55.20%	66.90%	68.70%	67.70%
J48	73.60%	73.60%	54.20%	73.60%	62.50%
OneR	65.00%	12.50%	85.00%	65.00%	66.50%
Decision Table	73.70%	73.30%	68.10%	73.70%	62.70%
JRip	73.70%	73.10%	70.50%	73.70%	62.90%

Above Table 5 shows the TP and FP rate of each classifier, the weighted average of Precision, Recall and F-Measure, obtained by using the 10-fold cross-validation approach.

Decision Table and JRip have the highest TP Rate (True Positive) by 73.7% and Recall values73.7%, followed by J48 having TP rate and recall value of 73.6%. OneR has greater precision when compared to other algorithms.

B. Prediction: Percentage Split

Another test option of split criteria available is also used to compare and evaluate the classifier outputs. In the percentage split method, the algorithm is trained in a certain percentage of

data first, and then the learning is tested on the remainder of the data. Table 6 presents the result of classifier output based on split criteria.

TABLE VI. RESULT OF CLASSIFIER ACCURACY BASED ON SPLIT CRITERION TEST MODE

Classifier	Train Data (%)	Test Data (%)	Correctly Classified (%)	Incorrectly Classified (%)
BayesNet	90	10	79.53	20.46
	80	20	78.59	21.40
	70	30	77.63	22.36
	60	40	76.79	23.20
	50	50	75.81	24.18
	40	60	74.63	25.36
	30	70	73.29	26.70
	20	80	72.42	27.57
	10	90	72.00	27.99
NaiveBayes	90	10	75.85	24.14
	80	20	76.18	23.81
	70	30	61.77	38.22
	60	40	61.92	38.07
	50	50	66.03	33.96
	40	60	61.48	38.51
	30	70	68.33	31.66
	20	80	30.04	69.95
	10	90	30.90	60.09
OneR	90	10	65.07	64.92
	80	20	63.02	36.97
	70	30	60.68	39.31
	60	40	57.92	42.07
	50	50	55.11	44.88
	40	60	51.40	48.59
	30	70	47.24	52.75
	20	80	41.93	58.06
	10	90	35.14	65.85
J48	90	10	73.61	26.38
	80	20	73.67	26.32
	70	30	73.62	26.37
	60	40	73.71	26.28
	50	50	73.68	26.31
	40	60	73.70	26.29
	30	70	73.61	26.38
	20	80	73.61	26.38
	10	90	73.64	26.35

Figures 3, 4, 5 and 6 demonstrate the graphical representation of the corresponding classifier output. Figures 3, 4 and 5 indicate Bayes net, NaiveBayes and OneR perform identically. When the percentage of data tested is less the results are more accurate. As the amount of test data increases the percentage of correct classification decreases as a result. This is because a number of data samples trained are less. As seen from Fig 6 it shows that J48 has correctly classified the higher number of instances when the test and trained data is almost equal, and lowest classification rate are when test data is either least or most.

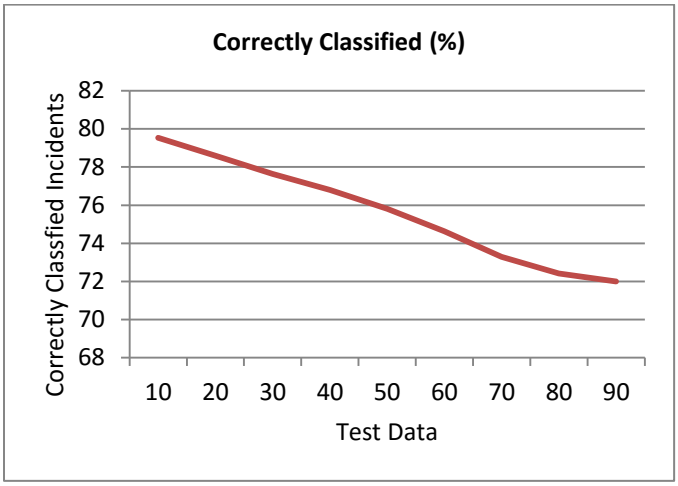


Fig. 3. Bayes net Classification using split percentage test option

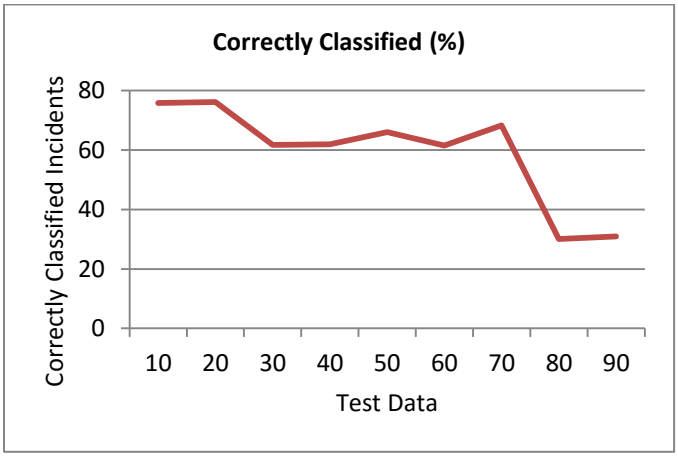


Fig. 4. NaiveBayes Classification using split percentage test option

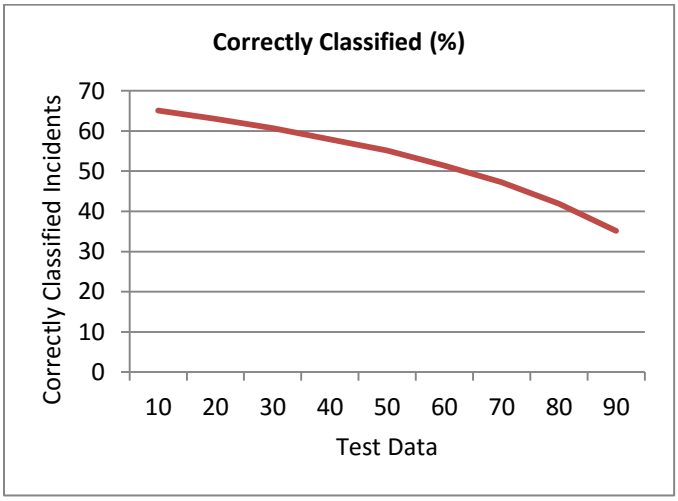


Fig. 5. OneR Classification using split percentage test option

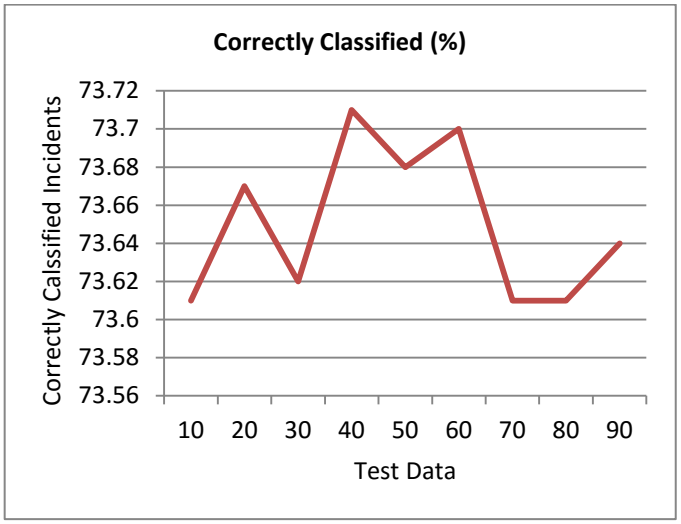


Fig. 6. J48 Classification using split percentage test option

Further analysis of data is performed based on different criteria's.

TABLE VII. CRIME AND ACCIDENT ON WEEKDAY/WEEKEND

	Accident	Crime	Total
Weekday	84,475	189,783	274,258
Weekend	25,106	73,028	98,134
Grand Total	**109,581**	**262,811**	**372,392**

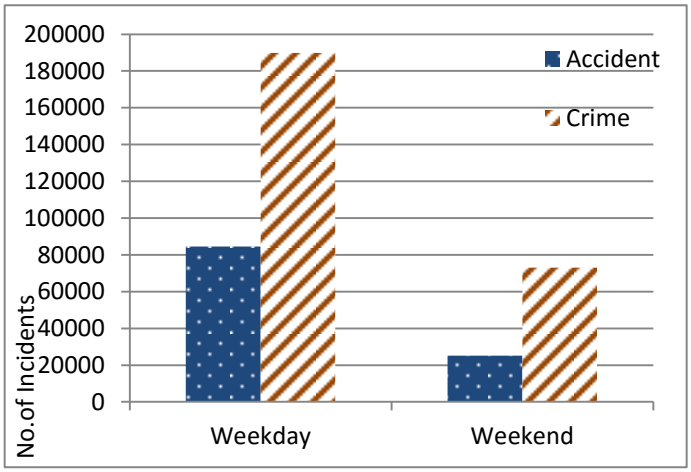


Fig. 7. Crime and accident based on weekday and weekend

TABLE VIII. COUNT OF INCIDENTS ON A MONTHLY BASIS

Month	Crime	Accident	Total
January	24,364	10,525	34,889
February	20,904	10,004	30,908
March	22,010	8927	30,937
April	19,018	8186	27,204
May	20,935	8708	29,643
June	22,085	8781	30,866
July	23,951	8887	32,838
August	24,322	9306	33,628
September	22,833	9203	32,036
October	22,477	9345	31,822
November	20,193	8528	28,721
December	19,719	9181	28,900
Grand Total	**262,811**	**109,581**	**372,392**

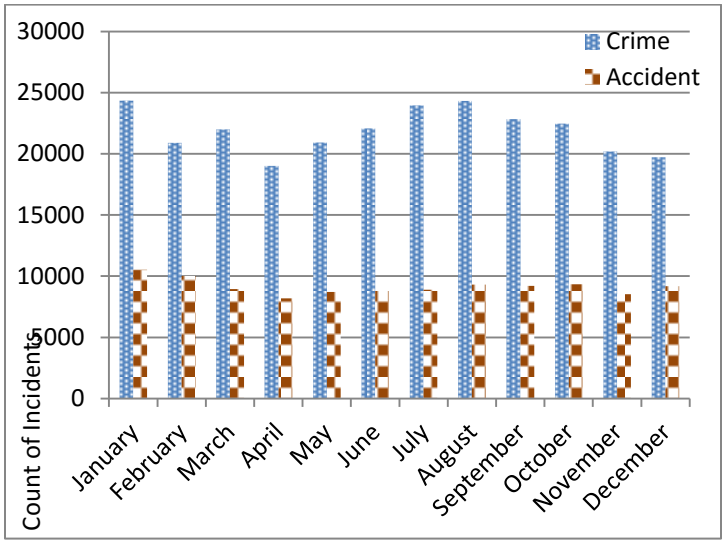


Fig. 8. Count of crime and accidents on a monthly basis

Figure 8 indicates that crime and accidents are more likely to occur during the months of January and February. This is because people start their daily routines after a long vacation of Christmas and New Year. As a result, more public is out in the traffic as people commute and drive to, schools, offices, and work. The trends show an increase of incidents that occur during July and August, as this is the start of the academic year for schools and colleges. During this time, accidents are 60% lower on the weekends when compared to weekdays due to less traffic and crowd on roads. Crime is 60% less on the weekends, as most people stay home relaxing; therefore, crimes such as murder, burglary, and robbery are less likely to occur.

TABLE IX. YEAR-WISE PRESENTATION OF CRIME AND ACCIDENTS

Year	Accident	Crime	Total
2011	20,722	36,419	57,141
2012	19,398	36,258	55,656
2013	19,588	51,820	71,408
2014	21,914	61,340	83,254
2015	23,245	63,632	86,877
2016	4714	13,342	18,056
Total	**109,581**	**262,811**	**372,392**

TABLE X. TYPES OF OFFENSES

Offense Type	No. of Offenses
Murder	210
Arson	533
White-collar-crime	5299
Robbery	5908
Aggravated-assault	8030
Other-crimes-against-persons	13,544
Auto-theft	19,271
Drug-alcohol	21,488
Burglary	24,571
Theft-from-motor-vehicle	32,998
Larceny	40,737
Public-disorder	41,712
All-other-crimes	48,510
Total	**372,392**

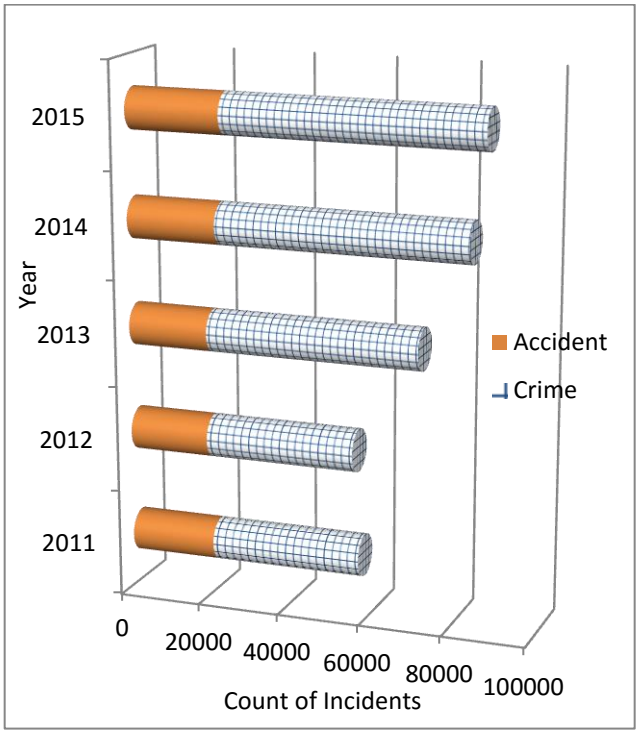


Fig. 9. Number of crime and accidents identified year-wise

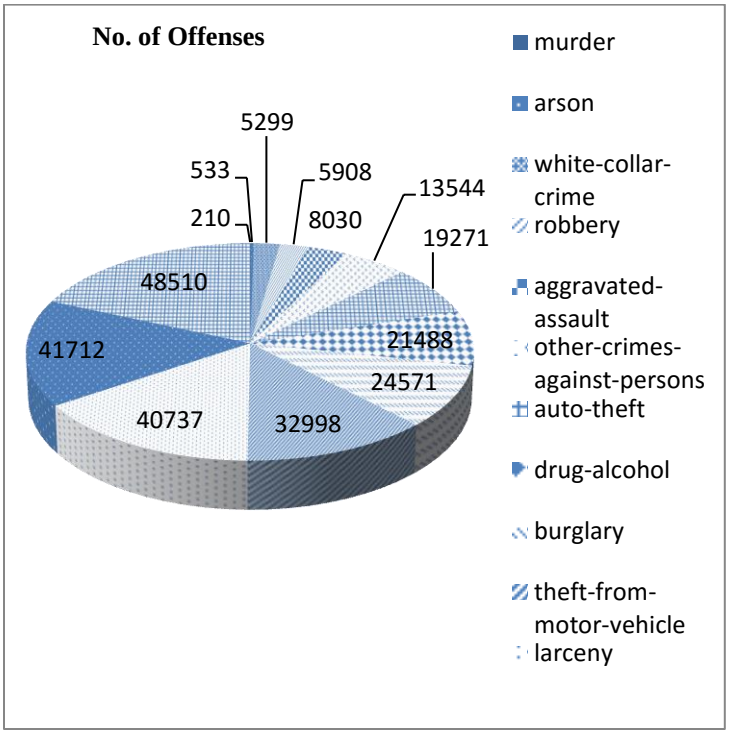


Fig. 10. Different types of offenses indicating number of incidents in each category

TABLE XI. COUNT OF INCIDENTS YEAR-WISE IN EACH OFFENSE TYPE

Offense Category	2011	2012	2013	2014	2015	2016	Total
Aggravated-assault	1314	1467	1522	1599	1755	373	8030
All-other-crimes	1843	1986	9920	15,491	15,589	3681	48,510
Arson	92	92	95	130	107	17	533
Auto-theft	3545	3421	3383	3514	4460	948	19,271
Burglary	4698	4711	4800	4553	4836	973	24,571
Drug-alcohol	1416	1714	4784	6061	6153	1360	21,488
Larceny	5959	6691	8350	9336	8778	1623	40,737
Murder	41	33	39	33	55	9	210
Other-crimes-Against-persons	1286	1427	2617	3649	3840	725	13,544
Public-disorder	6454	5948	8195	9728	9400	1987	41,712
Robbery	1133	1212	1058	1072	1188	245	5908
Theft-from-motor-vehicle	7575	6632	6222	5129	6226	1214	32,998
Traffic-accident	20,722	19,398	19,588	21,914	23,245	4714	109,581
White-collar-crime	1063	924	835	1045	1245	187	5299
Total	**57,141**	**55,656**	**71,408**	**83,254**	**86,877**	**18,056**	**372,392**

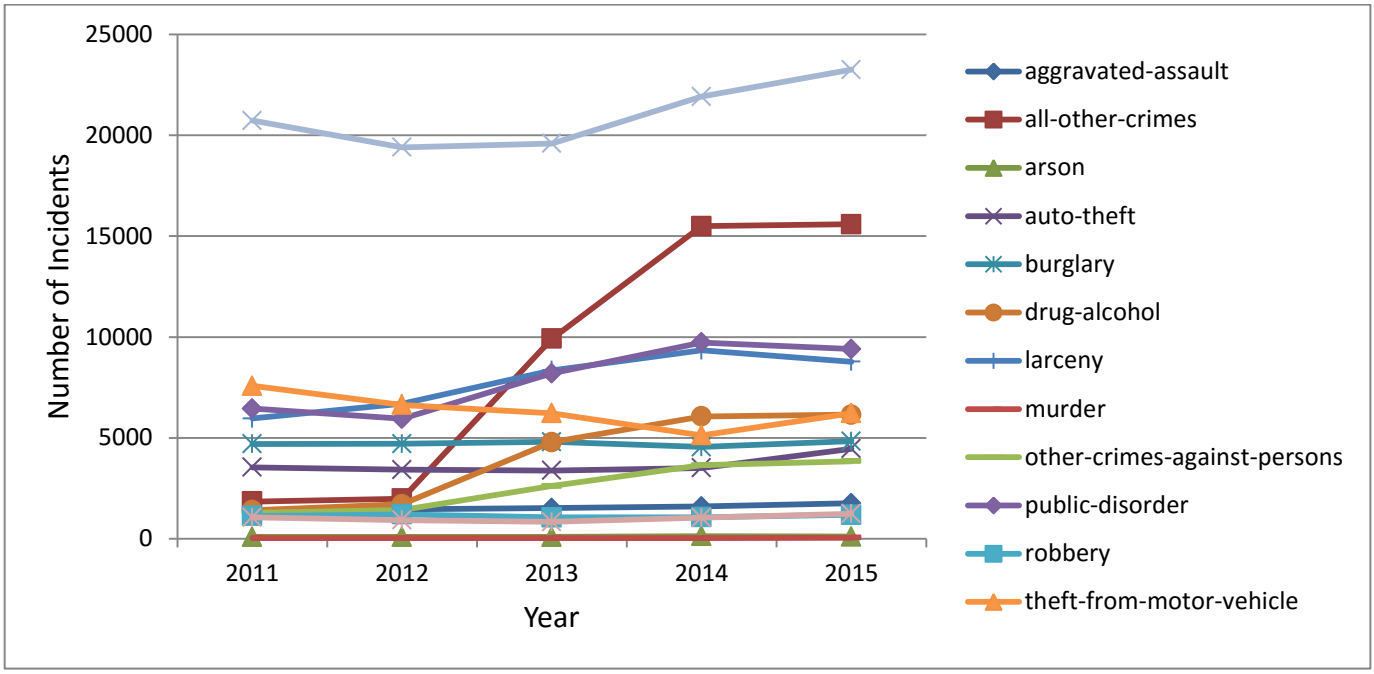


Fig. 11. Number of incidents occurring in each category of offense year-wise

Above Figure 11 shows that drug and alcohol consumption has been increasing year-by-year. In the year 2009, marijuana was legalized in many states of the US, it was allowed on the basis of certain medical conditions. However after a couple of years, it was legalized in Colorado as well. This legalization in 2012 has made the availability of it easier and since then the intake of this drug has been increasing continuously [15]. It is evident from the analysis results as per Fig. 11 from the year 2012-2013 there has been more than 100% increase in drug and alcohol consumption, nevertheless, no strong evidence has found that people consume marijuana truly for medical reasons.

IV. Conclusion

Data Mining techniques and tools have brought tremendous change in the way data is analyzed revealing useful information. This paper has analyzed the application and performance of six classification algorithms that produce different results. Different test methods were used to predict the outcomes for same classification methods. This study has found that various crime patterns have heightened in particular seasons. Results obtained for various classification methods show different outputs and performance measures. Our analysis indicates JRip and Decision Table classified the most number of correct incidents with 73.71% and 73.66%, whereas OneR classified showed the least number of correct incidents with 64.95%. Although JRip is the most accurate classifier, it took the maximum time building the model with 21.2 sec. NaiveBayes model builds the quickest time with 0.57 sec. This study is helpful for various agencies, police department and other organizations aiding them to foresee prediction rate of incidents and develop strategies, plans, and preventive measures for the purpose of crime reduction.

References

[1] J. H. Trevor, R. J. Tibshirani and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2011.

[2] C. C. Aggarwal, *Data Mining: The Textbook*. Springer, 2015.

[3] R. A. El-Deen Ahmeda, M. E. Shehaba, S. Morsya and N. Mekawiea, Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining. In*Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on* IEEE, pp. 1344-1349.

[4] S. Gnanapriya, R. Suganya, G. S. Devi and M. S. Kumar, Data Mining Concepts and Techniques. *Data Mining and Knowledge Engineering*, vol. 2, p. 256-263, 2010.

[5] K. B. Saran and G. Sreelekha, Traffic video surveillance: Vehicle detection and classification. In *2015 International Conference on Control Communication & Computing India (ICCC)* IEEE, pp. 516-521, November 2015.

[6] P. C. Kratcoski and M. Edelbacher, "Collaborative Policing: Police, Academics, Professionals, and Communities Working Together for Education, Training, and Program Implementation", CRC Press: 2015, vol. 25.

[7] S. García, J. Luengo and F. Herrera, *Data preprocessing in data mining*. Switzerland: Springer, 2015.

[8] R. Deb, A. W. C. Liew, Incorrect attribute value detection for traffic accident data. In *Neural Networks (IJCNN), 2015 International Joint Conference* IEEE, 2015, pp. 1-7.

[9] V. Veeralakshmi and D. Ramyachitra, Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset. *Issues*, vol 1, p. 79-85.

[10] Bayes Nets. Retrieved from http://www.bayesnets.com/

[11] I. H. Witten, E. Frank and M. A. Hall, Data Mining: Practical machine learning tools and techniques, 3rd ed., Morgan Kaufmann, 2011.

[12] S. Kalmegh, Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News, February 2015.

[13] C. Sitaula, "A Comparative Study of Data Mining Algorithms for Classification. *Journal of Computer Science and Control System"s*, vol. 7, 29.

[14] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi and A. I. Madbouly, A comparative analysis of classification algorithms for students college enrolment approval using data mining. In *Proceedings of the 2014 Workshop on InteractionDesign in Educational Environments*, 2014, ACM, p. 106.

[15] J. Schuermeyer, S. Salomonsen-Sautel, R. K. Price, S. Balan, C. Thurstone, S. J. Min and J. T. Sakai, Temporal trends in marijuana attitudes, availability and use in Colorado compared to non-medical marijuana states: 2003–11. *Drug and alcohol dependence*, 2014, vol 140, p. 145-155.

Simulation and Analysis of Optimum Golomb Ruler based 2D Codes for OCDMA System

Dr. Gurjit Kaur
Department of ECE
Gautam Buddha University, Greater Noida,
U. P., India

Aarti Bhardwaj
Department of ECE
Gautam Buddha University, Greater Noida,
U. P., India

Rajesh Yadav
Department of ECE
Gautam Buddha University, Greater Noida,
U. P., India

Manu Gangwar
Department of ECE
Gautam Buddha University, Greater Noida,
U. P., India

Disha Srivastava
Department of ECE,
Gautam Buddha University, Greater Noida,
U. P., India

Nidhi
Department of ECE,
Gautam Buddha University, Greater Noida,
U. P., India

***Abstract*—The need for high speed communications networks has led the research communities and industry to develop reliable, scalable transatlantic and transpacific fiber-optic communication links. In this paper the optimum Golomb ruler based 2D OCDMA codes has been demonstrated. An OCDMA system based on the discussed 2D codes is designed and simulated on Optisystem. The encoder and decoder structure of OCDMA system have been designed using filter and time delays. Further the performance is analysed for various parameter such as bit rate, number of users, BER (Bit Error Rate), quality factor, eye diagram and signal diagram. The system is analyzed for up to 18 users at 1 Gbps and 1.25 Gbps bit rate.**

Keywords*—OCDMA System; 2D Codes; OOC; Golomb Ruler; BER; Eye Diagram; MAI

I. INTRODUCTION

In Local Area Networks (LANs), since the traffic is bursty, it demands high speed and large capacity communication network. The optical fiber addresses these requirements because the bandwidth of optical fiber is enormous and it can provide higher carrier frequency and therefore greater information carrying capacity of the communication and higher transmission bandwidth for the communication [1]. There are various multiple access techniques [2] which are being used to accommodate the large number of users such as Optical Code Division Multiple Access (OCDMA). The OCDMA system plays an essential role in long haul and high speed communication where users share the same transmission media [3] as shown in Figure 1.

In OCDMA each user is assigned a unique signature code which is modulated by the data of the corresponding user. The signal from all the users is combined on a single optical fiber, which is broadcasted to each user in the network. Single-user decoding is achieved by correlating the aggregate signal and the signature sequence of the desired user. If the output of the decoder is in autocorrelation then the receiver can detect the signal sent to it. On the other hand, if the decoder is in cross correlation then the receiver cannot receive the signal. For OCDMA systems, optical codes should have maximum autocorrelation and minimum cross correlation property.

As the number of user increases the Multiple Access Interference (MAI) also increases and this is the main cause of performance degradation in OCDMA network. So cross correlation is needed to be kept less for maintaining probability of error low. Many codes have been proposed for the OCDMA system. Mendez et al. presented the one dimensional optical orthogonal code [4].

In one dimensional (1-D) codes, on increasing the number of users, the length of the codes also increases. And hence, the bit rate decreases for a given chip width [5]. To overcome this problem of 1-D codes in OCDMA, two dimensional (2-D) codes have been proposed such as Time-Space (T/S) and Wavelength-Time (W/T).

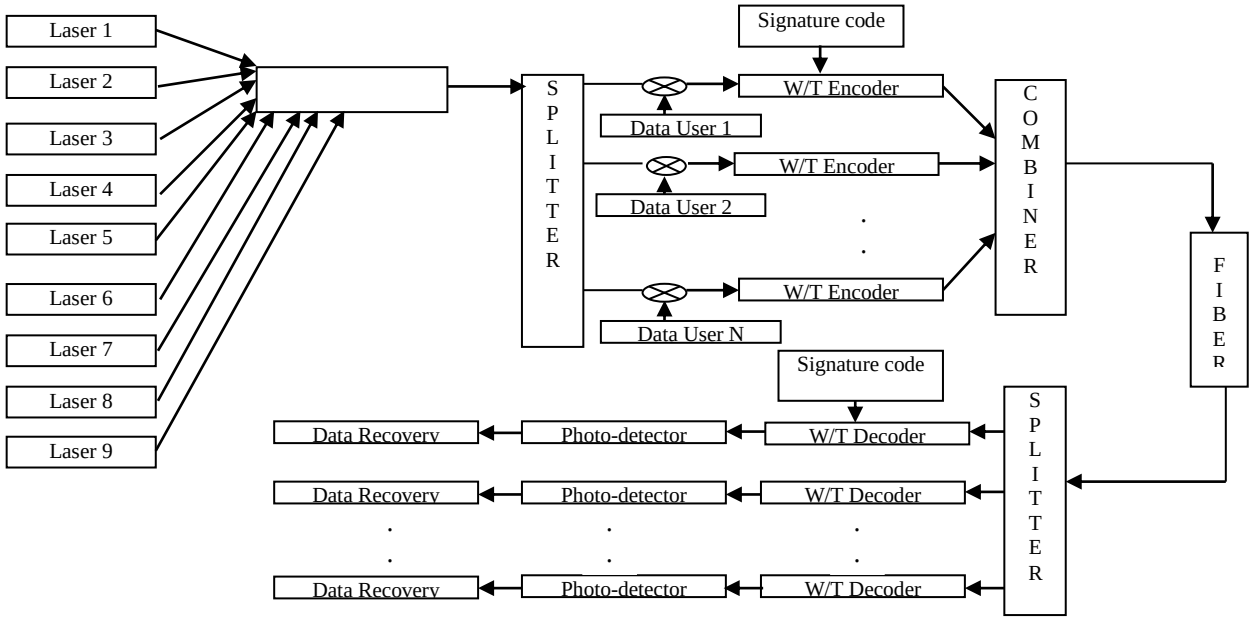


Fig. 1. OCDMA system Block Diagram

Heo have proposed the construction of 2-D wavelength-time codes by hybridization of the prime codes and the pseudo-random noise codes using differential detection with on-off keying [6]. Yeon have designed modified pseudo-random noise codes for W/T spreading with two sequences having different lengths [7]. Wan and Hu have constructed hybrid codes by concatenating the prime codes and optical orthogonal codes[8]. Mendez construct 2-D codes from 1-D Golomb rulers to increase the number of code set size [9]. In 2D optical codes, the length of the codes reduces and hence improves the BER performance. Large numbers of researchers are working on the design of 2D codes for OCDMA system [9 - 10]. In this paper, the design of Optical Orthogonal Codes (OOC) using optimum Golomb rulers has been demonstrated and the OCDMA system performance on Optisystem tool has been analyzed.

The rest of paper is organized as follows: Section II discusses the OCDMA coding theory that gives an insight of the optical codes for optical CDMA communication system. The mathematical modeling of proposed 2D optical code is described. Section III presents a concise introduction to optical OCDMA systems for simulation. Section IV discusses the result of the proposed optical prime codes and its performance estimation in terms of the autocorrelation and cross-correlation function. In section V the current findings along with the future directions are concluded. The paper ends with the references studied and cited in the paper.

II. Optical Coding Theory

A. Construction of codes

The W/T code can be represented as matrices with wavelength and time as axis. This matrix is known as Pseudo Orthogonal (PSO) matrix code. Total wavelength is divided into n different channels and total time is divided into m time slots. These PSO matrix codes are constructed with the help of spanning ruler or optimum Golomb ruler [11].

N-mark Golomb ruler is a set of n distinct non negative integers ($a_1, a_2, a_3, \ldots a_n$) called "marks" such that the positive differences $|a_i - a_j|$ computed over all possible pairs of different integers i, j = 1,2,3....n with $i \neq j$, are distinct.

A perfect Golomb ruler of order (mark) 4 and length 6 is shown in Figure 2. It is not possible to have another type of perfect ruler. So this ruler is Optimum Golomb ruler in which the distance between the two points is unique.The optimum Golomb ruler g(1, 7) of weight 7, length 26 and cardinality (number of user) 1 is shown in Figure 3. Figure 4 represents the construction of 4 code matrices M1...M4 from shifted version of Golomb ruler g (1,7) with filler zeros (shown) that increases the code dimension.

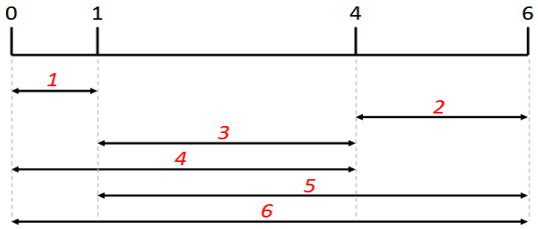


Fig. 2. Golomb ruler of order 4

The headings of table in Figure 4(a) define the column and row to which the table entries should be transposed. These matrices can be converted into 2D W/T codes by taking row as i^{th} wavelength and column as j^{th} time slot as shown in Figure 4(b). For example for matrix M4, the code will be $\{\boldsymbol{\lambda_4}:\boldsymbol{\lambda_2}:\_:\boldsymbol{\lambda_2}:\boldsymbol{\lambda_4}:\_:\boldsymbol{\lambda_1}:\boldsymbol{\lambda_1}\}$ which signifies fourth wavelength in first time slot, second wavelength in second time slot, second wavelength in fourth time slot, fourth wavelength in fifth time slot, first wavelength in seventh and eighth time slot.

1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1

Fig. 3. Optimum Golomb ruler g(1,7) of weight 7, length 26 and cardinality 1

	Column1				**Column2**				**Column3**				**Column4**				**Column5**				**Column6**				**Column7**				**Column8**			
M	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**	**R1**	**R2**	**R3**	**R4**
1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
2	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0
3	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0

Fig. 4. (a): Construction of code matrices

M1=

	C1	C2	C3	C4	C5	C6	C7	C8
r1	1	0	0	0	1	0	0	0
r2	0	0	0	0	0	1	1	0
r3	1	0	1	0	0	0	0	0
r4	1	0	0	0	0	0	0	0

M2 =

	C1	C2	C3	C4	C5	C6	C7	C8
r1	0	0	0	0	0	0	0	0
r2	1	0	0	0	1	0	0	0
r3	0	0	0	0	0	1	1	0
r4	1	0	1	0	0	0	0	0

M3=

	C1	C2	C3	C4	C5	C6	C7	C8
r1	0	1	1	0	1	0	0	0
r2	0	0	0	0	0	0	0	0
r3	1	0	0	0	0	1	0	0
r4	0	0	0	0	0	0	1	0

M4 =

	C1	C2	C3	C4	C5	C6	C7	C8
r1	0	0	0	0	0	0	1	1
r2	0	1	0	1	0	0	0	0
r3	0	0	0	0	0	0	0	0
r4	1	0	0	0	1	0	0	0

Fig. 4. (b): Construction of code matrices

Now the concept of folded optimum Golomb ruler has been expended by using more than one optimum Golomb ruler as shown in Figure 5 and design a matrix with eight wavelength (8 rows) and four time slots (4 columns) that can produce (8×4) = 32 Pseudo Orthogonal (PSO) codes as represented in Figure 6. It should be noted that the cardinality goes from 4 to 32.

1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0

$g_1(4,4)$

1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

$g_2(4,4)$

1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0

$g_3(4,4)$

1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0

$g_4(4,4)$

Fig. 5. Optimum Golomb rulers of weight 4, length 25 and cardinality 4

	Column 1								**Column 2**								**Column 3**								**Column 4**							
M	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**
1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0

Fig. 6. (a): Construction of code matrix for 8 users

$$M1=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&0\end{bmatrix}\quad M2=\begin{bmatrix}0&0&0&0\\1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&1&0\\0&1&0&0\end{bmatrix}\quad M3=\begin{bmatrix}0&0&1&0\\0&0&0&0\\1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&1&0\end{bmatrix}\quad M4=\begin{bmatrix}0&0&0&1\\0&0&1&0\\0&0&0&0\\1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}$$

$$M5=\begin{bmatrix}0&0&0&0\\0&0&0&1\\0&0&1&0\\0&0&0&0\\1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\quad M6=\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&1\\0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\\0&0&0&0\end{bmatrix}\quad M7=\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\\0&0&1&0\\0&0&0&0\\1&0&0&0\\0&1&0&0\end{bmatrix}\quad M8=\begin{bmatrix}0&0&1&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\\0&0&1&0\\0&0&0&0\\1&0&0&0\end{bmatrix}$$

Fig. 6. (b): Codes for 8 users

From Figure 6(b), the code for matrix M1 will be $\{\boldsymbol{\lambda}_1 ,(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_7): \boldsymbol{\lambda}_6\}$ and the code for matrix M2 will be $\{\boldsymbol{\lambda}_2 ,(\boldsymbol{\lambda}_3, \boldsymbol{\lambda}_8): \boldsymbol{\lambda}_7\}$. Similarly the code for matrices M3....M32 can be constructed.

The code dimension can be determine as: if r is the number of rows, c is the number of columns and L is the length of Golomb ruler then shifting of Golomb ruler is shown in figure 6(a). The dimension of matrix is r×c and there are r×c-L possible shifts. It shows that the number of shifts permitted is r×c. So, the following equation should hold true to assure that matrix code set size m is equal to the number of rows in the matrices [12].

r×c-L $\geq$ r-1

Here, the length of Golomb ruler is 25, r=8 (wavelengths) and c=4 (time slots). So the possible shifts are equal to 8×4-25=7.

B. Probability of error in W/T codes

One dimensional codes spread either in time or frequency. Several types of 1-D codes are OOC, ZCC, Walsh code and Hadamard code. These codes can be characterized by N (L_t, W, λ_a, λ_c). Where

- N is the number of code
- L_t is the temporal length of the code
- W is the weight of the code (number of ones in the code)
- λa is out of phase autocorrelation peak
- λc is Cross correlation peak

The autocorrelation of one dimensional code x(t) is defined as

$$Z_{x,x}(l) = \sum_{n=0}^{L_T-1} x_n\, x_{(n+l)modL_T} \qquad (1)$$

$Z_{x,x}(l)$ satisfies

$$Z_{x,x}(l)\begin{cases} = W & if\ l = 0 \\ \leq \lambda_a & if\ 1 \leq l \leq L_T - 1 \end{cases}$$

The cross correlation of one dimensional code x(t) and y(t) is defined as

$$Z_{x,y}(l) = \sum_{n=0}^{L_T-1} x_n\, y_{(n+l)modL_T} \qquad (2)$$

$Z_{x,y}(l)$ satisfies

$$Z_{x,y}(l) \leq \boldsymbol{\lambda_c} \quad if\ 0 \leq l \leq L_T - 1$$

The autocorrelation of 2-dimensional codes x(t) is defined as

$$Z_{x,x}(l) = \sum_{m=0}^{R-1} \left(\sum_{n=0}^{L_T-1} x_{m,n}\, x_{m,(n+l)modL_T}\right) \qquad (3)$$

$Z_{x,x}(l)$ satisfies

$$Z_{x,x}(l)\begin{cases} = W & if\ l = 0 \\ \leq \lambda_a & if\ 1 \leq l \leq L_T - 1 \end{cases}$$

The cross correlation of 2-dimensional codes x(t) and y(t) is defined as

$$Z_{x,y}(l) = \sum_{m=0}^{R-1} \left(\sum_{n=0}^{L_T-1} x_{m,n}\, y_{m,(n+l)modL_T}\right) \qquad (4)$$

$Z_{x,y}(l)$ satisfies

$$Z_{x,y}(l) \leq \lambda_c \quad if\ 0 \leq l \leq L_T - 1$$

The probability of error/bit P_e is given by

$$P_e = \frac{1}{2} \sum_{i=Th}^{N-1} \binom{N-1}{i} \left(\frac{W^2}{2L_T}\right)^i \left(1 - \frac{W^2}{2L_T}\right)^{N-1-i} \qquad (5)$$

III. System Simulation

The OCDMA system based on the optical codes is simulated by a commercial fiber optic simulation tool. The transmitter and receiver section of the OCDMA system based on W/T code is shown in Figure 7 and Figure 9 with their corresponding parameters in Table 1 and Table 2. In transmitter section, CW (Continuous Wave) laser is used as an optical source, Pseudo random Bit Sequence (PRBS) generator is used to generate random data and Mach-Zehnder Modulator to modulate the carrier signal generated by PRBS generator. Eight different wavelengths range from 1549.2 nm to 1554.8 nm with wavelength spacing 0.8 nm are multiplexed by WDM (Wavelength Division Multiplexer) from the laser array.

The modulated signal of each user is assigned a unique code by encoder which is shown in Figure 8. It consists of optical filters, time delays, splitter and combiner, the splitter splits the carrier signal and optical filters selects four specific wavelengths from the carrier signal to produce the encoded bit sequence. The time delay in the encoder places the selected pulses of specific wavelengths in appropriate time slot and combiner combines these four pulses to construct the encoded signal.

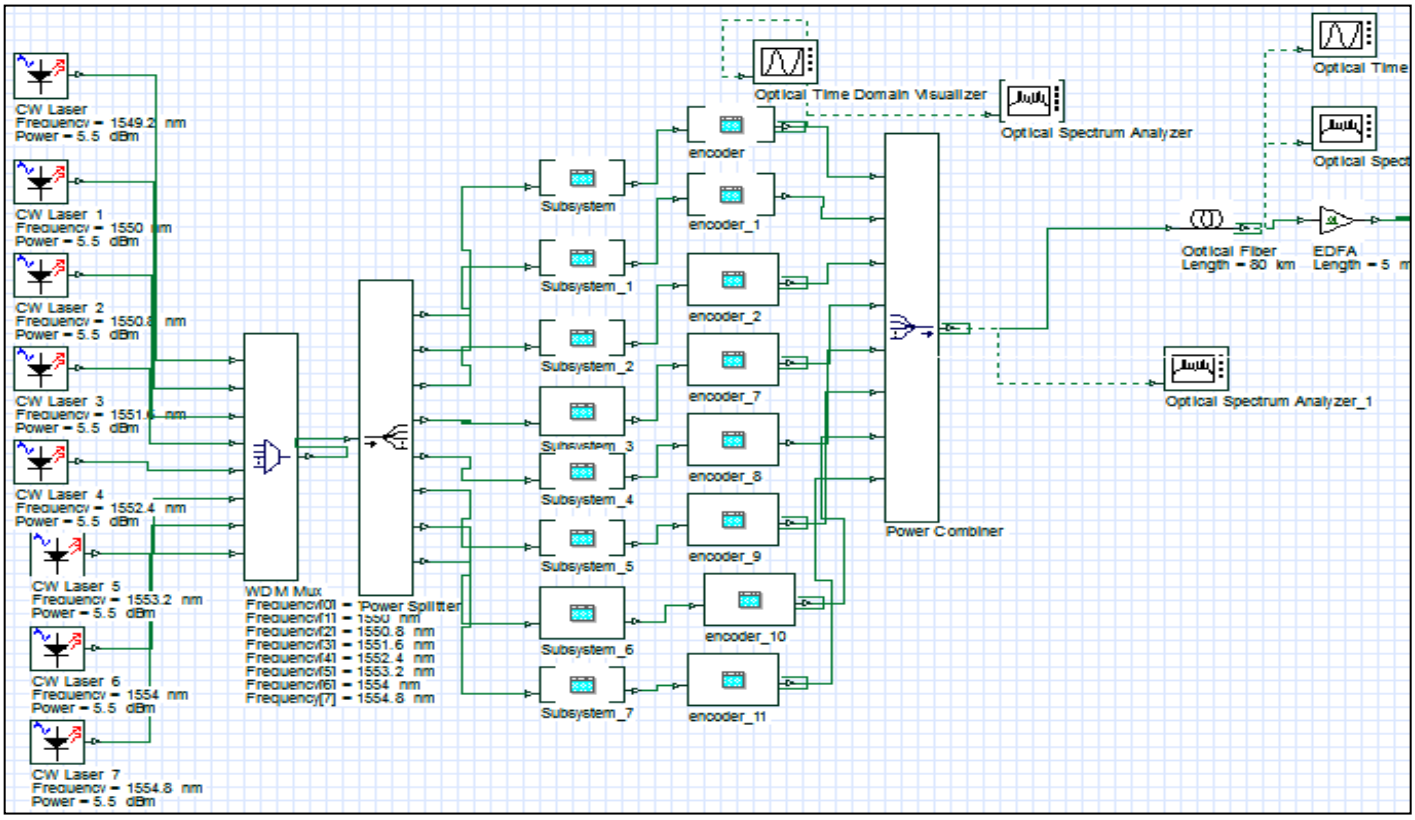


Fig. 7. OCDMA Transmitter block on Optisystem

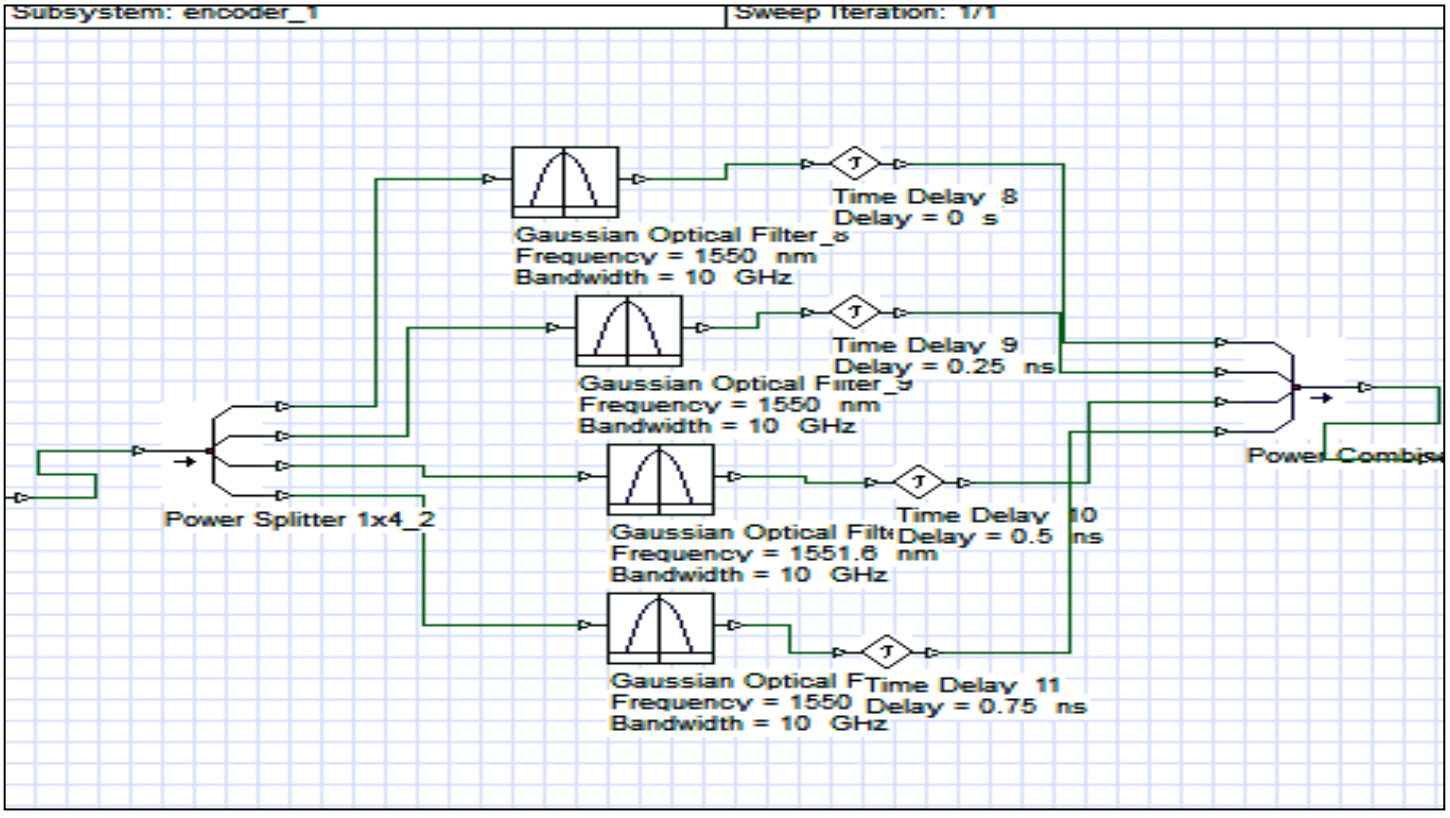


Fig. 8. Encoder Structure on Optisystem

TABLE I. TRANSMITTER DESIGN PARAMETER

Component	Parameter	Value
CW Laser	Wavelength	1549.2 nm
	Spacing	0.8 nm
	Power	5.5 dBm
	Line Width	10 MHz
	Initial Phase	0 degree
Pseudo Random Bit Sequence	Bit Rate	1 and 1.25 Gbps
Mach-Zehender Modulator	Excitation ratio	30 dB
Fiber	length	50 km
	Attenuation	0.2 dB/km
	Reference wavelength	1550 nm

The encoded data is combined and passed through single mode optical fiber (SMF) by considering length of 50 Km. Optical CDMA systems is designed which considers all practical impairments. The Table 3 represents time delay at 1 Gbps and 1.25 Gbps. If data rate is 1 Gbps that means the duration of 1 bit is 1 ns and since four time slots are taken, so the duration for each time slots will be 0.25 ns. For code M4 $\{\lambda_4 :\lambda_5 :\lambda_2:\lambda_1\}$, wavelength λ_4 is given in first time slot with zero delay, wavelength λ_5 is given in second time slot with 0.25 ns delay and wavelength λ_2 is given in third time slot with 0.5 ns delay and wavelength λ_1 is given in fourth time slot with 0.75 ns delay. Similarly delay for data rate 1.25 Gbps can also be calculated.

The optical signal is passed through the receiver section followed by decoder and photo detectors with low pass filter. The receiver extracts the information that is transmitted by transmitter. The decoder consists of optical filters and inverse time delays with respect to the transmitter that decodes a particular code as the corresponding encoder.

TABLE II. RECEIVER DESIGN PARAMETER

Chip Period	For 1Gbps bit rate (ns)	For 1.25 Gbps bit rate (ns)
T1	0	0
T2	0.25	0.2
T3	0.5	0.4
T4	0.75	0.6

TABLE III. TIME DELAY AT 1 GBPS AND 1.25 GBPS

Component	Parameter	Value
Photo-detector	Dark Current	10nA
	Center frequency	1552.5 nm
Low Pass Bessel Filter	Cutoff frequency	8 GHz
	Insertion loss	0 dB

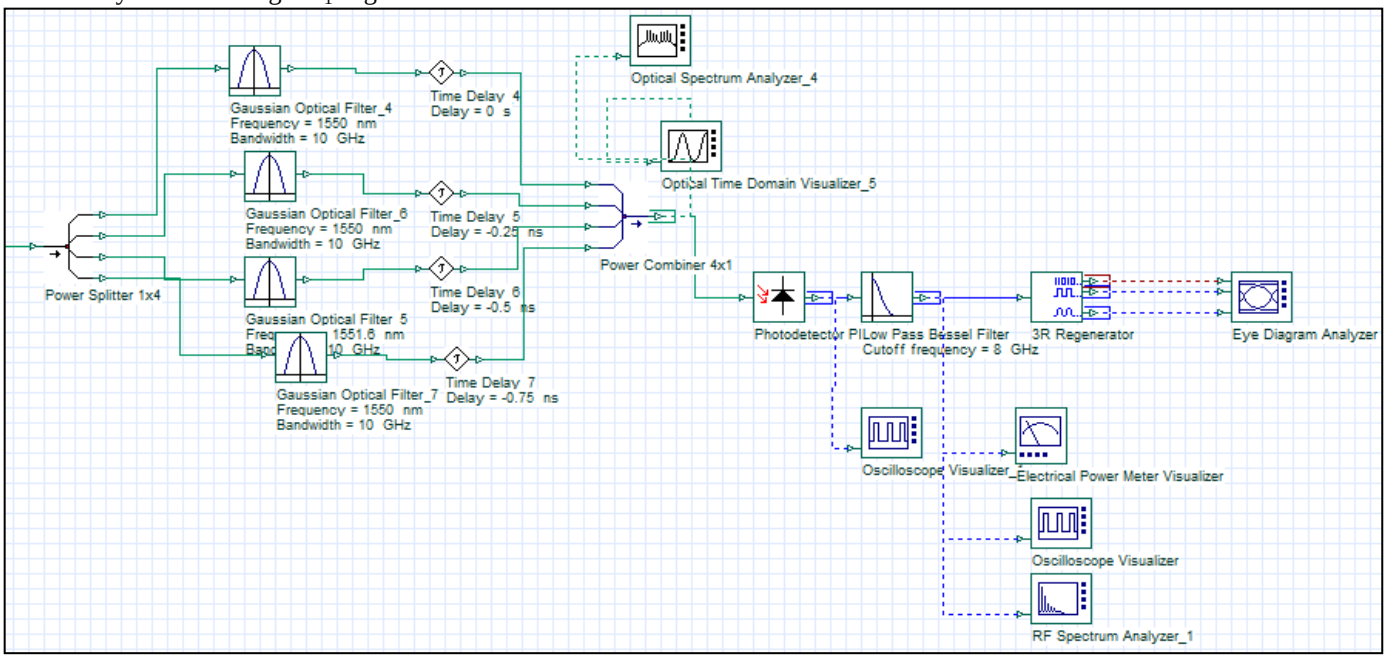


Fig. 9. OCDMA Receiver structure on Optisystem

V. Results and Discussions

The OCDMA system simulated on Optisystem uses CW laser centered at 1549.2 nm.

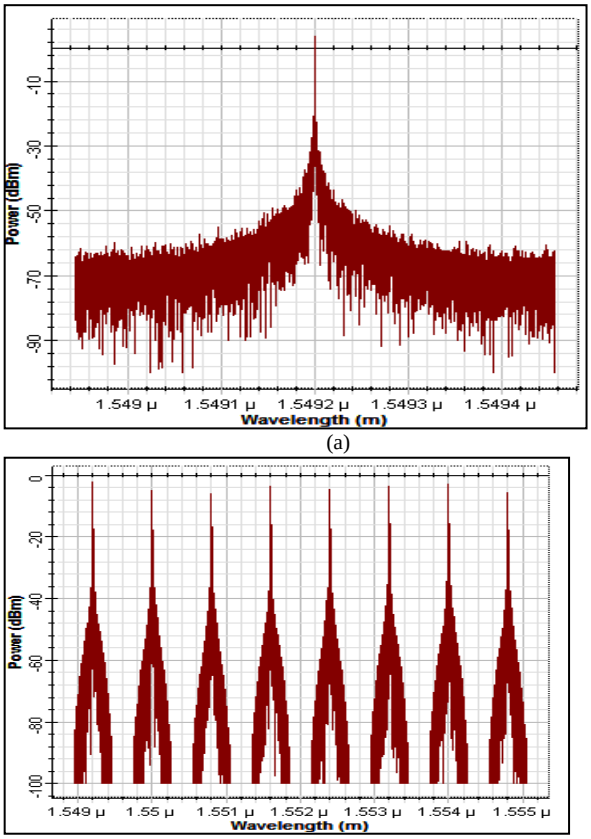
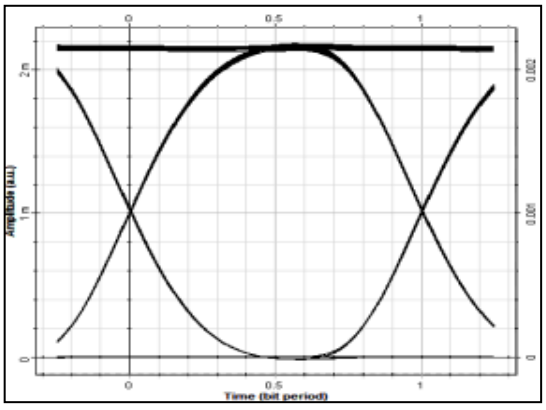


Fig. 10. CW Laser (a) Output of laser and (b) Multiplexed laser array spectrum

Figure 10 shows the CW laser output and eight multiplexed transmitted wavelengths. This multiplexed spectrum is modulated by user data with the help of MZM modulator. The output of MZM modulator is further encoded by different codes which are designed in Figure 6.



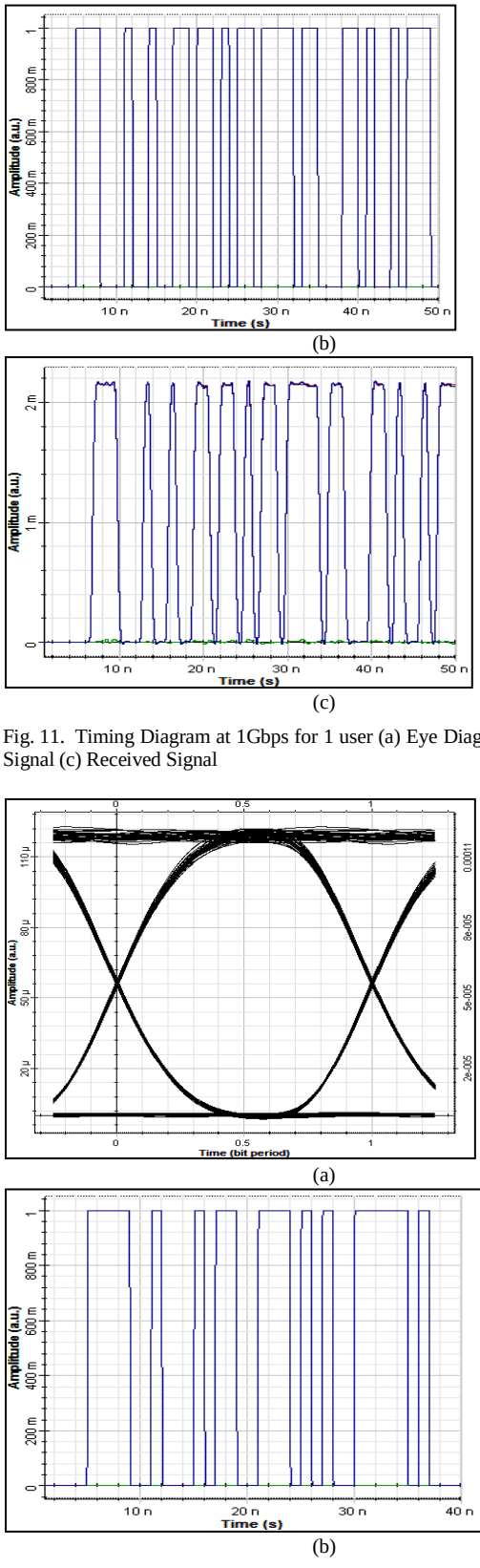
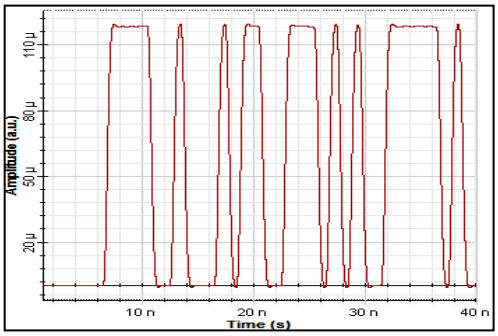
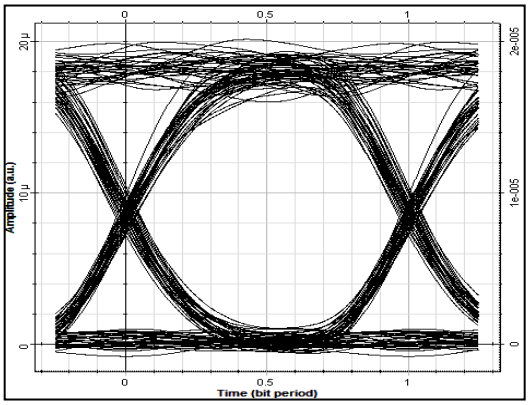


Fig. 11. Timing Diagram at 1Gbps for 1 user (a) Eye Diagram (b) Transmitted Signal (c) Received Signal

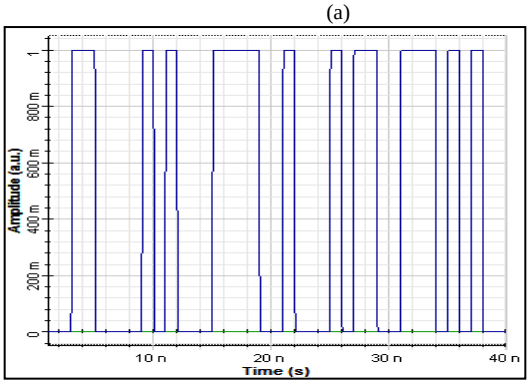


(c)

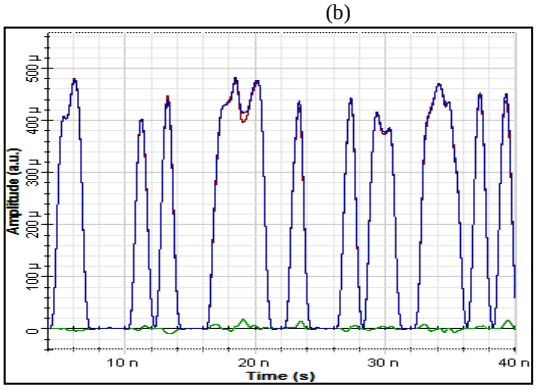
Fig. 12. Timing Diagram at 1Gbps for 3 users (a) Eye Diagram (b) Transmitted Signal (c) Received Signal



(a)

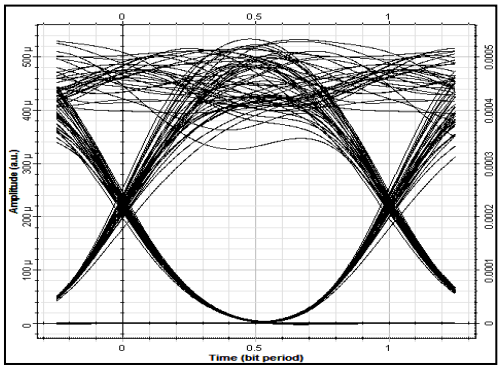


(b)

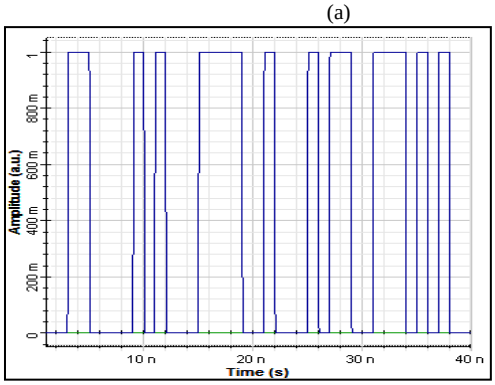


(c)

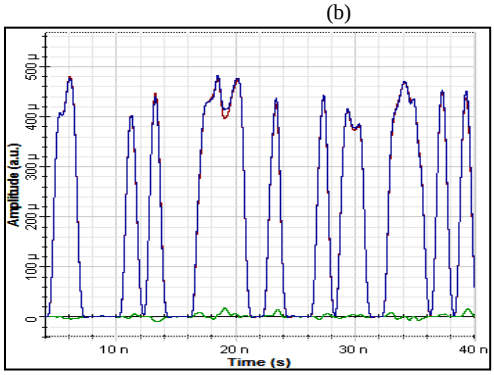
Fig. 13. Timing Diagram at 1Gbps for 8 users (a) Eye Diagram (b) Transmitted Signal (c) Received Signal



(a)



(b)



(c)

Fig. 14. Timing Diagram at 1Gbps for 12 users (a) Eye Diagram (b) Transmitted Signal (c) Received Signal

The demultiplexer demultiplexes the output which is given to the decoder section. The receiver receives the optical signal and converts it back to the electrical signal. The timing diagram at 1 Gbps for 1, 3, 8, and 12 users are shown in Figure 11 to Figure 14. The eye diagram allows visualizing the main parameters of electrical signal such as eye width, eye opening, quality factor, SNR etc. Figure 11- 14(a) represents the eye diagram. For an eye diagram measurement, noise on eye will cause the eye to close. Therefore the SNR is also directly indicated by amount of eye closer which represents that as the number of users increased the eye diagram starts closing means due to multiple access interference upper portion starts building noise. The Figure 11-14(b) represents the transmitted signal and Figure 11-14(c) represents the received signal. From Figure 11-14(a), it is revealed that as the number of users increase from 1 to 12, noise is added and the eye height start decreasing and from Figure 11-14(c), it is noted that there is some noise at the amplitude of the signal which result that the

amplitude decreases almost to 400µ (a.u.) which is initially at 2000µ (a.u.) and the received signal is also dispersed in time domain as it passes through the fiber. Similarly signal can be received at 1.5 Gbps but as the data rate has been increased from 1 Gbps to 1.25 Gbps, MAI further increases.

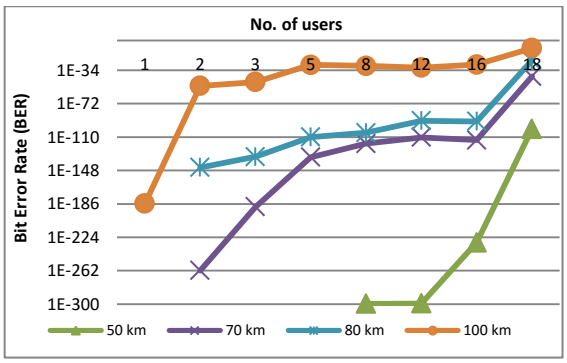


Fig. 15. BER vs. Number of users for different fiber length at 1 Gbps

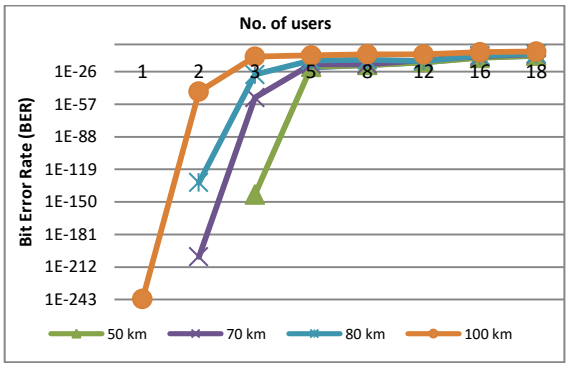


Fig. 16. BER vs. No. of users for different fiber length at 1.25 Gbps

The comparative analysis between BER and number of users for various lengths of fiber at 1 Gbps and 1.25 Gbps is shown in Figure 15 and Figure 16 respectively. It is observed that as the length of the fiber increases the bit error rate also increases. For three number of simultaneous users in 1Gbps system, the bit error rate increases from 0 to $1e^{-40}$ as the fiber length has been increased from 50 to 100 km whereas for 1.25Gbps system, the bit error rate increases from $1e^{-150}$ to $1e^{-15}$.

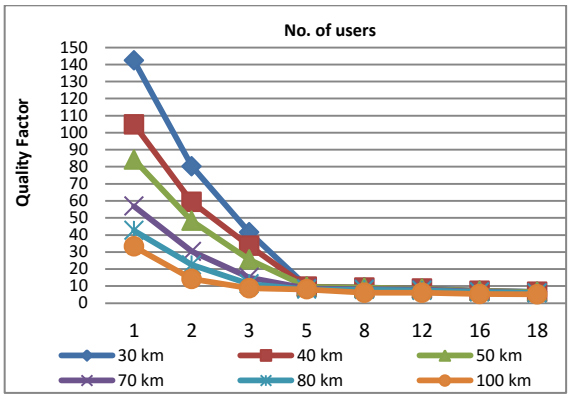


Fig. 17. Q factor vs. number of users for different fiber length at 1Gbps

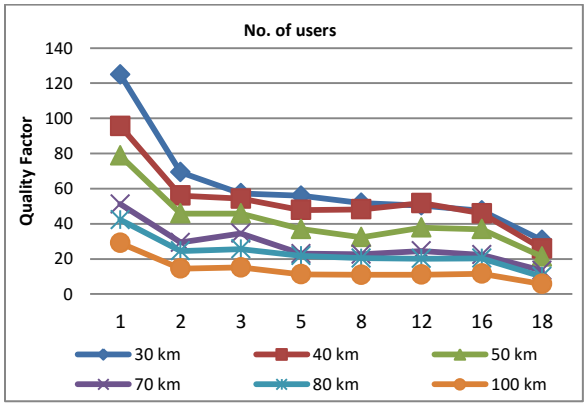


Fig. 18. Q factor vs. number of users for different fiber length at 1.25Gbps

The comparative analysis between quality factor and number of users for various lengths of fiber at 1 Gbps and 1.25 Gbps are presented in Figure 17 and Figure 18 respectively. It is observed that as the length of the fiber increases the quality factor decreases. For two simultaneous users in 1 Gbps system, the quality factor decreases from 80 to 10 as the fiber length increases from 30 to 100 Km whereas for 1.25 Gbps system, the quality factor decreases from 70 to 8. Similarly at data rate of 1.25 Gbps, the performance deteriorates. Hence, the analysis reveals that as the number of simultaneous users in the system increase, MAI become dominant with degrading in amplitude of received signal. For practical implementation of the optical CDMA based on the proposed codes, the optical fiber loop can be used instead of time delay line.

VI. Conclusion

One dimensional code have disadvantage of increasing code length with the increment of number of users. The simulated design of 2D W/T OCDMA code shows that cardinality of 2D code is high. The performance of OCDMA system has been evaluated with increasing number of users in the form of quality factor and BER .It has been observed that as the number of user's increases, the eye opening decreases which results in decrease in amplitude due to noise. At data rate of 1Gbps as the length of the fiber increases from 50 Km to 100 Km, the bit error rate shows the significant increment from 0 to e^{-40} with decrement of quality factor from 40 to 5 for 3 users.

References

[1] J. A. Salehi, "Code division multiple access techniques in optical fiber networks part I: Fundamental principles," IEEE Transactions Communication, vol. 37, no. 8, pp. 824–833, 1989.

[2] R. K. Fan Chung and J. A. Salehi, "Optical Orthogonal Codes: Design, Analysis and Applications," IEEE Transactions on Information Theory, May, vol. 35, no. 3, pp. 595-604, 1989.

[3] A. J. Mendez, R. M. Gagliardi, H. X. C. Feng, J. P. Heritage, and J. M. Morookian, "Strategies for realizing optical CDMA for dense, high speed, long span, optical network applications', Journal of Lightwave Technology, vol. 18, no. 12, pp. 1685–1696, 2000.

[4] A. J. Mendez, S. Kuroda, R. M. Gagliardi, and E. Garmire, "Generalized temporal code division multiple access (CDMA) for optical communication," SPIE Proceeding, vol. 1, no. 125, pp. 287-291, 1989.

[5] V. J. Hernandez, A. J. Mendez, V. Bennett, R. M.Gagliardi, and W. J. Lennon, "Design and performance analysis of wavelength/time (W/T) matrix codes for optical CDMA," Journal of Lightwave Technology, vol. 21, no. 11, pp. 2524–2533, 2003.

[6] H. Heo, S. Min, Y. H. Won, Y. Yeon, B. K. Kim and B. W. Kim, "A new family of 2-D wavelength-time spreading codes for optical code-division multiple-access system with balanced detection", IEEE Photon Technology Letter vol. 8, pp. 2189–2191, 2004.

[7] Y., Yeon, B. K. Kim, S. C. Cho, S. J. Park and B. W. Kim, "Two dimensional wavelength/time optical cdma system adopting balanced modified pseudo random noise matrix codes", U.S. Patent application, 0100338 A1, 2005.

[8] S. P. Wan, and Y. Hu, "Two-dimensional optical CDMA differential system with prime/OOC codes", IEEE Photon Technology Letter vol. 13, pp. 1373–1375, 2001.

[9] V. J. Hernandez, A. J. Mendez, C. V. Bennett, R. M. Gagliardi and W. J. Lennon, "Bit-Error-Rate Analysis of a 16-User Gigabit Ethernet Optical-CDMA (OCDMA) technology Demonstrator Using Wavelength/ Time Codes," IEEE Photonics Technology Letters, vol. 17, no. 12, pp. 2784-2786, 2005.

[10] J. Faucher, R. Adams, L. R. Chen and D. V. Plant, "Multiuser OCDMA system demonstrator with full CDR using a novel OCDMA receiver," IEEE Photon Technology Letter, vol. 17, no. 5, pp. 1115-1117, 2005.

[11] P. Patel, V. Baby, L. Xu, D. Rand, I. Glesk, and P. R. Prucna1, "A scalable wavelength hopping and time spreading optical CDMA system," IEEE LEOS-03 Proceeding, pp. 1048–1049, 2003.

[12] R. Poboril, J. Latal, P. Koudelka, J. Vitasek, P. Siska, J. Skapa, and V. Vasinek, "A Concept of a Hybrid WDM/TDM Topology using the Fabry-Perot Laser in the Optiwave Simulation Environment," Optics and Optoelectronics, vol.9, no.4, pp. 167-178, 2011.

Crowding Optimization Method to Improve Fractal Image Compressions Based Iterated Function Systems

Shaimaa S. Al-Bundi
Department of Mathematics-College of Education for pure Sciences- Ibn Al-Haitham-Baghdad University
Bagdad, Iraq

Nadia M. G. Al-Saidi
Applied Sciences Department-University of Technology
Baghdad, Iraq

Neseif J. Al-Jawari
Department of Mathematics-College of Sciences-Al-Mustansiriah University
Baghdad, Iraq

***Abstract*—Fractals are geometric patterns generated by Iterated Function System theory. A popular technique known as fractal image compression is based on this theory, which assumes that redundancy in an image can be exploited by block-wise self-similarity and that the original image can be approximated by a finite iteration of fractal codes. This technique offers high compression ratio among other image compression techniques. However, it presents several drawbacks, such as the inverse proportionality between image quality and computational cost. Numerous approaches have been proposed to find a compromise between quality and cost. As an efficient optimization approach, genetic algorithm is used for this purpose. In this paper, a crowding method, an improved genetic algorithm, is used to optimize the search space in the target image by good approximation to the global optimum in a single run. The experimental results for the proposed method show good efficiency by decreasing the encoding time while retaining a high quality image compared with the classical method of fractal image compression.**

Keywords*—Fractal; Iterated Function System (IFS); Genetic algorithm (GA); Crowding method; Fractal Image Compression (FIC)

I. INTRODUCTION

Fractal image compression (FIC) is produced from Barnsley's research IFS system [1] and the fractal image block coding suggested by Jacquin [2]. In 1988, Barnsley [3] used FIC based on the theoretical IFS system to represent computer graphics and compress the aerial image. Using this approach, Barnsley obtained a compression ratio of 1000:1, but the approach requires manual interference. Thereafter, Jacquin suggested a new FIC method that depends on image block and can behave automatically without manual interference. This method has become a perfect representation for this research direction, in which FIC theory is realized. At present, FIC has obtained extensive interest from the research community, because of its novel concept, high compression ratio, independent resolution, and fast image decoding. This technique is based on the fractal inverse problem and aims to find an IFS, in which the attractor is close to a query image.

The emerging technique for image compression that based on fractal theory is fully different form traditional image compression techniques. It is focused on two main problems: the first one is how to find the IFS mappings and the second is finding of an efficient algorithm to find those mappings, such that, they can approximate the original image. Toward solving these problems, Jacquin [2] proposed an efficient technique by partitioning of a given image M into non-overlapping range blocks and an overlapping domain blocks, the IFS parameters is achieved by finding the best corresponding domain block for each range block. Therefore, as a result of this encoding process, we obtain a different transformation for each range block. If we composed all the transformations of all range blocks and iterated starting with the initial image, the attractor (fractal) that approximate the original image is produced, it is also called the fixed point of the transformations. This type of representation is called partitioned IFS [4] or local IFS [5].

Many researchers have emphasized on overlapping of an efficient and reliable image compression technique based on fractal. It is firstly presented by Barnsley and Sloan [6] in 1988, when they introduced of finding an *IFS*, whose attractor approximate the given image and the *IFS* is sent instead of sending the image itself over the channel. In 1992, A. Jacquin [7] Barnsley's student improved *IFS* theory and introduced the concept of local *IFS* through presenting the concept of fractal image coding. In 1994, Y. Fisher [4] made many improvements on Barnsley's algorithm. He combined his idea in a very famous book in this field. Since Jacquine's publication of the original fractal coding scheme, several papers try to popularize his work both in practical and theoretical [8], among others, however none of these attempts in general have been proven to be efficient. Therefore, many efforts are highlighted towards employing of evaluative algorithms. Numerus optimization models have been proposed to represent a normal evolution mechanism [9]. Genetic algorithms [10,11] is one of these models. In these algorithms, the population represents as an IFS models and it is responsible of making adjustments toward the optimum through a random process that used for selection of genetic operators called crossover and mutation.

GAs that are used to address an optimization problem are required to solve multimodal and multidimensional problems, through which a large search space with different optima can be obtained. These problems do not have deterministic algorithms to obtain the global optimum; if they do exist, however, the algorithm is an inclusive search along the

solution space that, in turn, leads to exponential time and machine resources using algorithms of this kinds in solving the problem described above. Therefore, the algorithms used to solve various complex problems can show their respective capacities. The GAs work with population of individuals that are iteratively adjusted towards the optimum by means of a random operation of selection restructure and mutation [11].

Meanwhile, crowding is a technique that is applied in GAs to maintain variety in the population and prohibit early convergence to local optima. This technique involves the combination of both the offspring and the identical individual from the present population in this process, which is called coupling phase; determining which of the two will remain in the population is a process called alternation phase [12]. The current work depends on the alternation phase of crowding, which is applied by using one of the following three approaches: deterministic [12, 13], probabilistic [14, 15], and simulated annealing [16]. In our work, we used an improved crowding method to achieve the aim with a shorter time and good quality. We achieved our goal by selecting the chromosome for a maximum of three times to prevent repetitive selection and provide an opportunity to check another chromosome that may obtain a better result.

The rest of the paper is presented as follows; section 2 presents the theoretical background of fractal, fractal inverse problem, and *PIFS*. The detailed explanations on the fractal image compression, collage theorem, and Jacquin approach for fractal image coding are discussed in sections 3. The GA and its relationship with the fractal image compression in introduced in section 4. Crowding method and its improved version is introduced in section 5. The implementation and the analysis of the results is discussed in section 6. Finally, the work is concluded in section 7.

II. Basic Concepts of Fractal Image Coding

The theory of self-affine transformation and self-similarity is the bases that fractal image coding depends on. In this section, we introduce the theoretical basis for fractal image compression, such as the *IFS*, contraction mapping, and fixed point theorem.

A. Self-similarity Property

One of the base properties of fractal image is self-similarity. A typical image is said to be self-similar if the image looks "almost" the same on any scale. However, all images do not contain this kind of self-similarity found in fractals and actually contains different sort of similar parts (Distasi et al. [17], Truongx et al. [18]). Figure 1 shows an example of this fractal image.

Self-similar parts in the Lena image are shown in Figure 2, as can be seen in part of her shoulder and the reflection in the mirror with her hat [19]. In this type of image, only a portion of an image is self-similar, whereas, in Figure 1, the whole image is self-similar.

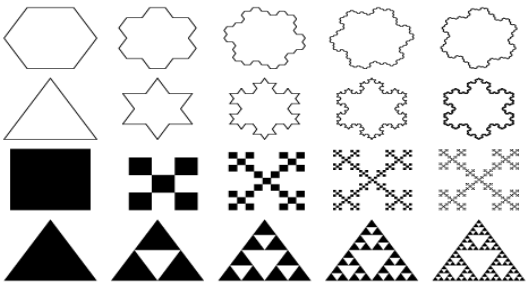


Fig. 1. Fractal image repeated at different locations

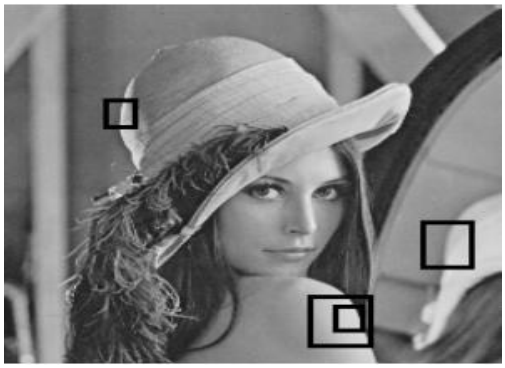


Fig. 2. Self-similarity in the Lena image

Now, let *(X, d)* be a metric space and a sequence (X_n) is called a Cauchy sequence, if for any given $\varepsilon > 0$, we have $d(X_m, X_n) < \epsilon$ for all $m, n \in N$ (natural numbers). *(X, d)* is called complete if every Cauchy sequence in X converges to an element of X.

Readers that are interested in greater detail can refer to [3,20].

Definition 1: Let $f: R^2 \to R^2$ be a transformation of the form $w(x_1, x_2) = (ax_1 + bx_2 + e, cx_1 + dx_2 + f)$, where a, b, c, d, e, and f are real numbers. This transformation is called a (two-dimensional) affine transformation. The following equivalent notations have been used:

$$w(x) = w_i \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} = Ax + l \qquad (1)$$

where $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a two-dimensional, 2 × 2 real matrix, and l is the column vector $\begin{pmatrix} e \\ f \end{pmatrix}$, such as $(e, f) \in R^2$.

The matrix A can always be written as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} r_1 \cos\theta_1 & -r_2 \sin\theta_2 \\ r_1 \sin\theta_1 & r_2 \cos\theta_2 \end{pmatrix}.$$

Definition 2: let (X, d) be a metric space, a transformation $f: X \to X$ is called contractive mapping if $((x), f(y)) \leq s.d(x, y)$ for all $x, y \in X$, where $0 \leq s < 1$ is called contractivity factor of f.

Definition 3: Let $f: X \to X$ be a transformation on a metric space *(X, d)*, a point $x_f \in X$, such that $f(x_f) = x_f$ is called the fixed point of the transformation f. The fixed point is highly important; it represents that the part of the shape in which we are interested *that* is not affected by the transformation.

The Hausdorff metric is an important concept in fractal theory. Therefore, many mathematicians have discussed and proven basic concepts and results of this space [21,22]. The Hausdorff metric is known as the space of fractals and is denoted by $H(X)$. It is generated from the complete metric space X comprising elements that are the compact sets in X. The distance that is defined in $H(X)$ is given as follows.

Definition 4: Let *(X, d)* be a complete metric space, for the space of fractal $H(X)$, The Hausdorff distance h is defined on this space as follows:-

$$h(A,B) = d(A,B) \vee d(B,A),$$

for any points A and $B \in H(X)$

where $d(A,B) = max\{d(x,B): x \in A\}$
and $d(x,B) = min\{d(a,B): b \in B\}$.

The *IFS* is the most important concept of fractal theory. The *IFS* was developed by Hutchison (1981), and then by Barnsley and other researchers [1, 6]. These systems of mapping have been widely discussed and used in many applications, such as image compression. The general formula for *IFS* is introduced as follows.

Definition 5: Let (*X*, *d*) be a complete metric space. An *IFS* on *l* is a finite set of contractive self-mappings $w_i: X \to X$ with respective contractivity factors s_i for $i = 1, 2, \dots, N$, such that $s = \max\{s_i, i = 1, 2, \dots, N\}$. *IFS* is based on the affine transformations given by

$$w(x) = w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r_1 \cos\theta_1 & -r_2 \sin\theta_2 \\ r_1 \sin\theta_1 & r_2 \cos\theta_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

Definition 6: In *IFS*, any compact subset (fixed point) $A \in H(X)$ is called an attractor for *IFS* if

$$A = \bigcup_{n=1}^{n} w_n(A).$$

The fixed point observes existence and uniqueness based on the contraction mapping theorem. The iteration process of the *IFS* based on any starting image the attractor, which is fully known by the parameters of W.

III. Fractal Image Compression

The *FIC* approach is an important search area with many possible application fields. This approach is focused on finding fractal code that generates given objects. Barnsley [3] introduced this concept with the well-known collage theorem. When the object is considered as an image, *FIC* is often involved, which is also known as fractal image coding. The foundation for *FIC* is the *IFS*. This problem has been studied by many authors, and a method has been proposed by Jacquine [7] to solve this type of inverse problem. The major problem of standard fractal image coding is its time consumption compared with other image coding methods. Some time is spent in searching for a similar domain block. Therefore, new techniques to solve this problem and accelerate this method are in great demand.

The problem of finding *IFS's* that used to generate fractal is called an inverse problem. However, if the given set is self-similar, then the required construction is almost straightforward. The IFS can easily be found by conducting mathematical translation of the property of self-similarity. This solution is verified in the collage theorem, which is the first step towards solving the inverse problem.

A. The Collage Theorem

This theorem states the process of obtaining the set of transformations that represent an accurate approximation of a fixed image. It is stated as follows.

Let (X, d) be a complete metric space. let $\{X; w_n, n = 1, 2, \dots, N\}$ be an *IFS* with contractivity factor s, $0 \leq s < 1$ and let *L* be a closed subset of *X* such that

$$h(L, \bigcup_{n=1}^{N} w_n(L)) < \epsilon,$$

for some $\epsilon > 0$, and *h* is the Hausdorff distance. Then

$$h(L, A) \leq \frac{\epsilon}{1-s},$$

where *A* is the attractor of the *IFS*s

B. Jacquin Approach for Fractal Image Compression

FIC depends on the self-similarity property in an image. The main idea comes from the partitioned iterated function system (*PIFS*), which is an expansion of *IFS* theory. The difference between the two concepts appears in the application domain. Thus, the main difference is that instead of dealing with the whole image, a specific part is used to obtain the *PIFS* parameters.

For an original image *M* of size *m*×*m*, it is partitioned into $(m/n)^2$ blocks which are non-overlapping to form a set of range blocks each of them is of size *m*×*m*. To comply with the contractive point theorem, the domain block is twice the size of the range block. Hence, a set of $(m\text{-}2n+1)^2$ elements, each of them is a block of size 2*n*×2*n* is constructed from *M* and known as domain blocks. In this case, the partitioned is overlapping. In each search for similarity between the range and domain blocks, two types of blocks emerge from the same image, as shown in Figure 3. As an example, for *M* of size 128×128 if the size of the each range block is 8×8, then we have $(128/8)^2 = 256$ range blocks and $(128\text{-}2\times8+1)^2 = 12{,}769$ domain blocks of size 16×16.

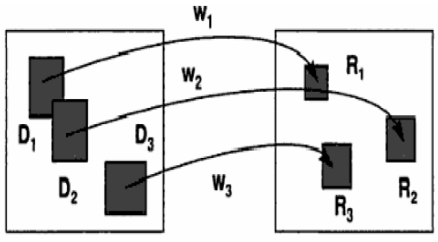


Fig. 3. Clarification of *PIFS*

The image is equally partitioned by the range blocks, which resulted that each pixel of the image is included in one of the range blocks. However, since the domain block is overlapping, this may cause losing of some pixels. The aim of

this process is to find an approximate domain block for each range block.

By the PIFS technique, the third dimension is appeared that represent the pixel z. after shrinking the domain block to the size of the range block, the eight transformation is applied to resulted in eight different blocks z_k, k =0, 1,…, 7. These transformationsT_k, k =0, 1,…, 7 can be represented in (2).

$$T_0 = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad T_1 = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}, \quad T_2 = \begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix},$$
$$T_3 = \begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}, \quad T_4 = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}, \quad T_5 = \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix},$$
$$T_6 = \begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}, T_7 = \begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix} \quad \ldots. \quad (2)$$

T_1 and T_2 correspond to the flips of z along the horizontal and vertical lines, respectively. The flip of z along the horizontal and the vertical lines is denoted by T_3, whereas, an additional flip along the line of the main diagonal is performed by the transformations T_4, T_5, T_6, and T_7 which are correspond to T_0, T_1, T_2, and T_3. Finally $T_0(z) = z$.

In fractal coding, a contrast scaling s and a brightness offset o on the transformed blocks occur, so the fractal affine transformation becomes three-dimensional as shown in (3).

$$W_i \begin{bmatrix}x\\ y\\ z\end{bmatrix} = \begin{bmatrix}a_i & b_i & 0\\ c_i & d_i & 0\\ 0 & 0 & s_i\end{bmatrix} \begin{bmatrix}x\\ y\\ z\end{bmatrix} + \begin{bmatrix}e_i\\ f_i\\ o_i\end{bmatrix} \quad \ldots. \quad (3)$$

We let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be two squares containing n pixel from D_i and R_i, respectively. Here, we minimize the quantity of s and o between of the coded range block X_D and its corresponding coordinates of the domain block Y_D, such as;

$$s = \frac{n^2 \sum_{i=1}^{n^2} a_i\, b_i - \left(\sum_{i=1}^{n^2} a_i\right)\left(\sum_{i=1}^{n^2} b_i\right)}{n^2 \sum_{i=1}^{n^2} a_i^2 - \left(\sum_{i=1}^{n^2} a_i\right)^2}$$

$$o = \frac{1}{n^2}\left[\sum_{i=1}^{n^2} b_i - s \sum_{i=1}^{n^2} a_i\right]$$

where the minimum distance between R_i and D_i is found by the RMS such that;

$$RMS = \frac{1}{n}\left[\sum_{i=1}^{n^2} b_i^2 + s\left(s.\sum_{i=1}^{n^2} a_i^2 - 2\sum_{i=1}^{n^2} a_i b_i + 2.o\sum_{i=1}^{n^2} a_i\right) + o.\left(o.n^2 - 2\sum_{i=1}^{n^2} b_i\right)\right] \quad \ldots\ldots\ldots \quad (4)$$

When this error is less than the predefined threshold, the search will be finished.

The execution time of fractal compression implementation is the main problem that most standard algorithms. Fisher's algorithm [4], involves a classification pattern that has been greatly accelerated; however, the resulting image quality is extremely poor because of the search space reduction from the classification used by Fisher. To overcome these problems, some evolutionary algorithms is used to serve in solving these problems, its methodology is introduced in the following section.

IV. Genetic Algorithm

Using a GA is important for obtaining solutions to complicated search problem. In this section, we discuss the relationship between the processing of GA and its operation when dealing with the fractal inverse problem. Holland's 1975 book [10] introduced GAs as a summary of biological evolution and showed the theoretical structure of GAs.

A. Genetic Algorithm for Fractal Image Compression

Searching processes begin after dividing the image into range and domain blocks as in standard Jacquine approach image compression. For each range block, the domain block and identical transformations that best cover the range block is specific. In general, for best correspondence, transformation codes are set, including contrast and brightness. Searching succeeds when the domain block is appropriate for the suitable range block. Eventually, mapping data are stored. Then, we use GAs to attain the intended outcome via fractal compression. Usually, GAs are employed to find the near optimal solution, and thus the GA for fractal compression of images is shown below [10].

1) Chromosomes

Given that, the GA works on the chromosomes, producing chromosomes from the range and domain blocks is a crucial step in using the GA for *FIC*. The transformation parameters obtained for each block are coded on a set of a fixed number of bits. These parameters are then stored as chromosomes. By encoding the parameters of an image, a chromosome comprises *N* genes that are equal to the number of the non-coded parts of an image. These genes are generated from parameters X_D and Y_D, which refer to the coordinates of the domain block, and the flip, refers to the transformation isometrics. Figure 4 show the chromosomes

Range Block									
Block1			Block2			…….	Block N		
X_D^1	Y_D^1	$Flip^1$	X_D^2	Y_D^2	$Flip^2$	…….	X_D^N	Y_D^N	$Flip^N$

Fig. 4. Image representing a chromosome

2) Fitness Function

Fitness function is a specific task for each chromosome, which refers to the capability of each chromosome to survive and proliferate. We denote fitness as the value of error between the coded range and domain blocks that are assigned by the transformation with analogous luminance and contrasting values. The error is computed using the root mean square equation (4).

3) Genetic Operators

Crossover and mutation are two basic operators that are used in all implementations of genetic algorithms. These operators are described as follows.

- *Crossover Operator*: The crossover operator selects two parents based on their fitness, and then attempts to produce a new child with the best possible quality. A high fitness value provides the crossover operator with a high probability of selection. The crossover operator changes the genes of the parent. Given that a random number a is produced in the interval [0, 1], the new coordinates are computed using the equation below.

First offspring $X_D = a * X_{D1} + (1-a) * X_{D2}$

$$Y_D = a * Y_{D1} + (1-a) * Y_{D2} \quad \ldots.(5)$$

Second offspring $X_D = (1-a) * X_{D1} + (1-a) * X_{D2}$

$$Y_D = (1-a) * Y_{D1} + (1-a) * Y_{D2}$$

- *Mutation Operator*: The mutation operator changes the value of one or more genes in the chromosome, thereby adding completely new gene values to the gene pool. The GA may achieve a better solution using these new values. The mutation operator also introduces the verity in the chromosomes. The information changes randomly based on the mutation rate.

B. Genetic Algorithm for Fractal Image Compression

Fractal image compression algorithm

1. Decompose the input image *M* into blocks according to Jacquine's technique
2. Begin with *FIC* parameters, such as range block size, fitness function, error limit, and number of iterations;
3. Begin with GA parameters, such as mutation rate and crossover rate;
4. Set *t* = some tolerance level;
5. Partition image *M* into non-overlapping ranges R_i's and overlapping domain D_i;s;

For each range block R_i in the range, do

- The transformations (a random population of chromosomes) is generated

while number of populations is not the maximum and the optimal domain is not found, Do

- The fitness value is computed for all individuals to be used for search for the optimal domain in the domain pool using the fitness function;
  - when the optimal domain block is found;
  - apply the crossover operator on individuals;
  - apply the mutation operator on individuals; and
  - generate the new population;

end *while*

The obtained transformation parameters from the search is written in the transformation *W*

end *for*

V. Crowding Method

De Jong [23] introduced crowding as a general technique for maintaining population variety and early convergence. Crowding is often used to determine survival of genetic algorithms in order to determine the individuals in the present population and identify the offspring that will pass to the next generation. It is divided into two principal phases, namely, coupling and alteration. In the coupling phase, the offspring individuals are coupled with individuals in the present population based on a likeness metric. Meanwhile, in the alteration phase, the pairs of offspring and individuals that will remain in the population are selected. The main crowding scheme of De Jong [23] involves the random selection of offspring individuals from the present population. The identical selected of individual is used to replace the selected offspring. Which makes crowding is an improved genetic algorithm is that;

1) In the crowding method, parent selection is not commonly used, therefore, the individuals are randomly paired in the present population. However, in the population, each individual becomes a parent.

2) In the crossover operator, for each pair (P_1, P_2), the parents are recombined with probability P_c. In the mutation operator, the two producing children (c_1, c_2) are mutated with probability P_m, where P_c denotes crossover probability, P_m denotes mutation probability, and M denotes population size [13].

3) The population of the next offspring includes one of the two parents that complete with each child.

4) The distance between two individuals i_1, i_2 is denoted by $d(i_1, i_2)$.

If $d(p_1,c_1) + d(p_2,c_2) < d(p_1,c_2) + d(p_2,c_1)$

$p_1 \leftarrow$ win the emulation between p_1 and c_1.

$p_2 \leftarrow$ win the emulation between p_2 and c_2.

Else

$p_1 \leftarrow$ win the emulation between p_1 and c_2.

$p_2 \leftarrow$ win the emulation between p_2 and c_1.

For survival, each offspring oriented to fight with its most identical parent. Other variants exist when more than two parents and children are selected before applying the resemblance metric [26]. This idea is the basis of several widely applied modern crowding approaches. The difference between these approaches is used to determine the winner in each competition.

B. The Proposed Crowding Algorithm for FIC

The crowding method [23] is proposed to eliminate the selection process and introduce a preselecting process. This will cause in a very fast GA to be used for multidimensional optimization problem. By reducing the selection process, the individuals are mutate randomly with any other population individuals. During the replacement process the paining between the offspring and one of the parents is performed first. This operation is done with probability P_c. This pairing process is happened according to the similarity between them. In the evaluation step, the fitness function which represented the least square error between the offspring and the parents is responsible for deciding about which individual of the population is allowed to stay.

In this section, we proposed an improved crowding method in order to be applied to improve *FIC*. With this method the diversity is preserved in the population with the opportunity for each individual to be a parent. What distinguish our proposed method from original crowding method [23] and Mahfoud method [13] is some technical differences in the main phases of the algorithm. The population set $\{T_1,T_2,\ldots,T_n\}$ is constructed by finding all contruction mapping T_i that resulted from the similarity measure between the range block and domain block of the query image. This set is calculated using Jacquine approach [7]. Each individual T_i is assigned a fitness value $f(T_i)$ as its weight, where $f_i \in \{f_1,f_2,\ldots,f_n\}$ represents the minimum distance

between R_i and D_j, j=1,…,*m*. This value is used in the selection process of the parents $\{P_1,P_2\}$ and controlled by a chosen factor known as crowding factor that determine the number of the maximum selection of this individual as shown in the following diagram.

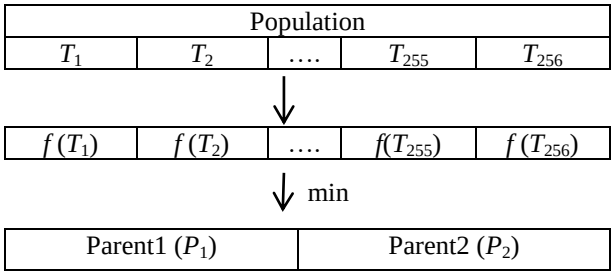


Fig. 5. Selection process

Crowding algorithm

Input: Image *I,M* % *M* is known as a crowding factor
p_c % crossover probability
p_m % mutation probability
Output: New optimized population $cp = \{T'_1, T'_2, \dots, T'_n\}$
Bagin
Genetic crowding population $cp = \{T_1, T_2, \dots, T_n\}$
Evaluate (*cp*) % calculate the fitness for *cp*
while terminated () do
P = selection (*cp, M*) % select parents *P* = {*T: T*$\in$ *TP*, such that, *RMS (T(i), i)* < ε}
Offspring = recombination (*cp*, p_c, *P*)
Offspring = mutation (offspring, p_m)
Compare offspring with parents and add the best one to the *cp*
end while
end.

In the selection phase, the parents are chosen according to the minimum fitness values, where the maximum number of selection of each individual is equal *M* (crowding factor).

Selection function

Input: c_p % crowding population
M % maximum number the selection
Output: *P* % parents
Begin
for *i* = 1 to *n* % *n* is the size of c_p
$p_i = \frac{f_i}{\sum_{j=1}^{n} f_j}$ % *f* is the fitness
end
while *z* = false do
Select the best two individual *P* from c_p according to p_i and call it *P*
If $p_c f <= M$ then
Z = true
end if
end while
end

In the recombination phase, the offspring is generated according to two logical values as shown in the following algorithm.

Recombination function

Input: c_p % crowding population
p_c % crossover probability
p_m % mutation probability
output: offspring
begin
select $\{p_1[x], p_2[x]\}, \{p_1[y], p_2[y]\}, \{p_1[f], p_2[f]\}$
case 0 0
$p_2[x] = p_1[x], p_2[y] = p_1[y], p_2[f] = p_1[f]$
case 0 1
$p_1[x] = p_2[x], p_1[y] = p_2[y], p_1[f] = p_2[f]$
case 1 0
$p_1[x], p_2[x], p_1[y] = p_2[y], p_1[f] = p_2[f]$
case 1 1
$p_2[x] = p_1[x], p_2[y] = p_1[y], p_2[f] = p_1[f]$
end select
end

After the recombination phase, the resulting offspring are competed with their parents the mutation phase for surviving. The decision of winning is taken based fitness value (the similarity measure between the offspring and the parents) in order to decide the one that should in the new population, such that: If *RMS*(*P*,*C*)<ε then *C* is the winner of the competition

else *P* is the winner of the competition

Mutation function

Input: P_m % mutation probability
Offspring
Output: offspring
begin
x, y, f = rand () % generate random number
if x > y and x > f then
x = rand()
else if y > x and y > f then
y = rand()
else if f > x and f > y then
f = rand()
end if
end

The termination value of the algorithm is deduced according to learning process on a sample of different images to determine the best that can satisfy the compromising between the optimum solution and the execution time.

VI. Implementation and Analysis

A. Implementation

The proposed system was established using Matlab Ver.8.2 and then tested on an pc with cor i7, 2.5 GHz and 8 GB RAM, windows 10 pro. The proposed system was tested on five 8-bit gray images of size 512×512. We tested the proposed system on three ranges, namely, 2, 4, and 8. The RMS of the decoded image partitioned by the 8×8, 4×4 and 2×2 block sizes. A smaller block size indicated a smaller

appropriate error for the affine transformation. The partitions of range blocks were calculated according to $(m/n)^2$, while the partitions of domain blocks were calculated according to $(m–2n+1)^2$. Table 3 illustrates the coding, decoding, time, and compression ratio of the selection images using the proposed technique, while Table 4 illustrates the peak signal-to-noise ratio (PSNR) and MSE of some selection images using the proposed technique.

B. Analysis

The results of the abovementioned algorithms in terms of compression ratio, quality, and implementation time (coding time) are compared for the genetic and crowding *FIC* algorithms with the standard *FIC* algorithm as shown in Tables 1-6. The comparison was performed on an image with a range pool containing 16384 range blocks of size 4×4 and a domain pool containing 255025 domain blocks of size 8×8. Table 5 presents the comparison results.

In the chosen images, the *PSNR* is inversely proportional to *MSE*, and the compression ratio is proportional to that value. A small range block size resulted in a higher compression ratio. The time for producing the image depends on how much error is allowed in the transformations. The employ of the image determines the required amount of compression and the image quality. The predefined number *M* is used as an indicator that determine the number of times for selecting the individual as a parent. This modification in the selection in new generation. This diversity in the population that achieved by the crowding method resulted in some advantages, which are:-

1) Through the search, different local maxima can be achieved.

2) The diversity is maintained.

3) For different crowding factor, the subpopulation is almost stable.

The replacement process is responsible about picking the new individual to construct new population.

By applying the proposed crowding method the following results is obtained. Tables 1-6 represent the analysis of the results for some chosen images.

TABLE III. CODING, DECODING TIME, AND COMPRESSION RATIO BASED ON THE PROPOSED METHOD FOR DIFFERENT RANGE SIZE

Images	Range	Coding Time	Decoding Time	Compression Ratio
	2×2	5.66	2.10	3.12
	4×4	0.99	0.26	7.04
	8×8	0.18	0.18	12.38
	2×2	5.98	2.84	2.33
	4×4	0.895	0.278	4.33
	8×8	0.26	0.19	7.41
	2×2	6.09	2.97	3.56
	4×4	0.878	0.265	5.72
	8×8	0.29	0.19	7.4
	2×2	5.90	2.02	2.29
	4×4	0.895	0.269	6.03
	8×8	0.17	0.17	10.1
	2×2	6.158	2.83	3.21
	4×4	0.920	0.261	7.16
	8×8	0.22	0.18	14.06

TABLE IV. PSNR AND MSE FOR DIFFERENT RANGE BLOCK

Images	Range	PSNR	MSE
	2×2	12.11	0.07
	4×4	13.18	0.026
	8×8	13.38	0.04
	2×2	7.13	0.38
	4×4	9.86	0.39
	8×8	9.86	0.51
	2×2	7.03	0.113
	4×4	12.48	0.113
	8×8	10.64	0.113
	2×2	9.06	0.039
	4×4	9.18	0.023
	8×8	10.08	0.049
	2×2	10.01	0.08
	4×4	12	0.039
	8×8	12.89	0.03

VII. CONCLUSIONS

Image compression technique is always in a continuous competition and challenge according to the fast developing of the technology fractal image compression is an emerging technology that based on the fast that most of real world images contain some redundant area that are similar to the other area in the same image. It is basic idea is how to express an image by a set of *IFSs*. The argumentative discussion about compromising between the compression ratio and the contracted image quality is motivation for new optimized technique towards this goal. Genetic algorithm is ….. to be appropriate used to solve of a multidimensional problem that have large search space with no exact solution exist. In this study, we improve this technique by omitting of the parent selection which resulted, each individual becomes a parent. However, the selection process is specified by a pre-defined value known as crowding factor that determine the number of selection of each individual. Therefore, each offspring is randomly selected from the population, and its most identical parent. Comparing the performance of the proposed technique is accomplished through some experiments which show best result over the standard fractal compression technique and standard genetic algorithm technique as shown in tables (6) and charts (1-3). From these figures are can see that RMS error is inversely proportional to the PRNS ratio. They show a good compromise value that resulted in good performances.

REFERENCES

[1] M.F. Barnsley and S. Demko. Iterated function systems and the global construction of fractals. In Proceedings of the Royal Society of London A399, 243 - 275, 1985.

[2] A. E. Jacquin, "A fractal Theory of Iterated Markov Operators with Applications to Digital Image Coding", PhD., Georgia Institute of Technology, 1989.

[3] M. F. Barnesly, "Fractal everywhere", Second edition, Academic Press, 20-100, 1988.

[4] Y. Fisher,: Fractal Image Compression: Theory and Application to digital images, Springer Verlag, New York, 1995.

[5] M.F. Barnsley and L. P. Hurd, fractal Image Compression. AK. Peters, Wellesley, Mass, 1993.

[6] M. F Barnsley, A.D Sloan, A better way to compress images, Byte Mag. 215–223, 1988

[7] A.E. Jacquin, Image coding based on a fractal theory of iterated contractive image transformations. Image Proc., IEEE Trans. 1, 18-30**,** 1992.

[8] T. Abiko, , M. Kawamata,: IFS coding of non-homogeneous fractal images using Gr¨obner basis. Proc. of the IEEE International Conference on Image Processing 25–29, 1999.

[9] Z. Michalewicz,: Genetic Algorithms + Data Structures = Evolution Programs, second edition, Springer-Verlag, 1994.

[10] Holland, J.H.: Adaptation in natural and artificial systems. The University o Michigan Press, 1975.

[11] D. E. Goldberg, "Genetic Algorithm in search, optimization and Machine Learning", Addison – Wesley, reading , MA, 1989.

[12] S. W. Mahfoud. Niching Methods for Genetic Algorithms. PhD thesis, Department of General Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 1995.

[13] S. W. Mahfoud. Crowding and preselection revisited. In R. MÄanner and B. Manderick, editors, Proceedings of the 2nd International Conference on Parallel Problem Solving from Nature (PPSN II), Elsevier, Amsterdam, The Netherlands, pages 27-36, Brussels, Belgium, 1992.

[14] O. J. Mengshoel. "Efficient Bayesian Network Inference: Genetic Algorithms", Stochastic Local Search, and Abstraction. PhD thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1999.

[15] O. J. Mengshoel and D. E. Goldberg. Probabilistic crowding: Deterministic crowding with probabilistic replacement. In W. Banzhaf, J. M. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. J. Jakiela, and R. E. Smith, editors, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), Morgan Kaufmann, San Francisco, CA., pp. 409-416, 1999.

[16] S. W. Mahfoud and D. E. Goldberg,"Parallel recombinative simulated annealing", A genetic algorithm, Parallel Computing, 21, 1-28, 1995.

[17] R. Distasi, m. Nappi, and d. Riccio, “R range/domain approximation error-based approach for fractal image compression,”IEEE trans. Image process., 15, 1, 89–97, 2006.

[18] T. K. Truong, j. H. Jeng, i. S. Reed, p. C. Lee, and a. Q. Li, “A Fast encoding algorithm for fractal image compression using the dct inner product,” IEEE trans. Image process., 9, 4, 529–535, 2000.

[19] Yancong, Y. and Ruidong, P. “Fast Fractal Coding Based on Dividing of Image”, 2010.

[20] J.E. Hutchinson, Fractals and self-similarity, Indiana University journal of mathematics, 30(5), 713-747, 1981.

[21] K. J. Falconar, "The Hausdorff dimension of self-affine fractals", Math. Proc. Comb Phil. Soc. 103, 339-350, 1988.

[22] K. J. Falconar, "Random fractals", Math. Proc. Comb Phil. Soc.100, pp:559-582, 1986.

[23] K. A. De Jong. An Analysis of the Behavior of a Class of Genetic Adaptive Systems. PhD thesis, Department of Computer and Communication Sciences, University of Michigan, Ann Arbor, MI, 1975.

[24] C. Chen, T. Liu, J. Chou," A Novel Crowding Genetic Algorithm and Its Applications to Manufacturing Robots"**,** IEEE Transection on Industrial Informatics, 10, 3, 2014.

TABLE I. STANDARD FRACTAL IMAGE COMPRESSION BY JACQUINE APPROACH [7]

Images					
Coding Time	2.03	2.03	2.99	2.98	2.97
Decoding Time	0.20	0.20	0.20	0.20	0.20
PSNR	11.15	7.13	11.74	9.01	11.18
MSE	0.81	0.89	0.94	0.91	0.83
Compression ratio	11.6	9.21	10.82	9.41	12.1

TABLE II. FRACTAL IMAGE COMPRESSION BASED ON GENETIC ALGORITHM

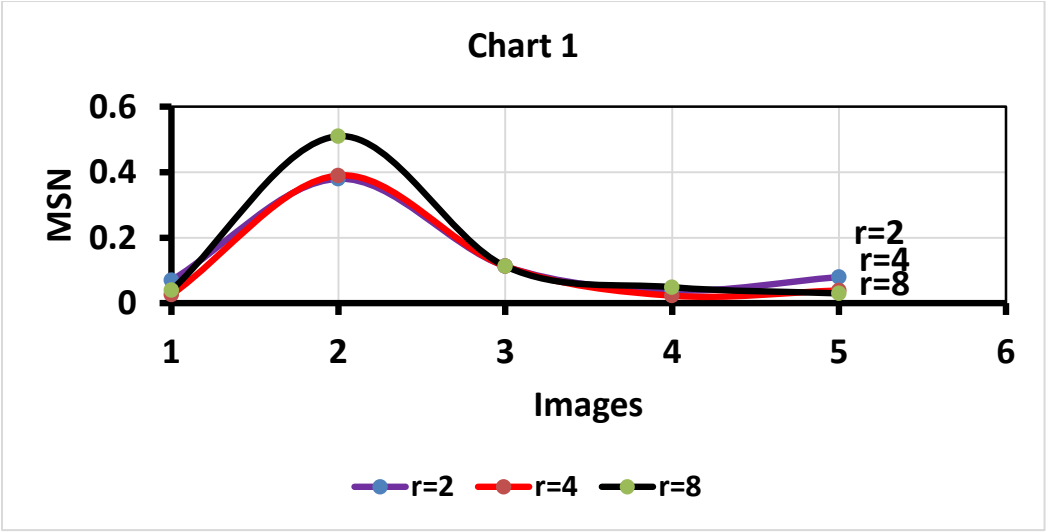
Images					
Coding Time	3.08	1.94	2.01	2.09	1.92
Decoding Time	0.61	0.27	0.93	0.21	0.57
PSNR	12	8.76	11.97	9.93	12.01
MSE	0.129	0.138	0.109	0.262	0.396
Compression ratio	12.6	7.33	8.88	11.53	14.44

TABLE III. CODING, DECODING, TIME, AND COMPRESSION RATIO OF THE SELECTION IMAGES USING THE SUGGESTED TECHNIQUE

Images					
Coding Time	0.99	0.89	0.87	0.89	0.92
Decoding Time	0.26	0.27	0.26	0.26	0.26
PSNR	13.18	9.86	12.48	9.14	12
MSE	0.026	0.39	0.113	0.03	0.039
Compression ratio	7.04	4.33	5.72	6.03	7.16

TABLE V. COMPARISON BETWEEN STANDARD, GENETIC AND PROPOSED CROWDING FOR RANGE BLOCK OF SIZE 4

	Images					
Fractal image compression based on Jacquine method	Coding Time	2.03	2.03	2.99	2.98	2.97
	MSE	0.81	0.89	0.94	0.91	0.83
	Compression Ratio	11.6	9.21	10.82	9.41	12.1
Fractal image compression based on genetic algorithm	Coding Time	3.08	1.94	2.01	2.09	1.92
	MSE	0.129	0.138	0.109	0.262	0.396
	Compression Ratio	12.6	7.33	8.88	11.53	14.44
Fractal image compression based on crowding method	Coding Time	0.99	0.89	0.87	0.89	0.92
	MSE	0.026	0.39	0.113	0.03	0.039
	Compression Ratio	7.04	4.33	5.72	6.03	7.16



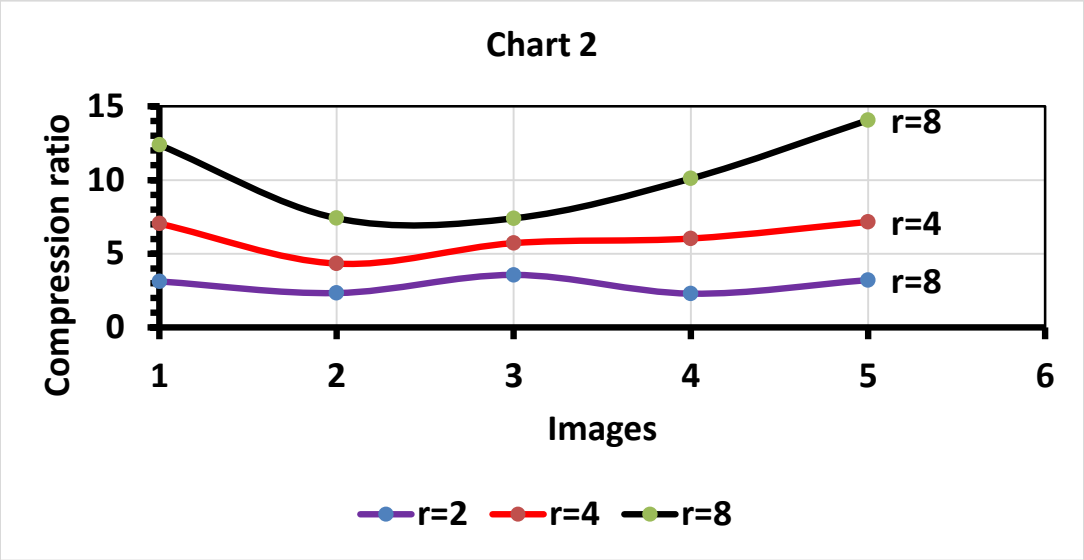


Chart 2
Compression ratio
15
10
5
0
1
2
3
4
5
6
Images
r=8
r=4
r=8
r=2
r=4
r=8

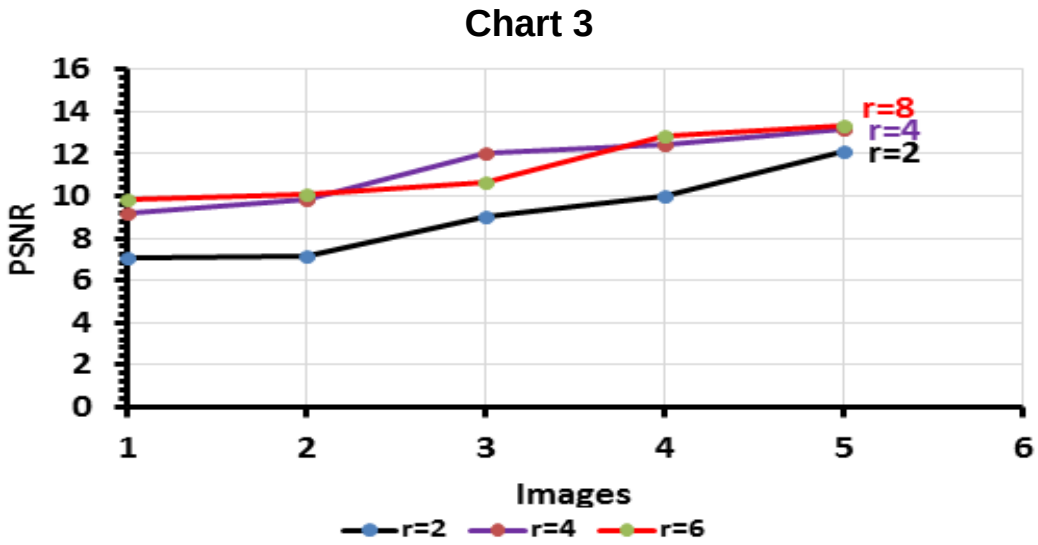


Chart 3
PSNR
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
Images
r=8
r=4
r=2
r=2
r=4
r=6

Current Trends and Research Challenges in Spectrum-Sensing for Cognitive Radios

Roopali Garg*,
UIET, Panjab University,
Chandigarh, India

Dr. Nitin Saluja
CURIN, Chitkara University,
Rajpura, Punjab, India

***Abstract*—The ever increasing demand of wireless communication systems has led to search of suitable spectrum bands for transmission of data. The research in the past has revealed that radio spectrum is under-utilized in most of the scenarios. This prompted the scientist to seek a solution to utilize the spectrum efficiently. Cognitive Radios provided an answer to the problem by sensing the idle (licensed) bands and allowing (secondary) users to transmit in these idle spaces. Spectrum sensing forms the main block of cognition cycle.**

This paper reviews the current trends in research in the domain of spectrum sensing. The author describes the type of channel being modelled, diversity combining schemes used, optimal algorithms applied at fusion centre, spectrum sensing techniques employed. Further, the research challenges are discussed. It is presented that various attributes like sensing time, throughput, rate reliability, optimum cooperative users, sensing frequency etc. needs to be addressed. A trade-off needs to be established to optimize two opposing parameters like sensing and throughput.

Keywords*—CR; cognitive radio; FC-PSO; fast-convergence particle swarm optimization; FC; fusion centre; KLMS; kernel least mean square; PU; primary use); ROC; receiver operating characteristic curves; SU; secondary user; soft combination; spectrum hole

I. INTRODUCTION

The scarcity of available spectrum and the inefficient usage of the same motivated the researchers to look for solutions in Dynamic Spectrum Access (DSA). DSA networking was introduced by Defense Advanced Research Projects Agency (DARPA) which caused a paradigm shift from traditional fixed spectrum access to dynamic spectrum access. [1] [2] [3].

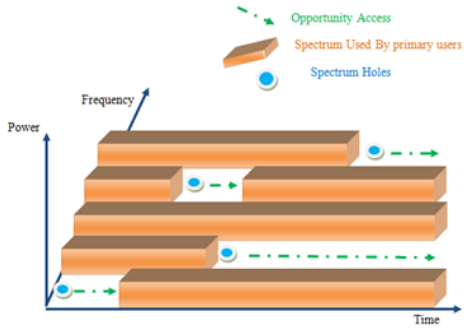


Fig. 1. Spectrum Hole

Cognitive Radio (CR) is the enabling technology for dynamic spectrum access. The idea of cognitive radio was first conceived by Joseph Mitola III. The latin word 'Cognoscere' forms the root of term 'Cognition'. It means 'to know' [4]. CRs have the capability to sense the radio environment. They can choose the portion of spectrum that is unused. These portions are referred to as Spectrum holes or white spaces (as shown in figure 1). Thereafter, the data can be transmitted on this chosen band without interference with the licensed user.

Primary Users (PU) are users having access rights to licensed spectrum band. On the other hand, Secondary Users (SU), have cognitive radio capabilities. They have the ability to sense the surroundings for availability of unused band. They request the PU to make use of this unused spectrum for wireless communication. PU have higher priority over SU. SUs ensures that they do not cause interference to PUs.

Cognition cycle involves the function of sensing the spectrum, making a decision about the hole and the licensed user, sharing the spectrum and mobility of secondary user in case a licensed user is detected. Accordingly, the four phases of cognition cycle can be described as: *Spectrum Sensing* is one of an important building block of cognitive radio. It involves the task of sensing the radio environment for the presence of spectrum holes and detection of PUs. *Spectrum decision* decides for the optimum selection of spectrum hole to transmit the data. As there are several CR users sharing the same spectrum, there is a need for a mechanism which coordinates the network access to all specified users. This can be defined under *Spectrum Sharing*. Under *Spectrum Mobility* if any primary licensed user is detected, then the CR should seamlessly switch over to some other suitable spectrum hole for further transmission [5]. Spectrum sensing encounters the issues like fading, shadowing and noise uncertainty. The scheme of cooperation has been suggested by researchers as an answer to these problems [6]. Here, CR users cooperate to share their sensing information for making a combined decision which is usually more accurate than individual decision. It reduces the probability of false alarm and mis-detection. Moreover, it solves the hidden primary user problem and reduces the sensing time [7].

The raw or processed data from each user is sent to a data fusion centre. It processes this collected data and finally makes a decision. The implementation of Cooperative sensing can be classified as *Centralized Sensing* [8], *Distributed Sensing* [9] , *External Sensing* and *Relay Assisted Sensing* [6] depending on presence of fusion centre or use of multi-hop for sensing.

The rest of the paper is divided into six sections. The hypothesis governing the absence or presence of licensed user is presented in Section II. Section III highlights the techniques employed for carrying spectrum sensing. Section IV presents the latest trends in the domain of study. The research challenges are described in Section V. Finally the paper is concluded in section VI and is appended with references.

II. Problem Formulation

The frame structure of a CR (as depicted in figure 2) consists of sensing time T_s followed by data transmission time T_t [10] i.e $T_s + T_t = T_f$; Where T_f is the frame period and the sensing frequency is $1/T_s$.

Spectrum sensing can be formulated with two hypotheses *:*

H_0: Channel is vacant temporarily i.e PU is absent

H_1: Channel is occupied i.e PU is present

Thus the spectrum sensing problem is to decide between

Null Hypothesis *H_0:*

$$Y[n] = U[n] \quad :PU\ is\ absent \tag{1}$$

Alternate Hypothesis *H_1:*

$$Y[n] = h.\ X[n] + U[n] \quad :PU\ is\ present \tag{2}$$

Here n = 1,2,3…..N, where N is the number of samples and h is the channel gain. It is considered equal to 0 under null hypothesis H_0 and 1 under alternate hypothesis H_1.

Y[n] is sensed signal by SU. X[n] represents the primary signal. It is assumed to be i.i.d (independent and identically distributed) random process with mean zero and variance $E[|X(n)|^2] = {\sigma_x}^2$. U[n] represents noise added by the wireless channel. It is assumed to be Gaussian and i.i.d random process with mean zero and variance $E[|U(n)|^2] = {\sigma_u}^2$

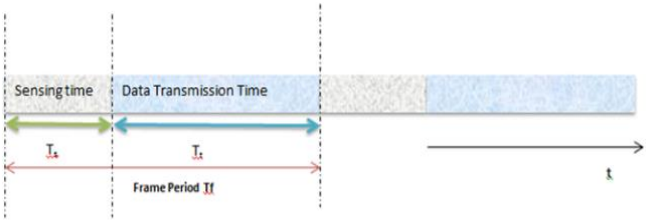


Fig. 2. Frame Structure of *Cognitive Radio*

A threshold is defined as per the specification of grade of services. If test value of Y[n] is greater than threshold then, alternate hypothesis H_1 is assumed to be accepted else null hypothesis H_0 is accepted.

Sensing accuracy is determined by Receiver Operating Characteristic (ROC) curves. These curves are the plot of probability of false alarm with probability of accurate detection or plot of probability of miss detection with probability of false alarm. These probabilities can be formulated as under:

Probability of detection: It is the probability that CR user will declare the presence of PU truly when it is present. A miss in detection of PU will lead to interference with the PU.

$$P_d = P_r\{Decision\ H_1|H_1\} \tag{3}$$

Probability of false alarm: is probability that CR user will declare the presence of Primary user when it is actually not present.

$$P_f = P_r\{Decision\ H_1|H_0\} \tag{4}$$

Probability of miss detection: Probability of missing the signal when it was actually present.

$$P_m = P_r\{Decision\ H_0|H_1\} \tag{5}$$

Probability of accurate detection: It is sum of probability of detection if PU is present and probability of no detection as PU is absent.

$$P_{ad} = P_r\{Decision\ H_1|H_1\} + P_r\{Decision\ H_0|H_0\} \tag{6}$$

III. Spectrum Sensing Techniques

The main objective of sensing methods is to detect the spectrum holes so that the SU can use these vacant bands. The methods for spectrum sensing are classified in figure 3 and the description of each is given in the following sub-section.

A. Prior Information Needing

1) Matched Filtering (MF)

In this methodology, the concept of matched filter detection technique is applied. Here an unknown signal x (t) is convolved with filter impulse signal h (t).Prior knowledge of bandwidth requirement, operating frequency, frame format, pulse shaping and modulation types etc is needed [11]. The advantage offered by Matched filtering technique is that it offers high probability of detection in less sensing time. It is considered as the best method in this category. Even with less signal samples detection is good. It is robust to noise uncertainty. Moreover, it offers good detection even at low SNR. On the other hand, the implementation is quite complex and involves large power consumption. In addition, there is a need for precise information about certain waveform patterns of PU [12].

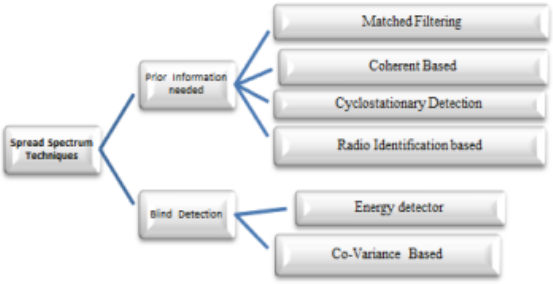


Fig. 3. Classification of Spectrum Sensing Techniques

2) Cyclostationary Detection (CSD)

This method utilizes the cyclostationary features like operating frequency, required bandwidth, frame format and modulation types from received signal of PU statistics like mean, cyclic correlation and autocorrelation. Periodicity in the signal causes cyclostationary features [13].

The robustness to uncertainty in noise power and propagation channel makes it attractive. However, there is need for high sampling rate. As large number of samples are

needed, it adds to the complexity. Also, the sensing time is high [14].

3) Coherent detection

In coherent detection, preambles, mid-ambles and pilot patterns are regularly transmitted. A preamble is a sequence (known already) which is sent before each slot. The sequence which is transmitted in the middle of the slot is known as mid-amble. When the information of these known patterns is available, received signal is correlated with its own known copy. Thus it assists in signal detection [10].

When there is long known primary signal pattern, the accuracy gets better. The stumbling block of this method is that huge amount of PU information is needed for the signal patterns to witness high performance. But it is not practically possible to achieve such large amount of information. Coherent detection technique is more reliable and has less convergence time than energy detector technique [15].

4) Radio Identification Based Detection

The main concept used in this technique is to identify the presence of some known technology and to achieve communication using this technology. If the technology used by primary user for transmission is known, then good knowledge can be derived about the spectrum characteristics. In addition it imparts higher accuracy. Eg: if the primary user is using bluetooth, then the CR can use this information to gain some knowledge in space dimension. Here, it uses the knowledge that bluetooth operates in the range of 10m. Then CR device may use bluetooth to communicate for some applications in the range supported by bluetooth. The detection process is highly accurate, has average sensing time. Also, it is robust to SNR. Never-the-less, high power consumption due high complexity makes it unattractive [16].

B. Blind Detection

1) Energy Detector

Energy detector is the most commonly used method commercially. The decision is made by comparing the output of signal detector with a fixed threshold value which helps in deciding the presence or absence of PU [16]. The attractive features of this methodology are that it is simple and easy to implement. It requires less sensing time and has low power consumption. Nonetheless, the noise uncertainty leads to increase in probability of false alarm. It is very unreliable in low SNR regime. Also, it is unable to distinguish between PU from other signal source. This technique has low accuracy as compared to other techniques [17] [18]. An energy detector is ineffective in detecting spread spectrum signals [19].

2) Co-variance Based Sensing

The spectrum sensing using co-variance technique works on the concept of comparison between covariance of the detected signal with the covariance of noise [20].The probability to differentiate between signal and noise is too high even at low values of SNR. The power consumed is low though increased complexity, increased computational overhead are the weakness of this methodology. Moreover, it has low detection performance and proves inefficient in case of spread spectrum signal [21].

IV. Current Trends

The current trends in spectrum sensing jointly deals with several attributes, addresses a trade-off between this parameters and suggests algorithms to converge to an optimal solution.

Rashid in [22] proposed a fast-convergence particle swarm optimization (FC-PSO) scheme. It is fine sensing scheme to find a trade-off between sensing time and throughput. The parameters considered in the study are detection performance, optimization time, and SU gain. The paper utilized energy detection scheme for in-band spectrum sensing. Detection performance, optimization time, and SU gain are some of the key parameters considered in this paper. A trade-off problem was formulated between the stopping threshold, sensing performance and the optimization time. Traditionally, the stopping criterion was based on measurement of error or limiting the maximum number of iterations. However, FC-PCO algorithm imposed a limit rule to stop the evaluation. When a certain number of particles reach a global optimum, the algorithm stopped. This is based on the fact that the remaining particles will not produce any new information and will be a waste of computational resources.

Additionally, the proposed algorithm was compared with Exhaustive Search Algorithm, Golden Section Search (GSS) algorithm, Artificial Immune System (AIS) algorithm, PSO algorithm. It was concluded that proposed FC-PSO algorithm has fastest convergence time, better computational complexity and was energy efficient.

However, as the frame duration is kept constant in this research, the future direction steer towards the study of frame optimization. Additionally, the factors affecting the choice of threshold, is unknown. So, methodologies can be devised to search for the best threshold value. Further, in order to detect a PU using cooperative spectrum sensing, a joint decision of an optimum number of users can be taken.

Serkan in [23**]** proposed a novel spectrum sensing scheme of SU and PU users without any prior knowledge. Here, a known amount of noise is intentionally added at the sensing side. Thus the sensing detector becomes unaffected to signal and noise types, thereby diminishing the problem of noise uncertainty. As this additional noise is removed after the sensing process, there is no effect on the transmission process. This scheme has been found to be robust under Rician and Nakagami fading models. The performance analysis of the algorithm was carried on real data and the experimental results showed probability of detection to be 0.90 for SNR values of -10dB.

Xiguang in [24] utilizes Kernel Least Mean Square (KLMS) algorithm for proposing a novel cooperative spectrum sensing scheme. Each SU uses energy detection technique to take a binary decision for spectrum sensing. This decision, based on KLMS algorithm, is sent to fusion center to make the final decision on the status of occupancy of the spectrum. The proposed technique can keep track of changing environment and increase the reliability of decisions in FC considerably because KLMS performs very well in judging a complex non-linear mapping in an online manner. The Monte-

Carlo simulations confirm the desirable performance of this innovative scheme.

In this paper, ROC curves are presented showing the comparison of proposed scheme with some of the existing schemes like Energy Detector, Cooperative Energy Detection, Anderson-Darling based detection and Kolmogorov-Smirnov based detection. It was shown that the new technique provided a maximum probability detection of 0.984 as compared to the above mentioned techniques for probability of false alarm of 0.1 and average SNR of -8dB.

Bagheri in [25] has derived analytical expressions for Maximal Ratio Combining (MRC) and Square Law Combining (SLC) schemes. Each secondary user senses the spectrum with Energy Detectors. The channels were modelled as Nakagami-m multipath fading and lognormal shadowing. K-out-of-n fusion decision rule was used. The Least Mean Square (LMS) algorithm was utilized at the fusion centre to detect presence or absence of PU.

Xin-Lin in [26] derived a mathematical model between spectrum sensing frequency and number of remaining packets that need to be sent; spectrum sensing frequency and the new channel availability time during which the cognitive radio networks is allowed to use a new channel (after the current channel is re-occupied by primary users) to continue to transmit the packet. Spectrum sensing frequency is how frequently a CR user detects the free spectrum.

Higher spectrum sensing frequency increases Media Access Control (MAC) layer processing overhead and delay. This can cause some multimedia packets to miss the receiving deadline. Thus the multimedia quality at the receiver side is decreased. A smaller number of remaining packets and a larger value of new channel availability time will help to transmit multimedia packets within a delay deadline. Hughes-Hartogs and DPSO algorithms are used to obtain the optimal solution in multi-channel case.

Herath in [27] studied the performance of the energy detector with diversity reception. The fading Channels used in the study were Nakagami-m and Rician fading channel. Maximal Ratio Combining (MRC), Equal Gain Combining (EGC) and Selection Combining (SC) diversity schemes were utilized. For the EGC diversity case, with Nakagami-m fading, P_d expressions are derived for the cases $L = 2, 3, 4$ and $L > 4$. For the SC diversity case, with Nakagami-m fading, P_d expressions are derived for the cases when $L >= 2$. Here, L is number of diversity branches.

Stotas in [28] proposed a novel method to address the sensing-throughput problem. Here, both sensing time and throughput were maximized. Spectrum sensing and transmission of data was done at the same time for whole frame duration. A time slot was allocated for spectrum sensing at the beginning of each frame. During this slot, data transmission was prohibited. This resulted in less false alarms and better probability of detection. Energy detectors were employed for spectrum sensing. The frame duration was fixed at 100ms. The received SNR from SU was kept at 20dB. The bandwidth of the channel and the sampling frequency was chosen as 6MHz. The probability that the frequency band is active was taken as 0.2.

Tevfik in [10] presented a survey and comparison of various techniques used for spectrum sensing such as energy detector based, waveform based, cyclostationary based, radio identification based and matched filtering based. It describes various aspects of spectrum sensing for cognitive radio. The concept of sensing in multi-dimension like frequency, time, space, code and angle was also introduced. Further, the various challenges associated with spectrum sensing were studied. Additionally, the fundamental behind cooperative sensing and its several types was explained

Liang in [29] mathematically formulated a sensing-throughput problem and validated the same with simulations. It considered the issue of modelling the sensing time while maximizing the throughput for the SU under the constraint that the interest of PUs was sufficiently guarded. The sensing scheme used for the same was energy detection. The frame duration was fixed at 100ms, SNR of PU at secondary receiver was taken as -20dB with 6Mhz Channel and $P_d = 0.9$. It was researched that optimal sensing time that gave highest throughput was 14.2 ms. On employing distributed spectrum sensing the optimal sensing time reduced to 9.5ms. For distributed spectrum sensing, 4 distributed SUs were utilized which worked cooperatively to carry out sensing using Logic-AND decision fusion rule.

Kim in [18] proposed an optimal in-band sensing scheduling algorithm. The sensing-time and sensing-frequency of energy and feature detection were optimized. The factors taken into consideration were noise uncertainty and inter-CRN interference. It was observed that energy detection above average Received Signal Strength ($aRSS_{threshold}$) incurred at most 0.385% of sensing overhead. This overhead was compared with three feature detectors: pilot-location detection, PN511 detection, cyclostationary detection. In this paper, $aRSS_{threshold}$ was varied from -114.6 dBm to -109.9 dBm. The noise uncertainty values were kept between 0.5 dB to 2 dB and -112.9 dBm to -110.5 dBm. The interfering Cognitive Radio Networks were in the range of 1 to 6.

V. Research Challenges

There are several factors inspiring the scientists across the globe to study the problems faced during sensing. These factors need to be addressed jointly to converge to an optimum solution.

- The information regarding the presence of PU is mostly missing in commercialized applications of wireless communication systems. Another hurdle is the inability to distinguish between PU and the noise. Therefore the need for blind algorithms to sense the PU is obvious as the algorithms perform without prior information about the channel or primary signals.
- As the SNR values at the sensing detector side can be as low as -1dB, hence the need is to have the detectors which are agile enough to detect under low SNR values.

- Depending on the type of application, the communication channel is prone to several types of noises. At the sensing detector, there is always a noise uncertainty. Thus detection process should be robust with constraint of noise uncertainty.
- The sensing time should be as small as possible so as to not compromise with the throughput. Thus the need is to have fast and less complex sensing operations.
- The selection of optimum threshold values for spectrum sensing techniques like Energy detection is critical because this forms the basis for decision for absence or presence of PU.
- Various types of feature detectors can be compared with the energy detectors.
- *Delay analysis in distributed schemes:* Distributed cooperative sensing schemes utilizes a repetitive procedure to make a final cooperative decision. This leads to cooperative sensing delay. If to converge to a decision, a large number of iteration is needed, then report delays would be large. Thus, delay analysis and the time taken to converge to a decision needs to be jointly examined [6] [30].
- *Cooperation-Processing Trade-Off:* If there are large number of cooperative users, probability of detection increase even if there are low-sensitive detectors possessed by each user. A detector with less sensitivity requirement leads to shorter sensing time and therefore less local processing.

However, large number of cooperative users causes large overhead as there is large volume of data that needs to be reported and centrally processed. This increased overhead causes an increase in the sensing time and the processing of data. Thus there is a need for a trade-off between them.

- *Reactive Vs. Proactive Sensing:* In reactive sensing scheme, the spectrum sensing is performed only when data needs to be transmitted. In contrast, proactive schemes maintain a list of idle licensed bands available for opportunistic access. The later scheme helps in reduction of sensing time. However, both the mentioned spectrum sensing techniques involve high sensing overhead. The reduction of sensing overhead thus is always a challenge. The applications which are very sensitive to delays may find proactive sensing attractive. For instance, while searching an over-crowded band with few spectrum holes may increase delay considerably if on-demand scheme is followed. While, the applications having energy efficiency constraints but are delay tolerant may prefer reactive sensing. Thus a CR needs to adapt to either of the sensing mode (reactive or proactive) in order to achieve optimum performance. The implementation of such system requires parameters to be optimized [31].
- *Rate-Reliability Trade-Off:* Whenever a PU is detected, the CR user needs to stop transmission and relocate to a new idle band. Though the delays associated with these relocations may be reduced using proactive scheme, yet cognitive users have to bear Quality of Service (QoS) degradation. This is due to the fact that communication peers have to coordinate for the frequency transition. The communication reliability can be increased by spreading the transmission data over wide spectrum. So that, if a PU reclaims this band, it affects only a small portion of cognitive user's bandwidth. At the receiver side, the data from several frequency chunks can be combined to form a reliable CR link. Orthogonal-Frequency-Division-Multiplexing (OFDM) can be utilized for this purpose due to its inherent flexibility in using non-adjacent frequency bands. The challenge posed by this method is that additional temporal or spectral resources have to be allocated for periodic sensing of these extra frequency bands which reduces the effective data rate of CR user [30].
- *Sensing-Throughput:* If the sensing time is increased, it increases the probability of detection $\mathbf{P_d}$ and decreases the probability of false alarm $\mathbf{P_f}$. However this results in decreased transmission time $\mathbf{T_t}$. This results in reduced throughput of CR. Thus, sensing time and throughput needs to be optimized [22] [29].
- *Reporting Time - Sensing Time*: Jaewoo in [32] formulated a problem to maximize the average capacity of SU by jointly taking into account the sensing time, reporting time, a fusion scheme and soft combination at fusion centre. Also, a limited reporting scheme for a multiband cooperative spectrum sensing was proposed. This scheme proved useful in increasing the average SU capacity by reducing the reporting overhead while providing desirable protection to primary user's interest. Probability of false alarm was also calculated. It was inferred that with 40 cooperative users, the proposed scheme gave a rise of 21% in average capacity of SU as compared to conventional schemes.

VI. Conclusion

Radio spectrum is scarce. Therefore there is a need to efficiently utilize it. Cognitive radios employ the concept of dynamic spectrum reuse. They intelligently identify the spectrum holes in licensed-bands which are used by primary users, and dynamically allow the secondary users to transmit in these spaces. The absence or presence of a licensed user is governed by null hypothesis or alternate hypothesis. The spectrum sensing techniques can sense with prior knowledge of signal and power or can use blind detection techniques. When a large number of cognitive users operate in cooperation, then the probability of false alarm reduces. Further, this provides answer to issues like hidden primary user problem, fading, shadowing and noise uncertainty. This article reviews the ongoing research in the domain of spectrum sensing. Here types of channel modelled, various spectrum sensing techniques employed, diversity combining schemes used, traditional and optimised algorithms at the fusion centre are discussed.

VII. Open Problem

The future directions in this domain hold the need for optimization of techniques used in spectrum sensing and cooperative spectrum sensing. The attributes that can be investigated are probability of false alarm, probability of detection, SNR values, throughput, sensing time etc. The issue like selection of threshold value to detect the PU needs to be studied. The sensing time should be as small as possible but not at the cost of throughput. Moreover, sensing can be done on demand or proactive basis.

Acknowledgments

The authors would like to thank Panjab University for allowing to conduct the research at University Institute of Engineering & Technology. The study could not have been possible without the constant support of Dr. Ashok K. Chitkara, Chancellor, Chitkara University. Deep gratitude to Mr. Abhishek Pal Garg, MBA (INSEAD, France), Deputy Secretary, Government of India for proof reading and providing valuable feedback. Many thanks to the anonymous reviewers, whose insightful comments made this a better paper.

References

[1] C. Xin, M. Song, L. Ma and C. C. Shen, "ROP: Near optimal rendezvous for dynamic spectrum access networks," *IEEE Transactions on Vehicular Technology,* vol. 62, no. 7, pp. 3383-3391, 2012.

[2] M. Song, C. Xin, Y. Zhao and X. Cheng, "Dynamic spectrum access: from cognitive radio to network radio," *IEEE Wireless Communications,* vol. 19, no. 1, pp. 23-29, 2012.

[3] B. Zhang, Y. Chen and K. Liu, "An indirect-reciprocity reputation game for cooperation in dynamic spectrum access networks," *IEEE Transactions on Wireless Communication,* vol. 11, no. 12, pp. 4328-4321, 2012.

[4] L. Giupponi, A. Galindo-Serrano, P. Blasco and M. Dohler, "Docitive networks: An emerging paradigm for dynamic spectrum management," *IEEE Wireless Communication,* vol. 17, no. 4, pp. 47-54, 2010.

[5] I. F. Akyildiz, W. Y. Lee, M. C. Vuran and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communication Magazine,* vol. 46, no. 40, pp. 40-48, 2008.

[6] I. F. Akyildiz and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Physical Communication,* vol. 4, no. 1, pp. 40-62, 2011.

[7] D. Cabric, A. Tkachenko and R. Brodersen, "Spectrum sensing measurement of pilot, energy and collaborative detection," in *IEEE Military Communication Conference,* Washington D.C, USA, 2006.

[8] L. Lu, X. Zhou, U. Onunkwo and G. Y. Li, "Ten years of research in spectrum sensing and sharing in cognitive radio," *EURASIP Journal of Wireless Communications and Networkings,* pp. 1-16, 2012.

[9] B. Wang and K. R. Liu, "Advances in cognitive radios: A survey," *IEEE Journal of Selected Topics in Signal Processing,* vol. 5, no. 1, pp. 5-23, 2011.

[10] T. Yücek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys & Tutorials,* vol. 11, no. 1, pp. 116-130, 2009.

[11] S. Shobana, R. Saravanan and R. Muthaiah, "Matched filter based spectrum sensing on cognitive radio for OFDM WLANs," *International Journal of Engineering and Technology,* vol. 5, no. 1, pp. 142-146, 2013.

[12] N. Giweli, S. Shahrestani and H. Cheung, "Spectrum sensing in cognitive radio networks: QoS considerations," in *Seventh International Conference on Networks & Communications NETCOM 2015,* Sydney, Australia, 2015.

[13] K. Kim, I. A. Akhbar, K. K. Bae, J. S. Um, C. M. Spooner and J. H. Reed, "cyclostationary approaches to signal detection and classification in cognitive radio," in *Symposium on New Frontiers in Dynamic Spectrum Access Networks,* Dublin, 2007.

[14] J. Lunden, A. Koivunen, A. Huttunen and H. V. Poor, "Spectrum sensing in cognitive radios on multiple cyclic frequencies," in *2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications CrownCom-2007,* Orlando, Florida, USA, 2007.

[15] H. Tang, "Some physical layer issues of wideband cognitive radio systems," in *International Symposium on New Frontiers in Dynamic Spectrum Access Networks,* Baltimore, USA, 2005.

[16] T. Yücek and H. Arslan, "Spectrum characterization for opportunistic cognitive radio systems," in *IEEE Military Communication Conference,* Washington D.C, USA, 2006.

[17] Y. Zeng, Y.C Liang, A.T. Hoang and R.Zhang;, "A review on spectrum sensing for cognitive radio challenges and solutions," *EURASIP Journal of Advances in Signal Processing,* vol. 10, 2010.

[18] H. Kim and K. G. Shin, "In-band spectrum sensing in cognitive radio networks: Energy detection or Feature detection?," in *14th ACM International Conference on Mobile Computing and Networking,* San Fransico, California, USA, 2008.

[19] D. Cabric, S. M. Mishra and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in *Thirthy-Eight Asilomar Conference on Signals, Systems and Computers,* Pacific Grove, California, USA, 2004.

[20] Y. Zeng and Y. C. Liang, "Spectrum-sensing algorithms for cognitive radio based on statistical covariances," *IEEE Transactions on Vehicular Technology,* vol. 58, no. 4, pp. 1804-1815, 2009.

[21] D. B. Rawat and G. Yan, "Spectrum sensing methods for dynamic spectrum sharing in cognitive radio networks: A survey," *International Journal of Research and Reviews in Wireless Sensor Networks,* vol. 1, no. 1, pp. 1-13, 2011.

[22] R. A. Rashid, A. Rashid, A. H. B. A. Hamid, N. Fisal, S. K. Syed-Yusof and H. Hosseini, "Efficient in-band spectrum sensing using swarm intelligence for cognitive radio network," *Canadian Journal of Electrical and Computer Engineering,* vol. 38, no. 2, pp. 106-115, 2015.

[23] S. Ozbay and E. Ercelebi, "A new wireless network scheme for spectrum sensing in cognitive radio," *Elektronika Ir Electrotechnika,* vol. 21, no. 6, pp. 90-95, 2015.

[24] X. Xu, R. Qu, J. Zhao and B. Chen, "Cooperative spectrum sensing in cognitive radio networks with kernel least mean square," in *5th International Conference on Information Science and Technology (ICIST),* China, 2015.

[25] A. Bagheri, A. Shahini and A. Shahzadi, "Analytical and learning-based spectrum sensing over channels with both fading and shadowing," in *International Conference on Connected Vehicles and Expo,* Nevada, USA, 2013.

[26] H. Xin-Lin, W. Gang, H. Fei and S. Kumar, "The impact of spectrum sensing frequency and packet-loading scheme on multimedia transmission over cognitive radio networks," *IEEE Transactions on Multimedia,* vol. 13, no. 4, pp. 748-761, 2011.

[27] S. P. Herath, N. Rajatheva and E. Tellambura, "Energy detection of unknown signals in fading and diversity reception," *IEEE Transactions on Communications,* vol. 59, no. 9, pp. 2443-2453, 2011.

[28] S. Stotas and A. Nallanathan, "Overcoming the sensing-throughput trade-off in cognitive radio network," in *IEEE International conference on Communication (ICC),* Cape Town, South Africa, 2010.

[29] Y. Liang, Y. Zeng, E. Peh and A. Hoang, "Sensing throughput tradeoff for cognitive radio networks," *IEEE Transaction Wireless Communication,* vol. 7, no. 4, pp. 1326-1337, 2008.

[30] A. Ghasemi, "Spectrum sensing in cognitive radio networks: Requirements, challenges and design trade-offs," *IEEE Communications Magazine,* vol. 46, no. 4, pp. 32-39, 2008.

[31] L. Wang and C. Wang, "Spectrum handoff for cognitive radio networks: Reactive-sensing or Proactive-sensing?," in *IEEE International Conference on Performance, Computing and Communications*, Austin, Texas, USA, 2008.

[32] J. So and T. Kwon, "Limited reporting-based cooperative spectrum sensing for multiband cognitive radio networks," *International Journal of Electronics and Communications,* vol. 70, no. 5, pp. 1-12, 2016.

Evaluation of a Behind-the-Ear ECG Device for Smartphone based Integrated Multiple Smart Sensor System in Health Applications

Numan Celik [1*], Nadarajah Manivannan [1], Wamadeva Balachandran [1]
[1] Department of Computer and Electronics Engineering
Brunel University London
London, United Kingdom

***Abstract*—In this paper, we present a wireless Multiple Smart Sensor System (MSSS) in conjunction with a smartphone to enable an unobtrusive monitoring of electrocardiogram (ear-lead ECG) integrated with multiple sensor system which includes core body temperature and blood oxygen saturation (SpO_2) for ambulatory patients. The proposed behind-the-ear device makes the system desirable to measure ECG data: technically less complex, physically attached to non-hair regions, hence more suitable for long term use, and user friendly as no need to undress the top garment. The proposed smart sensor device is similar to the hearing aid device and is wirelessly connected to a smartphone for physiological data transmission and displaying. This device not only gives access to the core temperature and ECG from the ear, but also the device can be controlled (removed and reapplied) by the patient at any time, thus increasing the usability of personal healthcare applications. A number of combination ECG electrodes, which are based on the area of the electrode and dry/non-dry nature of the surface of the electrodes are tested at various locations near behind the ear. The best ECG electrode is then chosen based on the Signal-to-Noise Ratio (SNR) of the measured ECG signals. These electrodes showed acceptable SNR ratio of ~20 db, which is comparable with existing tradition ECG electrodes. The developed ECG electrode systems is then integrated with commercially available PPG sensor (Amperor pulse oximeter) and core body temperature sensor (MLX90614) using a specialized micro controller (Arduino UNO) and the results monitored using a newly developed smartphone (android) application.**

***Keywords*—wireless body area networks; body-worn sensors; ECG; core body temperature; oxygen saturation level (SpO_2); biosensor integration; m-health**

I. INTRODUCTION

The rapid growth of wireless technologies brings new innovative ideas that enables continuous real-time remote patient monitoring in healthcare services using compact wireless body sensors. The services and technologies provide relatively uncontroversial, well-communicated and monitoring devices, developed to give more affordable solutions specifically for mobile healthcare, such as daily activity monitoring, personal healthcare and monitoring systems, and body sensor systems that can alert the clinicians via the patients' mobile phones. New trend in remote patient monitoring is moving toward the use of personal mobile devices compatible with multiple biomedical sensors using wireless communication, such as Bluetooth and Zigbee [1].

In practice, this recent mobile health (m-Health) technology enables to see the people's daily activity in their smartphones. In addition, these mobile-based portable embedded devices will provide platforms to monitor their critical physiological data continuously and remotely. An assessment report has been prepared for the European Union regarding the effectiveness of m-Health in biomedical applications and the diagnosis of the diseases in 2013. According to this report, m-Health applications could save €99 billion in healthcare costs in the EU and add €93 billion to the EU GDP in 2017, if its adoption is encouraged [2].

An integrated wearable monitoring system, which aim to bring compact body sensors such as electrocardiography (ECG), blood pressure, photoplethysmography (PPG), core body temperature (CBT), heart rate, pulse oximetry, and EEG together forms the concept of a wireless body area network (WBAN) or personal area network (PAN) and displays the physiological signals on a monitoring device. One of the benefits that PAN systems bring is the ability to integrate multiple intelligent sensors, wireless connectivity and a battery into a wearable patch unit that sends the physiological data to a mobile device. Figure 1 indicates the concept of a typical WBAN where the general tasks of the electronics designer are compactness, integration of body sensors and wireless connectivity, including a telemedicine system, which can alert a clinician when life-threatening changes occur or to provide a feedback to the patient to help maintain an optimal health status using Cloud health services.

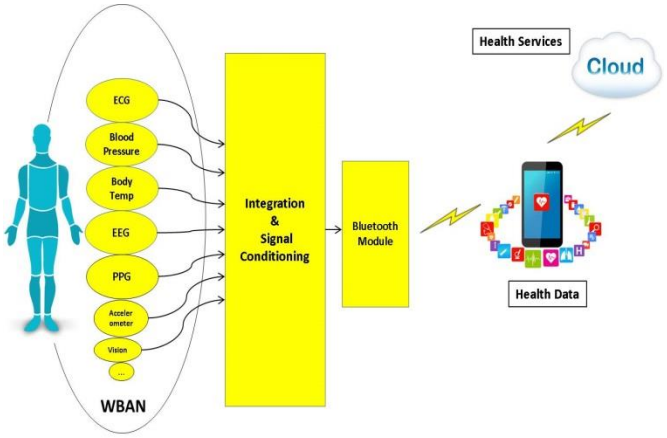


Fig. 1. Typical personal health system with integrated sensor network (wireless body area network)

Here, we examine a continuous, wearable and wireless critical-sign patient monitoring system which was focused on integrating ECG sensors placed behind the ear; CBT sensor placed in the ear; and PPG sensor clipped on the finger. The reason for choosing the ear as a location in these experiments is that it makes possible to measure ECG and CBT together with subject comfortless of ECG monitoring. Behind the ear region has less hair than chest and non-hair regions give more suitability for long term use of ECG monitoring. Moreover, the preparation of skin will not be a necessity and it will not be difficult to remove the sensors that bring a user-friendly perspective to the patients and clinicians.

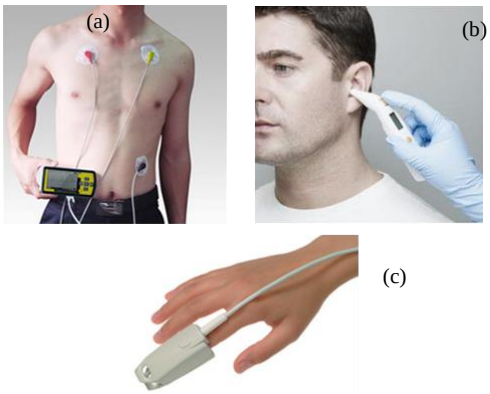


Fig. 2. (a) the typical 3-lead ECG monitoring system; (b) ear-tympanic device to measure core body temperature; (c) the reflective PPG sensor to measure blood oxygen saturation (SpO2) and heart rate

This paper aims to advance the ambulatory ECG treatments by proposing a wearable behind-the-ear smart sensor that transmits integrated physiological data wirelessly to a smartphone and monitors ECG, CBT, heart rate and SpO_2. Since most conventional systems take each data separately and from different locations on the body (see in Figure 2), the proposed work based on ear-lead ECG monitoring integrated with the multisensory system is important for many reasons. Firstly, this is the first attempt of ear-lead ECG monitoring to be integrated with two other body-sensors, PPG and tympanic sensors to measure heart rate, SpO_2 and CBT respectively. Secondly, a smart sensor system is developed that facilitates sensor integration by sealing into an ear mold. Thirdly, the collected health data in the microcontroller unit (MCU) are wirelessly transferred to an Android smartphone and an application is written to display ECG, CBT and PPG data. Another aim of this paper was to investigate ECG signals when different types of electrodes (gel and dry electrodes) have been placed on different locations on the body, such as chest, ear and arm regions. Moreover, this design is less disrupted by motion artifacts for ear-lead ECG monitoring when compared to the previous work which is done by Da He [27] using two ECG sensors. In the future, this option not only gives flexibility to the patient, but also can give the clinicians to access the physiological data in real time via online access from personal healthcare records.

The remainder of this paper is organized as follows: Section 2 includes related works on this topic and deals with the problem formulation; Section 3 explains the system description of the work; Section 4 illustrates experimental results and analysis of ear-lead ECG using different sensors onto the different locations of the body; and also the demonstrations of CBT and PPG sensors. Section 4 also draws the combining ECG sensing unit with the integrated CBT and PPG sensors, and demonstrates wireless transmission between integrated MCU unit and a smartphone; finally conclusion and discussion are presented in Section 5.

II. Related Work

When a ubiquitous smart sensor is developed, there should be basic components that should be considered: easy-to-use, mobility, accuracy, and security. There are currently many ongoing researches that investigate various solutions on the design of wireless personal healthcare monitoring devices [3-5]. Lee et al. developed a mobile phone based ECG monitoring application. The system describes the design and implementation of a prototype tele-health system which monitored physiological signals of patients in real-time [4]. Sanches et al. designed an electronic temperature sensor within a headset Bluetooth device that sends the temperature measurements to a mobile phone. The proposed system measures central body temperature continuously at the ear [6]. Jung et al. proposed a wireless body sensor platform called 'Virtual Cuff' that comprises PPG and ECG sensors to estimate systolic and diastolic blood pressure (BP). The presented work fuses data from various sensors, including ECG, PPG, accelerometer, and GPS, for extrapolating BP information using signal characteristics that are derived from PPG and ECG waveforms [7]. Do Valle et al. examined [8] a behind-the-ear device that records EEG measurements on smartphone continuously and then uploads the patients' data to a secure server. Song et al. developed a body monitoring system design based on android smartphone including three main functions such as brainwave capture (EEG), ECG and temperature. These data are gathered by hardware and sent to the Bluetooth receiving device of android smartphone [9]. Boano et al. managed to measure core body temperature (CBT) on ambulatory patients and exercising athletes using a wireless wearable device that measures the tympanic temperature at the ear. The CBT data is transmitted via ATmega128RFA1 chip based on ZigBee communication which is different from other studies [10].

The authors of [11] proposed an ultra-wearable smart sensor system which combines ECG, tri-axial accelerometer, and GPS sensors to measure normal or elderly person's daily activities. This device also encompasses voice biofeedback and data fusion technologies in order to accommodate future needs and to make the smart sensor much better. The hardware unit of the system consists of a TI MSP430 microprocessor, Bluetooth wireless transmission to PC client software, micro SD card storage, and LCD display. The embedded algorithm combines two sensors for noise reduction, and utilizes voice biofeedback for exercise overload warning. Hernandez et al. demonstrated [12] that motion sensors of a smartphone can be used to recover heart and breathing rates of the users during stationary positions and activities while the smartphone was being carried in a bag or pocket or even during listening on the phone. They developed these rates from accelerometer data and compared them with measurements obtained with FDA-cleared sensors by evaluating effective accuracy numbers. Hii et al. presented [13] a comprehensive ubiquitous healthcare solution which includes a real time ECG monitoring and analyzing system based on an Android mobile device and also provides medicine care assistance. Wireless sensor network (WSN) technology is used in this system in order to transmit ECG data wirelessly from the patient's body to a smartphone device. As for medicine care assistance, barcode technology is applied to assist out-patients in medication administration, by capturing and decoding the barcode on medicines using the smartphone's embedded camera.

Wahl and his co-workers designed [15] an eyeglass (WISEglass) that consists of inertial motion, environmental light, and pulse sensors, processing and wireless data transmission functionality and also a rechargeable battery. The users will be able to monitor their daily activity recognition, screen-use detection, and heart rate estimation, because of having accelerometer, gyroscope and pulse sensors in itself. Regarding the work in ear-worn devices, the authors in [16] attempted to develop probes for Heart-phone to make such an unobtrusive earphone to measure heart rate using PPG technology. According to their design, the reflective photo sensor is embedded into each earbud on a pair of regular earphones. To obtain measurements, the sensor earphones are inserted into the ear and positioned such that the reflective photo sensor is against the inner side of ear. Then they can measure the amount of the reflected light from the blood vessels in the region. Moron et al. studied [17] on the technical performance of medical wireless personal area network (WPANs) that are based on smartphones. According to their telemedicine prototype, an Android based smartphone acts as a gateway between a set of wireless medical sensors and a data server. They also wanted to see the differences while modifying the smartphone model, the type of sensors connected to the WPAN, the use of other peripherals such as GPS receiver, the impact of the use of the Wi-Fi interface. The authors of [18] presented a wearable monitoring system to measure the driver alertness, evaluated by a smartwatch device based on fusion of direct and indirect method. The driver chronic physiological state is monitored by adopting a PPG sensor on the driver's finger that is connected to a wrist-type wearable device. A Bluetooth low energy module connected to the wearable device transmits the PPG data to the smartwatch in real-time. Lin et al. [20] proposed a wearable PPG sensor module based on a Programmable System on a Chip (PSoC), in the course of driving. It transmits measured PPG signal from earlobe to a smartphone via Bluetooth. On the smartphone, a heart rate (HR) detection algorithm is implemented. When the abnormal HR is detected, the smartphone uses the sound and vibration to warn the driver using a magnetic ring. At the same time, physiological data and GPS location are also be transmitted to a data server (healthcare server system) via the 3G mobile network, so that the staff in the server system can monitor the recent information and monitor the driver's status.

The brief summary of this literature review finds the use of the modern communication technology for data exchange between ambulatory patients and mobile devices by combining several sensors in one typical system. This is one of the key factors to make m-Health platform much more desirable. However, conventional ECG monitoring systems, mentioned in the literature work, are measured from the chest and require more than 3 sensors in some cases. While trying to get ECG data from the chest, it is an undesirable method to integrate ECG sensors with other multiple sensors on the body, causing complexity for the whole system. This complexity would increase the noise level of each sensor and also the power requirement for the system. In the proposed sensor system, we are focusing the combination of an ear-lead ECG monitoring with integrated CBT and PPG multisensory system for m-health applications. In order to avoid complexities, ear-lead ECG method was used to avoid hairy region and easy-to -use perspectives. The proposed m-health application is not only a solution for the complexities involved in traditional of ECG monitoring system, but also enables an integrated and personalized smart device.

III. System Description

The smart sensor platform was developed to experimentally combine three different body sensors and monitor regarded vital signals on the smartphone. The three sensors of the proposed system are ECG with behind the ear electrode, core body temperature (tympanic sensor in the ear canal), and both heart rate and blood oxygen saturation (PPG from the finger), and then the detected signals are communicated to a mobile client by using Bluetooth connection. The ECG and CBT sensors for displaying the physiological data are integrated in the form of an earbud. PPG sensor was attached onto the finger and connected to the microcontroller unit (MCU). Afterwards, the physiological signals are monitored on the mobile phone. Figure 3 shows the block diagram of the smart sensor system.

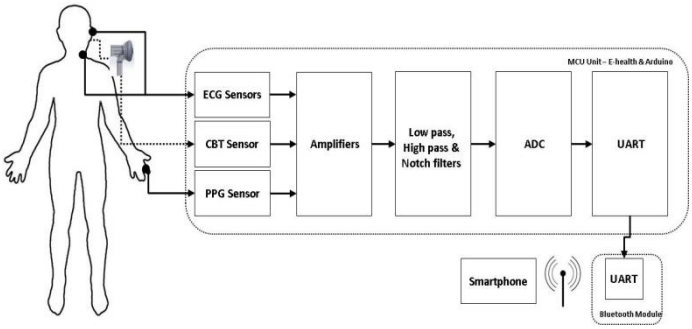


Fig. 3. System block diagram

A. ECG Sensing Unit

As can be seen from the block diagram, ECG and CBT sensors were deployed together because of leading sensors near the ear location, but PPG sensor is clipped onto the finger. Three ECG sensors in different sizes were used and tested using both gel and dry electrodes. Table 1 lists the properties of three commercially available electrodes, which were used in the experiments. In the proposed system, the first electrode of three-electrode setup was placed behind the ear; the second electrode was attached on the upper neck area, and the last electrode was placed on the arm. Initially, two types of gel (adhesive Ag-AgCl) electrodes were used to get ECG data from the ear. One of the gel electrodes was known as 'Covidien electrode', which has a diameter of 24 mm, and the second electrode was slightly bigger, which has a diameter of 38 mm. Both electrodes are of 1 mm of thickness. After using these two gel electrodes, a number of dry electrodes were applied to the skin to detect ECG signals. The dimensions of this dry electrode are 10 mm (diameter) and 2 mm (thickness). Figure 4 shows the photographs of used electrodes.

TABLE I. THE PROPERTIES OF ELECTRODES USED FOR THE EXPERIMENTS

Name	Type	Size – diameter x thickness (mm)
E1	Gel	38x1
E2	Gel	24x1
E3	Dry/Non-gel	10x2

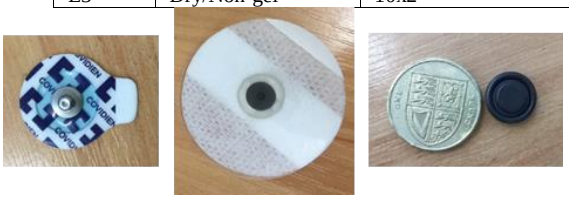


Fig. 4. Three ECG electrodes (from-left-to-right): E2, E1, and E3 (Table 1)

B. CBT Sensing Unit

In the proposed smart sensor system, we have used a thermopile sensor that is a non-contact sensor for measuring core body temperature (CBT). Because tympanic temperature directly reflects the core temperature of the carotid artery [28], we have proposed an ear-bud design of infrared tympanic sensor that can continuously measure the temperature of the tympanic membrane. This design connects with an MCU which is the core controller of the whole smart sensor system to perform signal processing. The thermopile sensor (MLX90614) [22] gives a very sensitive information regarding the core temperature with 17 bit ADC resolutions, thus this could be in some cases 0.0034 °C. Figure 5 illustrates the proposed design of CBT sensing unit and embedded thermopile sensor together.

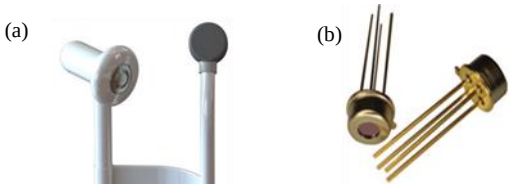


Fig. 5. CBT sensing components: (a) The proposed earphone-type infrared sensor, (b) infrared thermopile sensor

C. PPG Sensing Unit

The finger-worn PPG sensor consists of a transductor which initiates two LEDs and a photodiode detector. One of the LEDs emits red light (with a wavelength of λ = 660 nm) and other LED performs an infrared light (with a wavelength of λ = 880 nm). This technique is known as PPG (photoplethysmography) and the PPG sensor is based on the theory that the colour of blood range depends on the oxygen it contains. For instance, hemoglobin particularly reflects more red light and the PPG sensor detects oxygen saturation in the blood (SpO_2). For our test-bed application regarding PPG sensor unit, we measured both oxygen level in the blood (SpO_2) and heart rate together using Amperor Bluetooth Finger PPG sensor [24] which is shown in Figure 6. Proposed smart sensor system incorporates a Bluetooth connection to get heart rate and SpO_2 from PPG sensor.

Fig. 6. Bluetooth finger PPG sensor

The analog signals from the sensors are conditioned at the wearable hardware unit to levels suitable for digitization and processing. Two stage amplification units were used with gain of 10 and 100 to avoid the noises overriding the ECG signals, which is achieved by an instrumentation amplifier, and a micro-power operational amplifier, respectively. The ECG signals are restricted in bandwidth of 0.5-100 Hz using a high pass and low pass filters after the first and second steps of amplification, respectively. The power line interference in the ECG signal is filtered using a 50 Hz notch filter to avoid loss of 50 Hz component of the ECG signals. The PPG sensor probe has an infrared source at 880 nm and the photodetector giving current output, which is converted to voltage by an instrumentation amplifier with gain 10 and using high and low pass filters between 0.5 and 20 Hz. The CBT sensing block consists of a calibration circuit and a high gain amplifier with 10. The system runs from a lithium-battery (see in Figure7) which has a capacity of 500mAh, and lasts approximately 30 hours, depends on working with Bluetooth as it consumes a significant proportion of whole power assumption. The prototype ear-bud device was designed in SolidWorks and created using a 3D printer. The design is shown in Figure 5(a), measures 8.5 cm within cables and sits behind the patient's ear (mastoid area). The design was made using flexible unit to move around the ear and conform better to the ear so that it helps to secure the device.

IV. EXPERIMENTAL RESULTS

In this section, the experimental tests and results were demonstrated individually from ECG, CBT and PPG sensors

into a Matlab program. The impact of using different type of electrodes (gel and dry) was drawn for ECG monitoring system. Ear-lead ECG monitoring was compared to conventional chest-lead model and different scenarios were analysed as changing positions of ECG electrodes on the body to clearly see which scenario gives the best SNR and least noise. Furthermore, the integration of ECG, CBT and PPG sensors into wearable hardware unit and transmission the physiological data to the smartphone using Bluetooth radio were demonstrated in this section.

A. ECG Experimental Results

Figure 7 shows the experimental setup for ECG monitoring with wireless Android based smartphone. Figure 8a provides basic components of a typical ECG signal including various features (P, Q, R, S and T waves). A number of intervals can be measured and analyzed from an ECG recording. A normal ECG signal can give very important information regarding the heart status during a cardiac cycle. The time between the beginning of a particular point in a cardiac cycle and the beginning of another particular point in the next cardiac cycle is the interval between the beginning electrical responses of those particular points in the heart. For example, the heart rate can be measured by simply looking at R-R (beat-to-beat) interval, which indicates the time between two consecutive QRS complexes in an ECG recording. A sample ECG signal is shown in Figure 8b including two successive beats (R-R intervals and P-QRS-T complexes). This test signal is used to compare with corrupted ECG signals from the sensor platform. The SNRs of ECG signals in Table 2 are calculated by differentiating between each signal in Figure 7 and this test ECG signal, respectively.

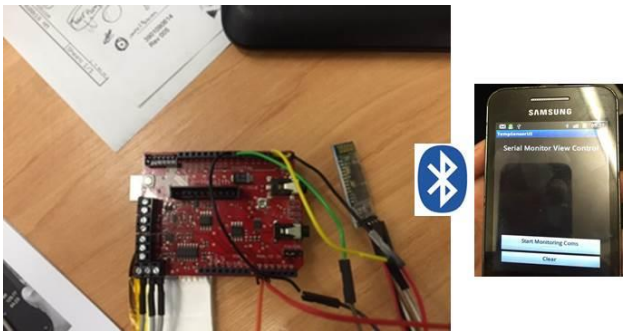
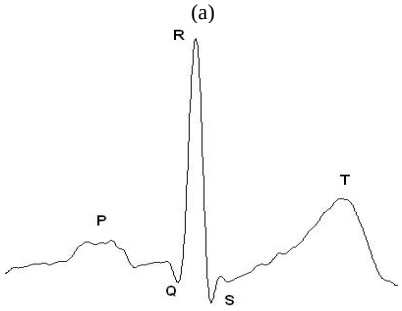


Fig. 7. Experimental device of Arduino based ECG measurement system with wireless smart phone monitoring



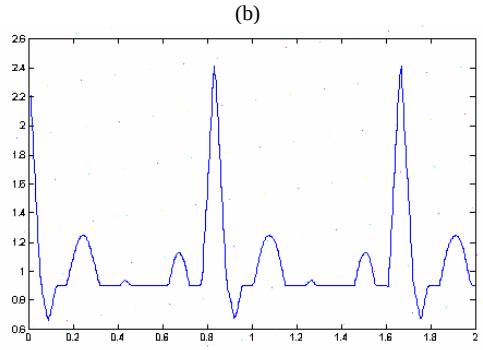


Fig. 8. ECG Signal Components: (a) A typical ECG signals including P-QRS-T morphology; (b) Test signal to be compared with the signals obtained from our experiments

Figure 9 shows the placements and types of electrodes for ECG measurements reported in this paper. Different scenarios are drawn in the figure in order to see the important changes and what kind of challenges there are. Another aim is also to select the best unobtrusive scenario from the figure, according to ECG signal qualities, and SNR. As can be seen from the figure, the electrodes were placed on different locations on the body such as ear, chest and arm. Both the standard three gel Ag/AgCl electrodes and dry electrodes with different positions were used to measure ECG signals.

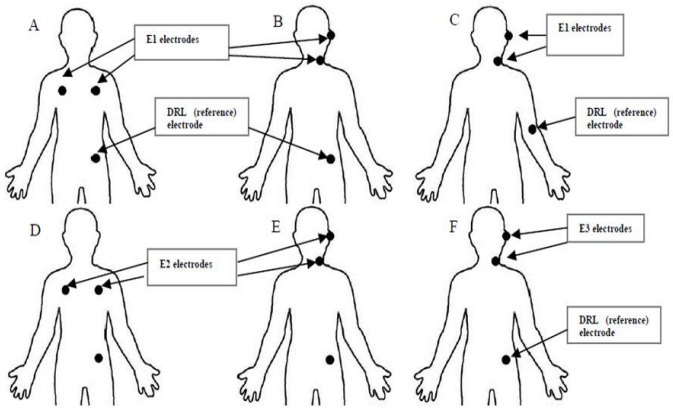


Fig. 9. The locations and type of the ECG electrodes for the measurements (Scenarios A – F)

All of ECG signals shown in Figure 10 were measured in lead I according to Einthoven's triangle [19] using ECG sampling module and rebuilt in Matlab without any software de-noise for further study. Power consumption is critical for such an application as this. The system runs from a lithium-battery (see in Figure 4) which has a capacity of 500mAh, and lasts approximately 30 hours, this depends on the use of Bluetooth module as it consumes a significant proportion of the available power.

Figure 10 indicates ECG results from each electrode and each placement according to the scenarios which are shown in Figure 9. The electrodes with larger array size exhibits less noises due to larger skin-electrode contact area. Even the dry electrode provided ECG signals from behind-the-ear that was comparable to Ag/AgCl electrode. The figures illustrate clear observations of the QRS complex and T-wave cardiac signs.

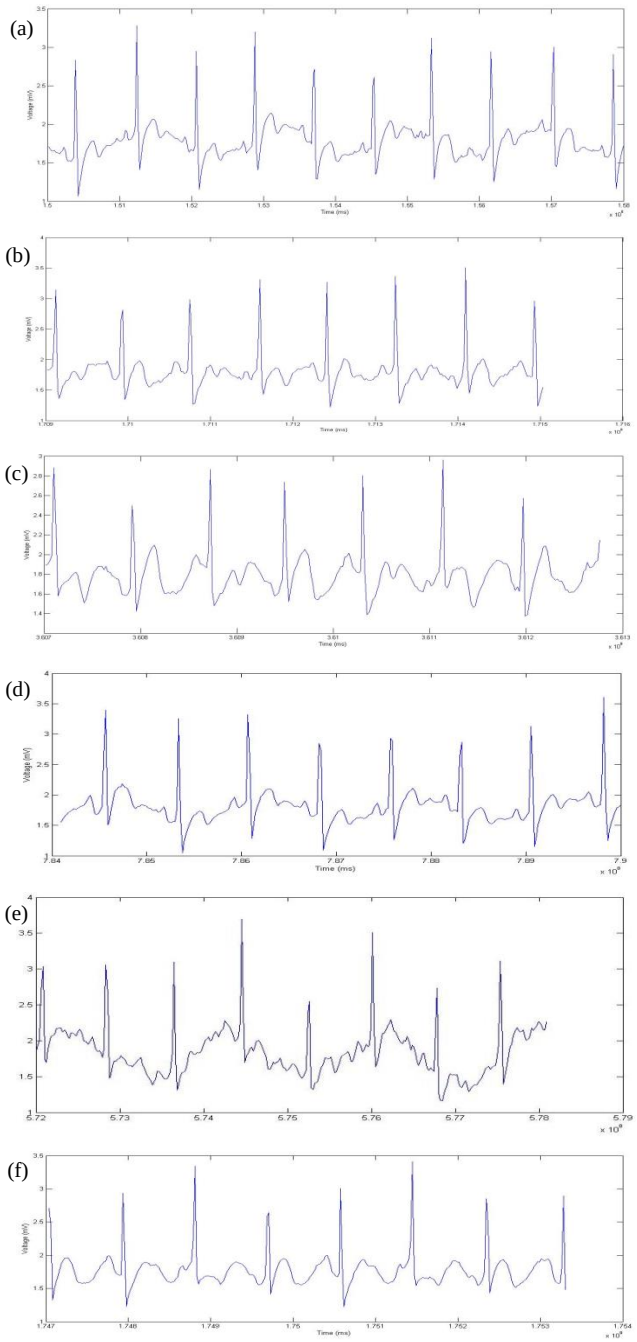


Fig. 10. The measured ECG signals at various locations as illustrated in Fig.9 (a) Scenario A; (b) Scenario B; (c) Scenario C; (d) Scenario D; (e) Scenario E; (f) Scenario F

As can be seen from Figure 9 and 10, E1 electrodes were applied in the measurements from (a) to (c). Figure 10(a) shows basic ECG measurement while two active electrodes were placed on the chest and one reference DRL (driven-right-leg) electrode was attached on the waist. This is a typical way of measuring ECG under the rule of 3-leads ECG recording [26]. Figure 10(b) indicates the ECG data taken from behind-the-ear and reference electrode was placed on the waist and Figure 10(c) shows the ECG signal for the same scenario as those in Figure 10(b) while reference electrode was placed on the arm instead of the waist. After that, we implemented E2 electrodes for measuring ECG and, Figure 10(d) illustrates ECG data taken from the chest and reference electrode was on the waist. Using the same type of electrodes, Figure 10(e) shows the ECG data coming from behind-the-ear and reference electrode was on the waist. Lastly, Figure 10(f) shows ECG signal using dry electrodes (E3 electrode) were placed behind the ear; and DRL electrode was on the waist. The received signal is accompanied with noise; however, R peaks (the most dominant feature in the ECG cycle) could be identified. Figure also illustrates that some distortions and fluctuations have occurred during ECG recording due to body movements and taking new position of ECG electrodes.

To further evaluate the performance of the electrodes, ECG signals were analyzed to calculate signal-to-noise ratio (SNR) using the following equation [14]:

$$SNR = 20 \log(S/(S` - S)) \quad (1)$$

where S is the filtered ECG signal with a frequency ranging from 0.5 Hz to 100 Hz, and S` is defined as ECG signal without filtering. Before calculation, the power line interference (50 Hz) was removed from both signals. Table 2 summarizes the SNR of 6 different ECG results which are represented in Figure 10 a-f, respectively.

TABLE II. SNR OF ECG SIGNALS WITH DIFFERENT ELECTRODES AND PLACEMENTS

Experiment Scenario (see Fig 9)	SNR (dB)	Response Time (s)
Scenario A	25.21	0.85
Scenario B	17.27	1.17
Scenario C	12.95	1.80
Scenario D	21.23	0.90
Scenario E	15.34	1.25
Scenario F	10.92	~ 35

As seen from the results of Figure 10 and Table 2, it is obvious the ECG signals from the chest (Scenario A and Scenario D) are the best, in terms of the detections of various parts of the ECG (P, Q, R, S and T) and they are less noisy too. Moreover, when we compared Scenario A with Scenario D or Scenario B with Scenario E, figures clearly show that E1 electrodes provides better ECG waveforms in terms of noise than attached E2 electrodes due to larger skin-contact size. On the other hand, after waiting some time (longer response time as given Table 2) to remove unwanted distortions, the P wave and QRS complex can be identified in Scenario F, which is the most effective method because of the use of the dry electrodes.

As the work is aimed at behind the ear electrodes for the benefits already mentioned previously, it is clear from the results, behind-ear electrodes still produces comparable signals and hence usable for the purpose of convenient ECG detection. While analyzing the concept of Scenario F, thus, dry electrodes are chosen for detection of ECG waveforms and integration with other body sensors together.

Furthermore, we monitored ECG signals (including QRS complex) on an Android based smartphone including heart rate (HR) as shown in Figure 11(b). Android was used because of its widespread use and development and the software is portable for coding within mobile devices. Additionally, it allowed simplicity of integration with ECG sensors via

Bluetooth. We also compared our ear-lead ECG results with recently proposed by Da He [27] who worked to get ECG data using behind-the-ear device (Figure 11a).

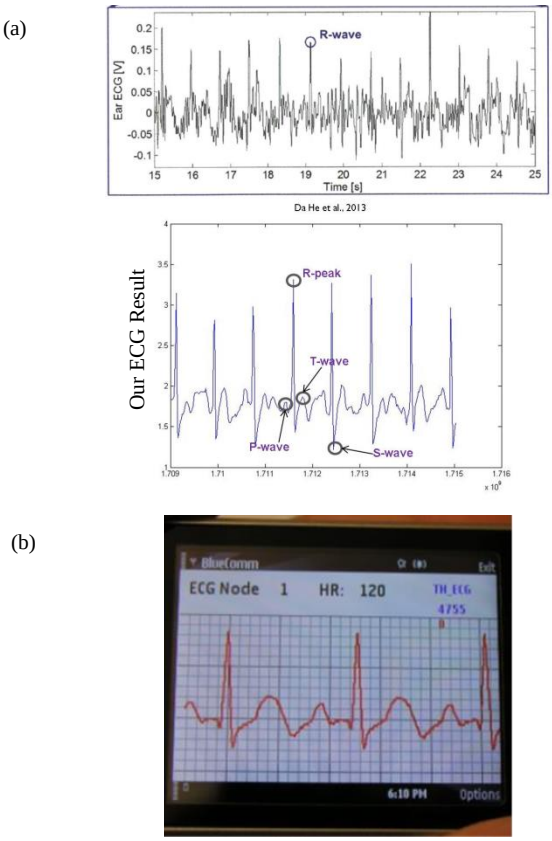


Fig. 11. (a) Compared our results with Da He's work on detecting ECG signals from the ear; (b) displaying our ECG data on the smartphone

Figure 11a shows a comparison of ECG results taken from the ear in Da He's work and ours. As can be seen, only R-peaks were detected including various noises in his results; however, our results identify PQRS-complex and T-waves with less noise. The critical point here is that Da He used two active electrodes because of the limited skin area near the ear, thus DRL (reference) electrode was omitted in his experiments.

The results reveal that the proposed gel electrode with reference electrode on the waist shows better signal quality and performance than other proposed electrodes. The table also indicates that once DRL electrodes placed on the waist, they have better SNR ratio than placed on the arm. Dry electrodes hold less signal quality than others, however, they still can be used to get ECG signal from inner ear area which is very useful for feasibility and interoperability issues. Response times of each proposed electrode shows that a similar advance with SNR ratio which means bigger size gel electrode gives a faster response.

B. CBT Experimental Results

Figure 12 shows the experimental device of CBT measurement and display of CBT results on the smartphone. Several analyses have been conducted on core body temperature during exercises to observe the changes in different environments (Figure 13). Researchers emphasize on analysis of core body temperature variations to diagnose or prevent serious diseases from heat illnesses. Heat illnesses can range from mild (e.g. heat rashes, heat cramps, etc.) to more severe health issues such as heat exhaustion and heat stroke caused by heat stress. Heat stress (hyperthermia – Temp > 37 ºC) is a condition when an individual is exposed to moderate to high temperatures from physical activity and some form of dehydration. Symptoms include blood pressure changes, increased heart rate and body temperature. The core body temperature is between 98.6°F (37°C) and 104°F (40°C). Symptoms include heavy sweating, rapid breathing, rapid yet weak heart rate and low blood pressure. The core body temperature exceeds 105°F (40.5°C), with symptoms including dizziness, lack of sweating, rapid and strong heart rate, high blood pressure [29]. Therefore, it is important to check out CBT regularly in healthcare for prevention of heat related diseases.

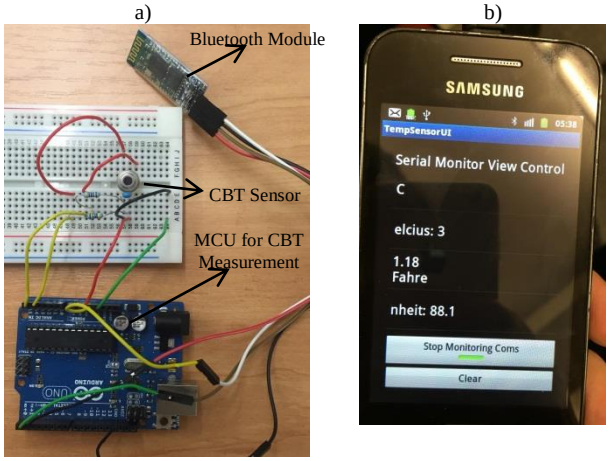


Fig. 12. (a) Experimental setup for Arduino based CBT measurement; (b) Monitoring of CBT results on the smartphone

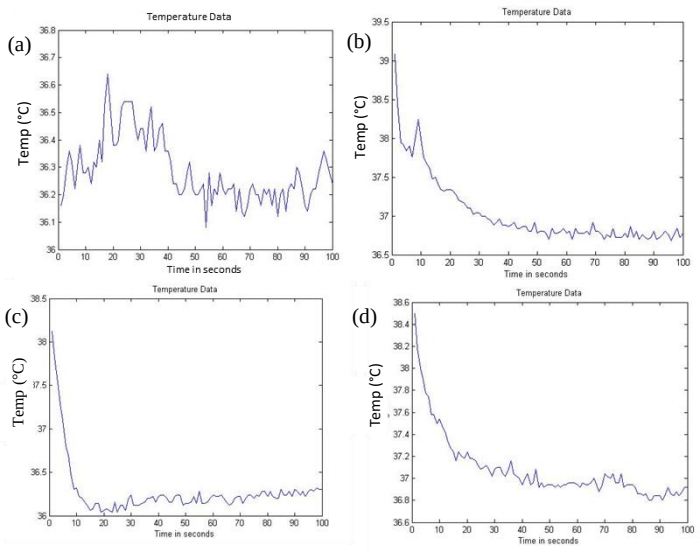
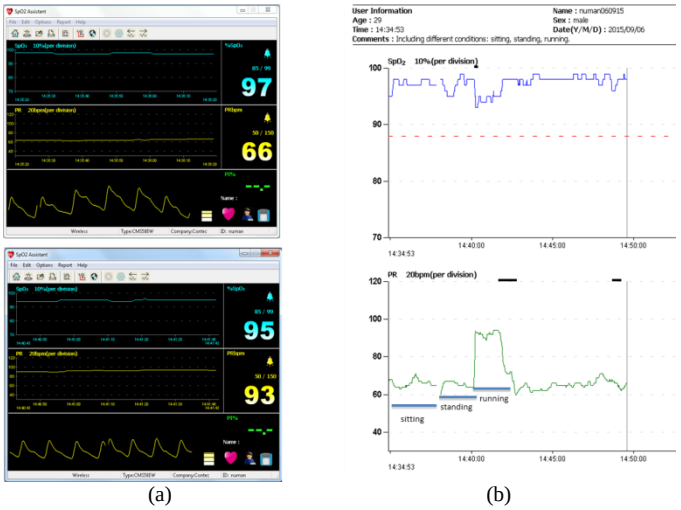


Fig. 13. (a) Raw CBT data taken from the ear; (b) CBT data just after exercise – 5 min running; (c) CBT data after walking outside in cold weather; and (d) final temperature data after sitting 5 mins with a thick jacket

Figure 13 shows four different results that were captured during the exercises in different ambient temperature values. Figure 10a indicates typical raw data of core body temperature which was taken from the ear without any exercise. It varies between 36 ºC and 36.7 ºC which is normal for a healthy person. We can certify that the core body temperature gets increased as the level of exercise gets harder, can be shown in Figure 13b. The temperature raised just over 37 ºC after 5 minutes of running exercise. Figure 13c illustrates the changes of core body temperature against the change of ambient temperature. The subject was walking outside in cold and windy weather that had around 14 ºC ambient temperature. Because of decreasing ambient temperature, the core body temperature gets decreased to nearly 36 ºC in this experiment. In Figure 13d, the core body temperature was captured when the subject was wearing a thick jacket in the office and ambient temperature was around 23 ºC. Thus, the core body temperature was elevated to almost 37.3 ºC in some cases (Figure 13d).

C. PPG Experimental Results

Figure 14a shows a person's PPG waveform, heart rate (HR) and blood oxygenation together using the finger-clipped PPG sensor, which is connected to a smart sensor unit via Bluetooth. According to the results, values of oxygen saturation in the blood were around 95%, and HR was around 93 beat per minute (bpm). Figure 14b demonstrates a PPG measurement on a 29 year-old subject with respect to SpO_2 and HR during the exercise. As can be seen from Figure 14, there were variations regarding PPG data during a 15-minute exercise including sitting, standing and running. Particularly, in the period of running, there is a sharp increasing on HR values, however, blood oxygenation values were decreasing. These data were captured and displayed on a PC. After PPG data sent to the smart sensor unit, it was also displayed on the smartphone via Bluetooth connection.



(a) (b)

Fig. 14. (a) Measuring heart rate and SpO2 using Amperor Bluetooth PPG sensor; (b) a PPG measurement on a subject during exercise using the same device

D. Proposed Sensor Integration

After digitizing and conditioning the analog signals from each sensor, the suitable digital physiological data are collected at the data acquisition hardware unit in an appropriate way. The whole data after being acquired by Arduino is converted in the form of packets and wirelessly transmitted to an Android based mobile phone [11]. Figure 15 indicates ECG electrodes placement and appearance of general prototype of performed body sensors, including a subject wearing the smart sensor on the ear.

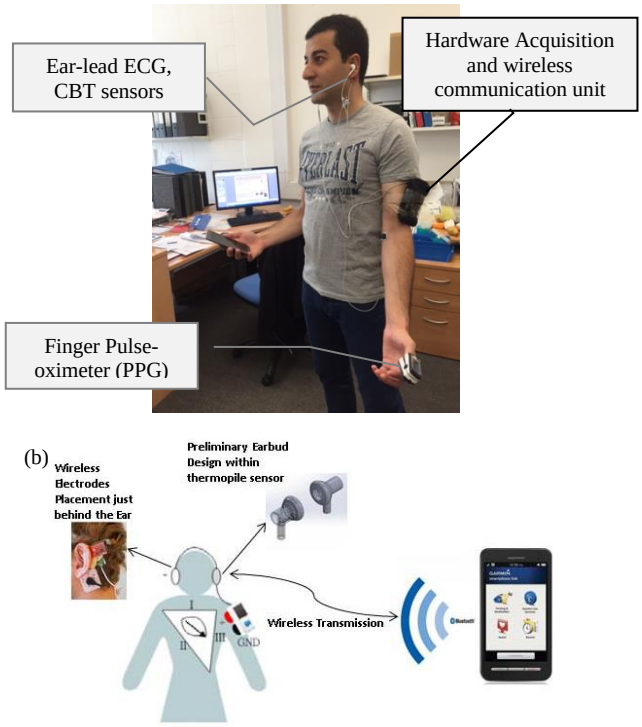


Fig. 15. (a) A typical set-up of the use of the proposed integrated wireless multiple sensors; CBT sensor was put into the earbud, and adhesive ECG sensors were attached onto the behind-the-ear and upper neck area; (b) prototype concept of Ear-lead wearable multiple sensor monitoring system

Figure 15 shows the developed experimental set-up and integration of protocol of sensors. Typically, both ECG adhesive Ag-AgCl and dry electrodes were attached to the skin, respectively, to measure ECG data. Two active electrodes were applied to near the ear (behind-the-ear, and upper neck area), and one reference electrode was placed on the arm. CBT sensor was put into the earbud design which is also shown in Figure 15(b). PPG sensor was clipped on thumb finger to measure oxygen saturation level and heart rate. After all, the whole biological data coming from these three sensors are combined together into the Arduino hardware acquisition unit (Figure 15a) in suitable data packets. In the final step, the collected physiological data are transferred via Bluetooth module to an Android based smartphone and monitored using a newly developed app.

E. Data Transmission and Bluetooth Connection

Bluetooth is a very useful technology to communicate wirelessly in short-range applications such as exchanging data from short distance fixed mobile devices. With the characteristics of synchronization with other Bluetooth devices, the wireless body area network can be successfully applied. Our propose system uses HC-05 Bluetooth module that

consists of different modes in its processing system. Figure 16 shows our Bluetooth serial module, which is paired to the smartphone.

Fig. 16. The view of Bluetooth module while pairing with the smartphone

As mentioned, there are different channels (ECG, CBT and PPG) that need to be transmitted through Bluetooth communication. The communication module has the capability to transmit data over the Bluetooth with maximum baud rate 115,200 bps. For our system, Core body temperature (CBT) measurement operates at 20 Hz sampling frequency (16-bit samples); ECG sensor works at 250 Hz frequency range (16-bit samples) with 3 electrodes and PPG operating at a 30 Hz (16-bit samples) with two photodiodes. Hence, the total data rate will comprise from aggregating each measured signal's bit rates. The total minimum bandwidth of the system will then be:

Baud Rate (BR) = bit rate of (CBT (kbps) + ECG (kbps) + PPG (kbps))

BR = (sampling frequency of CBT) * (Nyquist-criteria) * (sampling bit number) + (sampling frequency of ECG) * (Nyquist-criteria) * (sampling bit number) * (number of electrodes) + (sampling frequency of PPG) * (Nyquist-criteria) * (sampling bit number) * (number of electrodes)

BR = 20 * 2 (Nyquist-criteria) * 16 + 250 * 2 * 16 * 3 + 30 * 2 * 16 * 2 = 0.64 + 24 + 1.92

BR = 26.56 kbps.

The above baud rate is much less than the overall Bluetooth transmission baud rate, whose transmission capacity is 115.2 kbps. Thus, the residual bandwidth would be enough to perform two-way handshaking and sending the bio potentials successfully.

V. Discussion and Conclusion

Personal healthcare applications bring a growth area for wearable health monitoring systems. Wearable diagnostics and therapeutic systems contribute intelligent medical monitoring devices, which provide real-time feedback to the patients or remote monitoring servers. Wearable physiological monitoring applications, by integrating of body sensors are significantly important for patients with chronic health conditions, especially chronic neurological disorders, cardiovascular diseases and strokes that are leading causes of mortality worldwide. However, a number of ongoing research efforts target on various technical issues that need to be resolved in order to have much more sensitive, reliable, secure, and power-efficient wireless personal area network suitable in particular for mobile healthcare applications. Existing technological advances in remote monitoring systems are sometimes incapable of performing real-time patient monitoring systems because of the inability of traditional wet electrodes to perform a long term monitoring, and lack of easy-for-use design. Moreover, new designs or perspectives need to be improved in conventional Ag/AgCl electrodes for getting much more sensitive biological data due to lack of low-power microelectronics and miniaturization design in such applications. Wireless body sensors can be formed within a new technology perspective as well as the materials among mobile and wearable patient-monitoring devices to sense tiny biopotentials such as ECG and PPG from different locations on the body with very high reliability and accuracy. Otherwise, very noisy data will result using these conventional sensors, because such areas on the body (e.g. ear, upper neck) will be affected by motion artifacts.

The design and evaluation of an ear-lead multiple smart sensor system was presented in this paper. The system acquires different physiological information and continuously monitors an Android based smartphone, giving patients real-time control of data. This device includes non-intrusive sensors, specifically ECG, CBT and PPG with high accuracy. Moreover, we have also tested our Android based app, which combines the recording of all sensors together and displays ECG, CBT, and SpO_2 biological data on the smartphone of an ambulatory user. We also attempted to observe the influence of sensor positioning on signal quality using various types of ECG electrodes. Furthermore, it facilitates the difficulties of wearables giving patients significantly less restriction by eliminating the need for adapting intrusive equipment or using a laptop to see the biological data. Our results clearly demonstrate the feasibility of the concepts and interoperability of the sensors and solutions to the key technological and scientific problems. Despite of making a significant progress in addressing many of the issues, there are still considerable issues that need to be improved. Future studies will take into account the evolution of conventional body sensors and new perspectives for the improvement of the design of the particular ECG electrodes to reduce noisy data due to motion artifacts. We will further investigate the aspects of an m-Health service that the role of the smartphone can be used as a wearable physiological monitoring system including providing a real-time feedback to the patients from a central server. Therefore, the smartphones can be seen as a gateway in mobile healthcare monitoring applications. Recent advances in research are leading to this realization by involving the benefits of nanotechnology in biomedical science such as the large surface area or high electrical conductivity values of novel nanomaterials.

In this paper an ECG electrode set-up is demonstrated to pick up ECG signals from behind-the-ear in contrast to tradition chest-based ECG measurements. The results obtained are very promising and detection of the components of the ECG signal (P, Q and R) is highly possible. It was mentioned that the behind-the-ear ECG measurement is more user-friendly and gives extra convenience of using electrodes only when it is required (i.e.: no need to keep the electrodes attached all the time). A method of integrating ECG, PPG and co-body temperature using Arduino microcontroller and smart phone

for processing and displaying the data is also presented. In the future, more experiments will be performed to design better electrodes with increased signal-to-noise ratio to obtain the results similar to the traditional ECG signs and to demonstrate the integrated systems under various operating conditions.

ACKNOWLEDGMENT

This research was supported by Ministry of National Education of Turkey and the research group of DocLab in Brunel University London.

REFERENCES

[1] Bifulco, P.; Cesarelli, M.; Fratini, A.; Ruffo, M.; Pasquariello, G.; Gargiulo, G. A Wearable Device for Recording of Biopotentials and Body Movements. In Proceedings of the IEEE International Symposium on Medical Measurements and Applications, Bari, Italy, 30-31 May 2011; pp. 469-472.

[2] Socio-economic impact of mHealth – An assessment report for the European Union. Available online: http://www.gsma.com/connectedliving/wp-content/uploads/2013/06/Socio-economic_impact-of-mHealth_EU_14062013V2.pdf (accessed on 16 June 2015).

[3] Ren, Y.; Pazzi, R. W. N.; Boukerche, A. Monitoring patients via a secure and mobile healthcare system. *Wireless Communications, IEEE*, **2010**, *17*(1), 59-65.

[4] Lee, D. H.; Rabbi, A.; Choi, J.; Fazel-Rezai, R. Development of a mobile phone based e-health monitoring application. *Int. J. of Adv. Comput. Sci. and App*. **2012**, 3(3), 38-43.

[5] Nemati, E.; Deen, M. J.; Mondal, T. A wireless wearable ECG sensor for long-term applications. *Communications Magazine, IEEE*, **2012**, *50*(1), 36-43.

[6] Sanches, J. M.; Pereira, B.; Paiva, T.; Headset Bluetooth and cell phone based continuous central body temperature measurement system. In Proceedings of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, August 31 – Sept 4, 2010, pp. 2975-2978.

[7] Jung, K. H.; Tran, V.; Gabrielian, V.; Nahapetian, A. Virtual cuff: multisensory non-intrusive blood pressure monitoring. In *Proceedings of the 9th International Conference on Body Area Networks* (BodyNets '14), London, UK, 29 September – 1 October 2014; pp. 175-178.

[8] Do Valle, B. G.; Cash S. S.; Sodini C. G. Wireless behind-the-ear eeg recording device with wireless interface to a mobile device (iphone/ipod touch). In Proceedings of 36th IEEE Engineering in Medicine and Biology Society (EMBS), Chicago, US, 26-30 August 2014; pp. 5952-5955.

[9] Song, W.; Yu, H.; Liang, C.; Wang, Q.; Shi, Y.; Body monitoring system design based on android smartphone. In Proceedings of 2012 World Congress on Information and Communication Technologies WICT 2012, Trivandrum, India, 30 October – 2 November 2012; pp. 1147-1151.

[10] Boano, C. A.; Lasagni, M.; Romer, K.; Non-Invasive measurement of core body temperature in marathon runners. In Proceedings of the Body Sensor Networks 2013 (BSN '13), Cambridge, US, 6-9 May 2013; pp. 1-6.

[11] Shen, T. W.; Hsiao, T.; Liu, Y. T.; He, T. Y. An ear-lead ECG based smart sensor system with voice biofeedback for daily activity monitoring. In TENCON 2008-2008 IEEE Region 10 Conference, Hyderabad, India, 19-21 November 2008; pp. 1-6.

[12] Hernandez, J.; McDuff, D. J.; Picard, R. W. BioPhone: Physiology Monitoring from Peripheral Smartphone Motions. In Proceedings of 37th IEEE Engineering in Medicine and Biology Society (EMBS), Milano, Italy, 25-29 August 2015.

[13] Hii, P. C.; Chung, W. Y. A comprehensive ubiquitous healthcare solution on an Android™ mobile device. *Sensors*, **2011**, *11*(7), 6799-6815.

[14] Tranquillo, J. V. (2013). Biomedical signals and systems. *Synthesis Lectures on Biomedical Engineering*, *8*(3), 1-233.

[15] Wahl, F.; Freund, M.; Amft, O. WISEglass: Smart eyeglasses recognising context. In *Proceedings of the 10th International Conference on Body Area Networks* (BodyNets '15), Sydney, Australia, 28-30 September 2015.

[16] Poh, M. Z.; Kim, K.; Goessling, A.; Swenson, N.; Picard, R. Cardiovascular monitoring using earphones and a mobile device. *IEEE Pervasive Computing*, **2012**, (4), 18-26.

[17] Morón, M. J.; Luque, R.; Casilari, E. On the capability of smartphones to perform as communication gateways in medical wireless personal area networks. *Sensors*, **2014**, *14*(1), 575-594.

[18] Lee, B. G.; Lee, B. L; Chung, W. Y. Smartwatch-based driver alertness monitoring with wearable motion and physiological sensor. In Proceedings of 37th IEEE Engineering in Medicine and Biology Society (EMBS), Milano, Italy, 25-29 August 2015.

[19] J. Malmivuo and R. Plonsey. Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields . Oxford University Press, New York, 1995.

[20] Lin, Y. H.; Lin, C. F.; You, H. Z. A driver's physiological monitoring system based on a wearable PPG sensor and a smartphone. *Security-Enriched Urban Computing and Smart Grid*, **2011**, 223, 326-335.

[21] ArduinoUNO. Available online: https://www.arduino.cc/en/Main/ArduinoBoardUno (accessed on 16 May 2015).

[22] Infrared Thermometer – MLX90614. Available online: https://www.sparkfun.com/products/9570 (accessed on 18 May 2015).

[23] ECG Acquisition System. Available online: https://www.cooking-hacks.com/electrocardiogram-sensor-ecg-ehealth-medical (accessed on 12 May 2015).

[24] Pulse Oximeter Module. Available online: http://www.contecmed.com/ (accessed on 19 May 2015).

[25] HC Bluetooth Transceiver Module. Available online: http://www.gearbest.com/ (accessed on 21 May 2015).

[26] De Luna, A. B.; Batchvarov, V. N.; Malik, M. The Morphology of the Electrocardiogram. 2006. Available online: https://www.blackwellpublishing.com/content/BPL_Images/Content_store/Sample_chapter/9781405126953/9781405126953_4_001.pdf (accessed on 10 June 2015).

[27] He, D. D.; Winokur, E. S.; Sodini, C. G. An Ear-Worn Vital Signs Monitor. *Biomedical Engineering, IEEE Transactions*, **2015**, *62(11)*, 2547-2552.

[28] Bock, M.; Hohlfeld, U.; Von Engeln, K.; Meier, P. A.; Motsch, J.; Tasman, A. J. The accuracy of a new infrared ear thermometer in patients undergoing cardiac surgery. *Canadian Journal of Anesthesia*, **2005**, *52*(10), 1083-1087.

[29] Heat Illness: MedlinePlus. Available online: https://www.nlm.nih.gov/medlineplus/heatillness.html (accessed on 22 June 2015).

An Evaluation of Requirement Prioritization Techniques with ANP

Javed ali Khan
Department of Software Engineering
University of Science & Technology Bannu
Bannu, Pakistan

Izaz-ur-Rehman
Department of Software Engineering
University of Science & Technology Bannu
Bannu, Pakistan

Shah Poor Khan
Department of Software Engineering
University of Science & Technology Bannu
Bannu, Pakistan

Wasif Afzal
Department of Software Engineering
Bahria University
Islamabad, Pakistan

Iqbal Qasim
Department of Computer Science
University of Science and Technology Bannu
Bannu, Pakistan

Yawar Hayat Khan
Department of Software Engineering
Bahria University
Islamabad, Pakistan

***Abstract*—This article elaborates an evaluation of seven software requirements prioritization methods (ANP, binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort). Based on the case study of local project (automation of Mobilink franchise system), the experiment is conducted by students in the Requirement Engineering course in the department of Software Engineering at the University of Science and Technology Bannu, Khyber Pakhtunkhawa, Pakistan. Parameters/ measures used for the experiment are consistency indication, scale of measurement, interdependence, required number of decisions, total time consumption, time consumption per decision, ease of use, reliability of results and fault tolerance; on which requirements prioritization techniques are evaluated. The results of experiment show that ANP is the most successful prioritization methodology among all the available prioritization methodologies.**

Keywords—Requirement Engineering; Requirement prioritization; ANP; AHP; Software Engineering; Comparison

I. INTRODUCTION

While developing a software project, developers often face a situation where decision among several options has to be taken. Normally Software projects are composed of many requirements [1]. Requirement prioritization is an important and continuous process throughout software development.[20]

Usually all the requirements are not important for the user. Therefore there is a need of prioritizing the requirements according to the limited resources regarding time, budget, and to satisfy client regarding quality. Also developers' team does not have the information that which requirements are in the interest of users. When there is a single stakeholder it is easy to identify important and less important requirements but when the number of stakeholders becomes more than one then it becomes difficult to take decision in the development of the project, because different stakeholders have different views regarding the requirements. For software project development, software requirements prioritization is considered one of most urgent decision making process [2]. When developing a software, often comes a situation where decision among different options are made [21]. The software projects still have low success rates these days. Nowadays success rate of software projects is at the lower side. Software project goes into failure due to lesser customer involvement in software project, inadequate resources, unrealistic outcomes, dynamic software requirements and requirements specification [16]. Requirements prioritization helps by increasing user interaction with the system as it allows the stakeholder to specify such requirements which are of greater interest for customer. Requirements prioritization also helps to remove the disagreement among several stakeholders. Resources are assigned to requirements based on their priorities; it becomes possible due to requirement prioritization [17]. Requirements prioritization helps to know the problems in the requirements like misunderstanding or ambiguity among requirements so that future problems are prevented in advance in order to save resources in terms of cost and time [5]. With help of requirements prioritization conflicts amongst different stakeholders can be resolved. [22]

Hatton says that prioritizing requirements is very important when developing software product, which helps in minimizing project failure rate [8]. Shah nazir [18] used ANP to prioritized component selection using quality attributes and produce better results.

FBI Virtual Case File (VCF) project is a case study of huge software product. Project was completed in estimated cost of 170$ Million [3]. C.Huang and Mobasher performed detailed examination of Virtual Case File project and came to conclusion that project failed because of requirements mismanagement and not conducting prioritization of requirements [4]. Akmli [19] used ANP framework which is largely used MCDM approach for prioritization of quality and environmental criteria in generic case.

Nowadays software construction has become very fast. As many alternate options are easily available in software industry therefore it is necessary to complete the project

within assigned cost and time. For this purpose a requirements prioritization methodology must be used which is simple to utilize, simple to know and provides consistent & efficient results. This methodology should be able to prioritize interdependent requirements. The AHP and analytic network process (ANP) are two analytical approaches for Requirements prioritization. The AHP is applied to break down large unstructured decision problems into controllable and measurable modules. The ANP, as the general form of AHP, is powerful to deal with complex decisions where interdependence exists in a decision model. Despite the increasing number of applications of AHP in different fields that entail decision making, ANP has started to be engaged in Requirements prioritization in software engineering fields.

Still the field of requirements' prioritization using ANP is lacking in quality research papers.

This paper presents detailed assessment of seven requirements prioritization techniques which are: analytic network process (ANP), analytic hierarchy process (AHP), hierarchy AHP, spanning tree matrix, bubble sort, binary search tree and priority groups. In order to understand each prioritization technique, each technique is applied to prioritized Mobilink Franchise system.

These prioritization techniques are then evaluated against pre-defined criteria, which are taken from literature and software experts like ease of use, required completion time, reliability of results and measuring inter-dependency of requirements. ANP is found to be the most promising and reliable technique amongst all the prioritizing techniques despite of the fact that ANP takes greater time to complete prioritization process.

II. Requirments Prioritization Methods

In this section prioritization techniques are elaborated in detail, explaining how requirements are prioritized using each prioritization methodology in order to know limitation of each methodology.

A. Priority Groups

A number of studies mention the numerical assignment techniques such as [5],[6],[7],[8],[9],[10]. It is a basic requirements prioritization technique in which different prioritization groups are made and then requirements are mapped into these priority groups. Several prioritization groups may be varied but certain groups are same. For example the common groups are low, medium, high. When requirements are plotted to the specified requirement prioritization groups, then requirements inside each priority group have same priority.

B. Bubble Sort

Elements can be sorted using bubble sort technique. Bubble sort was mentioned by Hopcroft, Aho and Ullman [12]. Karlsson [5] first of all introduced bubble sort for the software requirements prioritization. In bubble sort requirements prioritization, the user takes first of all two requirements and compares these two requirements. If the two requirements are out of order then we take another requirement and compare it with the first requirement, this process continues until all the requirements are in order. The most important requirement is at the top and the less important is at the bottom.

C. Binary Search Tree

Hopcroft, Aho and Ullman [12] proposed another technique of binary search tree which is used for the sorting. In binary search tree all the nodes have at most two children. Binary search tree was first time introduced to requirements prioritization by Karlsson [5]. In binary search tree, each node shows a requirement. Less important requirements are placed to the left side of the node and more critical requirements are placed to the right side of the node of the binary search tree. In binary search tree requirements prioritization is done in the following way. Take one requirement and place it as a root node now take another requirement and compare it to the root node if that requirement is less significant than the root node, then compare it to the left child node of the root node, if that requirement is of greater significance than the base node, then compare it to the right side child of the root node. If the base node does have any child nodes then put that requirement as a new child of the root node. If the requirement have greater priority than root node, put that requirement as a child of root node on right side and if it is of less importance than the root node, put that requirement to left side node as a new child of the root node. This process is repeated until all the requirements are adjusted and placed in the binary search tree.

D. Analytic Hierarchy Process (AHP)

Analytic hierarchy process is a well-known requirements prioritization technique. Analytic hierarchy process was proposed by Saaty[13]. In AHP, first of all requirements are identified then criteria are identified in order to prioritize requirements against them. The Possible hierarchy made in AHP is pairwise comparison to each other. Relationship amongst hierarchies is identified. User will assign importance on the scale which is from 1 to 9. The scale is shown in the Fig 1. Now AHP changes the customer consideration to numeric values and numeric values are assigned to each element in the hierarchy. Redundancy might take place when prioritizing requirements with AHP, therefore consistency ratio must exists in order to know that legitimate prioritization has been achieved. AHP not only prioritizes requirements but also gives the knowledge that to what degree they are more prior. If there are n requirements to be compared by AHP then the number of pairwise comparisons will be n (n-1)/2.

E. Hierarchy AHP

The most abstract level software requirements are located at the top of the hierarchy and the more precise level requirements are located at the bottom of the hierarchy. Karlsson introduced hierarchy AHP to prioritize requirements which are placed at the same level [5]. In this technique, all unique pairs of requirements are placed at the same level. Now all requirements are not pair wise compared to each other, only those are compared which are placed at the same level.

Intensity	DEFINITION	Explanation
1	of equal value	Two requirements are of equal value
2	Slightly more value	Experience slightly favors one requirement over another
5	Essential or strong value	Experience strongly favors one requirement over another
7	Very strong value	A requirement is strongly favored and its dominance is demonstrated in practice
9	Extreme value	The evidence favoring one over another is of the highest possible order of affirmation
2,4,6,8	Intermediate values between adjacent judgments	When compromise is needed

Fig. 1. Scale used for the pairwise comparison in AHP[13]

F. Minimal Spanning Tree

This is another technique used to prioritize requirements which is proposed by Karlsson [5]. In minimal spanning tree prioritization method the idea is that, if the decision making is made absolutely constant, then the redundancy can be overcome. For example if requirement 1is known to be of greater priority than requirement 2 and requirement 2 is of greater priority than requirement 3, then requirement 1 must be of greater priority than requirement 3 but AHP allows the user perform pairwise comparison also, which is already done and hence increases the redundancy.

G. Analytic Network Process (ANP)

The ANP is "a multi criterion theory of measurement used to obtain relative priority scales of absolute numbers from individual judgments that also belong to fundamental scale of absolute numbers" [15]. The judgments show the comparative dependence of one or two elements in the network or cluster in a pair wise comparison method over the other element in the system, with respect to certain control criterion. In ANP, pair wise comparisons of each element in each level in the network are performed with admiration to their relative significance towards the control criterion. When in the network all the pair wise comparisons are finished, the vectors related to the highest Eigen values of the constructed matrices are computed and a priority vector is obtained. The wanted elements priority values are calculated by normalizing these vectors values. The super matrix is constructed from the output derived from the comparison method, where super matrix is contained of the collection of the matrices of column priorities.

ANP provides a common structure to deal with decision problems and to select a decision from a group of decisions. Major dissimilarity between ANP and AHP is that in AHP the elements are in a hierarchy, one cannot calculate its dependency on the criteria and on the same elements in the hierarchy. In AHP all the elements are independent. [15]. ANP is the broader form of AHP, its main similarity to AHP lies in the fundamental theory: both techniques have the idea of relative significance of influence as a major concept.ANP technique uses the same basic scale of the AHP for the measurement with the additional facility to answer two kind of questions as: for given criterion, which of the two elements have more dependence?, or for given criterion, which of the two elements have greater dependency? [15] The fundamental scale by both the techniques is depicted in the Fig.1.

III. PERFORMING EXPERIMENT

A. Goal Definition

The experiment was motivated by the need of determining the difference between the performance of requirements prioritization using ANP and other prioritization techniques; as ANP provides additional facility to prioritize the requirements which are interdependent.

Objective of the study: The objective of the study is the requirements prioritization through ANP and comparing the performance of ANP with other requirements prioritization techniques (binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort).

Purpose: The purpose of the experiment is to evaluate the performance of ANP on the basis of interdependent software requirements. This experiment also provides an overview to the requirement experts or requirement engineers and stakeholders about how to prioritize requirements techniques.

From researchers' perspectives: The researchers would like to know about the new requirements prioritization techniques and would look for more research in this area as to improve the performance of ANP while prioritizing requirements.

Quality Focus: The main effect studied in this experiment is the determination of priorities of interdependent requirements and the performance of ANP while prioritizing software requirements as compared to other requirements prioritization techniques on certain parameters/measurements.

1) Inherent Measures

Two inherent properties of the requirements prioritizing methods were identified:

Consistency indication: This property shows whether the requirements prioritization techniques are able to show consistency in the judgment of decision makers. This property needs redundancy in the decision making.

Scale of measurement: This property explains the scale which is used to obtain the final priorities of the requirements. Scaling the requirements is an important characteristic; through which we can get actual values and rank the requirements. The more powerful the scale the more reliable and accurate will be the result. The four (4) methods of scale which are nominal, ordinal, interval and ratio scales are used.

2) Objective Measures

Below objective measures were discussed while performing the comparison.

Required number of Decisions: For Analytic Network Process (ANP), AHP, Hierarchy AHP, spanning tree and bubble sort the number of decisions are already known, but for the binary search and priority groups, the number of decisions depend upon how the participants perform that session. This

measure shows how many comparisons are needed by the decision makers to solve the problem.

*Total time consumption***:** This measures the total time needed by the decision maker to complete the overall steps in prioritizing requirements. Total time consumption measure is different from the required number of decision measures, as each requirements prioritization method has different way to complete the comparison and therefore take different time.

Time consumption per decision: In time consumption per decision we note the time taken per decision.

3) Subjective measures

In order to get good understanding, the requirements prioritization techniques are compared with respect to usability. Scale from 1 to 6 is used to rank the requirements. If 1 is assigned to any requirement prioritization technique then it is considered as one of the best methods. The following features are considered to be judged.

Ease of Use: This measure shows how easily a specific requirements prioritization method can be used to prioritize the requirements.

*Reliability of Results***:** Reliability of results shows that how reliable the results are after judgment.

Fault Tolerance: Fault Tolerance shows how good a requirements prioritization method is to judge error while prioritizing requirements.

Context: The experiment is performed in the context of requirements prioritization. Moreover, the experiment is conducted in the Requirement Engineering course in the department of Software Engineering at the University of Science and Technology Bannu, Khyber Pakhtunkhawa, Pakistan. The experiment was performed on the basis of Karlsson et al [5] experiment. In Karlsson et al [5] experiment six (6) requirements prioritization techniques are compared on the basis of above mentioned parameters/measures. Karlsson et al [5] experiment was the base for this experiment. In this experiment a new proposed requirements prioritization technique is included along with the existing prioritization techniques, therefore the seven (7) requirements prioritization methods are compared with newly proposed prioritization technique. The comparison was done using the same parameters which were used in the Karlsson et al [5] experiment and were explained above. An additional parameter was included, to know that which requirements prioritization technique prioritize interdependent requirements. It was the main purpose of the experiment. For this purpose a quiz was taken from the students of software engineering in subject of requirement engineering of sixth (6^{th}) semester, while the topic was requirement prioritization. Among them top seven (7) students were selected for the experiment on the basis of their obtained highest marks in the quiz and their interest towards research.

4) Summary of scoping

Analyze the performance of ANP while prioritizing requirements.

For the purpose of comparison with other requirements prioritization techniques

With respect to evaluate interdependent requirements support

From the view point of researchers and industry.

In the context of requirements prioritization.

B. Planning

1) Context selection

The context of the experiment is requirements prioritization which is one of the topics of Requirement Engineering course studied at the Institute of Engineering and Computing Science, UST Bannu Khyberpakhtunkhawa, Pakistan, hence the experiment was run offline(not in the software industry). The experiment was conducted by graduate students of the Software Engineering who had taken Requirement Engineering as a subject in the 6^{th} semester. The experiment is specific, since it is focused on the requirements prioritization in an educational environment. The experiment shows the real problem; prioritization of interdependent requirements prioritization.

This experiment will provide good opportunities to other researchers to consider it in their research as it is well defined and it can help the requirements engineers/stakeholders how they can prioritize both dependent and independent requirements.

2) Hypothesis Formulation

An important part of the experiment is to understand and formally state what will be evaluated in the experiment. This goes to formulation of hypothesis/hypotheses. Below are the hypotheses chosen for the experiment.

a) Both dependent and independent requirements are prioritized, therefore it is expected that ANP will prioritize both dependent and independent requirements while other prioritization techniques will prioritize only independent.

b) The performance of ANP while prioritizing software requirements are expected to produce better prioritization results as compared to other prioritization techniques.

c) ANP is expected to produce more reliable results than that of the other prioritizing techniques.

d) While prioritizing the dependent or independent requirements by ANP, it is expected that there will be less chances of errors as compared to other prioritization techniques.

e) Required number of decisions to complete the prioritization process by ANP may be greater than that of the other prioritization techniques but Total time Consumption and Time Consumption per decision are expected to be less than that of AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

Hypotheses are formally stated and defined as below

a) Null hypothesis, $\mathbf{H_0}$: ANP and other prioritization methods, as binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort, prioritize both dependent and independent requirements.

H_0: ANP, binary search tree, AHP, hierarchy AHP, spanning tree matrix, Priority group and bubble sort prioritize both dependent and independent requirements.

Alternative hypothesis **H_1**: ANP prioritize both dependent and independent requirements while binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort prioritize only independent requirements.

b) Null hypothesis, **H_0**: There is no difference in the performance while prioritizing requirements by ANP, binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

H_0: Performance (ANP) = Performance (binary search tree), Performance (AHP) Performance (hierarchy AHP), Performance (spanning tree matrix), Performance (priority group) and Performance (bubble sort).

Alternative hypothesis **H_1**: Performance (ANP) ≠ Performance (binary search tree), Performance (AHP) Performance (hierarchy AHP), Performance (spanning tree matrix), Performance (priority group) and Performance (bubble sort).

c) Null hypothesis, **H_0**: There is no difference in the reliability of the results obtained from ANP, binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

H_0: Reliability (ANP) = Reliability (binary search tree), Reliability (AHP), Reliability (hierarchy AHP), Reliability (spanning tree matrix), Reliability (priority group) and Reliability (bubble sort).

Alternative hypothesis **H_1**: Reliability (ANP) ≠ Reliability (binary search tree), Reliability (AHP), Reliability (hierarchy AHP), Reliability (spanning tree matrix), Reliability (priority group) and Reliability (bubble sort).

d) Null hypothesis, **H_0**: There is less chance of error when prioritizing requirements by ANP, binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

H_0 : Errors, prioritizing requirements by ANP = Errors, prioritizing requirements by binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

Alternative hypothesis **H_1**: Errors, prioritizing requirements by ANP ≠ Errors, prioritizing requirements by binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

e) Null hypothesis, **H_0**: There is no difference in the total time consumption and time consumption per decision by ANP, AHP, hierarchy AHP, binary search tree, spanning tree matrix, priority group and bubble sort.

H_0 : Total time consumption and time consumption per decision by ANP = total time consumption and time consumption per decision by binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

Alternative hypothesis **H_1**: Total time consumption and time consumption per decision by ANP ≠ total time consumption and time consumption per decision by binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort.

3) Variable Selection

The independent variables are the requirements prioritization techniques. The dependent variables are performance, errors occurrence and reliable results

4) Selection of Subject

Subject is the students, taking part in experiment, some of them are selected as sample on the basis of performance in the quiz.

5) Experiment Design

The problem has been defined and the dependent and independent variables are identified. Thus now an experiment can be designed as below.

Randomization: A lecture was given on requirements prioritization to all the subjects of sixth (6th) semester of software engineering in the course of requirement engineering. After that a quiz was taken in the subject. Then those subjects having highest scores in the quiz were selected as participants for the experiment. The object was assigned randomly to subjects. Subjects selected for the experiment were selecting randomly as they were representing the whole class. The subjects were given an introduction also on the case study used for the experiment.

*Blocking***:** No order technique of blocking is applied. Seven (7) students participated in the experiment; all the samples from the participants were considered in the evaluations after prioritizing the requirements through requirements prioritizations techniques. No samples from the participants were blocked. Then all the results collected from the participants were analyzed to produce a generalized result.

Balancing: It would be better to have a balanced data set. But the experimental study is based on a topic of a subject for which participants get registered, therefore it was impossible to know the background of participants and to balance the data set.

6) Instrumentation

The background of the participants in requirements prioritization was found by taking quiz in the beginning. This data provides help in selection of the top participants for the experiment.

7) Validity Evaluation

Validity threats, having four levels, are considered for the experiment [14]. Internal validity is mainly concerned with the validity of actual study [14]. External validity is focused on the participants who are taking part in the experiment, their background related to requirements prioritization and requirement engineering in general. The conclusion validity is mainly related to the correspondence between the solutions and the results. Construct validity is about giving an overview of results of an experiment followed by theory.

The internal validity inside the course of requirement engineering may not be the problem, greater number of tests (equal to the number of participants in the experiment) make sure the internal validity is good.

While the external threats must ensure that similar results must be obtained when the same experiment is performed by other participants in the same course of requirement engineering. It is harder to generalize the results for other experiments because the students having no background of software engineering will not give good results. As if students from computer science are included as participants in the experiment then there will be a difference in the results. The results from the analysis of experiment can further be generalized to other experiments where the background of the participants is measured in terms of software engineering and computer science.

The main problem to conclusion validity is that how much quality data is collected for the experiment of requirements prioritization? We are comparing ANP with Six requirements prioritization techniques therefore specific data should be gathered to perform better experiment. The incorrect data does not belong to any specific background, therefore the problem is not related to background of participants. Thus conclusion validity is not considered to be that much critical [14].

The threats in construct validity are, that the measures that we have selected for the experiment are good enough to evaluate the requirements prioritization methods. For example whether the number of comparisons for prioritization methods are enough for the evaluation.

The results from the evaluation of prioritization of requirements techniques are likely to be used for other experiments in the area of software engineering where the backgrounds of the participants are considered from software engineering and computer science.

C. Operation

1) Preparation

To all subjects (participants) lecture was delivered on requirements prioritization techniques as ANP, binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort, were included in the experiment. Helping materials related to requirements prioritization techniques were handed over to subjects. The case study of Mobilink franchise computerization was used. Mobilink franchise system was described to all subjects. It was easily understood by all the participants because all subjects were familiar with Mobilink franchise computerization system. This is a local project, therefore the clients are easily available to discuss requirement. Project description is given below.

Introduction to Mobilink : Mobilink is Pakistan's leading cellular and Blackberry service provider. With more than 35 million subscribers, Mobilink maintains market leadership through cutting-edge, integrated technology, the strongest brands and the largest portfolio of value added services in the industry. It is a broadband carrier division providing next generation internet technology as well as the country's largest voice and data network with over 8,500 cell sites. Mobilink offers both postpaid (Indigo) and prepaid (JAZZ and JAZBA) solutions to the customers. Compared to competitors, both the postpaid (Indigo) and prepaid (JAZZ) brands are the largest brands of their kind in the Pakistan cellular industry.

*Franchise***:** Franchise is customer dealing office in which Mobilink representatives deal with their clients. Mobilink has many franchises in Pakistan; in every franchise there is a stock of subscriber identity module (SIM), scratch cards and OTTAR, which is a business name for easy load. Franchise also provides facility of blocking, renewal and purchasing SIMS.

Summary of Proposed System: Current system of the franchise was manually operated system. Owner needs to automate the manual system in order to make the system more dynamic. Current system consists five main modules which are jazz CDs or SIMs stock, jazz load or OTTAR, jazz cards, cash incoming and expenses.

The proposed system is related to franchise business as there is a stock of Jazz CDs (SIMs) in franchise. Every franchise has salesmen (DOs) which sell the SIMs to customers directly and to other mobile shops. Different types of SIMs are available according to the prices. DOs have specific percentage on the selling of SIMs. System also provides facility of replacing the old SIM with new SIM and blocking any SIM to Mobilink users.

Another module of the system is about easy load. First of all franchise demands balance from the company to their master SIM then DOs sell the load to different mobile shops and to customers directly. DOs have specific commission on selling the load.

There is a stock of jazz recharge cards in franchise. Again the DOs sell the cards to customers and to other mobile shops. On selling the cards a defined commission is given to DOs.

There is a module of cash incoming that is how much sell is done in a single month? , what is the total percentage of each DOs, bank transaction with the company and how much profit done by the company in a specific month? How much sale was done of cards, SIMs and easy load each month?

Lastly there is an expense module, where the calculations of all expenses in a month, like electricity bills, telephone bills, guest expenses and miscellaneous expenses are carried out. The net total income is calculate the by separating expenses from total sell. Daily and monthly sell reports are prepared.

Detailed Explanation of Project Business Requirements

MAIN UNITS

The manual Franchise system comprises the following five units…

- JAZZ CDs/SIMs
- JAZZ LOAD/OTTAR
- JAZZ CARD
- CASH INCOMING
- EXPENSES

a) Jazz CD/SIM

- It means the number of SIMs that are available i.e. the opening STOCK which is demanded by the Shah Noor company from the FRANCHISE.
- Basically the Salesmen (DOs) distribute these SIMs/CDs.
- There are different kinds of SIMs/CDs with respect to working and Rates e.g. JAZZ AWAMI, JAZZ RETAILER, JAZZ O301, 0300, GOLDEN CHARGES, PLATINUM CHARGES etc.
- The Peak Prices of SIM/CD, when the company demands them and the DOs sale them, are Rs.130 and Rs.180, these prices are dynamic because the sale can be reduced or can be increased. The peak price of CDs depends on the type of CDs i.e. JAZZ AWAMI, JAZZ RETAILER, JAZZ O301, 0300, GOLDEN CHARGES, PLATINUM CHARGES etc. and also sailing price may be different.
- The Commission rate on the basis of sailing SIM/CD is 5% for the DOs. It is also a dynamic rate because some DOs are on fixed pay and some are not.
- Therefore the profit and loss are calculated at the end of month.

b) Jazz load/OTTAR

- The available number of salesman/DOs is 1-6, keeping this entity also dynamic.
- Basically the company transfers Rupees to the Master SIM then given to DO'S to work on it.
- Following points are important in JAZZLOAD.
- How many loads are issued to the DOs?
- How much is sold?
- How many are remaining with the DOs.
- This remaining amount in the form of jazz load is the opening stock for the next day.
- The criteria of profit on RS.100 is 4%, if the load is RS.97 then the profit is 3% and on the 97% the profit is 1%.
- 4% is divided such that DOs get only 3% and the remaining is reserved for the manager or head of JAZZLOAD.
- Therefore the profit and loss are calculated at the end of month.

c) Jazz cards

- The JazzCards are also called the Scratched cards.
- At the start of the opening stock the number of available cards is noted.
- Cards of Rs. 100, 300,600 & 1000 are issued to DOs and the date is noted by DOs.
- How many cards are issued?
- How many of them are sold?
- How many cards are remaining?
- The peak price offered by the company is Rs. 96.75.
- The DOs sale them at Rs. 97, at profit 0.25%.
- In this profit, the DOs commission is 0.25% and the 0.5% is for company.
- If DOs sale them at Rs.96.75 (peak price) then 0.25% loss occurs to company but actually this is not the loss because it is recovered by the DOs.
- If it is not sold on Rs.96.75 which is the peak price then there is no gain and no loss in the case.
- The profit and loss are calculated at the end of month.

d) Cah incoming

It includes the following requirements.

- It is the total amount at the end of the month. It includes the Deposited date, Amount, cheque number, bank name and cheque date.
- The original amount + Profit are added together.
- Income of the company, DOs Total and all the commission records should be stored in a proper manner, so that it can be maintained at the end of every month.
- How much loss and net income are occurred?
- The record of the JazzCD, JazzLoad and Expenses is also kept.

e) Expenses

It includes the following requirements.

- Utility bills.
- Maintenance.
- Others (Guests).
- Rent of the building.
- Staff's salary.
- Calculation of profit and loss at the end of month.

To develop quality software, non-functional requirements are also considered. After discussion with customer, interest of the customer in non-functional requirements, Cost, performance, Quality, Reusability, Usability and Security, is noted. In order to develop a quality product, high budget is needed but our client has limited resources therefore within the limited budget a quality software has to be developed. Client needs high performance software with great reusability. Client is willing to upgrade the product in near future, therefore project must be developed with very reusable design and techniques. User demands for easy to use and operate software and with attractive interface.

Dependency among the Requirements

For scratch cards: The customers must have SIM then he/she can load different types of cards to his/her SIM e.g. of Rs.100, 300 or 1000 etc where cards will be available in franchise or with dealers. Franchiser will get cards from company.

For load (jazz Load): The customers must have SIM. First of all manager will get balance from company in his master SIM then it is transferred to the salesman (DO) SIM. After that DO transfers balance to shopkeepers (dealers) and dealer transfers balance to the customer SIM.

Easy Load: If loading cards are not available then the easy load is very beneficial in this case. If a customer wants to load only 40, 50 rupees then in this case the loading cards are not useful because mobile cards are available in multiples of 100 rupees.

Easy Load to Customer: Loading balance to customers through easy load depends on balance in master SIM balance. If there is no load in Master SIM then they can't do load to customers.

Master SIM: Master SIM of franchise depends upon Company as it is loaded from company, if there is no balance in master SIM then it can't load customer SIM.

Opening Stock: As for each new day there should be an opening stock if it is not available the owner of the franchise system should demand it from company and if the opening stock is available it should be added to the opening stock for next day.

Daily Summary: Daily summary depends upon selling of SIMs, Load or Cards.

Monthly Sale Report: Monthly Sale report depends upon everyday sale of SIMs, Load or Cards.

Salesman Salary: Salesman salary depends upon sell. The more salesman sales SIM, Cards, Easy Load the more salesman get the salary.

Net Profit: Net Profit for the Franchise in any month depends upon the total sale and expanses made throughout the month. Therefore after subtracting expanses from total sale the Net Profit can be calculated.

The non-functional requirements are interdependent and also dependent on functional requirements. Non-functional attributes can be applied to a single requirement as well as to a whole project.

The cost increases with the production or development of a quality product.

To ensure best security the cost will increase.

Reusability can affect performance.

Achieving usability will increase cost.

Glossary:

DO: DO is the business name of salesman.

OTTAR: OTTAR is the business name of loading balance from dealer SIM to the customer's SIM like easy load.

CD: CD is the business name of SIM (subscriber identification module)

Hypothesis of experiment were introduced to subjects during the lecture. Subjects were aware of the hypothesis of experiment.

Karlsson et al [5] experiment was distributed amongst the participants. This article was based on an experiment.

2) Execution

Experiment took two (2) weeks to complete. Karlsson et al [5] article was distributed among the subjects. After that an introductory lecture was given to participants on Karlsson et al [5] article as well as introduction of hypothesis to subjects. Karlsson et al [5] has compared six (6) requirements prioritization methods (binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort) in their experiment, participants were introduced to these six (6) requirements prioritization methods. Each participants took a week to read and understand the Karlsson et al [5] article. After one week all the participants were gathered in a class room and a lecture was delivered on all seven (7) requirements prioritization techniques (binary search tree, AHP, hierarchy AHP, spanning tree matrix, Priority group and bubble sort) including the new proposed requirements prioritization method ANP. On each prioritization method an example was carried out so that participants can easily understand the process of requirements prioritization.

Now participants knew hypothesis, had understood requirements prioritization methods and requirements were also clear to them.

Three (3) hours' time was given to participants to perform the experiment. All the participants returned the results according to hypothesis after performing the experiment

3) Data Validation

After performing the experiment, data was collected from all the seven participants. Data collected from participants was evaluated and analyzed. All the data collected, was considered for the analysis and evaluation, no data was dropped. Before going for actual experiment a test was taken from 30 students in the subject of requirement engineering. The topic to cover in the quiz was requirement prioritization. After collecting the quiz from the students, it was checked and evaluated and top seven students were selected as participants for the experiment on the basis of highest marks and their interest in the experiment.

D. Analysis and Interpretation

1) Descriptive Statistics

Industry/market analysis for the requirements prioritization methods is not included. Therefore descriptive statistics were not applied in our experiment. Parameters or measures used in experiment for the evaluation of requirements prioritization methods are likely to be measured by the participants themselves. Data set for the experiment was not too huge, as

only seven participants were included in the experiment which could be easily handled. Therefore this part of the experiment will be calculated statistically in future.

TABLE I. OBJECTIVE MEASURES FOR EXPERIMENT

Evaluation Criteria	ANP	AHP	Hierarchy AHP	Spanning Tree	Bubble Sort	Binary Search	Priority Groups
Consistency Index (Yes/No)	Yes	Yes	Yes	No	No	No	No
Scale of Measurement	Ratio	Ratio	Ratio	Ratio	Ordinal	Ordinal	Ordinal
Requirements Interdependence (Yes/No)	Yes	No	No	No	No	No	No

TABLE II. SUBJECTIVE MEASURES AFTER EVALUATION OF REQUIREMENTS

Evaluation Criteria	ANP	AHP	Hierarchy AHP	Spanning Tree	Bubble Sort	Binary Search	Priority Groups
Ease of use	4	3	5	4	1	6	3
Reliability of results	1	1	3	6	4	5	4
Fault tolerance	1	1	4	6	5	4	4

TABLE III. OBJECTIVE MEASURES AFTER EVALUATION OF REQUIREMENTS

Evaluation Criteria	ANP	AHP	Hierarchy AHP	Spanning Tree	Bubble Sort	Binary Search	Priority Groups
Required number of decisions	140	45	15	9	45	27	20
Total Time Consumption (Ordinal scale 1-6)	6	5	3	1	2	4	3
Time consumption par decision (Ordinal scale 1-6)	3	2	5	6	1	6	3

2) Data Reduction

Data reduction is an important part of experiment. The hypothesis of experiment was known to participants at the beginning of experiment. Also requirements that were going to be prioritized by requirements prioritization methods, were also described to them at the beginning. Dependency amongst functional requirements and non-functional requirements were also described before the beginning of experiment.

All the results from the participants were collected. Detailed analysis was done to calculate the final results.

All the results from the seven participants were valid and the individual results were considered in calculating final results. As data set was not huge therefore no data was excluded from the experiment.

After analysis of all the data collected from the participants, a generalized result was calculated which is discussed in the hypothesis.

3) Hypothesis Testing

First, hypothesis was whether requirements prioritization methods prioritize dependent or independent requirements or not? Each participant prioritized the described requirements on each prioritization method. Some information regarding this hypothesis was taken from literature, which was provided to participants while studying requirements prioritization techniques. Results are shown in Table I.

Now consider hypothesis 3 and 4, which were related to reliable results and less chance of errors while prioritizing software requirements. To prove hypothesis 3 and 4, three parameters/measures were evaluated against requirements prioritization techniques: ease of use, reliability of results and fault tolerance. To measure the hypothesis a scale was used from 1 to 6, where 1 represents highest value and 6 represents lowest value. Results are shown in Table II.

As it is known from the table, for the ease of use measures ANP gets 4 which means bit hard to use, and Bubble sort is the easiest method to use. In case of reliability and fault tolerance ANP gets maximum marks of 1, which means ANP produce most reliable result with less chance of error or errors can be identified easily. Now consider hypothesis 2, which is related to performance of requirements prioritization methods, again three (3) parameter/measures were evaluated to prove the performance of requirements prioritization methods. Again a scale of 1 to 6 was used to measure the hypothesis. Number of decisions needed to complete requirements prioritization are taken from literature while studying each requirements prioritization method. Each requirements prioritization technique has separate formula to calculate required numbers of comparisons. Details are shown in Table III.

IV. Advantanges and Disadvantages/Limitation of ANP

Below are the advantages and disadvantages of ANP derived from literature and current research.

TABLE IV. Advantages and Disadvantages of Anp

Advantages	***Disadvantages/Limitations***
ANP Can prioritize both dependent and independent requirements	Prioritization process in ANP is Complex
ANP provide Reliable and fault tolerant results	Tool support is need to minimize the complexity and time consumption while prioritizing requirements
ANP gives consistent results	
ANP gives results on ratio scale which further improves prioritization process	

V. Applications of ANP

ANP have a lot of applications almost in every field. ANP is derived from Analytic Hierarchy Process (AHP) with the additional feature of considering interdependencies amongst the criteria and alternatives. ANP is heavily used in multi-criteria decision analysis. ANP has been applied in many applications of social sciences where prioritization is needed. ANP is apparently new field in software engineering. AHP is very heavily used in software engineering domains but now ANP is taking over AHP. ANP is applied by Shah Nazir [20] in software design phase to select suitable software component based on quality criteria.

ANP is best suited for weight comparison also it is very influential when dealing complex network in decision making [23]. Babu et al also applied ANP in selection of architecture styles to optimize software architecture. [24]. ANP is applied by A.K Pandey et al in software testing phase in order to estimate the quality of software components [25].

The Proposed method is applied on Mobilink Franchise System, similarly it can be applied to any other case study of software engineering projects where interdependences exists amongst requirements.it can be noticed from the literature that ANP has been applied in different fields of software engineering.

ANP has started to be applied in the field of requirement prioritization of requirement engineering. Still there is lack of paper presentation in field of requirement and software engineering.

VI. Conclusion and Future Work

The study of Requirements prioritization techniques (binary search tree, AHP, hierarchy AHP, spanning tree matrix, Priority group and bubble sort) was carried out and the problems and limitations in these techniques, while prioritizing requirements, were noted. Some problems and limitations in requirements prioritization techniques were

identified. It is known that none of the requirements prioritization techniques prioritize dependent requirements and performance of requirements prioritization techniques is not good. The main problems are the delivery of reliable and fault tolerant results. Therefore the need of alternate technique of requirements prioritization, which can prioritize both dependent and independent requirements, was felt. Therefore new technique for prioritizing dependent and independent requirements is developed that is known as ANP. Steps of ANP are explained in details. The dependency amongst requirements is considered when requirements are prioritized with ANP due to which priority of interdependent requirements is calculated. It means that ANP prioritizes independent and dependent requirements.

An experiment was conducted to evaluate the performance of newly proposed requirements prioritization technique against existing requirements prioritization techniques (binary search tree, AHP, hierarchy AHP, spanning tree matrix, priority group and bubble sort). Experiment proves that main advantage of ANP is the prioritization of dependent requirements. The prioritization process of ANP is complex because a greater number of decisions is required for the completion of prioritization process. Reliable and fault tolerant results are the core characteristics of ANP.

Studies identify that more research and hard work is needed in the field of requirements prioritization to improve the performance of ANP. To deploy ANP to industry, is the core objective of future work. ANP will be used in some industrial projects of software engineering and hence its Performance will be evaluated. A user friendly tool will be developed so that users and requirements engineers can easily use it to prioritize requirements by ANP.

REFERENCES

[1] Firesmith, Prioritizing requirements, journal of object technology ed. Berlin, Germany: Springer-verlag, 2004.

[2] Ngo and G. Rhue, Decision Support in Requirement Engineering**,** In A. Aurum and C. Wohilin (Eds). Engineering and managing software requirements (pp.267-286) Springer Berlin Heidelberg, 2005.

[3] H.Goldstein, Who Killed the virtual case file? IEEE Spectrum, Tech.Rep.42, Sept 2005.

[4] C.Huang and Mobasher, Using data mining and recommender to scale up the requirement process, Proceedings of the 2nd International workshop on ultra large scale software intensive systems, 2008.

[5] J.karlsson, C.Wolin and B. Regnell, An evalution of methods for prioritizing software requirements, information and software technology, pp 939-947,2007.

[6] IEEE-STD 830-1998, "IEEE recommended practice for software requirement specifications.", IEEE computer society.

[7] S. Brender, Key words for use in RFC's to indicate requirements levels, RFC 2119.

[8] D. Leffingwell & D. widring, managing software requirements - A unified approach, upper Saddle River: Addison- Wesley.

[9] I. Sommerville & P. Sawyer, Requirements engineering, A good practice guide, Vhichester: John wiley and sons, May 5, 1997.

[10] S. Hatton, Early prioritization of goals, M.K Jean-Luchainaut, Elke A. Rundensteiner Ed., Springer Berlin Heidelberg, 2007.

[11] Dsdm Public version 4.2, from www.dsdm.org, Tech. Rep., Retrieved, 6 June, 2009.

[12] A.V. Aho, J.D. Ullman & J.E.Hopcroft, data structure and algrithems, Reading, MA: Addison-Wesley, January 11, 1983.

[13] Saaty, the analytic hierarchy process, McGraw-Hill, New York, 1980.

[14] C. Wohilin, P.Runeson, M.Host, M.C Ohlesson, B.Rgenell & A. Wesslen, Experimentation in software engineeringl, Springer, May, 2012.

[15] Saaty, Theory and applications of Analytic Network Process decision making with benefits, Opportunities, Costs and Risks,, USA RWS publications, 2005.

[16] New dawn technologies. Beat the oods, maiking IT projects a sucess, Retrieved 02 Aug 2009, form http://newdawntech.com/webinar/beattheoddspsf.pdf.

[17] J.Karlsson & K. Ryan, A Cost- Value approach for prioritizing requirements, IEEE Software, 14(5), 67-74.

[18] S.Nazir, S. Anwar, S. F. Khan, S. Shahzad, M. Ali, R.Amin, M.Nawaz, P.Lazaridis and J. Cosmas, Software Component Selection based on Quality Criteria using the analytic network process, Abstract and Applied Analysis, Hindawi Publishing corportation, 2014.

[19] S. Akınlı Koçak, G. Gonzales Calienes, G. Işıklar Alptekin, and A. Başar Bener, "Requirements prioritization framework for developing green and sustainable software using ANP - based decision making," In Proceedings of the Environmental Informatics Conference, 2013 (Hamburg, Germany, September 2-4, 2013).

[20] S.Khan, S. Ahmad, J.Ahmad and M.Haider, A Critical Survay on Requirment prioritization Techniques, ACEIT Conference Proceeding, 2016.

[21] J.A.Khan, I.U.Rehman, Y.H.Khan, I.J.Khan and S.Rashid, Comparison of requirement prioritization techniques to find best prioritization technique, I.J.Modern Education and Computer Science,2015,11,53-59

[22] J.A.Khan, I.U.Rehman, Y.H.Khan, S.A.Khan and W.Khan, Enhancement in agile methodologies using requirement engineering practices,, Science International Lahore,2016.

[23] P. Palanisamy, A. Zubar, and S. Kapoor, "A model for supplier selection using analytic network process," in Proceedings of the 10th International Conference on Operations and Quantitative Management (ICOQM '10), pp. 808–814, Nashik, India, 2011.

[24] K. D. Babu, P. G. Rajulu, A.R. Reddy and A.N. A. Kumari, Selection of Architecture Styles using Analytic Network Process for the Optimization of Software Architecture, International Journal of Computer Science and Information Security (IJCSIS), Vol. 8, No. 1, April 2010.

[25] A.K Pandey and C.P Agrawal, Analytical network process based model to estimate the quality of software components, International conference on,issues and chanllenges in intelligent computing techniques (ICICT),2014.

Cyber Profiling using Log Analysis and K-Means Clustering

A Case Study Higher Education in Indonesia

Muhammad Zulfadhilah
Departement of Informatics
Politeknik Hasnur
Banjarmasin, Indonesia

Yudi Prayudi
Departement of Informatics
Universitas Islam Indonesia
Yogyakarta, Indonesia

Imam Riadi
Department of Information Systems
Ahmad Dahlan University
Yogyakarta, Indonesia

***Abstract*—The Activities of Internet users are increasing from year to year and has had an impact on the behavior of the users themselves. Assessment of user behavior is often only based on interaction across the Internet without knowing any others activities. The log activity can be used as another way to study the behavior of the user. The Log Internet activity is one of the types of big data so that the use of data mining with K-Means technique can be used as a solution for the analysis of user behavior. This study has been carried out the process of clustering using K-Means algorithm is divided into three clusters, namely high, medium, and low. The results of the higher education institution show that each of these clusters produces websites that are frequented by the sequence: website search engine, social media, news, and information. This study also showed that the cyber profiling had been done strongly influenced by environmental factors and daily activities.**

Keywords—Clustering; K-Means; Log; Network; Cyber Profiling

I. INTRODUCTION

The increasing number of applications, hardware (device), and an Internet connection has affected the behavior of its users. In this case, APJI has been reported that in 2014 the order of the activities of Internet users in Indonesia is: users of social networks (social media), information search, chat (messaging), news search, video, email as a user internet activity in order of popularity. The data also indicate that the search for news and email usage is not a popular activity [1].

In general, cyber profiling studies is the exploration of data to determine what user activity at the time of internet access. One method that can be used to support the profiling process is a K-Means algorithm. Through these algorithms, the data can be grouped by the number of websites visited. This grouping aims to see what the user frequently accesses websites.

The data of internet users access at an institution can be categorized as a large data type so that the analysis can be done with data mining. In this case, the cluster algorithm as one of data mining techniques can be used to find groups (clusters) of a useful object, which the used are depends on the purpose of data analysis [2]. Clustering analysis is one of the most useful methods for the acquisition of knowledge and is used to find clusters that are a fundamental and important pattern for the distribution of the data itself [3].

Profiling is the process of collecting data from individuals and groups which can produce something interesting, surprising and significant, correlations that by using a machine that has good strength calculations to detect such data, while we as humans cannot [4]. Meanwhile, cyber profiling brings a good step in forensic computer science, based on the experience that has been achieved in the process of handling that have been made [5].

Educational institutions are one of the most likely group to conduct Internet activities. User behavior in educational institutions is also necessary to know the characteristics of user profiling and access to what is being done. Among the Indonesian has not been any research related to this issue, that's way cyber profiling would be very useful to know the behavior of Internet users in higher education in Indonesia.

Internet usage in higher education should be utilized by the user to support the educational process, but sometimes the facts obtained they used the Internet for the purpose outside of education, even less so there is an indication of such a user on educational institutions leading to cyber-crime. For that, we need to know more whether the use of the Internet in education is in line with the scope of activity in the education process activities.

II. CURRENT RESEARCH

A survey by APJI [1] showed that Internet users in Indonesia in 2014 reached 88 million. The survey was stated that there are three main reasons people use the Internet, namely access to social facilities/communications (72%), daily source (65%), and follow the development of the world (51%). The main reasons of internet access are practiced through four main activities, namely the use of social media (87%), searching for information (69%), instant messaging (60%) and search for the latest news (60%).

Research related to profiling, among others, performed by[6]. In these studies, [6] used machine learning to help the process of profiling to assist the experts in analyzing the crime.

Another study conducted by [4] and profiling results obtained knowledge of the risks of children and adolescents in accessing the internet. Based on these studies, [4] provide recommendations for caution in the use of personal data because the data will be accumulated and stored is likely to be used by parties who are not responsible.

In the study conducted by [7], the results of profiling can know the habits of Internet users and help network

administrators to improve the quality, security, and policy in the Internet network based on user behavior.

Meanwhile, [8] have also been doing profiling of Facebook users using the inductive method. However, the study also revealed that cyber profiling still has to use the deductive method because the cyber profiling process still requires additional data from the user completely. It to support the existence of differences in the behavior of individuals, because inductive generalizations extremely unreliable, and may cause misunderstanding in the analysis.

Another study conducted by [9] to use Twitter using ontology-based modeling OWL (Web Ontology Language), it is known that cyber profiling can be used to determine user interest based on URLs that have been shared via Twitter. The use of ontology also applied by [10], and the study revealed that cyber profiling using these methods could facilitate in providing information to the user when performing a search on a website.

III. BASIC THEORY

A. Data Mining

Data mining is an iterative and interactive process to find a new pattern or model valid, useful and understandable in a very large database. Data mining provides the search for patterns or trends that are desirable in a large database to help make decisions in the future. This pattern is recognized by a particular device that can provide a useful analysis and insightful data that can then be studied more carefully. The results of these patterns may be used in devices other decision support [2]. Data mining has stages like in Figure 1.

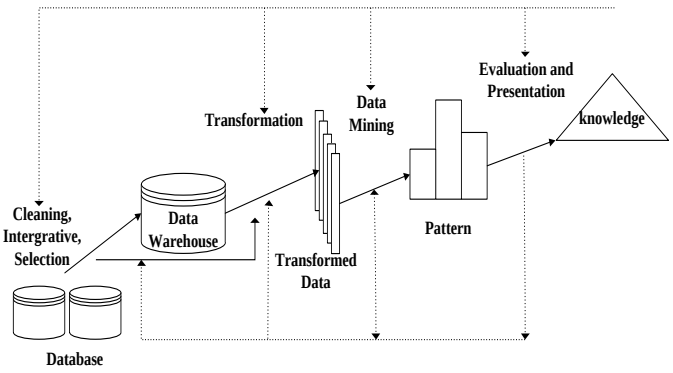


Fig. 1. Data Mining Process [10]

Data mining involves four tasks [11]:

1) Clustering – It is the task of finding a group and structure the data in some way or the "similar", without using known structures in the data.

2) Classification – It is the task of generalizing known structure to apply to new data. For example, an email program to attempt to classify an email as legitimate email or as spam.

3) Regression – Attempts to find a function which models the data with the least error.

4) Association rule learning – search the relationship between variables. For example, a supermarket might gather data on customer habits. Association rule learning can help supermarkets to determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

B. K-Means

Clustering is used to create a group (cluster) of the data so that it can easily find the necessary data. Clustering is a classification of similar objects into several different groups, it is usually applied in the analysis of statistical data which can be utilized in various fields, for example, machine learning, data mining, pattern recognition, image analysis and bioinformatics [11].

Clustering including supervised learning types. There are four types of clustering algorithms that have been compared based on performance, such as K-Means, hierarchical clustering, self-organization map (SOM) and expectation maximization (EM Clustering). Based on these test results can be concluded that the k-means algorithm performance and EM better than a hierarchical clustering algorithm. In general, partitioning algorithms such as K-Means and EM highly recommended for use in large-size data. This is different from a hierarchical clustering algorithm that has good performance when they are used in small size data [12].

The method of K-means algorithm as follows [13]:

1) Determine the number of clusters k as in shape. To determine the number of clusters K was done with some consideration as theoretical and conceptual considerations that may be proposed to determine how many clusters.

2) Generate K centroid (the center point of the cluster) beginning at random. Determination of initial centroid done at random from objects provided as K cluster, then to calculate the i cluster centroid next, use the following formula:

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \; ; \text{i=1,2,..,n} \qquad (1)$$

v : cluster centroid xi : the object to-i

n : the number of objects to be members of the cluster

3) Calculate the distance of each object to each centroid of each cluster. To calculate the distance between the object with the centroid author using Euclidian Distance.

$$d(x,y) = \|x - y\| = \sqrt{\sum_{j=1}^{n} (x_i - y_i)^2} \qquad (2)$$

xi= object x to-i n= the number of object

yi= object y to-i

4) Allocate each object into the nearest centroid. To perform the allocation of objects into each cluster during the iteration can generally be done in two ways, with a hard K-means, where it is explicitly every object is declared as a

member of the cluster by measuring the distance of the proximity of nature towards the center point of the cluster, another way to do with fuzzy C-Means.

5) Do iteration, then specify a new centroid position using equation (1).

6) Repeat step 3 if the new centroid position is not the same.

C. Log

Log (record keeping) is a file that records events in the computer program. Meanwhile, according to the definition of the log is a record of daily activities. Activities that are recorded directly called the transaction log. The log file can be used as a support in the process of cyber forensics to obtain digital evidence during the investigation stage [14].

The cleaning process must precede analysis of log data or preprocessing. Preprocessing is performed to remove duplication of data, check the data inconsistency, and correct errors in the data, such as print errors (typography) [15].

In Table 1 is an example of data on educational institutions

TABLE I. EXAMPLES OF DATA

Waktu		IP		Protokol		Website
1460509257	166960	192.168.15.	0	TCP_MISS/200	10018	a.tribalfusion.com:443
1460509207	115146	192.168.15.	0	TCP_MISS/200	5665	a248.e.akamai.net:443
1460624909	115780	192.168.15.	0	TCP_MISS/200	5945	accounts.google.com:443
1460509823	1425613	192.168.15.	0	TCP_MISS/200	84404	accounts.google.com:443
1460510173	115941	192.168.15.	0	TCP_MISS/200	3551	accounts.google.com:443
1460510343	115456	192.168.15.	0	TCP_MISS/200	3343	accounts.google.com:443
1460510463	116206	192.168.15.	0	TCP_MISS/200	3247	accounts.google.com:443
1460508537	5476	192.168.15.	0	TCP_MISS/200	5607	ad.turn.com:443
1460677224	115980	192.168.15.	0	TCP_MISS/200	21069	addons.cdn.mozilla.net:44
1460625101	121110	192.168.15.	0	TCP_MISS/200	21037	addons.cdn.mozilla.net:44

D. Cyber Profiling

The idea of cyber profiling is derived from criminal profiles, which provide information on the investigation division to classify the types of criminals who were at the crime scene. Profiling is more specifically based on what is known and not known about the criminal [8].

Profiling is information about an individual or group of individuals that are accumulated, stored, and used for various purposes, such as by monitoring their behavior through their internet activity [4].

Difficulties in implementing cyber profiling is on the diversity of user data and behavior when online is sometimes different from actual behavior. Given the privilege in personal behavior, inductive generalizations can be very reliable but can also lead to a misunderstanding of behavior analysis. Therefore the cyber-profiling process is via a combination of deductive and inductive methods [8].

For investigation, the cyber-profiling process gives a good, contributing to the field of forensic computer science. Cyber Profiling is one of the efforts made by the investigator, to know the alleged offenders through the analysis of data patterns that include aspects of technology, investigation, psychology, and sociology.

Cyber Profiling process can be directed to the benefit of:

- Identification of users of computers that have been used previously.
- Mapping the subject of family, social life, work, or network-based organizations, including those for whom he/she worked.
- Provision of information about the user regarding his ability, level of threat, and how vulnerable to threats
- Identify the suspected abuser

In a broader scope of cyber profiling can provide support information in a case, such as counterintelligence and counterterrorism [5].

The process of profiling against criminals often also known as cyber-criminal profiling criminal investigation or analysis. Criminal profiles generated in the form of data on personal traits, tendencies, habits, and geographic-demographic characteristics of the offender (for example: age, gender, socio-economic status, education, origin place of residence). Preparation of criminal profiling will relate to the analysis of physical evidence found at the crime scene, the process of extracting the understanding of the victim (victimology), looking for a modus operandi (whether the crime scene planned or unplanned), and the process of tracing the perpetrators were deliberately left out (signature) [16].

The new approach to cyber profiling is to use clustering techniques to classify the Web-based content through data user preferences. This preference can be interpreted as an initial grouping of the data so that the resulting cluster will show user profiles [17].

User profiling can be seen as the conclusion of the interests of users, intentions, characteristics, behavior and preferences [9]. User profiles are created for a description of the background knowledge of the user. User profile represents a concept model which is owned by the user when searching for information web [18].

IV. RESEARCH METHODS

To determine cyber-profiling of the higher educational institutions, so in this study the sample data is a log of Internet activities from one educational institution. Log data do not only contain any websites accessed by the user, but also includes packets received and sent over the network traffic. Data obtained containing the activities of network traffic for five days and produce data as much as 320.773 records.

In the early stages of research data collection, then do preprocessing that the data did not meet the criteria can be eliminated. Preliminary data obtained from 320.773 into a 1.638 record with the results of preprocessing. Furthermore, the mechanism of clustering using K-Means algorithm running on Rapid Miner and SPSS applications. The cluster data is then analyzed to make the process of profiling against internet users.

Figure 2 is a flow of the application of K-Means algorithm in the profiling process.

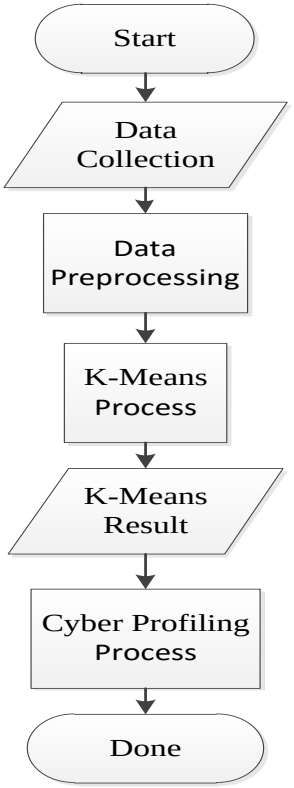


Fig. 2. Flow Research

Figure 3 is a flow of the algorithm K-Means:

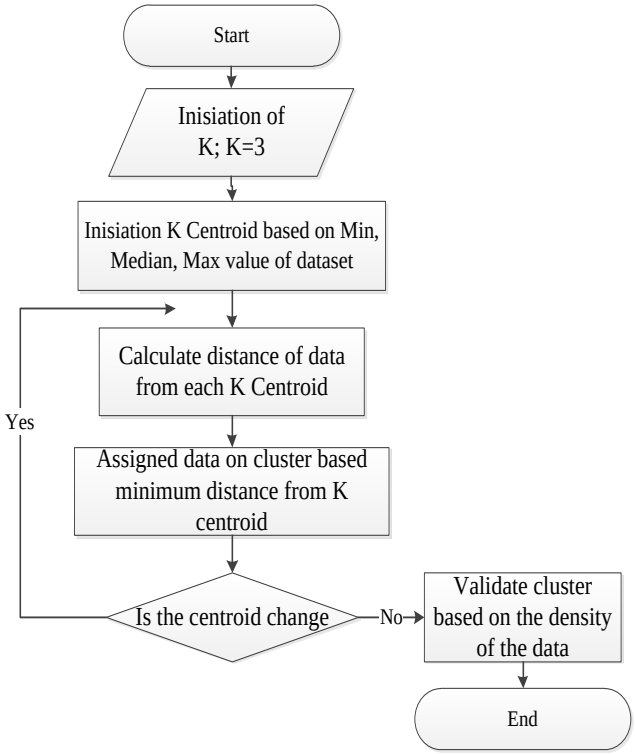


Fig. 3. K-Means Algorithm

V. RESULT

A. K-Means Clustering

Implementation of the K-Means algorithm, the result obtained is a level of visits to the website. The visit is divided into three groups: low, medium, and high.

Clustering by Rapid Miner and SPSS application indicates that the output produced has the same cluster of data. Based on the results of the cluster, it appears there are three clusters whose value is different, even on the first cluster value reached in 1479 (90.30%), the second is worth 126 (7.70%), and the third is worth 33 (2.00%). Those values represent the number of websites that have been divided in each cluster. Clustering results have shown this process has been running as expected to research.

Initialization of the initial cluster center in the clustering process can be seen in Table 2.

TABLE II. INITIALIZATION BEGINNING CLUSTER CENTER

	Cluster		
	1	2	3
Number of Visitors	1	37	71

An initial value is determined based on the data that have the highest value, the median value, and the smallest value. Those values are at the center of the initial cluster that will be followed in the process of K-Means.

The new centroid calculations will continue to do (iteration) until the discovery of iterations where centroid result is the same as the results of the previous centroid. In this study, there were eight iterations to determine the exact outcome of the 1638 cluster object. The iteration process can be seen in Table 3.

TABLE III. CLUSTERING PROCESS

Iteration	**Changes In Cluster Centers**		
	1	2	3
1	1,522	6,620	10,429
2	0,150	3,805	4,857
3	0,147	3,173	4,000
4	0.158	2,332	2,194
5	0,060	1,221	1,727
6	0,067	1,109	1,262
7	0,000	0,113	0,410
8	0,000	0,000	0,000

The result of the iteration process in determining the initial clustering center can be seen in Table 4.

TABLE IV. FINAL RESULT OF CLUSTER CENTER

	Cluster		
	1	2	3
Number of Visitors	2	19	46

The results of clustering details will be explained as follows:

- Cluster-1. On the results of clustering that has been done, the first cluster has as much data as 1467 records. The first cluster has the most members, but this cluster has a value which is below the overall average of the

data studied. In the first cluster has a data value in the range of 1-10, because in this cluster of existing data has a low level of traffic. Thus, cluster unity categorized on the website that has the least traffic from another cluster.

- Cluster-2. In the second cluster, members who entered at this cluster of some 126 records. The value of the results of the second cluster is in the range 11-31. This value indicates that the members of the second cluster have a medium level visits, because it has a higher value than the average value generated by clustering. Thus, the second cluster of clusters categorized as having moderate traffic levels.
- Cluster-3. On the results of the third cluster, cluster members who sign on as many as 33 records. The results of this third cluster have the fewest number of members in comparison with other clusters, but the members of this cluster have the highest value of the data that has been generated. The value in this cluster is in the 34-63 range, pointing to a result that the third cluster has a value far above average. Thus, the third cluster is categorized as a cluster that has the highest traffic levels.

The Results of clustering that has been done can be seen in Figure 4.

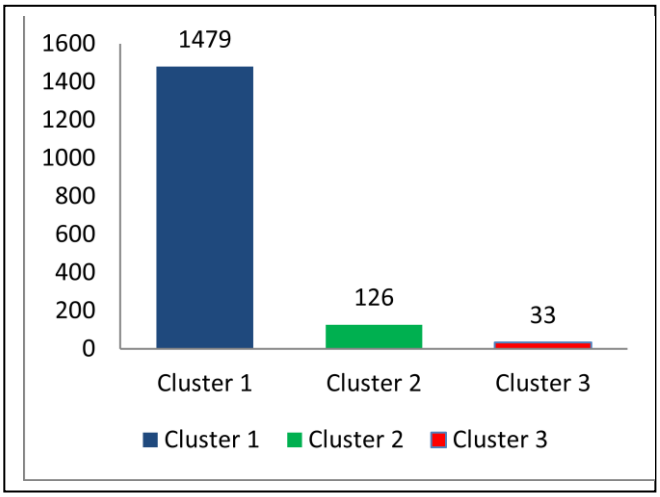


Fig. 4. The Result of Clustering

B. Analysis Results

In this study, the algorithm K-Means clustering has been implemented to perform in line with expectations. In the early stages of primary data obtained containing information about the websites accessed by users via the internet. In addition to the data contained informative website also contains data that updates to the operating system, the update of the web browser, and website advertising that usually appears as a pop-up.

Based on the results of the K-Means as shown in Figure 4 indicate that each cluster obtained having a different number of significant cluster members.

In the first cluster have shown low levels of traffic, but has some websites most. Data on the first cluster contains most of the advertising media website that coincided with a visit to a website activity. Meanwhile, in the second cluster that has moderate traffic levels, the data indicate a cluster member news sites that are in this cluster.

On the results of the third cluster is a group of websites with the highest traffic levels, but has the least number of websites. Data in this cluster shows that social media is a website with traffic levels were relatively high. Other data from the third cluster shows that Internet users access website search engine more frequently than from other websites including social media websites.

Although the research by [19] mention that the K-Means algorithm has shortcomings in central initialization beginning, in this study there were no such deficiencies. Clustering by two applications, Rapid Miner and SPSS produce the same data, this indicates that the determination of the value of the initial cluster center of these two applications has the same initial value.

Based on the results of clustering has been obtained that Internet users in educational institutions often make access to website search engine for information related to their field. This study also obtained results, that the social media website and streaming video sites accessed more frequently than the information and news website. These results caused by the delivery of information in the digital age now entering the realm of social media compared to other information websites.

In this study, social media included in the website frequently accessed by the user. This is according to research conducted by [1] [9] which states that social media becomes one of the surfing activity frequently accessed.

Cyber profiling process that has been done shows that the search for information more frequently accessed by users coming from educational institutions. This indicates that environmental factors and daily activities affect on what is accessed by the user. These results refer to the study [8] which states that the process of cyber profiling to predict based on the demographic information has a high degree of accuracy.

Based on the above data profiling results to higher educational institutions indicate that the use of the Internet has been used to support the educational process. The source of the data obtained showed no user activity in the area of higher education that leads to cybercrime.

In this case was supposed to complete the profiling results as mentioned by [5], the source of data should contain log activity on a computer that had been used, but in this research, the data has only been in the form of log data network traffic.

In this study cyber profiling can only provide information about the Internet activities performed by users, but for the threat of crime and profiling based on data from computers that have been used as mentioned by [5] are still not exist.

The results of this research have been able to meet the definition of cyber profiling, because it provides information about the user based on the current activity connected with the internet. The results of this study can be used by network administrator to improve the quality of services, policies, and security as mentioned by [7].

VI. CONCLUSION

The results of log analysis datasets using the K-Means algorithm to cyber profiling process show that the algorithm has to group activity based on the data of internet users visited the website. This grouping is divided into three, namely the visit low, medium, and high.

In this study, the K-Means algorithm is used as an algorithm for the cyber profiling process. K-Means algorithm being used is in line with expectations from this study, because it has a simple algorithmic process with a good degree of accuracy. But the K-Means algorithm has disadvantages, namely the process of making an initial value initial random center. This can lead to differences in the results of the cluster.

The results of this study indicate that Internet users in higher educational institutions are more accessible to website for searching information. The results also show that social media has a high-level visit after website search engine.

This study has limitations in the source of data for the profiling process. For the perfection of the profiling, the process should contain the data of any computer activities. Therefore, further research is expected to perform better cyber profiling with the more complete data source.

REFERENCES

[1] APJII, "Indonesian Internet User Profile 2014," 2015.

[2] Fajar Astuti Hermawati, Data Mining. Yogyakarta: CV. Andi Offset, 2013.

[3] H. Chunchun, L. Nianxue, Y. Xiaohong, and S. Wenzhong, "Traffic Flow Data Mining and Evaluation Based on Fuzzy Clustering Techniques," vol. 13, no. 4, pp. 344–349, 2011.

[4] D. B. van den Berg, P. dr. A. de Vries, P. dr. S. van der Hof, M. Kakaris, and A. Theocharidis, "Online Identities , Profiling and Cyber Bullying," no. March, 2013.

[5] J. J. Irvine, "Digital Forensic Analysis & Cyber Profiling," no. 703, pp. 1–32, 2010.

[6] A. S. N. Chakravarthy, "Analysis of cyber-criminal profiling and cyber-attacks : A comprehensive study," no. September, 2014.

[7] T. Bakhshi and B. Ghita, "Traffic Profiling : Evaluating Stability in Multi-Device User Environments," 2016.

[8] S. Yu, "Behavioral Evidence Analysis on Facebook: a Test of Cyber-Profiling," Defendologija, vol. 16, no. 33, pp. 19–30, 2013.

[9] P. Peña, R. Hoyo, J. Vea-murguía, C. González, and S. Mayo, "Collective Knowledge Ontology User Profiling for Twitter Automatic User Profiling," pp. 439–444, 2013.

[10] C. J. Lei Xu, J. Wang, J. Yuan, and Y. Ren, "Information Security in Big Data : Privacy and Data Mining," pp. 1149–1176, 2014.

[11] A. Chauhan, G. Mishra, and G. Kumar, "Survey on Data Mining Techniques in Intrusion Detection," vol. 2, no. 7, pp. 2–5, 2011.

[12] I. Riadi, J. E. Istiyanto, and Su. S. Saleh, "Internet Forensics Framework Based-on Internet Forensics Framework Based-on Clustering," no. January, 2013.

[13] N. S. Ediyanto, Muhlasah Novitasari Mara, "Characteristics classification by Method K-Means Cluster Analysis," Bul. Ilm., vol. 02, no. 2, pp. 133–136, 2013.

[14] A. Iswardani and I. Riadi, "Denial Of Service Log Analysis Using Density K-Mans Method," vol. 83, no. 2, pp. 299–302, 2016.

[15] Universitas Sumatera Utara, "Decision Tree," Repos. II.pdf, 2012.

[16] Margaretha, "Criminal Profiling dan Psychological Autopsy," http//psikologiforensik.com/2013/04/22/criminal-profiling-dan-psychological- Autops., 2015.

[17] P. B. Costa, S. Oliveira, and L. Nunes, "Profiling Web Users Preferences with Text Mining," pp. 1–4, 2013.

[18] P. Jayakumar and P.Shobana, "Creating Ontology Based User Profile for Searching Web Information," no. 978, 2014.

[19] S. Andayani, "Formation of clusters in Knowledge Discovery in Databases by Algorithm K-Means," 2007.

Enhancement in System Schedulability by Controlling Task Releases

Basharat Mahmood
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Saif ur Rehman Malik
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Naveed Ahmad
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Adeel Anjum
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

***Abstract*—In real-time systems fixed priority scheduling techniques are considered superior than the dynamic priority counterparts from implementation perspectives; however the dynamic priority assignments dominate the fixed priority mechanism when it comes to system utilization. Considering this gap, a number of results are added to real-time system literature recently that achieve higher utilization at the cost of tuning task parameters. We further investigate this problem by proposing a novel fixed priority scheduling technique that keeps task parameters intact. The proposed technique favors the lower priority tasks by blocking the release of higher priority tasks without hurting their deadlines. The aforementioned strategy helps in creating some extra space that is utilized by a lower priority task to complete its execution. It is proved that the proposed technique dominates pure preemptive scheduling. Furthermore the results obtained are applied to an example task set which is not schedulable with preemption threshold scheduling and quantum based scheduling but it is schedulable with proposed technique. The analyses show the supremacy of our work over existing fixed priority alternatives from utilization perspective.**

***Keywords*—Real-time Systems; Fixed Priority Scheduling; RM Scheduling; Priority Inversion**

I. INTRODUCTION

Real-time systems are built to execute temporally constrained tasks. On such platforms, the accuracy of a system depends not only upon the correctness of response, but also the time these results are obtained. Missing a task deadline may result in serious damage, especially in hard real-time systems [8]. It is not a must for a real-time system to be very fast, but it must be enough capable to execute its tasks within a specified time. Priority assignment to real-time tasks, is the process of deciding the order in which different tasks are executed. In real-time scheduling, the scheduler is responsible to allocate tasks on the processor in such a way that all timing constraints are satisfied.

Fulfilling the timing constraints of tasks is essential to real-time systems and hence the scheduling problem plays an important role in real-time systems theory. The fixed priority scheduling technique is widely used in real-time systems due to its simplicity and predictability. Under fixed priority class, Rate-monotonic (RM) algorithm is a well-known fixed priority assignment algorithm. It is an optimal algorithm for the implicit deadline model [10] [2]. The RM scheduling algorithm is subdivided into two main streams, preemptive scheduling and non-preemptive scheduling. In preemptive scheduling, a lower priority task is preempted when a higher priority task is released, while non-preemptive scheduling does not allow such preemptions. Generally, preemptive scheduling provides better schedulability than non-preemptive scheduling, but it is not always the case. Both preemptive and non-preemptive scheduling fail to guarantee 100 % CPU utilization.

Lot of efforts have been made recently to improve the schedulability of fixed priority scheduling. Different variants of preemptive scheduling have been proposed, which use the concept of priority inversion in order to improve the schedulability. These techniques allow a lower priority task to block a higher priority task. Deferred preemption [19], Preemption threshold scheduling [2] and Quantum based scheduling [5] are examples of such techniques, which improve the schedulability of fixed priority scheduling by priority inversion at runtime.

In this paper a new fixed priority scheduling technique named as CTR is proposed. The CTR technique blocks the task releases for a predefined interval of time without hurting their deadline in order to create some extra space for the currently executing task. In CTR technique, each task is assigned a feasible release block time. At runtime, tasks are kept in block state at their actual release times and are released after their assigned block time. In this way, some extra space could be created for the lower priority tasks to execute. It is proved that such blockage of task releases, does not hurt the deadline of tasks. It is also proved that the CTR technique dominates the RM preemptive scheduling in terms of schedulability. A task set is also given which is not schedulable with preemption threshold scheduling and quantum based scheduling but, it is schedulable with the CTR technique. This shows that the CTR technique has at least an incomparable relation with these techniques.

A. Related Work

In 1973, Liu and Layland did the pioneer and the most influential work in real-time scheduling theory. In their seminal paper [10], they proposed an optimal fixed priority assignment algorithm called rate-monotonic algorithm for the implicit deadline task model. In the same paper, they derived a sufficient schedulability test called LL-bound to predict the feasibility of the system. After that, a lot of work has been done to improve the system feasibility prediction. This work is mainly of two types, the exact schedulability tests [16][15][3][9] and sufficient schedulability tests[10][12].

To reduce the run-time overhead due to task preemptions, limited-preemptions model has been proposed [21][22]. In this model each task is divided into a number of non-preemptive regions and is considered non-preemptive within those regions. These regions may be either fixed or floating [22]. Under fixed pre-emption point model, the non-preemptive regions are predefined while in floating preemption point model the location of non-preemptive regions are un-known.

Different variants of RM preemptive scheduling have been proposed in literature, to improve the schedulability [20][19] [2][5]. In [20] dual priority scheduling model is presented. In this model each task is executed in dual phases with different static priorities. The transition from one phase to another is made at fixed points. This model dominates the RM scheduling but, is considered not viable due to its complexity.

The deferred preemptions scheduling technique [19] assigns each task τ_i an interval q_i for which it remains non-pre-emptible. At runtime when a higher priority task is released, it is kept blocked if the lower priority task is in non-preemptible section otherwise it is preempted.

Preemption threshold scheduling [2] is a dual priority scheduling technique. It assigns each task a regular priority and a preemption threshold value which is greater or equal to its priority. At runtime when a task is executed, its priority is raised to its preemption threshold value. In this way a task can block those higher priority tasks whose priority is less than its preemption threshold value. Preemption threshold scheduling dominates preemptive and non-preemptive scheduling in terms of schedulability.

Another variant of preemptive scheduling is the quantum based scheduling [5]. In quantum based scheduling, CPU time is divided into discrete units called quanta. At runtime, the CPU time is allocated to tasks in the form of quantum. When a quantum is allocated to the task, that task cannot be preempted until the quantum expires or task is completed. Both preemption threshold scheduling and quantum based scheduling improve the schedulability of fixed priority scheduling, but still fail to guarantee 100 % utilization.

In [23] ready-Q locking mechanism is proposed. Ready-Q locking improves the schedulability by locking the ready queue at runtime in order to reduce the interference from higher priority task during the execution of a lower priority task. For further improvement in schedulability, preemption threshold scheduling is also merged with ready-Q locking mechanism [23].

B. Contribution of the Paper:

This work has the following contributions

- A novel fixed priority scheduling technique CTR scheduling is proposed in this paper. The CTR scheduling controls the task releases in order to enhance schedulability
- It is proved that the CTR scheduling dominates RM preemptive scheduling in terms of schedulability
- It is also proved that the CTR scheduling has at-least an incomparable relation with preemption threshold scheduling and quantum-based scheduling
- The improvement in schedulability is also shown by experiments on synthetic task sets

C. Paper Organization:

The rest of the paper is organized as follows. In section II, the system model and basic terminologies are discussed. In section III, the proposed technique is explained in detail with examples. The proposed technique is compared with existing techniques in section IV and finally our work is concluded in section V.

II. System Model, Assumptions and Notations

A. Task Model

We consider the classical periodic real-time task model. Each task τ_i is defined by the tuple (C_i, P_i, D_i). Each task consists of a sequence of infinite jobs. The time at which the first job of a task is released is called its phase. If the phase of a task τ_i is Φ_i then the k^{th} job of τ_i is released at $J_{i,k} = \varphi_i+(k-1)*P_i$. The absolute deadline of the k^{th} job of task τ_i is $d_{i,k} = r_{i,k} + D_i$. The portion of CPU time used by a task τ_i is called its utilization and can be calculated by C_i/P_i. The total utilization of the system can be determined by $U_p = \sum_{i=1}^{n} U_i$.

B. System Model:

We consider a real-time system which consists of a single processor. The workload for the system is defined by set τ of n tasks. A fixed priority scheduler is used to schedule tasks. The scheduler assigns priorities to tasks according to RM algorithm. Each task τ_i is assigned a feasible release block time (Δr_i). When the k^{th} job of τ_i is released at $r_{i,k}$, it is considered in blocked state for the interval ($r_{i,k}$, $r_{i,k} + \Delta r_i$) and is released at $r_{i,k} + \Delta r_i$ time. During block state a task can only be executed if the CPU is free.

C. Assumptions:

We consider the following assumption for our technique

(A1) Task sets follow the implicit deadline model. It means that for any task (τ_i), relative deadline is equal to its period i.e. $D_i = P_i$ therefore, a task can be simply defined by (C_i, D_i)

(A2) Workload is defined by a set τ of n tasks and all tasks in τ are periodic

(A3) It is assumed that only computational resources are required to execute a task and all other resources are negligible

(A4) Any task can be preempted at any time and no task has any non-pre-emptible part

(A5) All tasks are independent and no precedence constraints exist among them

(A6) All runtime costs are negligible

The notations used are shown in TABLE I

TABLE I. NOTATIONS USED AND MEANINGS

Notation	*Meaning*
τ	Set of tasks
τ_i	Task i, $\tau_i \in \tau$
C_i	Worst case execution time of τ_i
P_i	Period of τ_i
R_i	Worst case response time of τ_i under RM scheduling
E_i	The time period during which a task τ_i is released and completes its execution under CTR scheduling
U_i	Utilization of τ_i
D_i	Relative deadline of τ_i
$J_{i,k}$	K_{th} job of τ_i
$d_{j,t}$	Absolute deadline of τ_j at time t
t	Current time
Φ_i	Release time of first job of τ_i
r_i	Feasible release block time of τ_i
$r_{i,k}$	Release time of k_{th} job of τ_i
$t_{i,t}^{avl}$	Time available to τ_i at time t
$t_{i,t}^{req}$	Time required to τ_i at time t
τ_t^{exe}	Task executing at time t
τ_t^{rel}	Task released at time t
$\tau_{r,t}^{h}$	Highest priority ready task at time t
$C_{i,t}^{rem}$	Remaining execution time of task τ_i at time t
R[ ]	Queue of released tasks
J[]	Queue of tasks
P (τ_i)	Priority of τ_i
$I_{(t,d_{i,t})}^{hp(i)}$	Interference from tasks having higher priority than τ_i during (t, $d_{i,t}$)
$e_{j,t}$	Extra space created for τ_j at time t due to the delay in release of higher priority tasks

III. CONTROLLED-TASK-RELEASES (CTR) REAL-TIME SCHEDULING

The CTR scheduling technique is discussed in following subsections in detail

A. Overview of the Technique

If we analyze the runtime behavior of RM preemptive scheduling, it is observed that when a higher priority task τ_i is released at time t, currently executing job of lower priority task τ_j is preempted immediately. Now, τ_j misses its deadline if

$$d_{j,t} - t \leq C_{j,t}^{rem} + I_{(t,d_{j,t})}^{hp(j)}$$

where $C_{j,t}^{rem}$ is the remaining execution time of τ_j at current time t, $d_{j,t}$ is the absolute deadline of currently preempted job of τ_j , t is the current time and $I_{(t,d_{j,t})}^{hp(j)}$ is the interference from tasks with higher priority than τ_j during (t, $d_{j,t}$) interval.

Such tasks can be made schedulable if the release of higher priority task is delayed without hurting its deadline. Such delays create some extra space for the lower priority tasks to complete their execution. The CTR technique utilizes this idea to improve the schedulability.

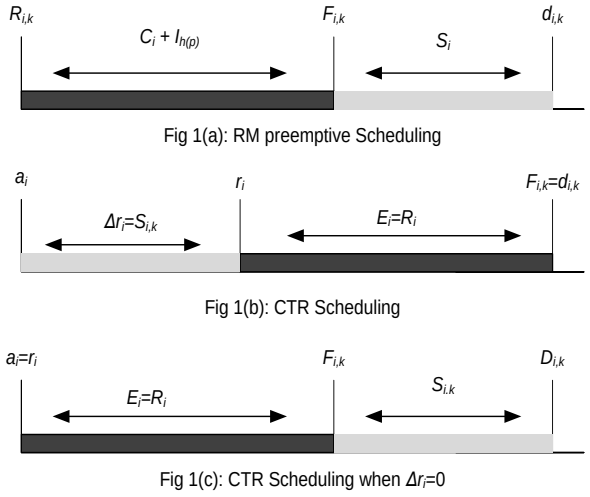


Fig. 1. Comparison of runtime behavior of RM preemptive and CTR scheduling (a) RM scheduling (b)CTR scheduling (c) CTR scheduling with Δri = 0

The CTR scheduling assigns each task τ_i a feasible release block time (Δr_i). At runtime, a task τ_i is considered in blocked state at its actual release time and remains in this state for Δr_i time. After Δr_i time τ_i is released and its priority is compared with the currently executing task. If τ_i has higher priority than the currently executing task, then the task is preempted otherwise it continues. At runtime, if the CPU is free and no task is executing then the lowest priority blocked task is executed.

In Fig 1, the runtime behavior of CTR scheduling is compared with RM preemptive scheduling. Fig 1(a) shows the execution of τ_i with RM scheduling. The k_{th} job of τ_i is released at $r_{i,k}$ and completes its execution at $F_{i,k}$ ahead of its deadline $d_{i,k}$. This difference is shown by $S_{i,k}$ where $S_{i,k} = d_{i,k} - (C_{j,t}^{rem} + I_{(t,d_{j,t})}^{hp(j)})$. On the other hand, the proposed technique blocks the release of task τ_i for Δr_i (or $S_{i,k}$) amount of time therefor, it completes at its deadline as shown in Fig 2(b). The CTR scheduling behaves similarly to RM preemptive scheduling if the release of the task τ_i is not blocked i.e. in $\Delta r_i = 0$ (shown in Fig 1(c)).

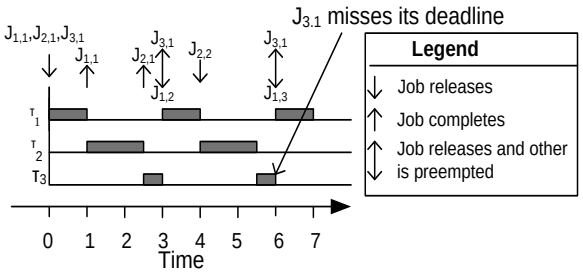


Fig. 2. RM preemptive scheduling of task set given in TABLE II, τ_3 misses its deadline

B. Motivational Example

The following example illustrates the benefits of CTR scheduling. The tasks and their attributes for this example are given in TABLE II.

TABLE II. EXAMPLE TASK SET

Task(τ_i)	WCET(C_i)	Period(P_i)	Release-Block time (Δr_i)
τ_1	1	3	2
τ_2	1.5	4	0.5
τ_3	1.5	6	0

If we apply the RM preemptive scheduling on the example task set given in TABLE II, then the task set is not schedulable because the deadline of τ_3 is missed. The scheduling of task set with RM preemptive scheduling is shown in Fig 2.

Now, if we apply the CTR scheduling on the same task set by assigning each task a feasible task release-block time as given in TABLE II, then the task is schedulable (The method of assigning release-block time is discussed in the following subsection). TABLE III shows the sequence of jobs in which they are activated and released under the task release mechanism of the CTR scheduling. The runtime behavior with the CTR technique is shown in Fig 3.

TABLE III. ACTIVATION AND RELEASE SEQUENCE OF TASKS UNDER CTR SCHEDULING

Task Jobs($J_{i,k}$)	Activation time	Release time
$J_{1,1}$	0	2
$J_{2,1}$	0	0.5
$J_{3,1}$	0	0
$J_{1,2}$	3	5
$J_{2,2}$	4	4.5
$J_{1,3}$	6	8
$J_{3,2}$	6	6
$J_{2,3}$	8	8.5
$J_{1,4}$	9	11

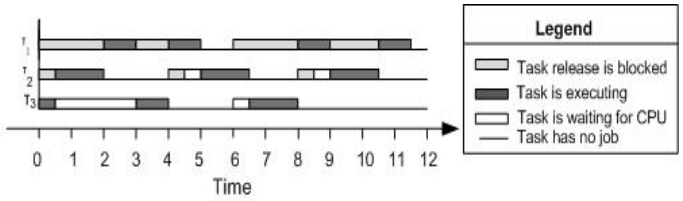


Fig. 3. CTR scheduling of task set given in TABLE II, task set is schedulable

The execution sequence with the proposed technique is explained below

- At t=0, τ_1 and τ_2 are active, but τ_3 is released. Therefore τ_3 is started
- At t=0.5, τ_2 is released. As τ_2 has higher priority than τ_3, therefore τ_3 is preempted and τ_2 starts
- At t=2, not only τ_2 is completed, but also τ_1 is released. So τ_1 starts
- At t=3, τ_1 completes its execution. As at t=3, only τ_3 is ready which was preempted earlier , therefore τ_3 starts and is completed at t=4
- At t=4, no task is ready, but τ_1 and τ_2 are active. As the CPU is free therefore the highest priority active task (τ_1) is assigned to the CPU
- At t=5, τ_1 completes its execution and τ_2 starts

The process continues in a similar way.

C. Assignment of Feasible Release-block Time

The feasible release-block time (Δr_i) for a task τ_i is the difference between the available time to it and the maximum time it may require in worst case. The available time to a task τ_i released at time t is

$$t_{i,t}^{avail} = t - d_{i,t}$$

$$\Longrightarrow t_{i,t}^{avail} = D_i \quad (1)$$

The maximum required time to a task τ_i in the worst case is

$$t_{i,t}^{req} = C_i + \sum_{j=1}^{i-1} \left\lceil \frac{D_i}{C_j} \right\rceil \times C_j \quad (2)$$

Now, the feasible release-block time (Δr_i) value for a task τ_i is

$$\Delta r_i = t_{i,t}^{avail} - t_{i,t}^{req}$$

$$\Rightarrow \Delta r_i = D_i - C_i + \sum_{j=1}^{i-1} \left\lceil \frac{D_i}{C_j} \right\rceil \times C_j \quad (3)$$

The Algorithm 1 to assign Δr_i values to tasks is given below

Algorithm 1: assigning feasible release-block time (Δr_i)

```
Require: Set of n tasks (τ)
for i = 1 → n − 1
do
    Δr_i = D_i − C_i + Σ_{j=1}^{i−1} ⌈D_i / C_j⌉ × C_j
    if Δr_i < 0 then
        Δr_i = 0
    End if
end for
```

Theorem1: *Given a task $\tau_i \in \tau$ released at time t, if τ_i is schedulable under the RM preemptive scheduling then it remains schedulable even if its release is blocked for Δr_i time.*

Proof: The given task τ_i released at time t is RM schedulable then

$$t_{i,t}^{avl} \geq t_{i,t}^{req}$$

$$\Rightarrow t_{i,t}^{avl} \geq R_i$$

Where R_i is the worst case response time of τ_i under RM preemptive scheduling and can be calculated by using Response-Time Analysis [3]. When the release of τ_i is blocked for Δr_i time, then the available time to τi is reduced to

$$t_{i,t}^{avl} = D_i - \Delta r_i$$

Now, τ_i remains schedulable if following inequality is satisfied

$$D_i - \Delta r_i \geq R_i$$

We solve the above inequality

$$D_i - (D_i - C_i + \sum_{j=1}^{i-1} \left\lceil \frac{D_i}{C_j} \right\rceil \times C_j) \geq R_i$$

$$\Longrightarrow D_i - D_i + C_i + \sum_{j=1}^{i-1} \left\lceil \frac{D_i}{C_j} \right\rceil \times C_j \geq R_i$$

$$\Rightarrow C_i + \sum_{j=1}^{i-1} \left\lceil \frac{D_i}{C_j} \right\rceil \times C_j \geq R_i$$

The above condition always remains true. Hence, it is proved that blocking the release of a RM task schedulable for Δr_i time does not hurt its schedulability.

D. Scheduling Algorithm

Now, we present the CTR scheduling algorithm. Initially, priorities are assigned to tasks according to RM algorithm. Then each task is assigned a feasible release-block time. At the start, all tasks with positive Δr value remain in blocked state and the highest priority ready task with zero Δr is assigned to the processor. When a task τ_i is released after remaining in blocked state for Δr_i time, its priority is compared with the priority of the currently executing task τ_j and if τ_i has higher priority than τ_j, then τ_j is preempted otherwise it continues. During runtime if a task is completed and there is no ready task, then the first task in blocked state is assigned to the processor. The scheduling process under the CTR scheduling is shown by algorithm 2.

E. Implementation of the CTR Scheduling

In this section, we discuss the implementation of CTR scheduling and the performance overheads associated with it.

1) Implementation

The implementation of mechanism to block the release of tasks at runtime is the core issue in the implementation of CTR scheduling. For this, the scheduler is required to distinguish the active and ready states of a task at runtime. In order to have better synchronization of task parameters, the job table is required to keep information about activation time and release time of jobs. The CTR scheduler requires two task queues named as job queue and ready queue. The job queue keeps the jobs which are next to release. These jobs are ordered by earliest to release first basis. The ready queue keeps the jobs which are ready to execute, but waiting for CPU due to the execution of a high priority job. These jobs are kept by descending order of their priorities. At run time, the scheduler sets the timing hardware to interrupt the CPU at the release time of the job at the head of the job queue. When an interrupt is generated the scheduler moves all the jobs with the same release time as of the interrupt time, to the ready queue and updates the timing hardware. Under CTR scheduling, when the ready queue is empty the scheduler is required to move an active job to the ready queue. To do this, the scheduler finds the first job in the job queue whose activation time is before or the same as of the current time and moves it to the ready queue.

2) Overheads

Here we discuss the overheads associated with CTR scheduling and compare with other scheduling techniques. The first major overhead associated with CTR scheduling is the assignment of release block time to tasks. This assignment is done off-line and has an O(n) complexity. As this assignment is made off-line therefore it does not cause any performance cut at runtime. No such assignment is required in RM preemptive scheduling. On the other hand, preemption

threshold scheduling has a more complex mechanism to assign preemption threshold values which has an $O(n^2)$ complexity. Similarly, in quantum-based scheduling the assignment of feasible quantum size also has and $O(n^2)$ complexity.

Algorithm 2: *The CTR scheduling*

```
Require: Set of n tasks (τ)
R[ ] = φ   // Queue that holds the released tasks
J[ ] = φ   // Queue that holds the active or blocked tasks
i, k=0

RM(τ )        // Assign priorities to tasks according to RM algorithm

Assign-Delays(τ)     //Assign feasible task release-block times
At t=o:
 CPU ⟷ τ^h_{r,t}  //Highest priority ready task gets the CPU
Upon task activation:
    if (Δr_cur > 0)
       J[ i + 1] ⟷ τ_t^cur
    End if
Upon task completion:
if (R[ ] ≠ φ)   // if ready queue is not empty
  CPU ⟷ τ^h_{r,t}
End if
else if (J[ ] ≠ φ)  //if the queue holding the blocked tasks is not
empty
  CPU ⟷ J[1]   // Assign first task from the blocked task to the CPU
End else if
    else
         wait for task release
    End else

Upon task release:
    if (P(τ_t^exe)<P(τ_t^rel)  //compare the priority of currently
released task with the executing task
         CPU ⟷ τ_t^rel
    End if
    else
         CPU ⟷ τ_t^exe
    End else
```

The second major overhead, which incurs at runtime is that, in CTR scheduling when there is no task in ready queue the scheduler is required to search the job queue to find an active task to execute. It does not affect the performance much because in heavily loaded systems it is very rare to have an empty ready queue.

IV. Evaluation Of The CTR Scheduling

In this section we compare the CTR scheduling with other techniques. We have proved the dominance of CTR scheduling over RM preemptive scheduling in schedulability perspective. This dominance is also validated by experiments on synthetic task sets. The incomparable relation of CTR scheduling with Preemption threshold scheduling and Quantum-based scheduling has also been proved and validated by experiments.

A. Dominance of CTR scheduling over RM preemptive scheduling

RM preemptive scheduling favors high priority task because it immediately preempts the lower priority task when a higher priority task is released. Such early preemptions can cause deadline to miss for lower priority tasks. It is tried in the CTR scheduling to overcome this deficiency by delaying preemptions feasibly. The CTR scheduling creates extra space for lower priority tasks by delaying the release of higher priority task. As a result, it provides better schedulability than RM preemptive scheduling. In following theorem, we prove that the CTR scheduling dominates RM preemptive scheduling in schedulability perspective. It means that CTR scheduling can feasibly schedule all those task sets which are schedulable with RM preemptive scheduling, but it is not guaranteed that RM preemptive scheduling can schedule all the task sets which are schedulable with CTR scheduling.

Theorem 2: *Given a task set τ consisting of n independent, periodic tasks whose deadlines are equal to their periods. If τ is RM schedulable then it is always schedulable with the CTR scheduling technique while the vice versa is not true always.*

Proof: Suppose a lower priority task τ_j is executing at time t and a higher priority task τ_i is released. We discuss the schedulability of both tasks under RM scheduling and CTR scheduling

Case 1: (Schedulability of τ_i) If τ_i is schedulable with RM preemptive scheduling then it is also schedulable with CTR scheduling (by Theorem 1).

Case 2: (Schedulability of τ_j) If τ_j is schedulable with RM preemptive scheduling then it gets its required time before the deadline of its current job. It can be written as

$$d_{j,t} - \mathrm{t} \geq C_{j,t}^{rem} + I_{t,d_{j,t}}^{h(j)}$$

As under the CTR scheduling the release of τ_i is blocked for therefore Δr_i time, it creates some extra space $e_{j,t}$ for τ_j, therefore the above in-equality can be written as

$$d_{j,t} - \mathrm{t} + e_{j,t} \geq C_{j,t}^{rem} + I_{t,d_{j,t}}^{h(j)}$$

Where $e_{j,t} \geq 0$. The above in-equality always remains true because

$$d_{j,t} - \mathrm{t} + e_{j,t} \geq d_{j,t} - \mathrm{t}$$

It shows that if τ_j is schedulable with RM preemptive scheduling then it is also schedulable with CTR scheduling. Now, if τ_j is not schedulable with RM preemptive scheduling then

$$d_{j,t} - \mathrm{t} < C_{j,t}^{rem} + I_{t,d_{j,t}}^{h(j)}$$

Now in similar situation the CTR scheduling creates some extra space $e_{j,t}$ for τ_j by blocking higher priority tasks. Now if

$$d_{j,t} - \mathrm{t} + e_{j,t} \geq C_{j,t}^{rem} + I_{t,d_{j,t}}^{h(j)}$$

then the task is schedulable with the CTR technique and if it holds true for all such situations, then the whole task set is schedulable with the CTR technique. It shows that the CTR technique can schedule some tasks which are not schedulable with the RM technique.

*1) **Experimental Evaluation:***

The CTR Scheduling has been compared with RM preemptive and non-preemptive scheduling, to evaluate the schedulability improvement. For this purpose, we have generated 10^4 task sets. Each task set consists of periodic tasks with $D_i = P_i$. The size of task sets is $n \in \{2, 3, 4, 5, 6, 7, 8, 9\}$ and their period ranges {2, 500}. The utilization of the system was kept from 88% to 100%. Priorities are assigned to tasks by RM algorithm. The performance of different techniques has been evaluated by the percentage of feasible tasks. The results are discussed below.

Fig4 summarizes the experimental results for the CTR, RM preemptive (RMP) and RM non-preemptive (RMNP) scheduling techniques in schedulability perspective. X-axis represents the system's utilization while the Y-axis shows the percentage of feasible task sets. Fig 4(a) shows the results of task sets with n=2 or 3. It can be seen clearly that the CTR scheduling surpasses both RMP and RMNP in schedulability perspective. At lower system utilization levels (88% to 90%) the performance gap is less (less than 10%) but it goes on increasing as we move towards higher system utilization levels. At 100% system utilization, RMP schedule 67% task sets feasibly while the RMNP schedules 58% and CTR scheduling schedules 92% task sets. This decrease in performance of RMP and RMNP occurs due to their extreme behavior towards preemptions which result in deadline miss for low priority tasks. On the other hand CTR scheduling performs better by adopting a more sensible behavior towards preemptions.

In Fig 4(b), the schedulability results are shown for task sets with n=4 or 5. The dominance of CTR scheduling over RMP and RMPNP is clearly observable. At lower system utilization levels, RMP and RMNP perform reasonably well but as the system utilization increases their performance decreases hugely. On the other hand, CTR scheduling out classes both RMP and RMNP techniques. It schedules 10% more tasks than RMP and RMNP at 88% utilization while this gap increases to 30% at 100% system utilization. Fig 4(c) and Fig 4(d) summarize the results for the task set with n= 6 or 7 and n=8 or 9. The dominating performance of CTR scheduling as compared to RMP and RMNP scheduling is again observable.

B. Incomparable relation of CTR scheduling with Preemption threshold scheduling and Quantum-based scheduling

In this section we have compared the CTR scheduling with Preemption threshold scheduling and Quantum-based scheduling. We have shown that CTR scheduling has at-least an incomparable relation with these techniques. It means, there exists at-least one task set which is not schedulable with Preemption threshold scheduling and Quantum-based scheduling but CTR scheduling feasibly schedules it. Consider a task set given in TABLE IV. The given taskset is not schedulable with Preemption threshold scheduling. If the preemption threshold value of τ_3 is 1, then its WCRT is 8 and its deadline is missed. Similarly, at higher preemption threshold values of 2 and 3, τ_3 remains un-schedulable. On the other hand, if we apply CTR scheduling on the same task sets by assigning $\Delta r_1 = 2$, $\Delta r_2 = 0$ and $\Delta r_3 = 0$, the task set becomes schedulable as shown in TABLE IV.

TABLE IV. EXAMPLE TASK SET AND COMPARISON OF WCRT

Task (τ_i)	**WCET(C_i)**	**Period(P_i)**	**WCRT(Preemption threshold)**	**WCRT(CTR Scheduling)**
τ_1	1	3	1	3
τ_2	2	4	2	3
τ_2	1	6	8	4

Quantum-based scheduling also fails to schedule the task set given in TABLE IV. For the give task set, the upper bound and lower bound on quantum size is 2. TABLE V shows that the WCRT of τ_3 is 8 and its deadline is missed, when quantum size is 2.

TABLE V. WCRT OF TASK SET GIVEN IN TABLE IV UNDER QUANTUM-BASED SCHEDULING

Quantum Size	**Task(τ_i)**	**WCRT**
2	τ_1	1
	τ_2	3
	τ_3	8

Therefore, by this example, the at-least incomparable relation between the CTR scheduling and Quantum-based scheduling and Preemption threshold scheduling is proved.

*1) **Experimental Evaluation:***

To evaluate the performance of Preemption threshold scheduling (PTS) and Quantum-based scheduling (QBS) against CTR scheduling, we repeated the experiments given in previous section for these techniques. When PTS and QBS techniques are applied on same task sets, the obtained results are shown in Figure 5. It can be seen that both PTS and QBS performed better than RM preemptive and non-preemptive scheduling but at higher system utilization (95% and above) CTR scheduling dominates.

Fig 5(a) shows the experimental results of task sets with n = {2, 3}. It can be seen that at lower system utilization levels

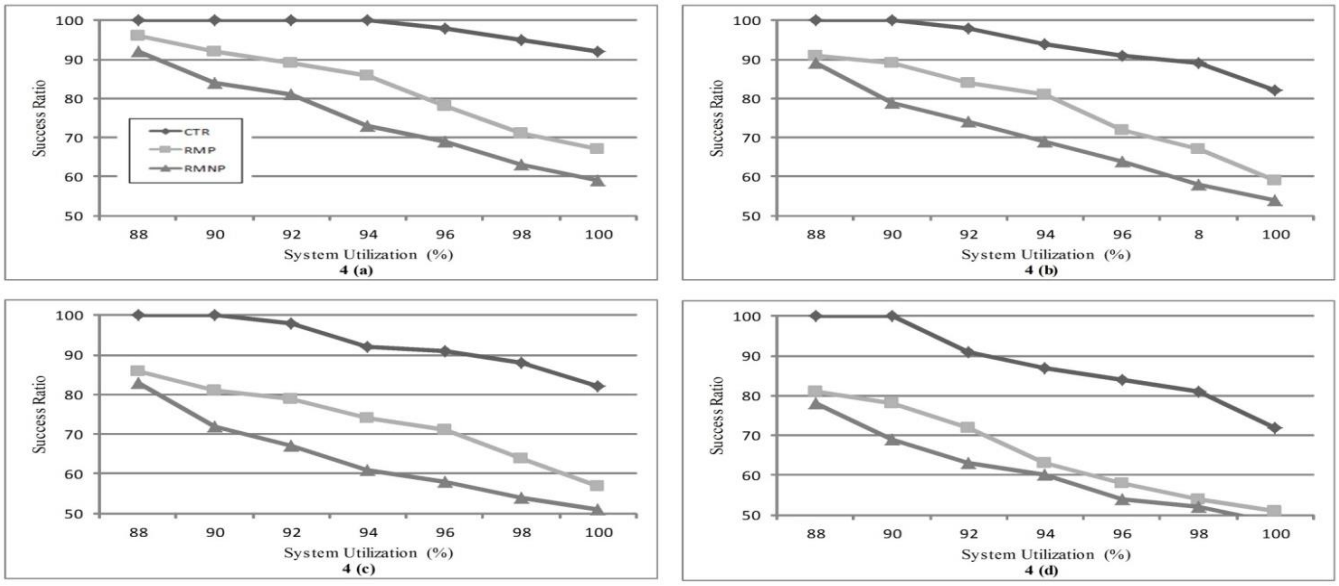


Fig. 4. Performance analysis of RM preemptive (RMP), RM non-preemptive (RMNP) and CTR scheduling on synthetic data sets in schedulability perspective (a) n={2,3} (b) n={4,5} (c) n={6,7} (d) n={8,9}

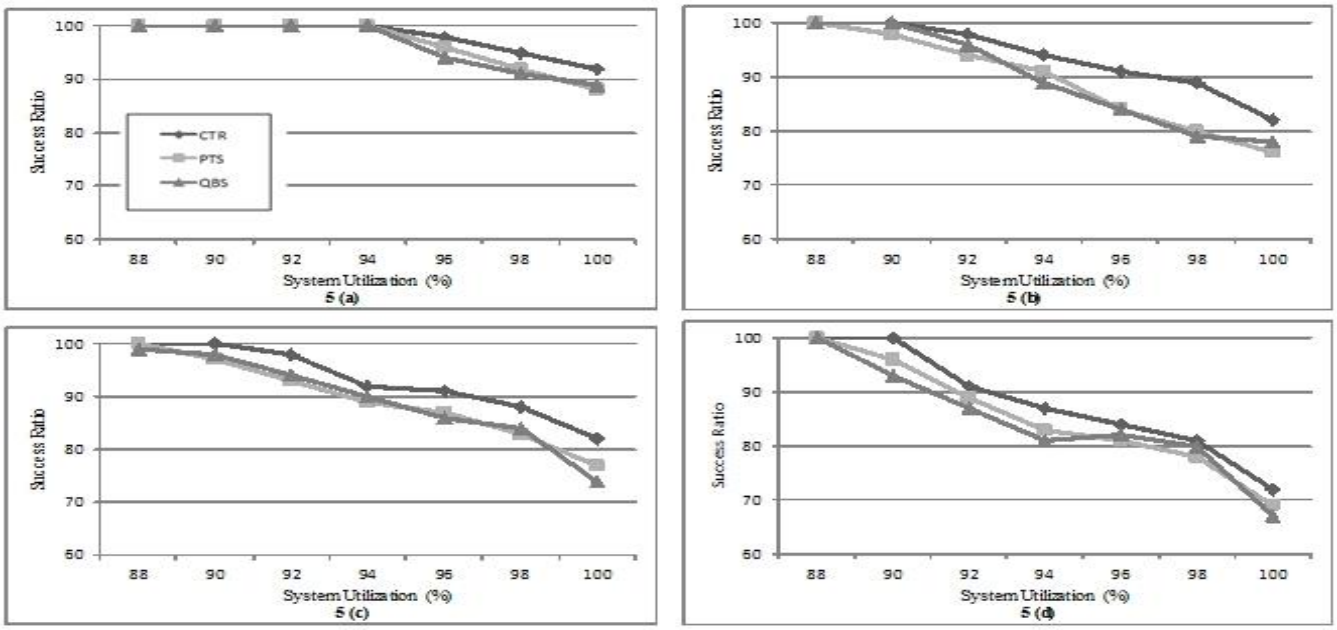


Fig. 5. Performance analysis of CTR, PTS and QBS scheduling on synthetic data sets in schedulability perspective (a) n={2,3} (b) n={4,5} (c) n={6,7} (d) n={8,9}

(88% to 94%) PTS, QBS and CTR scheduling performed very well but as we move further towards higher system utilization the CTR scheduling performs slightly better. At 100% system utilization the CTR scheduling schedules 92% tasks sets feasibly while PTS schedules 88% and QBS schedules 87% task sets. For task sets having 4 or 5 tasks, the obtained results are shown in Fig 5 (b). The similar performance of CTR, PTS and QBS scheduling at lower system utilization levels is easy to observe. At higher system utilization levels, again CTR scheduling performs better than PTS and QBS scheduling. Similar result are also obtained for task sets with n=6, 7 and n=8, 9 and are summarized in Fig 5(c) and Fig 5 (d).

As compared to RM preemptive and non-preemptive scheduling, preemption threshold scheduling and quantum-based scheduling adopt more flexible and wise behavior in preemptions perspective. As a result, preemption threshold scheduling and quantum based scheduling performed better

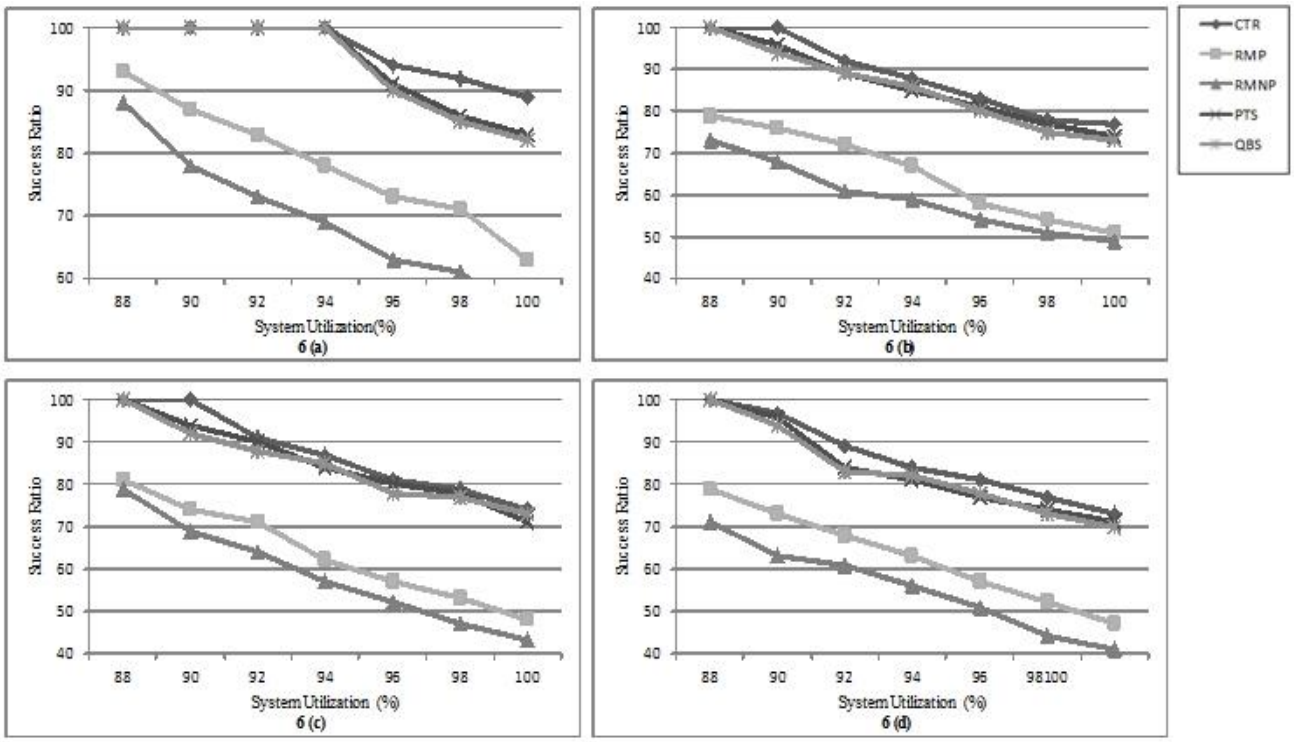


Fig. 6. Performance summary on synthetic constrained and arbitrary deadline task sets in schedulability perspective

but still could not attain the optimal system utilization. On the other hand, CTR scheduling achieves highest system utilization among all because it avoids preemptions till last moment and hence allows low priority tasks to complete, which results in higher utilization. Up-to 95% system utilization Preemption threshold, Quantum based and CTR scheduling showed similar performance but, at higher system utilization [96%, 100%] CTR scheduling achieves highest system utilization.

C. Scheduling constrained and arbitrary deadline tasks

The analysis given in previous sub-sections demonstrates the primacy of CTR scheduling in terms of schedulability over its alternatives but, these results are obtained only for implicit deadline tasks. In this section, we extend the analysis to handle tasks with constrained and arbitrary deadlines. Under the implicit deadline task model, for any task τ_i, D_i is always equal to P_i. This assumption makes the schedulability analysis very simple. However, many practical circumstances require relaxing this assumption. The constrained deadline task model permits D_i to be less than or equal to P_i while in arbitrary deadline model, D_i may be equal to or greater than P_i.

The mechanism to control the release of a task under CTR scheduling does not distress the RM schedulability of a task (by Theorem 1). This result is not specific to the implicit deadline model and remains true for constrained deadline tasks and arbitrary deadline tasks. Furthermore, the dominance of CTR scheduling over RM preemptive scheduling also holds for constrained deadline tasks and arbitrary deadline tasks. Because when a task set is schedulable with RM preemptive scheduling, it is also schedulable with CTR scheduling while a task set which is not schedulable with RM preemptive scheduling may be schedulable with CTR scheduling due to the gain achieved by delaying the task releases.

Corollary 1: *A RM schedulable task set τ, consisting of n periodic, independent, constrained deadline tasks is always schedulable with CTR scheduling but the vice-versa is not always true.*

Corollary 2: *A RM schedulable task set τ , consisting of n periodic, independent, arbitrary deadline tasks is always schedulable with CTR scheduling but the vice-versa is not always true.*

*1) **Experimental Evaluation:***

To evaluate the performance of CTR scheduling for the constrained task model and arbitrary deadline task model, we have created 10^4 task sets for each category. These tasks are generated in a similar way as explained in section IV. The Figure 6 summarizes the results. It can be seen clearly that, RM non-preemptive scheduling performs fine at low system utilization but the instant system utilization exceeds 91% the percentage of feasible task sets under RM non-preemptive scheduling starts decreasing and it tapers down to less than 53% at 100% system utilization. The performance of RM preemptive scheduling is better as compared to RM non-preemptive scheduling due to permitting preemption, but still at high system utilization the percentage of feasible tasks are low. At 100% system utilization, RM preemptive scheduling succeeds to schedule 62% task set feasibly. The performance of preemption threshold scheduling and quantum-based scheduling is comparatively better than RM scheduling, but below than the CTR scheduling at higher system utilization

level.

V. Conclusion and Future Work

Novel results are established for fixed priority scheduling by controlled task releases. The controlled release timings are exploited to improve the schedulability of fixed priority scheduling. It is proved that the pro-posed technique dominates RM preemptive technique in the sense that it schedules all task set that are schedulable with RM preemptive but vice versa is not true. As an example, it is shown that the proposed technique successfully schedule a given task system where preemption threshold scheduling or quantum based scheduling techniques fail. In this paper tasks are restricted to be only periodic; however, as a future work more interesting are expected when applied to sporadic tasks systems.

References

[1] Davis, R.I., Burns, A., Baruah, S., Rothvoss, T., George, L., and Gettings, O., Exact Comparison of Fixed Priority and EDF Scheduling based on Speedup Factors for both Pre-emptive and Non-pre-emptive Paradigms, Real-Time Systems Journal, 51(5), pp566-601, 2015.

[2] Wang, Y., and Saksena, M.: 'Scheduling fixed priority tasks with preemption threshold'. In proceedings of the 6th international conference on real time computing systems and applications, Hong Kong, China, Dec 1999, pp. 328-335

[3] Audsley, N.C., Burns, A., Tindell, K., and Wellings, A.: 'Applying new scheduling theory to static priority preemptive scheduling', Software Engineering Journal, 1993, 8, (2), pp. 80-89

[4] Bini, E., and Buttazo, G.C.: 'The space for Rate Monotonic schedulability'. 23rd IEEE Real-Time Systems symposium, Austin, TX, USA, Feb 2002, pp. 169-178

[5] Park, M.,Yoo, H.J., and Chae, J.: 'Analysis on Quantum-Based fixed priority scheduling of Real-Time tasks'. Proceedings of the 3rd international conference on ubiquitous information management and communication, Suwon, SKKU, Korea, Jan 2009, pp. 627-634

[6] Davis, R.I, A review of fixed priority and EDF scheduling for hard real-time uniprocessor systems, ACM SIGBED review, 11(1), pp. 8-19, 2014.

[7] George, L., Riverre, N., Spuri, M.: 'Preemptive and Non-preemptive Real-Time Uniprocessor Scheduling', Research Report RR-2966, INRIA, France, 1996

[8] Huhang, W.H., Chen, J., Zhou, H., and Liu, C., PASS: Priority Assignment of Real-Time Tasks with Dynamic Suspending Behavior under Fixed-Priority Scheduling, Technical Reports in Computer Science, Dortmund University of Technology, 2015.

[9] Bini, E., and Buttazzo, G.C.: 'Schedulability Analysis of Periodic Fixed Priority Systems', IEEE Transactions on Computers, 2004, 53, (11), pp. 1462-1473

[10] Liu, C.L., and Layland, J.W.: 'Scheduling algorithms for multiprogramming in a hard real-time environment', Journal of the ACM, 1973, 20,(1), pp. 40-61

[11] Tindell, K.W., Bums, A., Wellings, A.: 'An extendible approach for analyzing fixed priority hard real-time tasks', Real-Time Systems Journal, 1994, 6, pp. 133-151

[12] Bini, E., and Buttazzo, G.C.: 'A Hyperbolic Bound for the Rate Monotonic Algorithm'. In Proceedings of the 13th Euromicro Conference on Real-Time Systems, Delft, Netherlands, June 2001, pp. 59-66

[13] Min-Allah, N., Ali, I. , Jian-Sheng, X., Yong-Ji, W.: 'Online Feasibility Analysis with Composite-Deadline'. In Proceedings of the 4th International Conference on Innovations in Information Technology, Dubai, UAE, Nov 2007, pp. 357-361

[14] S. Baruah, V. Bonifaci, G. D'angelo, H. Li, A. Marchetti-Spaccamela, S. van der Ster, and L. Stougie. Preemptive uniprocessor scheduling of mixed-criticality sporadic task systems. Journal of the ACM, 62(2):14:1–14:33, 2015.

[15] Lehoczky, J.P., Sha, L., Ding, Y.: 'The Rate Monotonic Scheduling Algorithm: Exact Characterization and Average Case Behavior'. In Proceedings of the IEEE Real-Time System Symposium, California, USA, Dec 1989, pp. 166-171

[16] Kim, J.E., Abdelzaher, T., and Sha, L., Budgeted Generalized Rate Monotonic Analysis for the Partitioned, yet Globally Scheduled Uniprocessor Model, Real-Time and Embedded Technology and Applications Symposium (RTAS), 2015 IEEE, pp. 221-231, 2015.

[17] Sha, L., Abdelzaher, T., Erzen, K., Cervin, A., Baker, T., Burns, A., Buttazzo, G.C., Caccamo, M.,Lehoczky, J., Mok, A.K.: 'Real-Time Scheduling Theory: A Historical Perspective', Real-Time Systems, 28 (2), pp. 101–155, 2004.

[18] Wan-Chen, L., Kwei-Jay, L., Hsin-Wen, W., Wei-Kuan, S.: 'Rate monotonic schedulability tests using period-dependent conditions', Real-Time Systems journal, 37 (2), pp. 123-138, 2007.

[19] Thekkilakattil, A., Dobrin, R., and Punnekkat, S., The limited-preemptive feasibility of real-time tasks on uniprocessors. *Real-Time Syst*, 51(3), pp. 247-273, 2015.

[20] Burns, A., and Wellings, A.: 'Dual Priority Assignment: A Practical Method for Increasing Processor Utilization'. In Proceedings of 5th Euromicro Workshop on Real-Time Systems, Oulu, Finland, June 1993, pp. 48-55

[21] Baruah, S.: 'The limited-preemption uniprocessor scheduling of sporadic systems'. In ECRTS 05, Pro-ceedings of Euromicro Conference on Real-Time Systems, Balearic Islands, Spain, July 2005, pp. 137-144

[22] Buttazzo, G., Bertonga, M., and Yao, G., Limited Preemptive Scheduling for Real-Time Systems. A Survey, IEEE transactions on industrial informatics, 9(1), pp. 3-15, 2013.

[23] Marinho.J, Petters, S.M, and Bertogna, M.: 'Extending Fixed Task-Priority Schedulability by Interference Limitation'. In Proceedings of the 20th International Conference on Real-Time and Network Systems, pp. 191-200, 2012.

An Emergency Unit Support System to Diagnose Chronic Heart Failure Embedded with SWRL and Bayesian Network

Baydaa Al-Hamadani
Department of Computer Science
Zarqa University
Zarqa, Jordan

***Abstract*—In all the regions of the world, heart failure is common and on raise caused by several aetiologies. Although the development of the treatment is fast, there are still lots of cases that lose their lives in emergence sections because of slow response to treat these cases. In this paper we propose an expert system that can help the practitioners in the emergency rooms to fast diagnose the disease and advise them with the appropriate operations that should be taken to save the patient's life. Based on the mostly binary information given to the system, Bayesian Network model was selected to support the process of reasoning under uncertain or missing information. The domain concepts and the relations between them were building by using ontology supported by the Semantic Web Rule Language to code the rules. The system was tested on 105 patients and several classification functions were tested and showed remarkable results in the accuracy and sensitivity of the system.**

Keywords*—Ontology Engineering; Bayesian Network; Heart Failure; Expert System; Validation Test

I. INTRODUCTION

In dealing with real world applications, one inescapably has to deal with uncertain or missing information. In diagnosing Expert System (ES), it is impossible to model all the conditions and variables in specific values, and because of the large number of these variables, probabilistic models are not suitable, while Bayesian Network (BN) does [1, 2].

A Bayesian Network is a graphical representation of probabilistic information expressed as directed acyclic graph with nodes that represent the variables with uncertain or missing values, and edges between then to represent the probabilistic value that represent the influence of these variables [3]. For several reasons BN plays a significant role in modelling uncertainty in ES and Decision Support Systems. The graphical representation that shows the conditional independencies between the nodes are easy to be understand by the user of the system. Moreover, since BN defines a unique joint between two specific nodes, the consistency and correctness of inference are guaranteed due to the mathematical calculation dependencies [4].

There are lots of efforts in recent years into designing ESs that can assist experts in different fields to make their decision. The reasons behind this are reducing clinical errors the patient's waiting time, and unnecessary medical and laboratory tests. Heart failure diagnosing require special attention from the ES builders since the patients of this disease need to be diagnosed, treated and monitored continuously and with fast response. The main symptoms are breathlessness in *specific cases*, *extreme* tiredness and ankle swelling, which *may* extend up the leg and get *worst* at night. As noticed, the description of the symptoms requires specifying some probability values to represent the level of relations between variables.

Using ontology in the design of expert systems is a hot issue. In the field of AI, ontology is a collection of classes, attributes, and the relationships between them. It represents the vocabulary for transferring thought and performing reasoning in a domain [5]. The main reason behind using ontology to represent the knowledge and the relationships in expert systems is its ability to reuse the domain knowledge by sharing the common understanding in a specific field. Its interoperability feature allows the ontology systems to be spread and developed more powerfully.

This paper presents a framework for designing and implementing an expert system to be used in emergency units to help practitioners to diagnose heart failure disease in the presence of uncertain or missing information. The design depends on two different technologies, Bayesian network to model uncertain values for some required variables, and ontology to model the concepts of the ES and the relationships between them.

II. BACKGROUND

A. Bayesian Network

A Bayesian network is a probabilistic model P showed at a directed acyclic graph (DAG). In another word, BN of n variables consists of a DAG of n nodes and a number of arcs. Each node in a DAG represents a random variable X_i; and a directed arc between two nodes X_i , X_i represents the direct influence or causal from X_i to X_i . There is a probability distribution P associated with each node i, such that: $P(X_i|\pi(X_i))$, where $\pi(X_i)$ is the parent set of X_i . All the probability factors of a given DAG are listed in a Conditional Probability Table (CPT) and the joint probability distribution or probability inference of a BN is the product of its CPT:

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P\left(x_n | x_{\pi(i)}\right) \quad \dots\dots\dots (1)$$

Using joint probability distribution, BNs support probabilistic inference in the joint space. To relax a BN, Bayesian classifiers required. In the designing of our model we adopted three types of classifiers. The first one, which is the simplest and most used one, is the naïve Bayesian classifier which assumes that the independence features are conditionally independent. The second classifier is the Tree Augmented Network (TAN) which is performed by adding directional edge between variables that are not belong to the same class node [6]. The last classifier used is the Forest Augmented Network (FAN) to perform better ranking by joining several trees together [7].

B. Expert Systems

Since the beginning of the using computers, scientist and physicians started to think of ways to utilize computers in assisting them to do their work. The first article appeared in the field of diagnosing process was in 1959 [8] that suggested a technique that helps physicians in diagnosing diseases and focused on pointing the light on the benefits of using computers in medical fields. Moreover, several early systems appeared. The most well-known system was developed in Leeds University in 1972 [9] to diagnose abdomen pain using Bayesian probability theory. Later in 1976, another well-known medical diagnosing system appeared, MYCIN [10]. MYCIN embedded the field of Artificial Intelligent (AI), which uses abstract symbols rather than numerical calculations, to build the production rules and to strength the reasoning process to identify bacteria causing severe infections. In 1991, A Dynamic Hospital Information System (HELP) appeared [11] that has the ability to generate alerts when abnormal signs in the patient record are noted.

There are five main components in any rule-based expert system. The ***knowledge base*** which has the rules and any other form of information collected from the human experts in the field. Knowledge can be either abstract or concrete. While abstract knowledge can be represented by rule and probability distribution, concrete knowledge refers to the information related to a specific abstract knowledge (facts). The heart of every expert system is the ***Inference Engine*** which draws conclusions by applying abstract knowledge to concrete knowledge. These conclusions can be based on either deterministic knowledge (knowledge about certain facts), probabilistic knowledge (knowledge about uncertain facts), or nondeterministic knowledge (fuzzy knowledge) [12]. Another component in the building blocks of an expert system is the ***Explanation Mechanism*** which provides the user with the necessary explanation about the way a specific conclusion drawn and the reasons behind using specific facts. When the expert system deals with users, it should be accompanied with a ***User Interface*** component which should be user friendly and easy to use.

C. Ontology and its Engineering

"An ontology is a formal explicit representation of concepts in a domain, properties of each concept describes characteristics and attributes of the concept known as slots and constrains on these slots" [13]. In the field of computer science and information technology, Ontology is the process of representing knowledge in a specific domain [14]. Ontology is used to represent the sharable knowledge in terms of concepts which represented by classes, relations which represent the relations between the concepts, instances which are the objects represented by the concepts, and axioms which represent the rules that tie the concepts to the instances [15].

Fonseca [16] defines ontology as "an ontology refers to an engineering artefact, constituted by a specific vocabulary used to describe a certain reality". He identified the difference between ontology and information systems in modelling and reasoning about information, as ontology deals with the information in conceptual level, while information systems do this task in implementation time.

In their work, Rousey et.al. [17] gave different perspectives of the word "Ontology" in the field of Compute Science. "For example, ontology can be: a thesaurus in the field of information retrieval, a model represented in OWL in the field of linked-data, or a XML schema in the context of databases".

On the other hand, knowledge Engineering is the steps that should be followed to build ontology. There are several methods use in this aspect [18], In this paper Toronto Virtual Enterprise (TOVE) method was used [19]. The reason behind choosing this methodology and its steps are illustrated in the forthcoming sections.

D. Web Ontology Language (OWL)

Among several ontology languages, OWL is the most used one. It is a World Wide Web Consortium (W3C) recommendation in 2004 to be "used by applications that need to process the content of information instead of just presenting information to humans" [20]. It has several features over XML and RDF by providing additional vocabulary along maintain their properties. It is then used to define instances (individuals) and maintain their properties, and then it is used to reason about these classes and individuals [17]. OWL has three sub-languages:

1) OWL-Lite: This sub-language intended for users who need simple modelling and constraints. Although it provides quick path to thesauri and other taxonomy, its cardinality is limited to either 0 or 1.

2) OWL-DL: To fill the shortage of OWL-Lite, this sub-language comes with features that enrich the use of OWL. Class Boolean combinations and class property restrictions are some of the added features. Other properties in describing a class in term of other disjoint classes are another new feature. With all these features, OWL-DL becomes the most used language since it provides the user with full expressiveness [20].

3) OWL-Full: this sub-language offers to its users maximum expressiveness and syntactic freedom of RDF[17]. As instance, OWL-Full treats a class as a set of individuals and as an individual at the same time. Its data type property generalizes to include inverse functional property.

E. Semantic Web Rule Language (SWRL)

Based on the combination between OWL-DL and OWL-Lite sublanguages, SWRL was developed to be the rule language of the semantic web. It allows the users to write the first-order logic rules required to reason about the specified OWL individuals. The semantic rules of SWRL are the same as the description logic of Owl to make the reasoning process easier and stronger.

Each SWRL rule has an antecedent and a consequent, each of which could be a disjunction of several atoms. There are several atom types that are supported y SWRL, such as class atoms, individual property atoms, data value property atoms, and data range atoms. The most powerful atoms are built-in atoms, where SWRL provides several types of existing built-in and allow the user to design and use his own built-ins [21].

III. Related Work

Several BN expert systems were proposed to provide the experts with the required decision especially in medical field. Some of these systems were based on building their knowledge base using ontology. In diagnosing heart diseases, [22] proposed a Decision Support System to be used by the cardiovascular experts and the data obtained from the Rapid Access Chest Pain Clinic in England. The medical conditions were modelled as binary clauses, either yes or no. No other values were used in this system to model the uncertain variables.

Jayanta and Marco [3] presets a framework for an expert ES that assists the expert to assess the several minerals levels in the patient's body. The data of the study was dedicated to elderly people over 65 years old and no ontology engineering used.

To the best of our knowledge, all the proposed Expert Systems which focus on heart diseases were designed to be used by the expert in the field. There is only one proposal for a system that can be used in emergence units. In their work, Joan et al. [23] proposed a Bayesian-based ES to be used in the emergency units to assist practitioners to diagnose unstable angina. In this paper we propose an ES that is supplemented by BN and Ontology engineering to diagnose heart failure in the emergency units.

IV. Methodology for Ontology Engineering

A. System Architecture

The architecture of the proposed system consists of several modules. Figure 1 illustrates these modules and the interactions between them. As any other expert system, the core of our system is the knowledge base module which consists of the fact base and the rule base. The facts are extracted, using the user interface, from the user as the patient's symptoms in addition to the laboratory and clinical test results.

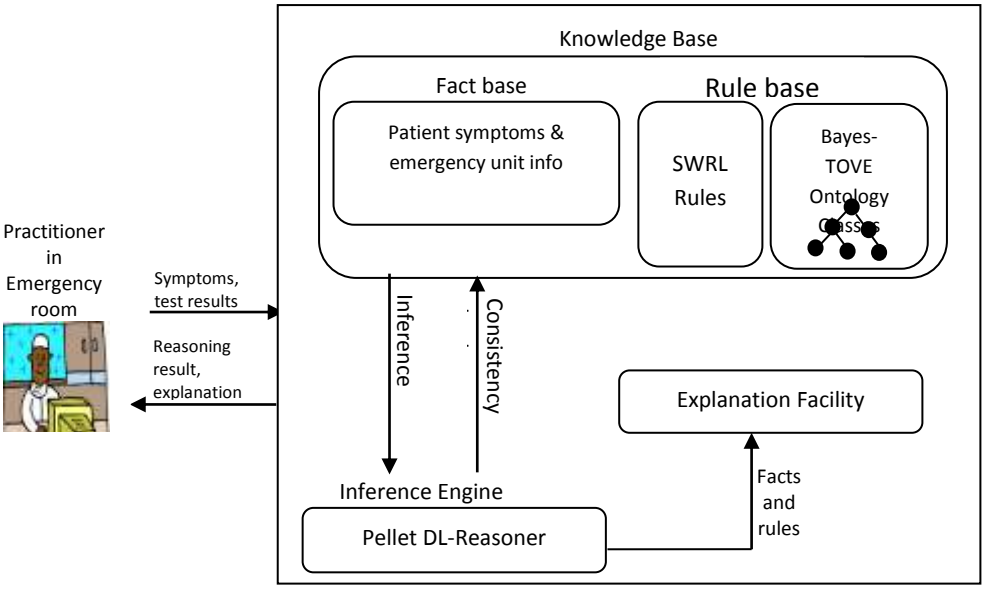


Fig. 1. The system architecture

The rules base consists of SWRL decision rules and the ontology structured classes along with the relationships between these classes. The decision rules were inferred from technical guidelines published by the National Institute of Health and Care Excellence in the UK [24, 25]. While the ontology classes were formed using Protégé Ontology editor.

The inference engine is the core of any expert system which depends on the facts and the rules to reason the required decision. In our work we use Pellet [26], which considered to be one of the best OWL-DL reasoned with several features such as data-type reasoning and debugging, rules integration, and reasoning conjunctive queries. In this stage more decision rules could be inferred and added to the list of available rule base. The final decision results will be introduced to the user through the user interface alongside with the explanation about this decision inferred from the explanation module.

B. Ontology Engineering

The methodology used to build our ontology-based ES is TOVE, designed by Gruninger and Fox [19]. The main goal of TOVE is to develop a set of integrated ontologies and it has several characteristics that make it widely used as ontology engineering. It provide the ability to create a sharable representation of the ontology, to define the meaning of each semantic in first-order logic, to reason about the semantics automatically, and to depict the context in graphical context. According to several literatures [19], TOVE has six stages to be followed and these stages are modified in this paper to depict the reasoning under uncertain or insufficient information. Bayes-TOVE stages are illustrated in Figure 2.

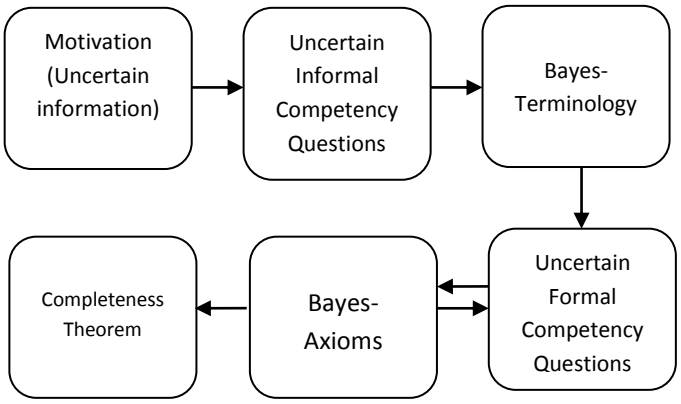


Fig. 2. Bayes-TOVE

1) Motivation

In the motivation stage, the requirement of the system should be set either as questions, story problem, or examples. The motivation scenarios of our proposed Ontology are:

a) The lack of ES that can be used in the emergency units to diagnose heart failure or deal with patients already diagnosed to have the disease.

b) Since heart failure should be diagnosed fast and since its symptoms are similar to other diseases, the things that make the practitioners get confused in the diagnosing process.

c) Patients with heart failure and are in emergency room should be treated taking into consideration the existing disease and the new symptoms that require them to be entered in emergency unit.

d) When a patient get entered to an emergency room, she or her carer may not get all the information that are required in the diagnosing process such as laboratory and clinical test results. This led to use BN to represent some uncertain or missing variables.

2) Uncertain Informal Competency Questions

In this stage the motivation scenario is changed into informal competency queries that the ontology should answer. The resultant queries provide a clear idea about the new designed ontology and represent the first evaluation step to determine the importance of the ontology and if it can be replaced by existing ones [27]. Some of the informal questions of our system are:

a) Does the patient already diagnosed to have a heart failure?

b) If the answer to question1 is yes, then:

—What are the results of the patient's laboratory test results?

—What drugs are taking by the patient?

—What are the new symptoms?

—Does the new symptoms relate to heart failure?

—If the patient requires new drug, what are the conflict effects with the current drugs.

—What treatment should be taken?

c) If the answer to question1 is no, then:

—What are the patient's symptoms?

—What are the similarities between the existing symptom and the heart failure symptoms?

—What treatments should be taken?

—What drug should be given to the patient?

These queries emphasise the concepts and the relationship between these concepts that are going to be embedded in the ontology. Several of these questions cannot be answered in emergency unit accurately but the reasoning still necessary to be completed.

3) Bayes-Terminology

In this stage the proposed informal questions from the previous stage should be described through either First-Order-Logic or through Knowledge Interchange Format (KIF) axioms. In this project we used Standards Upper Ontology KIF (SUO-KIF). The reason behind authoring this language was to understand the meanings of expressions without the need for a manipulating interpreter [28]. As instance, the following rule:

"Refer patients with suspected heart failure and high BNP level or high NTproBNP level, to have transthoracic Doppler 2D echocardiography and specialist assessment within 2 weeks". [24]

Could be written in SUO-KIF as the following macro-like structure:

$$\Rightarrow$$

$$\begin{pmatrix} and \\ (exist\ (?x\ ?d\ ?t1\ ?t2) \\ (instance\ ?x\ patient) \\ (instance\ ?d\ HF) \\ \begin{pmatrix} ((instance\ ?t1\ BNP\ v1)\ and\ (value\ v1\ high)) \\ or\ \begin{pmatrix} (instance\ ?t2\ NTproBNP\ v2) \\ and\ (value\ v1\ high) \end{pmatrix} \end{pmatrix} \\ (and \\ (treatment\ ?x\ ?m\ ?t) \\ (instance\ ?m\ Dopler2D) \\ (instance\ ?t\ 2Weeks)) \end{pmatrix}$$

where, $treatment$ is considered as a function, while

instance and *exist* are relations. The SUO-KIF structure can be represented as a tree in which its nodes are the instances and the edges are the functions, relations, or operations on these instances. This stage is necessary for ontology reusability since the KIF representation is generic across many domains. This example emphasises the demand need for reasoning under uncertain information. The actual values of "high BNP" should be 100 and 400 pg/ml and it is not necessarily refer to HF, it could refer to other diseases such as diabetes if the patient age is over 70. Figure 3 shows the BN acyclic graph for the given rule.

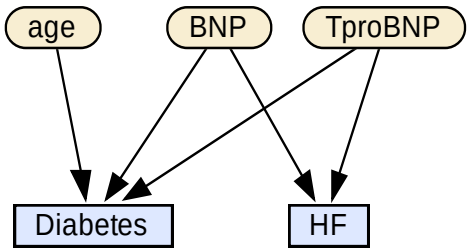


Fig. 3. Example of BN graph

4) Uncertain Formal Competency Questions

The main aim of this stage is to make sure that the ontology system is consistent based on the axioms of the ontology [19]. This stage should specify the following: the set of terminologies based on the axioms in the proposed ontology ($T_{Ontology}$), the set of instances (T_{Ground}), the Bayesian network (T_{BN}), and it determines the following:

$$T_{Ontology} \cup T_{Ground} \cup T_{BN}\ \Phi Con\ Q \ldots\ldots (2)$$

such that ΦCon checks the consistency between the terminologies and instances on one side and first-order sentences in the language on the other side (Q).

Using Formal Competency Questions, the ontology can be distinguished and the relationships with other ontologies can be specified. Several approached, however, have been proposed to use the Competency questions as an ontology evaluator and mechanisms were proposed to check if a given ontology meets its competency questions. [29]

5) Bayes Axioms

Axioms in the ontology are the definition of the concepts and relations and constraints between them [19]. Moreover, axioms should represent the semantic of the objects and their relations. Although it is considered to be the most difficult process in building ontologies, axioms are considered to be the most important and significant part as well. In this paper, axioms are going to be represented as a tree-like structure rather than first-order logic statements, as suggested in [30]. Figure 5 shows the structure of the axioms used in the proposed ES.

V. System Implementation

The architecture of the system, which is depicted in Figure 1, has several stages. The first stage is to implement the ontology classes, object properties, data properties and their characteristics alongside with the relations between them (see Figure 4). OWL-DL is used as an ontology language since it has several features listed in section 2. Protégé is used to create the ontology system. Embedded with some reasoners, Protégé is an open source, W3C recommendation, with several features that enables its users to run and check consistency of the system.

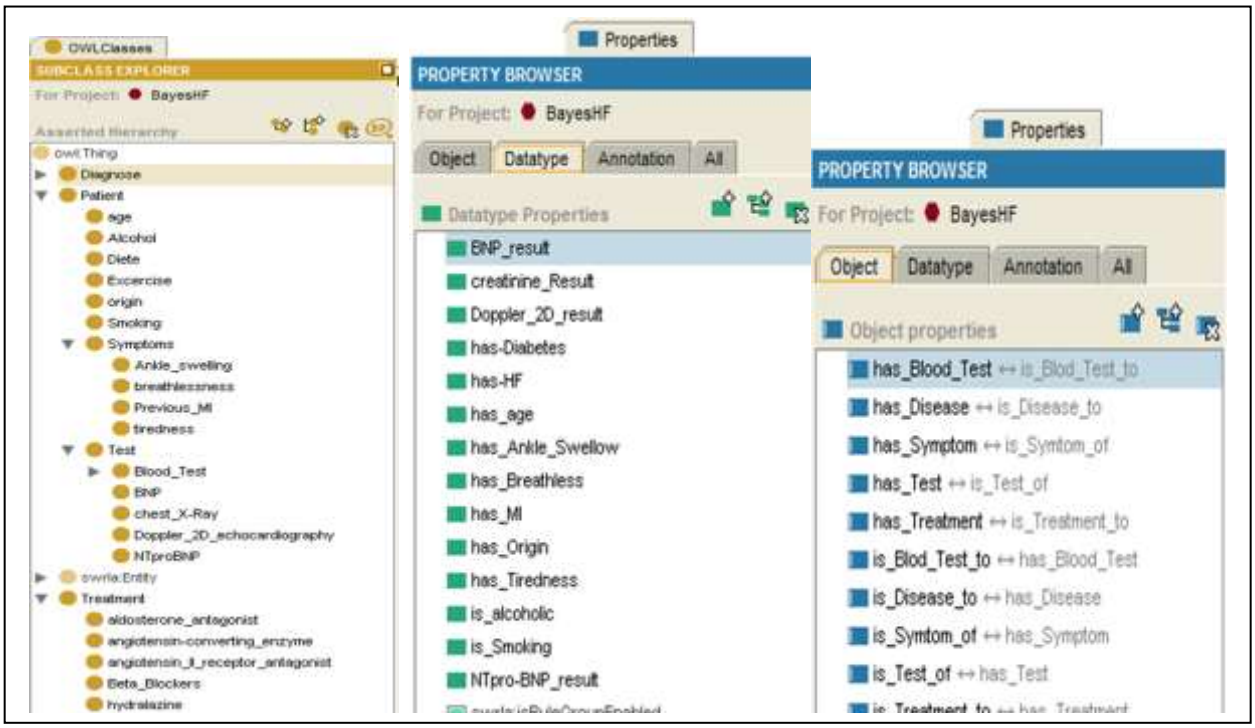


Fig. 4. Sample of OWL classes, data properties, and object properties for the ES

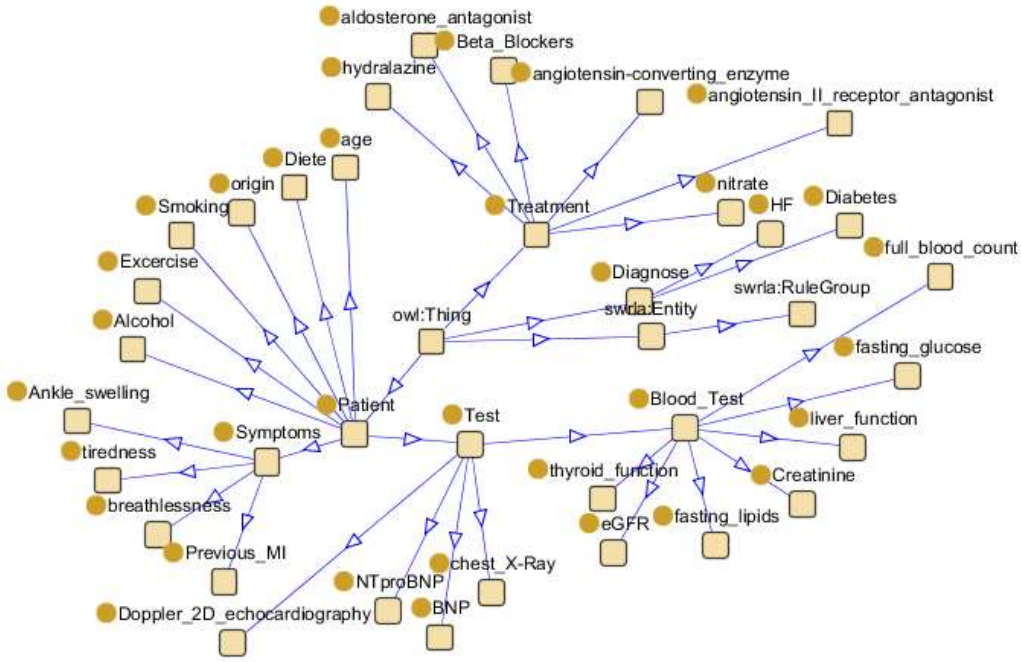


Fig. 5. Bayes axioms for the proposed system

The second stage was to build the inference rules for the Expert System using SWRL rules. All the rule were built depending on the guideline published by the National Institute for Health and Care Excellence [24]. Figure 6 illustrates sample of the inference rules used and an inferred axiom from implementing the rules. Figure 7 shows the result of implementing the inference rules on one of the patients. The red circle in the figure represent the recommendation to the practitioner that the patient should take the Doppler-2D-Echocardiography to make sure if she has HF or not.

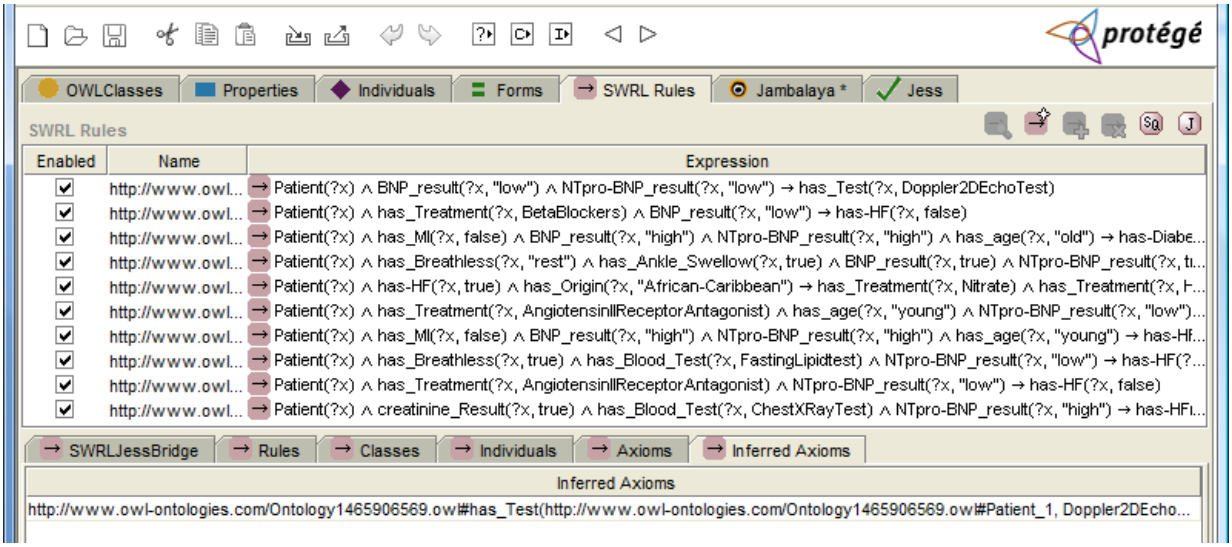


Fig. 6. Sample of SWRL rules used in the ES

The rule engine used was Java Expert System Shell (jess), a small, light, and one of the fastest rule engines available [31]. It is a powerful program developed by one of the team members at Sandia Laboratories in Canada.Netica API [32] was used to construct the junction tree algorithm and belief propagation of the BN. It was used to calculate the probabilities of occurrence of HF disease according to the given symptoms and laboratory tests.

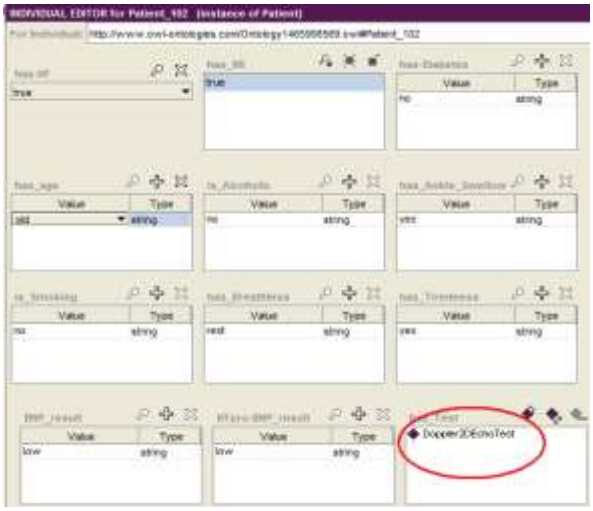


Fig. 7. Example of an inferred axiom

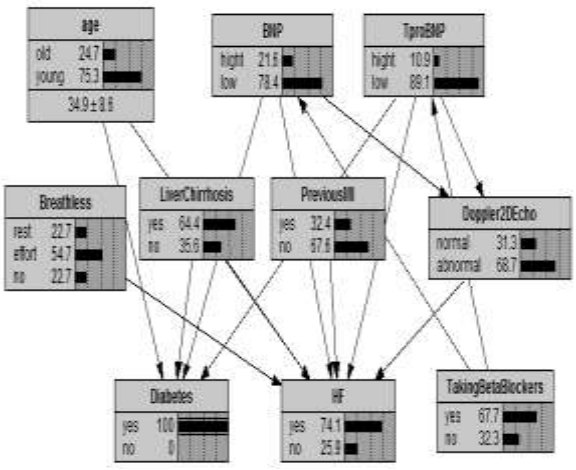


Fig. 8. Part of the BN to reason under uncertainty

Figure 8 shows sample of the BN acyclic directed graph used to reason under uncertain information. The reasoning process depends on the Conditional Probability Table (CPT) build using Netica and Figure 9 illustrates part of it.

BNP	TproBNP	Breathless	LiverChirrhosis	PreviousMI	Doppler2DEcho	age	yes	no
hight	hight	rest	yes	yes	normal	old	55	45
hight	hight	rest	yes	yes	normal	young	45	55
hight	hight	rest	yes	yes	abnormal	old	98	2
hight	hight	rest	yes	yes	abnormal	young	95	5
hight	hight	rest	yes	no	normal	old	40	60
hight	hight	rest	yes	no	normal	young	38	62
hight	hight	rest	yes	no	abnormal	old	98	2
hight	hight	rest	yes	no	abnormal	young	95	5
hight	hight	rest	no	yes	normal	old	55	45
hight	hight	rest	no	yes	normal	young	45	55
hight	hight	rest	no	yes	abnormal	old	98	2
hight	hight	rest	no	yes	abnormal	young	95	5
hight	hight	rest	no	no	normal	old	40	60
hight	hight	rest	no	no	normal	young	38	62
hight	hight	rest	no	no	abnormal	old	98	2

Fig. 9. Part of CPT Creating using Netica API

VI. Testing and Validation

The dataset used to test the ES was belonged to 105 patients collected from Emergency units in three private hospitals in Jordan and the system has been used by several practitioners. The performance of the system was calculated using the area under the curve (AUC) of Receiver Operating Characteristics (ROC) graph and compute the factors associated to it [33]. AUC reflects the percentage of correct classification and its value ranged between 1, indicating optimum classification of all cases, and 0, indicating completely random classifications. Keep in mind that no realistic system should have AUC value less than 0.5 representing 50% random classification [33]. Depending on the tested data, AUC value of implementing the ES was 0.7164, which considered to be equalized with other similar systems. Moreover several classification functions were used to test the ES (see Table 1) depending on the values of TP, FP, TN, and FN:

TABLE I. Achievement Factors of Testing the ES

Factor	Value
Sensitivity	0.833
Specificity	0.666
Accuracy	0.761
Positive prediction	0.769
Negative prediction	0.75

1) Sensitivity: refers to the ability of the system to correctly identify the patients having HF based on the given symptoms and laboratory test results. For our ES, this value is high comparing with other related systems, which makes the system reliable.

2) Specificity: refers to the ability of the system to correctly identify the patient do not having HF. The result of testing this function (Table 1) shows that only 33% of the patients were incorrectly identified having HF and this is due to the using of BN that lowered the error rate.

3) Accuracy: this classification function measures the statistical bias of the system, i.e., how close the results are to the true values. More than 75% of the tested patients had correct results, either having HF or do not and with correct identification.

4) Positive and negative predictive values (PPV) and (NPV): the PPV is the probability that a patient correctly diagnosed to have HF and NPV is the probability of a healthy person correctly diagnosed not to have HF. The resulting ratios of these functions on the data set showed that around 75% of the tested patients were correctly diagnosed.

VII. Conclusion and Future Work

This paper presents an expert system to help physicians in hospital's Emergency units to diagnose Heart Failure disease. The system is based on BN with discrete nodes to enable the users to reason under uncertain or incomplete information. The SWRL rules are used to build the inference rules and Jess is used as an inference engine. The validation tests are done using 105 cases and the results of the classification functions show that the system has high level of validity. The model achieves more than 75% of PPV, NPV, and accuracy; while it

achieves 83% in sensitivity function. To make the system usable even more easily, a web application is going to be constructed as a future work and to link the system with the data base systems that are pre-prepared by the hospitals to increase the accuracy of the system and to get benefit of previous information stored for each patient.

ACKNOWLEDGEMENT

This research is funded by the Deanship of Research in Zarqa University /Jordan.

REFERENCES

[1] Russell, S.J. and P. Norvig, Artificial Intelligence: A Modern Approach. 3rd ed: Prentice Hall. 2010,

[2] Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference: Morgan Kauffmann Publisher. 1988,

[3] Jayanta, K.G. and V. Marco, Building a Bayesian network model of heart disease, in Proceedings of the 38th annual on Southeast regional conference. 2000, ACM: Clemson, South Carolina.

[4] Wiegerinck, W., B. Kappen, and W. Burgers," Bayesian Networks for Expert Systems: Theory and Practical Applications". Handbook on Neural Information Processing, vol. 49 no: pp. 401-431.2013

[5] http://www.slideshare.net/jcmorris/the-role-of-ontology-in-modern-expert-systems-dallas-2008-presentation.

[6] Cerquides, J. and R.L.o.d. M`antaras, Tractable Bayesian Learning of Tree Augmented Naive Bayes Classifiers. 2003, INSTITUT D'INVESTIGACIO´ EN INTEL.LIGENCIA ARTIFICIAL (CSIC)

[7] Ziebart, B., A. Dey, and J.A. Bagnell. "Learning Selectively Conditioned Forest Structures with Applications to DBNs and Classification, ". in Proceeding of the The 23rd Conference on Uncertainty in Artificial Intelligence. pp. 2007

[8] Ledley, R.S. and L.B. Lusted," The Use of Electronic Computers to Aid in Medical Diagnosis". Proceedings of the IRE, vol. 47 no 11: pp. 1970-1977.1959

[9] Dombal, F.T.d., D.J. Leaper, J.R. Staniland, A.P. McCann, and J.C. Horrocks," Computer-aided Diagnosis of Acute Abdominal Pain". Br Med J., vol. 2 no 5804: pp. Pages: 9-13.1972

[10] Shortliffe, E., Computer-Based Medical Consultations: MYCIN: Elsevier. 264 pages. 1976,

[11] Gilad, J.K., M.G. Reed, and T.A. Pryor, HELP: A Dynamic Hospital Information System: Springer Publishing Company, Incorporated. 357. 2012,

[12] Castillo, E., Expert Systems and Probabilistic Network Models: Springer Science & Business Media. 1997,

[13] Noy, N.F., R.W. Fergerson, and M.A. Musen, The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility, in Knowledge Engineering and Knowledge Management Methods, Models, and Tools. Springer Berlin Heidelberg. pp. 17-32.2000

[14] Gruber, T.R.," A Translation Approach to Portable Ontology Specifications". Knowledge Acquisition, vol. 5 no 2: pp. Pages 199-220.1993

[15] Davies, J., Lightweight Ontologies, in Theory and Applications of Ontology: Computer Applications, M. Healy, R. Poli, and A. Kameas, Editors., Springer Science & Business Media.2010

[16] Fonseca, F.," The Double Role of Ontologies in Information Science Research". Journal of the American Society for Information Science and Technology, vol. 58 no 6: pp. Pages 786-793.2007

[17] Roussey, C., F. Pinet, M.A. Kang, and O. Corcho, An Introduction to Ontologies and Ontology Engineering, in Ontologies in Urban Development Projects (Advanced Information and Knowledge Processing), G. Falquet, C. Métral, and J. Teller, Editors., Springer-Verlag London Limited.2011

[18] Uschold, M. and M. Gruninger," Ontologies: principles, methods and applications". The Knowledge Engineering Review, vol. 11 no 02: pp. 93-136.1996

[19] Gruninger, M. and M.S. Fox. "Methodology for the Design and Evaluation of Ontologies". in Proceeding of the In Workshop on basic ontological issues in knowledge sharing (IJCAI-95). pp. 1995

[20] http://www.w3.org/TR/owl-features/.

[21] Gil, Y., E. Motta, V.R. Benjamins, M. Musen, M. O'Connor, H. Knublauch, S. Tu, B. Grosof, M. Dean, W. Grosso, and M. Musen, Supporting Rule System Interoperability on the Semantic Web with SWRL, in The Semantic Web - ISWC 2005. Springer Berlin Heidelberg. pp. 974-986.2005

[22] Farooq, K., A. Hussain, S. Leslie, C. Eckl, C. MacRae, and W. Slack, An Ontology Driven and Bayesian Network Based Cardiovascular Decision Support Framework, in Advances in Brain Inspired Cognitive Systems. Springer Berlin Heidelberg. pp. 31-41.2012

[23] Vila-FrancÃs, J., J. SanchÃs, E. Soria-Olivas, A.J. Serrano, M. MartÃnez-Sober, C. Bonanad, and S. Ventura," Expert system for predicting unstable angina based on Bayesian networks". Expert Systems with Applications, vol. 40 no 12: pp. 5004-5010.2013

[24] NICE, C.G., Management of chronic heart failure in adults in primary and secondary care. 2010, National Institute for Health and Care Excellence UK.

[25] NICE, C.G., Management of Stable Angina. 2012, National Institute for Health and Care Excellence: UK.

[26] Sirin, E., B. Parsia, B.C. Grau, A. Kalyanpur, and Y. Katz," Pellet: A practical OWL-DL reasoner". Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5 no 2: pp. 51-53.2007

[27] Gruninger, M. and M.S. Fox. "The Role of Competency Questions in Enterprise Engineering". in Proceeding of the Workshop on Benchmarking - Theory and Practice, Trondheim, Norway. pp. 1994

[28] Ian, N. and P. Adam, Towards a standard upper ontology, in Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001. 2001, ACM: Ogunquit, Maine, USA.

[29] Bezerra, C., C.d. Cienc, E.e. Tecnol, F. Freitas, and F. Santana, Evaluating Ontologies with Competency Questions, in Web Intelligence (WI) and Intelligent Agent Technologies (IAT). 2013, IEEE. pp. 284 - 285.

[30] Staab, S. and A. Maedche. "Axioms are Objects, too — Ontology Engineering beyond the Modeling of Concepts and Relations". in Proceeding of the Seventeenth International Joint Conference on Artificial Intelligence(IJCAI-01). pp. 2000

[31] http://herzberg.ca.sandia.gov/.

[32] https://www.norsys.com/netica.html.

[33] Fawcett, T. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers". in Proceeding of the ReCALL. pp 1-38. 2004

Analyzing Data Reusability of Raytrace Application in Splash2 Benchmark

Hao Do-Duc[1,2], Vinh Ngo-Quang[3]
[1]Division of Computational Mathematics and Engineering (CME), Institute for Computational Science (INCOS), Ton Duc Thang University, Ho Chi Minh City, Vietnam
[2]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[3]IC Design Research and Education Center, Vietnam National University, Ho Chi Minh City, Vietnam

***Abstract*—[1]When designing a chip multiprocessors, we use Splash2 to estimate its performance. This benchmark contains eleven applications. The performance when running them is similar, except Raytrace. We analyse it to clarity why the performance is not good. We discover, in theory, Raytrace never reuses data. This leads the fact that the performance is not good due to the low hit ratio in data cache.**

Keywords—Chip multiprocessors; benchmark; ray tracing; reflection; intensity; ray-Tree

I. INTRODUCTION

When designing Chip Multi-Processors (CMP), we always use one or many benchmarks to evaluate our products. One of the most used benchmark is Splash2 [1]. It contains many applications such as: FFT (Fast Fourier Transform) [2], Cholesky factorization [3], Barnes (N-Body problem) [4], etc. They are the most popular classical problems in parallel computing, the main applied field of CMP. Each of these problems requires its unique kind of data and the way to solve it. The complexity problems in real life, in general, are the combination of some basic problems, which are included in Splash2. If a CMP solves the basic problem well, it will solve the real problem well too.

Many CMPs are regularly used to process computer graphics. For this, Splash2 provides three relevant applications namely Radiosity, Raytrace, and Volrend. In many experiments, however, the performance when running Raytrace is not good while the others are better. This inspires us to study and analyze Raytrace application and explain why the CMP do not solve it well.

Our paper is organized as follows. It begins with the introduction of CMP benchmark and questions why the performance of CMP when running Raytrace is not good. Section II presents clearly about Raytrace applications. This section begins with the rendering problem in computer graphics and analyzes the ray tracing method after that. Section III shows us how to use parallel computing to do ray tracing method. The experiments are presented in section IV. We run Raytrace in many CMPs to evaluate the performance, and then, we show its performance in comparison with other applications. The final section is the conclusion. It is presented in section V.

[1]This work was supported by Vietnam National University - Ho Chi Minh city grant number C2015-40-01.

II. RENDERING IN COMPUTER GRAPHICS

We are living in a 3-Dimension (3D) space, but our eyes only observe 2-Dimension (2D) of the world. How we impress the real world by observation? That is based on our hobbies; we can change the view points to get more information about the locations of many objects. So that, we can image exactly where an object is. But we can not change our viewpoint when using a monitor such as a computer monitor. From a 2D image in the monitor, how can we identify the location of any object? That is up to the way we present the image in the monitor. How we present a 2D image, which helps us to impress the location of each object, is called rendering.

A. What is rendering?

Rendering is the way to present a 2D image, which helps us to image about its 3D sense or the location of each object. An object in 3D space is identified by three information: height, width, and depth. 2D image presents the height and the width, and the rendering problem presents the depth of the object. There are two main types of rendering: local illumination and global illumination. Both of them use the intensity, which is from the light, to present the deep of each object. But they use the lights in different ways. The local illumination is very simple. We only use the light coming directly from a light source for the image. That means we do not use others kind of light such as light reflected from a mirror to present the object. Its advantage is simple in both idea and coding, but it is not really a good method. Global illumination is more complexity and efficient than local illumination. We consider all of the lights while presenting an object: directed light, reflected light and shadow light. From these lights, we can create many effects such as reflection, shadows. This approach is the main method for this problem in modern graphics.

B. Raytrace method

Global illumination contains many methods such as ray casting, ray tracing, etc. They use many kinds of light to present an object. Ray tracing is popularly used in both industry and personal applications. Its idea is simple: tracing a path from a point of view through each pixel in a virtual screen, then calculating the light intensity and the color of the object which is visible through it[5]. First, we need to define the main problem: we have a set of light sources and a set of objects, and their location in 3D space. We want

to compute the intensity and colors of each pixel, which is used to present the 3D space including the mentioned light sources and objects, on the screen. Figure 1 shows us an illustration for ray tracing

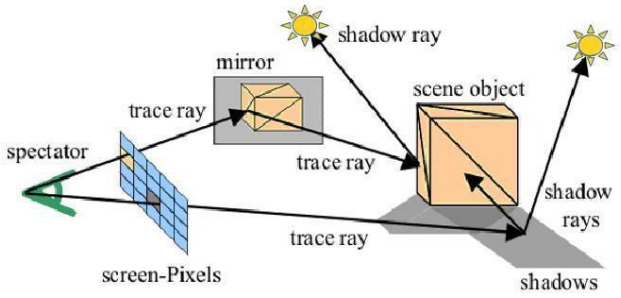


Fig. 1. Identifying the value for every pixel using ray tracing. (Source: ICE RWTH Aachen University)

We want to compute the value at the pixel in the intersection point between the line, that connects our eye and the object, and the screen. Besides, the connecting line contains not only the directed light but also the reflected light from the other surfaces. This leads us that we can observe both the scene in the viewport and the scene which is reflected by the objects in the viewport. Imaging, from our eye, a ray is released. It meets a surface and is reflected. In the real world, almost every object do not have a pure smooth surface, so the reflected rays are spread or diffuse like the illustration in figure 2.

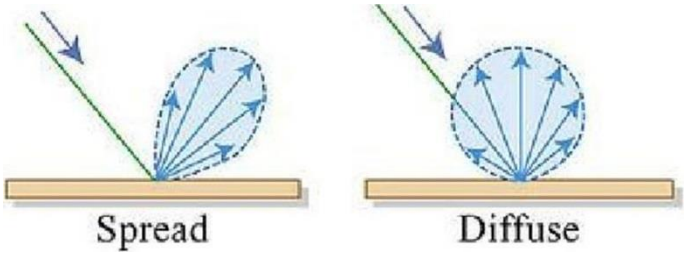


Fig. 2. When a ray meets a non-pure smooth surface, we receive the spread or diffuse reflection. (Source:MIT Open-CourseWare)

The corollary of spreading or diffusing reflection is that the reflected rays will meet many objects, and that process will repeat. But there is an important note: the power of a light ray is reduced after each reflecting point. In other words, we can say there are many rays from many objects have the contribution to the value of a pixel, the less reflecting time, the more contribution. When considering the ray from our point of view to a pixel on the monitor, we need to compute a group of rays reflecting between many objects. If we choose a sequence of objects and identify the reflecting ray between them, we will receive a ray path. The destination of a ray path is often the light source. An object can also be a destination when a ray reflects many times and ends up at that object. Fig.3 shows us a ray path as an example. From the view point to the bulb-light source, the path connects three other objects. Figure 4 shows us a ray-tree [6] or a group of ray paths when extending the contribution to one pixel.

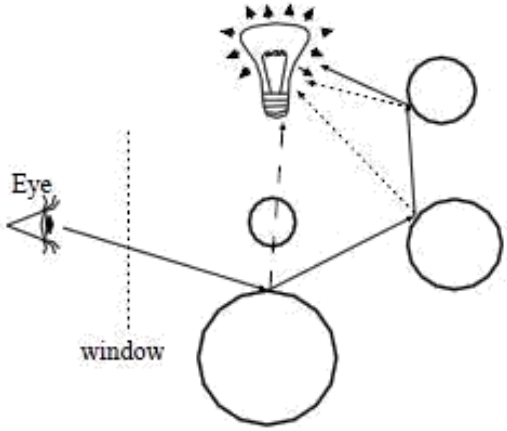


Fig. 3. One ray path in Ray tracing process

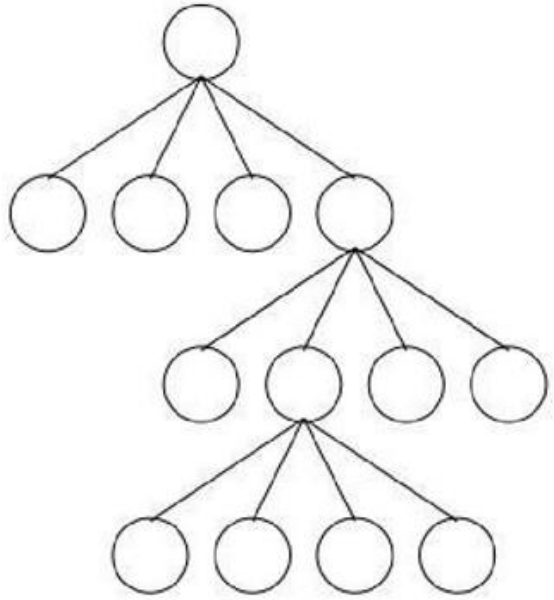


Fig. 4. The ray-tree when extending all of the contribution to one pixel from all objects

The color and intensity of one pixel are decided by one specific point of one specific object, and the value of that specific point is decided by many other points from many other objects. This loop is stopped when the number of reflections is large enough or, in other words, the contribution of a point or an object to the value of the pixel is small enough. We can use a hierarchical tree to illustrate this.

III. Using Parallel Computing to Solve the Rendering Problem by Ray Tracing Method

We presented the Ray tracing method in section II. Raytrace is the parallel version of that method. Our expected output is the color and intensity of all pixels on the screen. These value of one pixel are computed based on the pixel's ray-tree. In parallel method, multiple processor cores can

compute the value of multiple pixels simultaneously [6]. Because of our purpose, we analyze the property of the input data for Raytrace applications. The data is a set of light sources and objects. While calculating the value of a pixel through its ray-tree, an object can contribute different values for the computing process at different time because of the reflection. So we can consider their contributions are from different objects. On the other hand, if one object exists in two ray-trees or two pixels, its contributing values are neither the same because the view angles from two different pixels to the same object are different. Thus, we reach a conclusion that all nodes serving ray tracing process are not reused. This means that each node from each ray-tree is used just one time during their life. So, in theory, we can not reuse any node or any data for our computation.

IV. Performance of Cmp when Running Raytrace Application

This section presents two experiments focus on L1-Data cache. In the first experiment, we run a CMP using Raytrace and three other random selected applications as the workload. We will show the performance of CMP for each application in comparison with the others. In the second experiment, we show the performance of different CMP configurations when running Raytrace. This demonstrates that the negative properties of Raytrace are caused by theory, and they can not be solved by changing CMP. In this section, we use hit ratio in L1-Data cache as the measurement for estimating CMP's performance. This information is an important parameter of a CMP.

A. Experiment 1

We use three random applications in Splash2 to compare with Raytrace. The results are shown in figure 5. As we can see, in 4 cores the hit Ratios at L1-Data cache of Raytrace are significantly lower in comparison with the others. Its hit ratios are high, over 70%, because of the technique of coding. Three others have more positive properties, so the data is reused efficiently, and the hit ratio is nearly 99%, obviously higher than Raytrace.

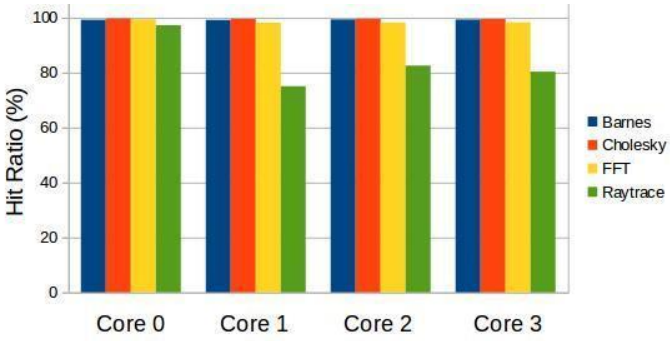


Fig. 5. Hit Ratio in L1-Data cache of CMP when running four applications

B. Experiment 2

We run Raytrace in 4 CMPs to estimate the average performance of reusing data. Our CMPs contain four processor cores. The sizes of L1-Data cache for the four CMPs are 4 KB, 16 KB, 64 KB and 256 KB, respectively. The results are presented in figure 6. When the cache size is too large, total 1 MB for L1-Data cache, the hit ratio is not high, just over 80%. With this result, we infer that the L1 data cache hit ratio or the performance of CMP can not be improved by increasing L1 cache size. We need to change the method or approach instead of changing CMP.

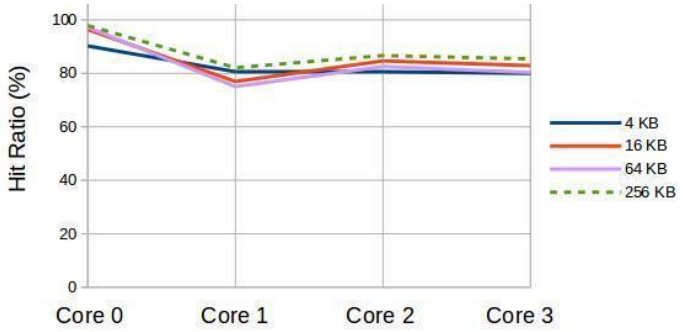


Fig. 6. Hit Ratio in L1-Data cache of 4 CMPs when running Raytrace

V. Conclusion

Our analysis proves that Ray tracing is a good method for the effects used in applications, but it is not well-fit for parallel computing hardware. Because the data is not reused, and each node is used just one time. This led the hit ratio in L1-Data cache is too low, and the performance is not good. We need a new parallel algorithm for this problem instead of increasing L1 cache size of the CMP.

References

[1] Steven Cameron Woo et al. "The SPLASH-2 Programs: Characterization and Methodological Considerations". 22nd Annual International Symposium on Computer Architecture. 1995

[2] Somasundaram Meiyappan. "Implementation and performance evaluation of parallel FFT algorithms". Technical report. 14 pp

[3] Rothberg and Gupta. "An efficient block-oriented approach to parallel sparse Cholesky factorization". ACM. 1993

[4] Jaswinder Pal Singh, John L. Hennessy and Anoop Gupta. "Implications of Hierarchical N-body Methods for Multiprocessor Architecture". ACM Transactions on Computer Systems. Pages 141-202. 1995

[5] Sid Delmar Leach. "3D Rendering". 224 pp. 2011.

[6] J. P. Singh et al. "Parallel Visualization Algorithms: Performance and Architectural Implications". 45-55. Journal of Computer. 1994

Management Information Systems in Public Institutions in Jordan

An Eye on Implementation Success Factors and their Relationship with Organizational Performance

Ahmad A. Al-Tit
Department of Business Administration
Qassim University
Buraidah, KSA

***Abstract*—Six constructs were utilized in this study to explore the factors affecting MIS implementation in Jordanian public institutions and to investigate the impact of MIS implementation on organizational (operational) performance. They were human factors, organizational factors, technological factors, environmental factors, MIS implementation components and organizational performance. The required data were collected using a valid and reliable questionnaire developed based on the literature review. Human factors were conceptualized as users' computer skills and experience, IS usefulness and IS ease of use. Organizational factors were assessed using three sub-indicators, which were top-management support, user training and IS confidentiality. Technological factors were evaluated by systematic quality, information quality and service quality. The overall industry, industry environment and external pressure were three indicators used to measure the environmental factors. Two variables were selected to measure MIS implementation: IT/IS capability and technological aspects related to information service quality. Since the current study tackled public institutions, the indicators of organizational performance were limited to operational ones. The questionnaire was distributed to 125 informants from IT/IS departments. The findings of the study indicated the acceptance of the hypothesis that the factors in question are significantly and positively related to MIS implementation, which in turn, when measured by IT/IS capability and information service quality, significantly and positively affect organizational performance. The main contribution provided by this study is that MIS implementation is not limited to information technology and systems capabilities and usefulness. Other factors should be considered, particularly when examining the impact of MIS implementation on organizational performance.**

Keywords*—*management information systems; adoption success factors; organizational performance; public institutions

I. INTRODUCTION

Researchers have propounded a number of reasons behind the importance of management information systems (MISs). Lipaj and Davidavičienė [1] and Kharuddin et al [2] indicated that one of these reasons is related to the role that MISs play in business performance enhancement. Although many organizations have adopted MISs, not all of them have achieved the presumed benefits [3]. Hence, considerable attention has been paid to the factors that play a critical role in the successful implementation of MIS. Two lines of research have been merged. The first one focuses on the factors affecting the implementation of MISs in different industries, while the other addresses the relationship between MISs and organizational performance.

On the one side, Al-Mamary et al [4] performed a study to explore the factors affecting the successful implementation of MISs in Yemeni organizations. They categorized these factors into technological factors, people factors and organizational factors. In 2015 Al-Mamary et al [5] found a positive relationship between these factors and organizational performance (OP). Using a sample consisting of 100 French organizations, Bacha [6] highlighted the significance of top-management and employee attitudes in the implementation of MISs. In Kuwait Alshawaf and Khalil [7] identified four success factors of information systems (ISs): information systems' strategy and resources, end-user support, information systems' sophistication and information systems' organizational level and user involvement.

In the United States, Kearns [8] studied the relationship between two major factors' impacts on IS planning and implementation, namely top-management support of ISs and management participation in IS planning. The results indicated that these two variables significantly reduced IS implementation problems. Farzandipur et al [3] sorted the factors affecting the implementation of MISs in hospitals into human factors (computer skills, IS usefulness and IS ease of use), managerial and organizational factors (IS project management, IS cost, training, user participation and IS confidentiality) and technological factors (support, safety, development and communication). Rahimi et al [9] brought user participation in the development stages of IS to light as a critical factor that affects the development of ISs in hospitals. According to them, users can take part in four stages of IS development: analysis, design, implementation and evaluation.

Fu et al [10] listed three main factors that influence the adoption of ISs by small and medium-sized enterprises in Taiwan. Those factors are technological factors, organizational factors and environmental factors. Each group of them relates to three types of objectives. Specifically, technological factors are related to the system function, technology trust and cognition benefit; organizational factors are interconnected to organizational characteristics, the organization's readiness and the partners' willingness and

abilities; and, finally, environmental factors are linked to the overall industry, industry environment and external pressure. Detailed criteria levels for these factors can be seen in Table 1. However, the authors deemed six out of these factors to be critical success factors.

On the other side, Al-Gharaibeh and Malkawi [11] carried out a case study of the Ministry of Planning to investigate the relationship between MISs and OP. Three dimensions of MISs were used: hardware and software components, networks, and individuals and procedures. According to their results, MISs have an impact on organizational performance in Jordanian public settings. Analysing data collected from thirteen countries, DA Silveira and Cagliano [12] explored and confirmed the relationship between inter-organizational information systems (computerized networks used for information exchange) and operational performance. Batra [13] hypothesized an impact of information technology (IT) on organizational effectiveness. The findings pointed out that IT has an impact on the overall organizational flexibility, which in turn influences the organizational performance of organizations and hence their organizational effectiveness.

Building on the above-mentioned literature, the purpose of this study is twofold: first, to explore the factors affecting the implementation of MISs in Jordanian governmental institutions; and second, to explore the relationship between MIS components and organizational performance in those institutions.

II. Literature Review and Hypothesis Development

A. MIS Definition, Requirements and Dimensions

Management information systems (MISs) are one of the five types of information systems. The other four types are office information systems (OISs), transaction-processing systems (TPSs), decision support systems (DSSs) and executive support systems (ESS) [1]. MISs have been defined by researchers in terms of their ability to provide information with good characteristics on which organizations depend to enhance their performance [5]. Other definitions have tackled MISs with regard to their functions, such as collecting, recording, storing and rearranging data [14]. Given that the first major aim of this study is to identify the factors affecting MIS implementation in Jordanian governmental institutions, a literature review was conducted. Examples of those factors are presented in Table 1.

TABLE I. Factors that Affect MIS Adoption as Depicted in the Literature

MIS requirements	Reference (s)
• Technological factors: - System quality - Information quality - Service quality • Organizational factors: - Top management support - User training • People factors: - Computer self-efficacy - User experience.	Al-Mamary et al. [**Error! Bookmark not defined.**]
• Information systems strategy and resources. • End user support. • Information systems sophistication. • IS organizational level and user involvement.	Alshawaf and Khalil [**Error! Bookmark not defined.**]
• Internal environment factors - Top-management support - Managers' participation in IS planning	Kearns [**Error! Bookmark not defined.**]
• User participation in IS development: - Analysis. - Design. - Implementation. - Evaluation.	Rahimi et at. [**Error! Bookmark not defined.**]
• Human factors: - Computer skills. - IS usefulness. - IS ease to use. • Managerial and organizational factors: - IS project management. - IS cost. - Training. - User participation. - IS confidentiality. • Technological factors: - Support. - Safety. - Development. - Communication.	Farzandipur, et al. [**Error! Bookmark not defined.**]
• Technological factors: System function, technology trust, and cognition benefits. • Organizational factors: organization characteristics and readiness, and partners' willingness and abilities. • Environmental factors: overall industry, industry environment, and external pressure.	Fu et al. [**Error! Bookmark not defined.**]

Consequently, the current study categorized the factors that have an influence on the adoption of ISs in organizations into four groups: human factors, organizational factors, technological factors and environmental factors. Regarding MIS components, Zhu and Nakata [15] argued that the most important components of MISs are IT capability and information service quality. Benitez-Amado and Walczuch [16] conceptualized IT capability in their study as a dependent variable that represents an organization's ability to use IT resources. Table 2 shows the major components of MISs in the literature. The current study focuses on IT capability and information service quality in addition to hardware and software components.

TABLE II. Major Components of Miss Found in the Literature

MIS components	Reference (s)
• IT capability: - Information storage. - Information processing - Information communication • Information services quality: - Service timeliness. - Service appropriateness - Information reliability.	Zhu and Nakata [15]
• IT capability: organizations ability to use IT resources	Benitez-Amado and Walczuch [16]
• Hardware and software components • Networks • Individuals and procedures	AL-Gharaibeh and Malkawi [11]
• Timeliness • Scope • Aggregation	Naranjo-Gil [17]

• Integration	

B. Organizational Performance

According to Al-Tit and Hunitie [18], OP can be defined as a measure employed to identify organizations' efficiency and effectiveness in achieving their goals. In general, two types of measures were used to evaluate organizational performance: financial and non-financial measures [19]. Table 3 presents the different indicators used in the literature to measure OP. Given that this study was conducted on public institutions, OP was measured in terms of operational dimensions, that is, non-financial measurements.

TABLE III. OP DIMENSIONS DEPICTED IN THE LITERATURE

OP dimensions	Reference (s)
• Internal process performance: - Internal process simplification. - Data validity improvement. - Internal communication efficiency. • Financial performance: - Sales increase. - Inventory turnover reduction. - Receivable turnover increase. - Profit margin growth.	Lipaj and Davidavičienė [1]
• Financial measures: - Cost of funds - Non-interest income - Earnings per share - Capital structure - Return on investment - Loan yield - Market ratios - Liquidity - Cash flow from operations - Relative market share and position - Operating income - Revenues - Customers' profitability • Non-financial measures: - Customer-employee-based performance: Responsiveness, personnel development, no. of customer's complaints, accessibility, delivery speed flexibility, customer satisfaction, on-time service, employee skills, communication, competence, productivity, efficiency, availability, courtesy and quality. - Innovation-based performance: Performance of individual innovations, performance of the innovation process, research and development, new product development, volume flexibility, and specification flexibility.	Salleh et al. [20]
• Work efficiency • Work effectiveness • Decision making	Alshawaf and Khalil [7]
• Sectoral excellence	Benitez-Amado and Walczuch [16]
• Satisfaction of employees	Gil-Padilla and Espino-Rodríguez [19]

C. Factors Affecting the Adoption of MISs

Fu et al [10], Al-Mamary et al [5] and Farzandipur et al [3] suggested four groups of factors that have an influence on the adoption of MISs: human factors, organizational factors, technological factors and environmental factors. Following these recent studies, the current study applied the same factors. Therefore, the following hypotheses were posed:

H01: Human factors significantly advance MIS implementation.

H02: Organizational factors significantly elevate MIS implementation.

H03: Technological factors significantly support MIS implementation.

H04: Environmental factors significantly improve MIS implementation.

D. Relationship between MISs and OP

Al-Mamary et al [5] carried out a study on the relationship between the success factors of MISs and the organizational performance in the telecommunication industry in Yemen. Their hypotheses were supported. That is, technological (system quality, information quality and service quality), organizational (top-management support and user training) and people factors (computer self-efficacy and user experience) were positively related to organizational performance. In their work on information systems' success factors and the organizational performance of public and private organizations, Alshawaf and Khalil [7] found significant differences between public and private organizations with regard to end-user support, top management and information systems management in IS financial decisions in favour of public organizations. They also found significant differences in terms of IS resource availability, top-management involvement in the IS strategy, end-user involvement in IS development and end-user training on information technology in favour of private organizations. The study revealed no significant differences between private and public organizations in Kuwait with respect to the age of IS units, IS organizational levels, IS sophistication or the perceived obviousness of the IS strategy. Ravichandran and Lertwomgsatien [21] found a positive relationship between IS human capital (IS skills and specificity), IT infrastructure flexibility (networks' and applications' sophistication), IS partnership quality (internal and external partnership quality) and organizational performance (operating and market-based performance) of different organizations from numerous industries such as banking, insurance, financial services, retail, manufacturing and services, transportation and utilities in the United States. As a result, the following hypothesis was postulated:

H05: MIS implementation has a positive impact on organizational performance.

III. STUDY MEASUREMENT MODEL

Figure 1 displays the measurement model of the study, in which four constructs (human factors, organizational factors, technological factors and environmental factors) were

assumed to have an impact on OP. Hypotheses 1–4 postulated significant relationships between those factors and MIS implementation in Jordanian public institutions. Hypothesis 5 presumed that MIS adoption has a significant impact on the overall OP of public institutions as measured by internal process performance, customer satisfaction, employee satisfaction and work efficiency and effectiveness.

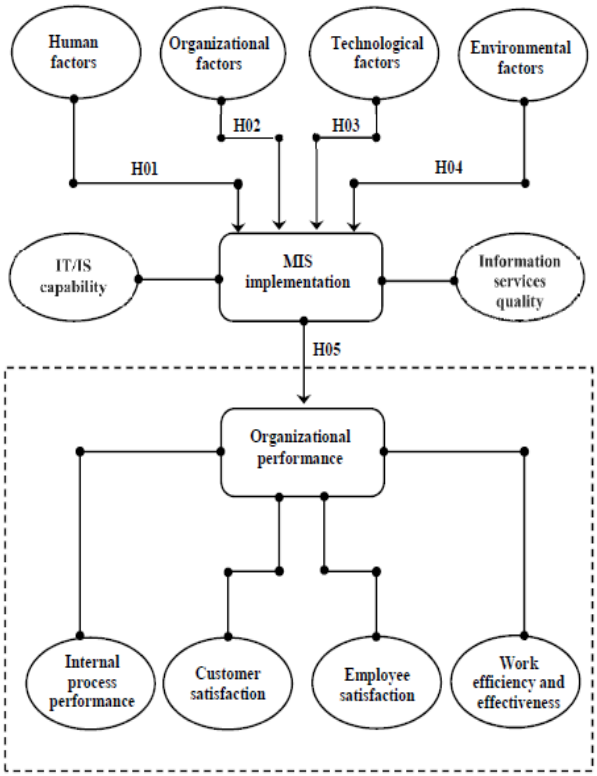


Fig. 1. Study measurement model

IV. Methodology

A. Sample and Data Collection

The current study was conducted in Jordanian public institutions. A sample consisting of 25 governmental institutions in Amman was selected to collect the required data. The analysis unit used comprises managers as well as employees working in IT/IS departments. A 5-point questionnaire was developed based on related work on IS implementation and organizational performance. It was anchored at "strongly disagree" for responses of 1 and "strongly agree" for responses of 5. A total of 125 questionnaires were distributed to the participants, out of which 73 were returned, which means a 58 per cent response rate. This rate is judged to be high, since the responses are limited to IT/IS departments.

B. Measures

The factors that might affect MIS implementation in public institutions were measured by the human factors, organizational factors, technological factors and environmental factors adapted from Fu et al [10], Al-Mamary et al [5] and Farzandipur et al [3]. Seven items based on Zhu and Nakata [15] and Al-Gharaibeh and Malkawi [11] were used to measure MIS implementation (IT/IS capability and information service quality). Organizational performance was measured using eight items concerning internal process performance, customer satisfaction, employee satisfaction, work efficiency and effectiveness. The dimensions used to measure OP were adapted from Lipaj and Davidavičienė [1], Salleh et al [20], Alshawaf and Khalil [7] and Gil-Padilla and Espino-Rodríguez [19]. Table 4 shows the study constructs, codes, indicators and number of items.

TABLE IV. Study Constructs, Codes, Indicators and Number of Items

Constructs	Code	Indicators	No. of items
• **MIS Factors**	**MISF**		**12**
- Human factors	HUF	HUF1 - HUF3	3
- Organizational factors	ORF	ORF1 - ORF3	3
- Technological factors	TEF	TEF1 - TEF3	3
- Environmental factors	ENF	ENF1 - ENF3	3
• **MIS implementation**	**MISI**		**7**
- IT/IS capability	ISC	ISC1 - ISC4	4
- Information service quality	ISQ	ISQ1 - ISQ3	3
• **Organizational performance**	**ORP**		**8**
- Internal process performance	IPP	IPP1 & IPP2	2
- Customer satisfaction	CST	CST1 & CST2	2
- Employee satisfaction	EMS	EMS1 & EMS2	2
- Work efficiency and effectiveness.	WEE	WEE1 & WEE2	2

C. Validity and Reliability

Two types of validity are tested in this section: content validity and convergent validity. Five academic experts evaluated the content validity. The convergent validity was assessed using the average variance extracted (AVE). On the other hand, two coefficients were used to rate reliability: Cronbach's alpha coefficients and composite reliability coefficients. The results of the validity and reliability tests summarized in Table 5 indicate that the scale used in this study is valid and reliable, as all the values of AVE are greater than 0.6 [ ], all the Cronbach's alpha coefficients are above 0.7 [8] and all the coefficients of composite reliability are above 0.6 [ ].

TABLE V. Results of Validity and Reliability Tests

Variable	AVE	Cronbach's alpha	Composite reliability
Human factors	0.64	0.73	0.65
Organizational factors	0.66	0.81	0.77
Technological factors	0.73	0.87	0.73
Environmental factors	0.69	0.76	0.80
MIS implementation	0.71	0.89	0.69
Organizational performance	0.74	0.88	0.82

D. Pearson's Product-Moment Correlation Coefficient

The Pearson's matrix shown in Table 6 illustrates the significant relationships between the factors affecting MIS implementation and MIS implementation as measured by IT/IS capability (ISC) and information service quality (ISQ). It appears that human factors (HUFs) are significantly correlated with both dimensions of MIS implementation ($r = 0.57$ and $r = 0.61$, $p < 0.05$). Additionally, organizational

factors (ORFs) are significantly correlated with both dimensions of MIS implementation ($r = 0.50$ and $r = 0.55$, $p < 0.05$), along with technological factors ($r = 0.61$ and $r = 0.43$, $p < 0.05$). Finally, environmental factors (ENFs) are significantly correlated with ISC ($r = 0.39$, $p < 0.05$) and ISQ ($r = 0.31$). The results also revealed a significant correlation between the two dimensions of MIS implementation.

TABLE VI. CORRELATIONS BETWEEN MIS FACTORS AND MIS IMPLEMENTATION

	HUF	ORF	TEF	ENF	ISC	ISQ
HUF	-					
ORF	0.44	-				
TEF	0.51	0.48	-			
ENF	0.46	0.37	0.45	-		
ISC	0.57	0.50	0.61	0.39	-	
ISQ	0.61	0.55	0.43	0.31	0.47	-

V. DATA ANALYSIS

A. Descriptive Statistics of the Factors Affecting MIS Implementation

Frequencies, percentages, mean scores and standard deviations were extracted, as shown in Table 7, to identify the frequencies and percentages of the responses to the scale points. The results obtained were used to categorize the factors affecting MIS implementation according to their importance.

TABLE VII. MEAN SCORES OF THE FACTORS AFFECTING MIS IMPLEMENTATION

MISF	N (%)	5	4	3	2	1	Mean	SD
• **HUF**	**73(100)**	-	-	-	-	-	**4.04**	**1.070**
- HUF1		33(45)	17(23)	11(15)	7(.09)	5(.07)	4.15	0.877
- HUF2		29(40)	16(22)	9(12)	11(15)	8(0.1)	3.99	0.965
- HUF3		27(39)	18(25)	12(16)	9(12)	7(.09)	3.97	1.000
• **ORF**	**72(99)**	-	-	-	-	-	**3.92**	**0.682**
- ORF1		24(33)	11(15)	28(39)	4(.06)	5(.07)	3.94	0.714
- ORF2		30(42)	18(25)	17(24)	6(.08)	1(.01)	3.92	1.100
- ORF3		22(31)	17(24)	10(14)	13(18)	10(14)	3.90	0.594
• **TEF**	**71(97)**	-	-	-	-	-	**3.72**	**0.416**
- TEF1		20(27)	21(29)	13(18)	12(16)	7(.09)	3.89	0.947
- TEF2		19(26)	24(33)	11(15)	7(.09)	12(16)	3.85	0.721
- TEF3		15(21)	19(26)	20(27)	12(16)	7(.09)	3.77	0.605
• **ENF**	**73(100)**	-	-	-	-	-	**3.27**	**0.819**
- ENF1		11(15)	16(22)	20(27)	18(25)	8(0.1)	3.67	0.700
- ENF2		15(21)	37(51)	2(.03)	9(12)	10(14)	3.66	0.601
- ENF3		25(34)	14(19)	16(22)	7(.09)	11(15)	3.37	0.814

It was concluded, based on the results in Table 7, that human factors are the most important factors in MIS implementation ($M = 4.04$, $SD = 1.070$), followed by organizational factors ($M = 3.92$, $SD = 0.682$), then technological factors ($M = 3.72$, $SD = 0.416$) and finally environmental factors ($M = 3.27$, $SD = 0.819$).

B. Structural Model

The results of the confirmatory factor analysis (CFA) established the goodness of fit of the data: the comparative fit index (CFI) = 0.931, the normalized chi-square ($\chi 2/df$) = 1.66, the goodness of fit index (GFI) = 0.913 and the root mean square error of approximation (RMSEA) = 0.051. Consequently, the overall fit was supported, as illustrated in Figure 2. Grounded on the path coefficients of the structural model, the associations between human factors (H01), organizational factors (H02), technological factors (H03) and environmental factors (H04) and MIS implementation are significant and positive. In other words, the model supported all the concerning factors affecting MIS implantation. Still, for hypothesis 5 a significant impact of MIS implementation, measured by IT/IS capability and information service quality, on the organizational performance was found.

C. Multiple Regression Analysis

Hypothesis 5 supposed that MIS implementation has a significant impact on organizational performance. Multiple regression analysis was conducted to test this hypothesis. The independent variable was MIS implementation and the independent variable was organizational performance. The regression findings displayed in Table 8 indicate that the MIS implementation dimensions have a positive and significant impact on the organizational performance of public institutions. MIS implementation explained 40% of the variance in the organizational performance. The F (33.16), β (0.514), t (5.106) and P values (0.000) verify this result.

TABLE VIII. REGRESSION RESULTS FOR MIS IMPLEMENTATION AND ORGANIZATIONAL PERFORMANCE

Model summary		ANOVA		Coefficients		
r	R^2	F	P	β	t	P
0.631	0.40	33.16	0.000	0.514	5.106	0.000

D. Final Model

Founded on the previously mentioned results, the final model of the study shown in Figure 2 demonstrates a positive correlation between human factors ($r = 0.59$), organizational factors ($r = 0.53$), technological factors ($r = 0.52$) and environmental factors ($r = 0.35$). Human factors ranked first as the most correlated factors in MIS implementation from the respondents' perspective ($M = 4.04$), followed by organizational factors ($M = 3.92$), then technological factors ($M = 3.72$) and environmental factors ($M = 3.27$). The significant and positive impact of MIS implementation on organizational performance was supported using the current data ($\beta = 0.514$, $t = 5.106$, $P = 0.000$).

VI. DISCUSSION AND CONCLUSION

The aim of this study was to explore the factors affecting MIS implementation in Jordanian public institutions. Four major factors were identified based on the literature: human factors, organizational factors, technological factors and environmental factors. On the other hand, the study aimed to investigate the impact of MIS implementation on organizational performance.

The results revealed that human factors, organizational factors, technological factors and environmental factors are significantly related to MIS implementation. That is, users' skills and experience, IS usefulness, IS ease of use, top-management support, user training, IS confidentiality, system quality, information quality, service quality, overall environment, institutional environment and external pressure are all factors that contribute to the success of MIS implementation. In line with these findings, Al-Mamary et al [4], Bacha [6], Alshawaf and Khalil [7] and Kearns [8] found similar results.

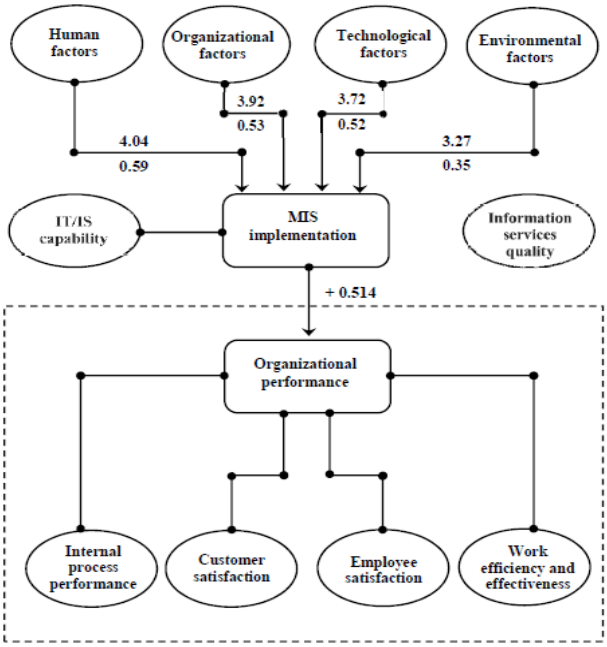


Fig. 2. Study final model

Concerning the relationship between MIS implementation and organizational performance, the findings pointed out that MIS implementation, when measured by IT/IS capability and information service quality, has a significant and positive impact on organizational performance, when measured by internal process performance, customer satisfaction, employee satisfaction, work efficiency and effectiveness. Ravichandran and Lertwomgsatien [21] found similar results. In conclusion, four factors representing twelve characteristics identified by this study play a central role in MIS implementation. Those factors have a direct effect on MIS implementation in Jordanian public institutions by enhancing the utilization of IT/IS capabilities and the quality of both information and services. MIS implementation, in turn, plays a positive role in improving organizational performance.

A. Implications for Management and Research

This study contributes to both management and research by exploring the factors affecting MIS implementation as well as the impact of MIS implementation on organizational performance. Information technology capabilities related to information storage, processing, communication and their attendant aspects are insufficient in the absence of human, organizational and technological factors, since these factors are in charge of management. The current study concluded that the investigation of the relationship between MIS implementation and organizational performance should consider the factors that might affect MIS components. Future studies can extend the proposed model using new factors and MIS constructs to understand the potential mediation role of MIS components in the relationship between MIS implementation and MIS performance. The study was conducted using a small sample selected from public institutions; the responses were restricted to managers and employees of IT/IS departments. Hence, a larger sample size and more informants might result in more generalizable results.

REFERENCES

[1] Lipaj, D., and Davidaviciene, V. "Influence of information systems on business performance," Science: Future of Lithuania, 5(10), pp. 38-45, 2013.

[2] Kharuddin, S., Ashhari, Z. and Nassir, A. "Information System and Firms' Performance: The Case of Malaysian Small Medium Enterprises," International Business Research, 3(4), pp. 28-35, 2010.

[3] Farzandipur, F., Jeddi, F. and Azimi, E. "Factors affecting successful implementation of hospital information systems," ACTA INFORM MED, 24(1), pp. 51-55, 2016.

[4] Al-Mamary, Y., Shamsuddin, A. and Aziati, N. "Factors affecting successful adoption of management information systems in organizations towards enhancing organizational performance," American Journal of Systems and Software, 2(5), pp. 121-126, 2014.

[5] Al-Mamary, Y., Shamsuddin, A. and Aziati, N. "The pilot test study of relationship between management information systems success factors and organizational performance at Sabafon Company in Yemen," International Journal of u- and e- Service, Science and Technology, 8(2), pp. 337-346, 2015.

[6] Bacha, E. "The impact of information systems on the performance of the core competence and supporting activities of a firm," Journal of Management Development, 31(8), pp. 752-763, 2012.

[7] Alshawaf, A. and Khalil, O. "IS Success factors and IS organizational Impact: Does ownership type Matter in Kuwait?," International Journal of Enterprise Information Systems, 4(2), pp. 13-33, 2008.

[8] Kearns, G. "How the internal environment impacts information systems project success: An investigation of exploitative and explorative firms," The Journal of Computer Information Systems, 48(1), pp. 63-75, 2007.

[9] Rahimi, B., Safdari, R. and Jebraeily, M. "Development of hospital information systems: User participation and factors affecting it," ACTA INFORM MED., 22(6), pp. 398-401, 2014.

[10] Fu, H-P., Chang, T-H., Ku, C-Y., Chang, T-S. and Huang, C-T "The critical success factors affecting the adoption of inter-organization systems by SMEs," Journal of Business & Industrial Marketing, 29(5), pp. 400-416, 2014.

[11] AL-Gharaibeh, Sh. and Malkawi, N. "The impact of management information systems on the performance of governmental organizations: Study at Jordanian ministry of planning," International Journal of Business and Social Science, 4(17), pp. 101-109, 2013.

[12] DA Silveira, G. and Cagliano, R. "The relationship between interorganizational information systems and operations performance," International Journal of Operations & Production Management, 26(3/4), pp. 232-253, 2006.

[13] Batra, S. "Impact of information technology on organizational effectiveness: A conceptual framework incorporating organizational flexibility," Global Journal of Flexible Systems Management, 7(1/2), pp. 15-25, 2006.

[14] Claver, E., Llopis, J., Gonzalez, M. and Gasco, J. "The performance of information systems through organizational culture," Information Technology & People, 14(3), pp. 247-260, 2001.

[15] Zhu, Z. and Nakata, C. "Reexamining the link between customer orientation and business performance: The role of information systems," Journal of Marketing Theory and Practice, 15(3), pp. 187-203, 2007.

[16] Benitez-Amado, J. and Walczuch, R. "Information technology, the organizational capability of proactive corporate environmental strategy and firm performance: a resource based analysis," European Journal of Information Systems, 21, pp. 664–679, 2012.

[17] Naranjo-Gil, D. "Managerial styles and management information systems for improving organizational performance," Journal of Positive Management, 1(1), pp. 3-10, 2010.

[18] Al-Tit, A. and Hunitie, M. "The mediating effect of employee engagement between its antecedents and consequences," Journal of Management Research, 7(5), pp. 47-62, 2015.

[19] Gil-Padilla, A. and Espino-Rodríguez, T. "Strategic value and resources and capabilities of the information systems area and their impact on

organizational performance in the hotel sector," Tourism review, 63(3), pp. 21-47, 2008.

[20] Salleh, N., Jusoh, R. and Isa, C. "Relationship between information systems sophistication and performance measurement," Industrial Management & Data Systems, 110(7), pp. 993-1017, 2010.

[21] Ravichandran, T. and Lertwomgsatien, C. "Effect of information systems resources and capabilities on firm performance: A resource-based perspective," Journal of Management Information Systems, 21(4), pp. 237-276, 2005.

[22] Al-Tit, A. "The effect of service and food quality on customer satisfaction and hence customer retention," Asian Social Science, 11(23), pp. 129-139, 2015.

[23] Wijaya, A. and Akbar, R. "The influence of information, organizational objectives and targets, and external pressure towards the adopting of performance measurement system in public sector," Journal of Indonesian Economy and Business, 28(1), pp. 62-83, 2013.

A Novel Method in Two-Step-Ahead Weight Adjustment of Recurrent Neural Networks: Application in Market Forecasting

Narges Talebi Motlagh
Control Engineering Department
Faculty of Electrical and
Computer Engineering University of Tabriz
Tabriz, Iran

Amir RikhtehGar Ghiasi
and Farzad Hashemzadeh and Sahraneh Ghaemi
Control Engineering Department
Faculty of Electrical and
Computer Engineering University of Tabriz
Tabriz, Iran

***Abstract*—Gold price prediction is a very complex nonlinear problem which is severely difficult. Real-time price prediction, as a principle of many economic models, is one of the most challenging tasks for economists since the context of the financial agents are often dynamic. Since in financial time series, direction prediction is important, in this work, an innovative Recurrent Neural Network (RNN) is utilized to obtain accurate Two-Step-Ahead (2SA) prediction results and ameliorate forecasting performances of gold market. The training method of the proposed network has been combined with an adaptive learning rate algorithm and a linear combination of Directional Symmetry (DS) is utilized in the training phase. The proposed method has been developed for online and offline applications. Simulations and experiments on the daily Gold market data and the benchmark time series of Lorenz and Rossler shows the high efficiency of proposed method which could forecast future gold price precisely.**

***Keywords*—Recurrent Neural Network; Two Step Ahead Prediction; Reinforcement Learning; Directional Statistics; Gold Market**

I. Introduction

A. Motivation

Customers have to arrange their lifetime work and consumption orientation, while organizations choose on how to develop up upcoming manufacturing abilities based on their expectations of future events. Thus, an accurate forecast is crucial in decision making. In either situation, the providers want to know how the uncertain upcoming market may open up. Lately, predicting the gold price is becoming progressively essential. For a long time, gold has been exchanged definitely on worldwide marketplaces. Many types of gold trading are also exchanged, such as gold futures trading, gold options and gold forward contracts[1], [2]. Moreover, since the cost of gold differs within a restricted range, it is able to decrease the effect of rising prices, control the increase of cost and help carry out a constrictive financial plan. Gold performs a crucial part in international markets and traders give more attention to it. Gold offers a way of secured risk and it can also be saved without devaluation. Besides, investors could save more money with an accurate forecast of the gold price.

B. Contribution

Forecasting a sequence of values in a time series is named multi-step ahead prediction. Applying a predictive model step-by-step and using the expected value of the moment phase to determine its value in the next time step is a common approach, known as multi-level forecast. Since small prediction error at the beginning may propagate into the future, the widely used recursive applications of one-step-ahead predictions have been demonstrated to have disadvantages in real life programs[3]. In this paper a new approach based on the two step ahead prediction of the gold market is proposed. This model shows high convergence rate, low prediction error and efficiency in gold market forecasting. The proposed method could be widely applied to similar management and decision making problems which depends on expectations of future events.

C. Related Works

The evolution of financial markets is a complicated phenomenon that is at the top in terms of difficulty of the modeling and prediction. One reason for this difficulty is the complex nonlinearity that is inherent at work. A reliable forecast of future events possesses great value. The basic idea for forecasts is usually the case that additional information from antecedent observed values and/or model outputs will be beneficial to forecasts. Then the forecast results tend to be closer to the true values as the forecast model iteratively adjusted through model performance with error reduction. The findings of recent literature confirm that stock markets are predictable from past outputs and other macroeconomics and financial variables. The predictability of stock market led the researchers to investigate the sources of this predictability[4]. On the other hand, prediction of the stock price is a highly complicated and very difficult task because there are too many factors such as political events, economic conditions, traders expectations and other environmental factors that may influence stock prices. In addition, stock price series are generally quite noisy, dynamic, nonlinear, complicated, nonparametric, and chaotic by nature[5]. The main goal of this study is to explore the predictability of the gold market. Predicting the gold market is important and of great interest because

successful prediction of gold prices may promise attractive benefits. It usually affects a financial traders decision to buy or sell gold. These tasks are highly complicated and very difficult because there are too many factors that may influence gold price. Soft computing techniques have been successfully applied to solve the problems of stock markets including gold market. Soft computing techniques are commonly used methods for stock problems. While stock markets and catching their non-linear actions are based on the noisy environment, Soft computing techniques offer useful tools in predicting stock market[6]. Neural networks, fuzzy systems and genetic algorithms are the intelligent systems that have been widely used for predicting of financial systems. In order to deal with varying environments, neural networks recognize patterns; the primary objective of this work is an accurate estimation of the gold market pattern using historical information about the gold price. Many companies, economic experts, individual traders and other inventory stock traders believe that they can estimate stock market trend and make profits. Thus, many models have been proposed to predict the stock market trend. In general, existing techniques to estimate stock market prices are classified in two types, fundamental and technical analysis[7], [8]. Fundamental analysis which is based on Efficient Market Hypothesis (EMH) and Rational Expectation (RE) methods are based on macroeconomic information, such as exports and imports, money supply, interest levels and inflation prices. For instance, Narayan et al., Wang et al. and Zhang et al. are trying to discuss gold market efficiency through different methods and try to propose a method to forecast gold price[9], [10], [11]. However, many experiments contradict the RE hypothesis[12]. Technical analysis that completely disregards the EMH, is based on the fact that history will repeat itself. Prediction occurs by taking advantage of effects that are hidden in the past trading activities, and by examining patterns shown in price series[13], [14], [15]. Outstanding property of Artificial Neural Networks (ANNs) is their ability in the estimation of nonlinear features, which make them useful tools for various issues[16], [17], [18], [19], [20]. Although, static neural networks may not be successful in setting up efficient nonlinear models for dynamical systems. Many static neural networks were developed to fix different issues, such as rainfall and stream flow forecasting[21], [22], reservoir flood control[23] and financial predictions[24], [25]. But they are not able to well retain the time variation characteristics of time series and can identify the short-term memory components. Besides, the performance of dynamic neural networks is better than of static neural networks and dynamic networks can effectively extract the dynamic characteristics of systems[26]. Lately, RNNs have drawn much attention for getting dynamic features of systems[27], [28], [29], [30]. Because of their powerful characteristics, RNNs have been efficiently used for a wide range of problems such as time series predicting[31], [32], [33]. RNNs are also capable of enhancing prediction precision[34]. However, the training of an RNN is challenging and could take a lot of time[35]. The Real Time Recurrent Learning (RTRL) algorithm is an efficient and effective algorithm for training repeated systems suggested by Williams and Zipser, uses internal feedback loops to improve the performance[36]. In accordance with the one-step-ahead RTRL criteria, 2SA RTRL criteria was developed. Basically a more than one-step-ahead forecast is more difficult to achieve satisfactorily due to the lack of measurements in the forecast. Many forecasting problems, however, require styles for providing multistep-ahead forecasts[37].

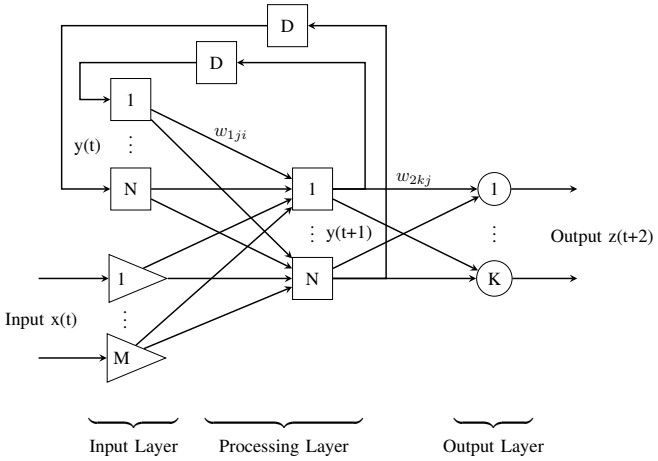


Fig. 1. Architecture of 2SA-RNN

II. Improved 2SA R-RTRL

In this paper the Reinforced-RTRL (R-RTRL) algorithm for RNNs is applied for 2SA gold price forecast. Reinforced Recurrent Neural Network(R-RNN) proposed by Chang is shown in figure 1 and it is consists of two layers and has M external inputs and K outputs[26].

As shown in figure 1, $x(t)$ denotes the $M \times 1$ discrete time varying input vector and $y(t+1)$ is the $N \times 1$ output of the corresponding processing layer. Network input is formed by two vectors $x(t)$ and $y(t)$ as illustrated in 1

$$\mu(t) = [x(t); y(t)] \tag{1}$$

On the other hand, $y(t+1)$ is the input of the second layer and $z(t+2)$ denotes the corresponding $k1$ output. The output of neuron j in the processing layer is given by 2

$$y_j(t+1) = f\Big(\sum_{i \in A \cup B} w_{1ji}(t)\mu_i(t)\Big) \tag{2}$$

where $f(.)$ is a nonlinear activation function of a neuron. In output layer, the net output of neuron k in the output layer at time $t+2$ and is computed by 3

$$z_k(t+2) = f\Big(\sum_j w_{2kj}(t+1)y_j(t+1)\Big) \tag{3}$$

Where , $w_{lji}(t)$ is network weight in 2 and 3. We now could define the kth element of time-varying $K \times 1$ error vector $e_k(t+2)$ in 4 where $d_k(t+2)$ denote the target value of neuron k at time $t+2$.

$$e_k(t+2) = d_k(t+2) - z_k(t+2) \tag{4}$$

The instantaneous total network error is defined as 5

$$E(t+2) = \frac{1}{2}\sum_{k=1}^{K} e_k^2(t+2) \tag{5}$$

The learning rate of a particular weight w_{lmn} can be proposed by 6

$$\Delta w_{lmin} = -\eta \frac{\partial E(t+2)}{\partial w_{lmin}} \tag{6}$$

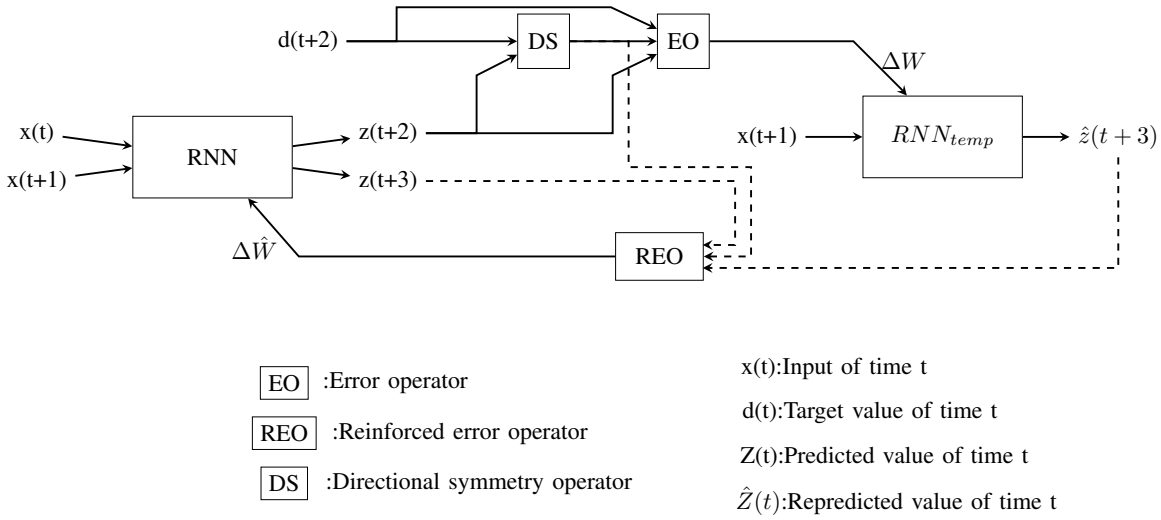


Fig. 2. Improved reinforced 2SA weight adjustment procedure for RNN

Reinforced 2SA weight adjustment procedure for RNN is represented completely in figure 2. For the reinforcement learning stage, following equations are proposed in this paper.

$$\begin{aligned}
\hat{y}_j(t+2) &= f\left(\sum_{i\in A\cup B}\Big(w_{1ji}(t)+\Delta w_{1ji}(t)\Big)\mu_i(t+1)\right)\\
\hat{z}_k(t+3) &= f\left(\sum_{j}\Big(w_{2ji}(t+1)+\Delta w_{2ji}(t+1)\Big)\hat{y}_j(t+1)\right)\\
\hat{e}_k(t+3) &= \hat{z}_k(t+3)-z_k(t+3)\\
\hat{E}(t+3) &= \frac{1}{2}\sum_{k=1}^{K}\hat{e}_k^2(t+3)\\
\Delta\hat{w}_{lmin} &= -\eta\frac{\partial E(t+2)}{\partial w_{lmin}}
\end{aligned} \tag{7}$$

Finally, the updating process is given by 8

$$w_{lmin}^{new} = w_{lmin}+\Delta w_{lmin}+\Delta\hat{w}_{lmin} \tag{8}$$

where, w_{lmn} denotes the weight matrix for processing layer in case of $l=1$ and for output layer in case of $l=2$, w_{lmn}^{new} is weight adjustment of IR-RTRL algorithm, Δw_{lmn} is weight change and Δw_{lmn} is reinforced weight change.

A. Improved Reinforced Method

The proposed reinforced function in 7, has very effective results for forecasting reservoir inflow during a typhoon. Even though minimizing the prediction error and making an accurate forecast is very important, predicting the direction of movement of financial time series has higher importance. Moreover, as discussed before, customers have to arrange their decision of trading which affect their benefits and total wealth. Furthermore, correct forecasting directions or turning points between the actual and predicted values could lead them toward improved decisions of trading. Thus, based on the fact that prediction of direction plays an essential role in efficiency of market forecasting methods, an improved method of learning is presented in which an improved punishment function is proposed in 9 which include a linear coefficient depending on the DS that is tend to be used as an evaluation criterion of direction prediction thus far[38], [39]. The DS is a statistical measure of a model's performance in predicting the direction of change, positive or negative, of a time series from one time period to the next.

$$\begin{aligned}
\tilde{E}_1 &= (\alpha+\beta.DS)E(t+2)\\
\tilde{E}_2 &= (\alpha+\beta.DS)\hat{E}(t+3)
\end{aligned} \tag{9}$$

where α and β are constant coefficient and DS is directional symmetry which is defined in 10.

$$DS = \frac{1}{M}\sum_{t=1}^{N}a(t)\times 100\% \tag{10}$$

where, M is the length of input signal and parameter $a(t)$ is defined by 11.

$$a(t) = \begin{cases} 1 & if\big(d(t)-d(t-1)\big)\big(z(t)-z(t-1)\big)>0\\ o & otherwise\end{cases} \tag{11}$$

In other words, DS is introduced to measure the performance of a model in predicting the direction of changes and DS=100 percent indicates the fact that the change of the direction has been perfectly predicted for time series from one period to the next.

B. Adaptive Learning Rate

The fixed learning rate may speed up the convergence of the error which could cause divergence from data trend. On the other hand, the convergence rate could not be adjusted by the fixed learning rate. In this section, to overcome shortcomings of fixed learning rate, an online adaptive learning rate $\eta(t)$ is proposed for updating weights, which improves the learning speed and performance effectiveness. The online adaptive learning rate, $\eta(t)$, updates via meta learning rate through time[38]. The adaptive updating of network weights is proposed in 12 and the learning rate updates through 13.

$$\Delta w_{lmin}(t) = -\eta_l\frac{\partial E(t+2)}{\partial w_{lmin}(t)} \tag{12}$$

$$\eta_l(t) = \eta_l(t-1) + \frac{\partial E(t+2)}{\partial \eta_l(t)} = \eta_l(t-1) + \frac{\partial E(t+2)}{\partial w_{lmin}(t)} \times \frac{\partial E(t+2)}{\partial w_{lmin}} \quad (13)$$

The training phase algorithm for the Improved R-RTRL(IR-RTRL) which is illustrated in figure 2, could be wrapped up as follows:

1) Initialize the network(RNN)
2) Apply input $x(t)$ to the RNN and get corresponding output z(t+2)
3) Compare z(t+2) with desired output $d(t+2)$ and get $E(t+2)$
4) Calculate DS
5) Update weights based on gradient method over punishment function $(\alpha+\beta.DS)(E(t+2))$ with the adaptive learning rate and get the temporal neural network, RNN_{temp}
6) Apply input x(t+1) to RNN_{temp} and get corresponding output $\hat{z}(t+3)$
7) Compare $\hat{z}(t+3)$ with the desired output $d(t+3)$ and get $E(t+3)$
8) Update weights based on gradient method over punishment function $(\alpha+\beta.dsf)\big(\hat{E}(t+2)\big)$ with the adaptive learning rate and get RNN for next iteration
9) Go to step 2

III. Model Test and Verification

We apply the proposed learning algorithm to a RNN and compare it with R-RTRL and Back Propagation Neural Network(BPNN) learning algorithms for the same network. Chaotic behavior is commonly observed in economic systems[40] and many studies validate chaotic behavior of economic systems[41], [42]. The forecast ability of the IR-RTRL networks is become manifest by making a comparison between two chaotic benchmark time series of Lorenz and Rossler, which number of researchers widely used and reported while comparing the learning ability of different neural networks and they both display chaotic behavior like financial time series. In order to compare the results, the Mean Square Error (MSE) and the Normalized Mean Square Error (NMSE) criterions given by 14 and 15 are employed.

$$MSE = \frac{1}{M}(z(t) - d(t))^2 \quad (14)$$

$$NMSE = \frac{(z(t) - d(t))^2}{(d(t) - \bar{d})^2} \quad (15)$$

where, M is the length of input signal and $\bar{d}$ is the average of the observed values. The Rossler system is a system of three non-linear ordinary differential equations. These differential equations define a continuous-time dynamical system that exhibits chaotic dynamics associated with the fractal properties of the attractor. The Rossler time series is given by 16.

$$\begin{aligned} dx/dt &= -z - y \\ dy/dt &= x + a \times y \\ dz/dt &= b + z \times (x - c) \end{aligned} \quad (16)$$

TABLE I: MODEL PERFORMANCE OF 2SA FORECASTING FOR ROSSLER TIME SERIES

	Training		Testing	
	MSE	NMSE	MSE	NMSE
IR-RTRL	5.50E-05	3.14E-03	3.86E-04	2.72E-02
R-RTRL	6.95E-05	4.00E-03	1.68E-03	8.12E-02
BPNN	2.72E-04	1.63E-02	2.44E-03	1.48E-01

TABLE II: MODEL PERFORMANCE OF 2SA FORECASTING FOR Lorenz TIME SERIES

	Training		Testing	
	MSE	NMSE	MSE	NMSE
IR-RTRL	4.59E-05	4.65E-03	1.67E-04	2.08E-02
R-RTRL	3.25E-04	3.45E-02	5.12E-04	7.74E-02
BPNN	1.91E-03	3.04E-01	2.27E-03	7.86E-01

The Lorenz series is also a system of three non-linear ordinary differential equations given by 17.

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma[y(t) - x(t)] \\ \frac{dy(t)}{dt} &= x(t)[r - z(t)] - y(t) \\ \frac{dz(t)}{dt} &= x(t)y(t) - bz(t) \end{aligned} \quad (17)$$

where x,y and z are the system state, t is time, and σ,b are the system parameters. As depicted in figure 3, training accuracy improves as the number of processing cycle increases. On the other hand, as the number of iteration increases during neural network training, the validation test error becomes higher and the network appears to become over trained when the number of iterations reaches a specific numbers.

Over training occurs when training data, which are already well modeled by the algorithm, continues to be iterated through the model and the number of epochs continues to increase. Over training is caused by the network memorizing the input-output pairs and becoming less able to recognize similar unknown input-output patterns which could be used as a validation tool of a neural network. The accuracy of predictions decreases when unknown test data are presented to the over trained network[43]. Additionally, the figure 3 is shown so as to make the fact that after special number of epochs, the network gets over trained.

Figure 4 and figure 5 and also tables I and table II demonstrate the fact that the BPNN has the worst performance in training and testing phase and the proposed IR-RTRL has the best testing and training performance in both Rossler and Lorenz time series which reveals the fact that the proposed method could appropriately adjust weights that leads to a reliable and accurate 2SA forecast.

IV. Application

Daily gold price data from the database of Bloomberg, which is the open access database including historical data of gold market is used to perform simulations. In this work, spanning data from 21 December 2012 to 12 July 2013 a total of 792 data are used for training and spanning data from 14 July 2013 to 15 August 2014 a total of 339 data for evaluating the model. Besides, all spanning data from 21 December 2012

TABLE III: Model performance of 2SA forecast of gold market price for online and offline learning

	Online Learning		Offline Learning			
			Testing		Training	
	NMSE	MSE	NMSE	MSE	NMSE	MSE
IR-RTRL	8.26E-03	2.64E+02	1.82E-01	3.32E+03	3.57E-03	5.29E+01
R-RTRL	1.44E-02	4.58E+02	3.85E+00	2.06E+04	9.38E-03	1.41E+02
BPNN	2.76E-02	8.59E+02	4.83E+00	2.2404e+04	1.38E-02	2.02E+02

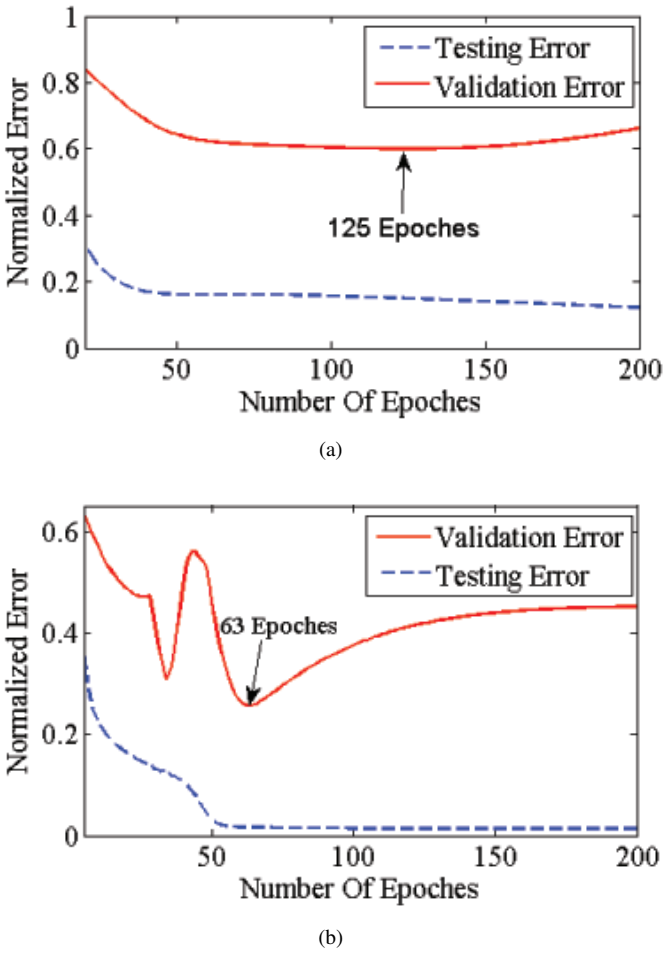


Fig. 3. Effect of increasing the number of epochs on total system error of 3(a) Rossler system and 3(b) Lorenz system respectively.

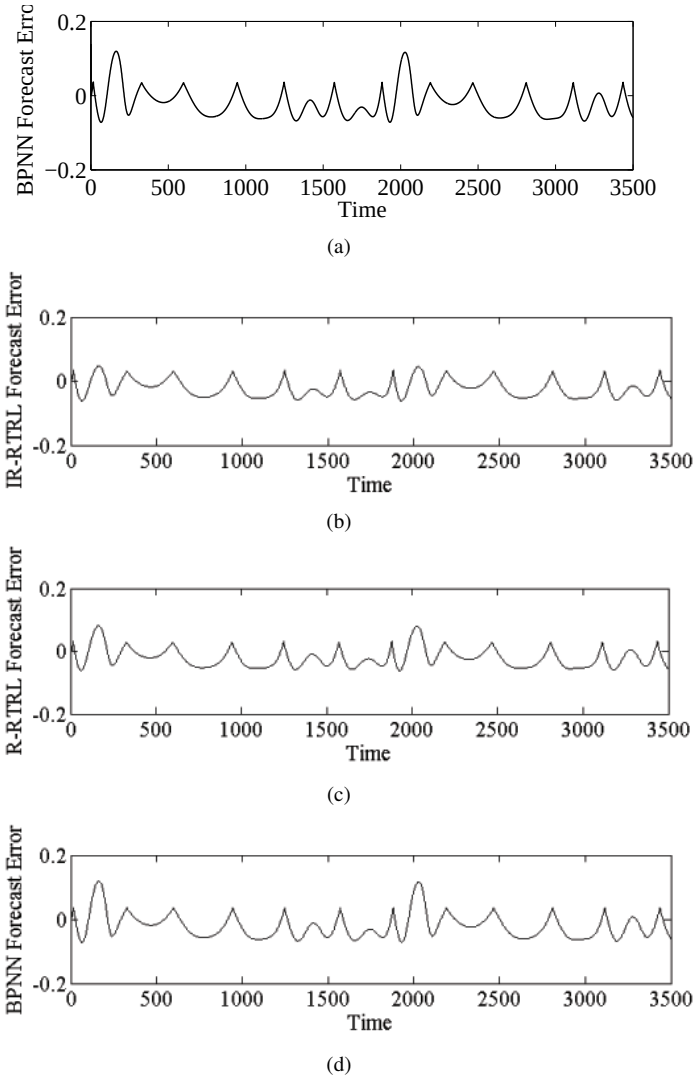


Fig. 4. 4(a)Rossler time series and online 2SA forecast error of it based on 4(b) Proposed IR-RTRL, 4(c) R-RTRL and 4(d) BPNN, respectively.

to 15 August 2014, a total of 1131 observations are used in online training of the network and more importance is given to the new data by weakening the influence of older data points. In each level of online learning, the modification of weights is performed on a time window which contains the specified number of data till considered point. Figure 6 depicts flow chart of the online learning scheme for the proposed IR-RTRL. As shown in the figure, after initializing network and parameters, a specified number of past data forms a time window of data which is used as network input. After training network for a specified input, to modify the structure of the training algorithm, the correctness of direction is being checked and in case of wrong detection, the training stage reiterates and the algorithm continues with the input data of next iteration. We depict the predominance of the proposed method by comparing the performance of it with R-RTRL network and BPNN. Number of neurons in the processing layer for all three neural networks is taken equal to have a better comparison. Thus, 2SA forecasting is performed based on nominated methods for gold price. As depicted in table III, results demonstrate that the proposed algorithm has smaller NMSE and MSE. According to the table, MSE of online R-RTRL learning is larger than twice the MSE of online IR-RTRL learning and for offline learning process, the priority of proposed method is more evident. Considering the results and comparing them, the proposed method has decreased 2SA prediction errors and it can forecast 2SA gold price values

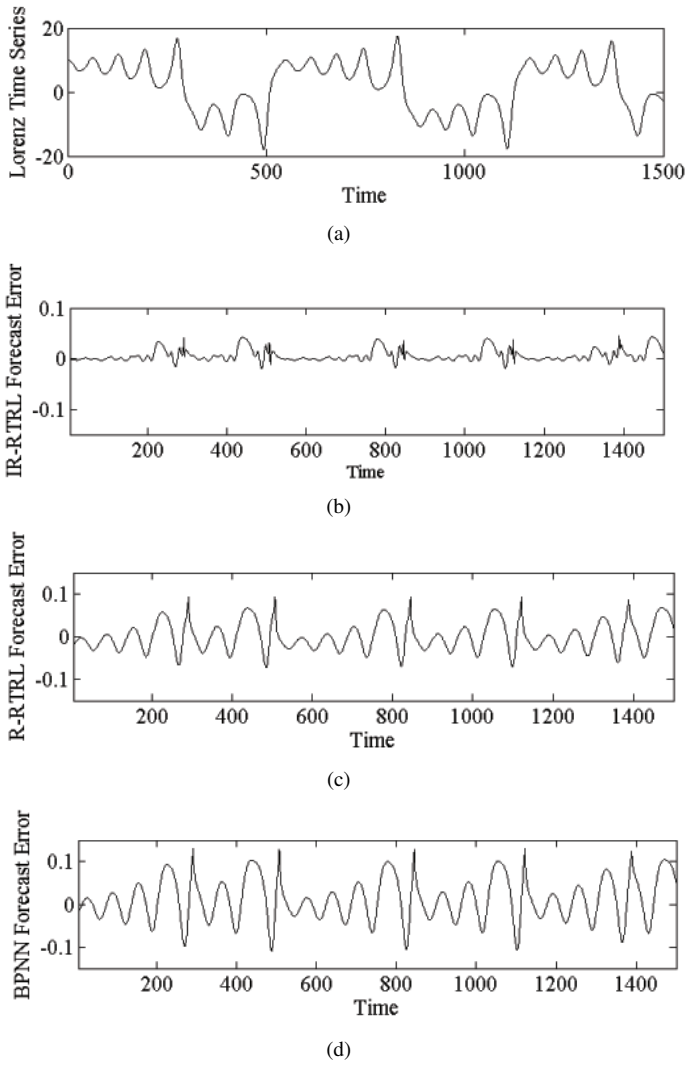


Fig. 5. 5(a)Lorenz time series and online 2SA forecast error of it based on 5(b) Proposed IR-RTRL, 5(c) R-RTRL and 5(d) BPNN, respectively.

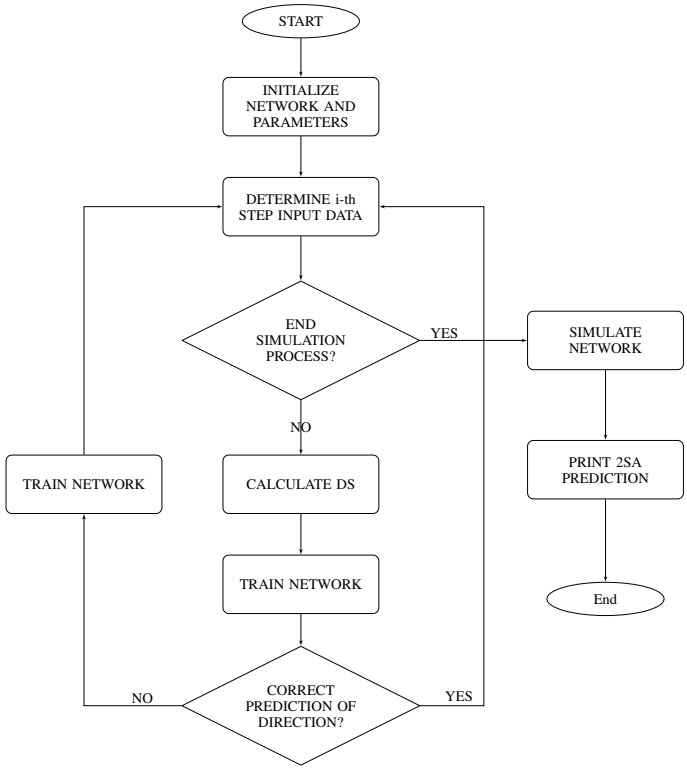


Fig. 6. Flow chart of the online learning algorithm for I-RRNN

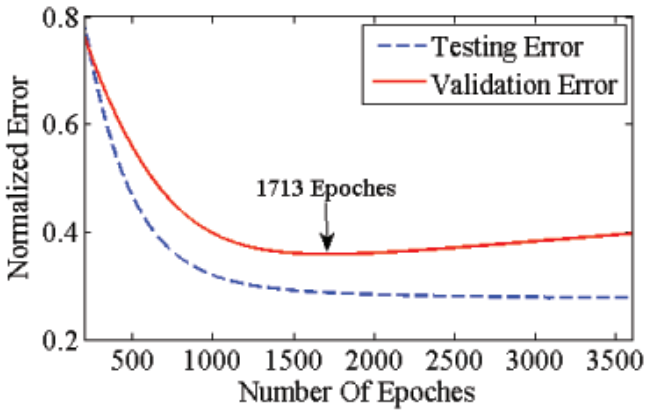


Fig. 7. Effect of increasing the number of epochs on total system error of Gold data

well.

According to figure 7, training accuracy improves as the number of processing cycle increases and yet, after the number of iterations reaches a specific numbers, the network appears to become over trained. The figure is shown as to make the overtraining edge and the intention of this analysis is to avoid over training of the proposed network. Figure figure 8 depicts the offline forecast errors of the Gold market based on the proposed IR-RTRL, R-RTRL and BPNN. Furthermore, figure 9 depicts the online forecast errors of Gold market based on the proposed IR-RTRL, R-RTRL and BPNN. Clearly, the forecast error of the proposed IR-RTRL is significantly smaller than the forecast error of R-RTRL and BPNN and one could draw the conclusion that the proposed learning algorithm is more efficient for online and offline learning. As a conclusion, the IR-RTRL network performs an efficient and precise prediction of future gold price for both online and offline forecasting issues.

V. Conclusion

An efficient and precise prediction of future price can be very valuable. The RTRL systems can successfully design powerful and complicated model with high precision for 2SA predictions. In this paper, the IR-RTRL is proposed to train a new neural network model for 2SA gold price forecasting. Empirical results obtained demonstrate the ability of effective learning and precise predicting of the model and reveal that the prediction using IR-RTRL model is better than results obtained through other models presented in this study. The extended model could be used for multi-step ahead prediction which may cause a lower prediction error and higher efficiency for predicting gold price. Moreover, the result of this prediction

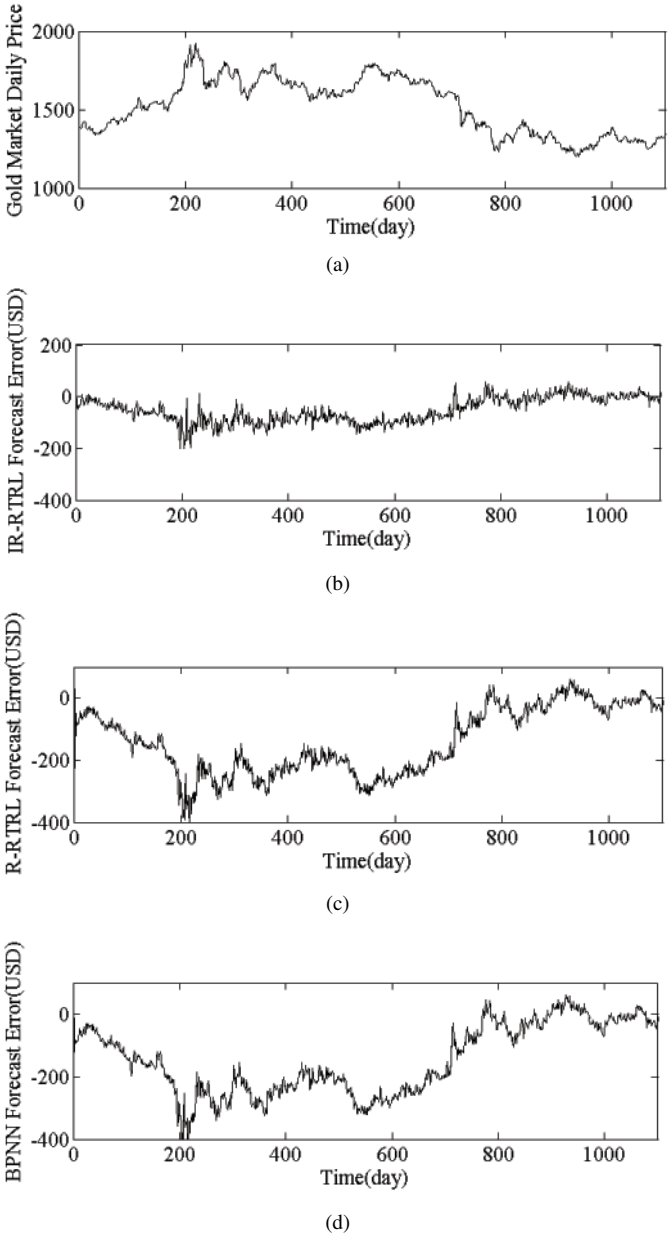


Fig. 8. 8(a)Gold market price and Offline 2SA forecast error of it based on 8(b) Proposed IR-RTRL, 8(c) R-RTRL and 8(d) BPNN, respectively.

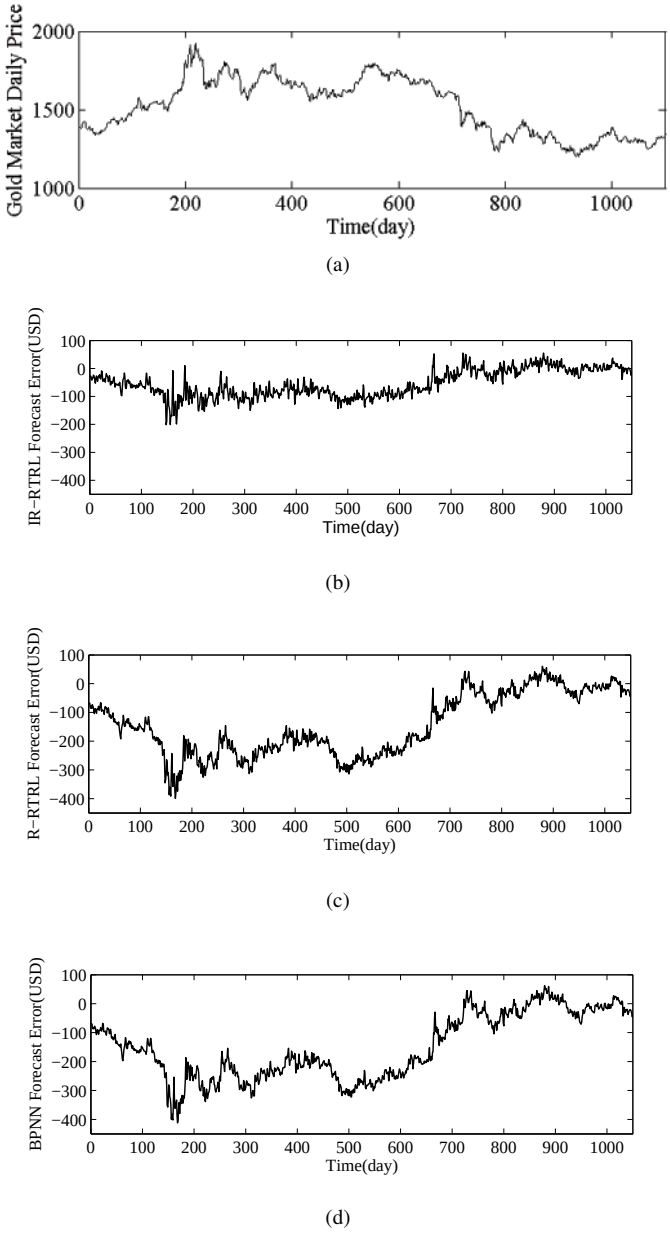


Fig. 9. 9(a)Gold market price and online 2SA forecast error of it based on 9(b) Proposed IR-RTRL, 9(c) R-RTRL and 9(d) BPNN, respectively.

could be used as 2SA price expectation in heterogeneous agent models for gold market modeling.

REFERENCES

[1] Gary Grudnitski and Larry Osburn. Forecasting s&p and gold futures prices: an application of neural networks. *Journal of Futures Markets*, 13(6):631–643, 1993.

[2] Shahriar Shafiee and Erkan Topal. An overview of global gold market and gold price forecasting. *Resources Policy*, 35(3):178–189, 2010.

[3] Haibin Cheng, Pang-Ning Tan, Jing Gao, and Jerry Scripps. Multistep-ahead time series prediction. In *Advances in Knowledge Discovery and Data Mining*, pages 765–774. Springer, 2006.

[4] Ramazan Gençay, Michel Dacorogna, Ulrich A Muller, Olivier Pictet, and Richard Olsen. *An introduction to high-frequency finance*. Academic Press, 2001.

[5] Yudong Zhang and Lenan Wu. Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. *Expert systems with applications*, 36(5):8849–8854, 2009.

[6] George S Atsalakis and Kimon P Valavanis. Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications*, 36(7):10696–10707, 2009.

[7] Mikhail Anufriev, Cars Hommes, and Tomasz Makarewicz. Learning to forecast with genetic algorithm. Technical report, Tech. rep.(February 2013), 2013.

[8] Blake LeBaron. Building the santa fe artificial stock market. *Physica A*, 2002.

[9] Paresh Kumar Narayan, Seema Narayan, and Xinwei Zheng. Gold and oil futures markets: Are markets efficient? *Applied energy*, 87(10):3299–3303, 2010.

[10] Yudong Wang, Yu Wei, and Chongfeng Wu. Analysis of the efficiency and multifractality of gold markets based on multifractal detrended fluc-

tuation analysis. *Physica A: Statistical Mechanics and its Applications*, 390(5):817–827, 2011.

[11] Yue-Jun Zhang and Yi-Ming Wei. The crude oil market and the gold market: Evidence for cointegration, causality and price discovery. *Resources Policy*, 35(3):168–177, 2010.

[12] Thomas W Epps and Mary Lee Epps. The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica: Journal of the Econometric Society*, pages 305–321, 1976.

[13] Robert D Edwards, John Magee, and WHC Bassetti. *Technical analysis of stock trends*. CRC Press, 2012.

[14] Rick Martinelli and Barry Hyman. Cup-with-handle and the computerized approach. *TECHNICAL ANALYSIS OF STOCKS AND COMMODITIES-MAGAZINE EDITION-*, 16:63–66, 1998.

[15] Jack L Treynor and Robert Ferguson. In defense of technical analysis. *The Journal of Finance*, 40(3):757–773, 1985.

[16] Soni Chaturvedi, Neha Sondhiya, et al. Review of handwritten pattern recognition of digits and special characters using feed forward neural network and izhikevich neural model. In *Electronic Systems, Signal Processing and Computing Technologies (ICESC), 2014 International Conference on*, pages 425–428. IEEE, 2014.

[17] Jabal Raval and Bhushan Jagyasi. Distributed detection in neural network based multihop wireless sensor network. In *Sensors Applications Symposium (SAS), 2014 IEEE*, pages 65–68. IEEE, 2014.

[18] Shuai Li and Yangming Li. Nonlinearly activated neural network for solving time-varying complex sylvester equation. 2013.

[19] Tianzhu Wen, Aiqiang Xu, Chunxia Liu, and Nan Li. Application of rbf neural network based on enn2 clustering in fault diagnosis. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on*, volume 2, pages 71–74. IEEE, 2014.

[20] Vahid Nourani and Mina Sayyah Fard. Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. *Advances in Engineering Software*, 47(1):127–146, 2012.

[21] Li-Chiu Chang, Fi-John Chang, and Yuan-Peng Wang. Auto-configuring radial basis function networks for chaotic time series and flood forecasting. *Hydrological processes*, 23(17):2450–2459, 2009.

[22] Yung-hsiang Chen and Fi-John Chang. Evolutionary artificial neural networks for hydrological systems forecasting. *Journal of Hydrology*, 367(1):125–137, 2009.

[23] Li-Chiu Chang, Fi-John Chang, and Hung-Cheng Hsu. Real-time reservoir operation for flood control using artificial intelligent techniques. *International Journal of Nonlinear Sciences and Numerical Simulation*, 11(11):887–902, 2010.

[24] Saeed Moshiri, Norman E Cameron, and David Scuse. Static, dynamic, and hybrid neural networks in forecasting inflation. *Computational Economics*, 14(3):219–235, 1999.

[25] Vassilis Kodogiannis and A Lolis. Forecasting financial time series using neural network and fuzzy system-based techniques. *Neural computing & applications*, 11(2):90–102, 2002.

[26] Li-Chiu Chang, Pin-An Chen, and F-J Chang. Reinforced two-step-ahead weight adjustment technique for online training of recurrent neural networks. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(8):1269–1278, 2012.

[27] Pratik Roy, GS Mahapatra, Pooja Rani, SK Pandey, and KN Dey. Robust feedforward and recurrent neural network based dynamic weighted combination models for software reliability prediction. *Applied Soft Computing*, 2014.

[28] Han Wang and Gangbing Song. Innovative narx recurrent neural network model for ultra-thin shape memory alloy wire. *Neurocomputing*, 134:289–295, 2014.

[29] John Covert and David L Livingston. A vacuum-tube guitar amplifier model using a recurrent neural network. In *Southeastcon, 2013 Proceedings of IEEE*, pages 1–5. IEEE, 2013.

[30] Qili Chen, Wei Chai, and Junfei Qiao. Modeling of wastewater treatment process using recurrent neural network. In *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, pages 5872–5876. IEEE, 2010.

[31] V Sharma and D Srinivasan. A hybrid intelligent model based on recurrent neural networks and excitable dynamics for price prediction in deregulated electricity market. *Engineering Applications of Artificial Intelligence*, 26(5):1562–1574, 2013.

[32] Derrick T Mirikitani and Nikolay Nikolaev. Recursive bayesian recurrent neural networks for time-series modeling. *Neural Networks, IEEE Transactions on*, 21(2):262–274, 2010.

[33] SV Barai, AK Dikshit, and Sameer Sharma. Neural network models for air quality prediction: a comparative study. In *Soft Computing in Industrial Applications*, pages 290–305. Springer, 2007.

[34] Sofien Chtourou, Mohamed Chtourou, and Omar Hammami. A hybrid approach for training recurrent neural networks: application to multi-step-ahead prediction of noisy and large data sets. *Neural Computing and Applications*, 17(3):245–254, 2008.

[35] Jing-Xin Xie, Chun-Tian Cheng, Kwok-Wing Chau, and Yong-Zhen Pei. A hybrid adaptive time-delay neural network model for multi-step-ahead prediction of sunspot activity. *International Journal of Environment and Pollution*, 28(3):364–381, 2006.

[36] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[37] Li-Chiu Chang, Fi-John Chang, and Yen-Ming Chiang. A two-step-ahead recurrent neural network for stream-flow forecasting. *Hydrological Processes*, 18(1):81–92, 2004.

[38] Lean Yu, Shouyang Wang, and Kin Keung Lai. An emd-based neural network ensemble learning model for world crude oil spot price forecasting. In *Soft Computing Applications in Business*, pages 261–271. Springer, 2008.

[39] Shifei Zhou, Kin Keung Lai, and Jerome Yen. A dynamic meta-learning rate-based model for gold market forecasting. *Expert Systems with Applications*, 39(6):6168–6173, 2012.

[40] Roderick V Jensen and Robin Urban. Chaotic price behavior in a nonlinear cobweb model. *Economics Letters*, 15(3):235–240, 1984.

[41] Richard Murphey Goodwin. *Chaotic economic dynamics*. Clarendon Press Oxford, 1990.

[42] Cars H Hommes. Heterogeneous agent models in economics and finance. *Handbook of computational economics*, 2:1109–1186, 2006.

[43] S Hykin. Neural networks: A comprehensive foundation. printice-hall. *Inc., New Jersey*, 1999.

Applications of Multi-criteria Decision Making in Software Engineering

Sumeet Kaur Sehra
Research Scholar(I.K.G.P.T.U., Jalandhar)
Assistant Professor
Guru Nanak Dev Engg. College, Ludhiana

Yadwinder Singh Brar
Professor
Guru Nanak Dev Engg. College
Ludhiana

Navdeep Kaur
Associate Professor
Shri Guru Granth Sahib World University
Fatehgarh Sahib

***Abstract*—Every complex problem now days require multi-criteria decision making to get to the desired solution. Numerous Multi-criteria decision making (MCDM) approaches have evolved over recent time to accommodate various application areas and have been recently explored as alternative to solve complex software engineering problems. Most widely used approach is Analytic Hierarchy Process that combines mathematics and expert judgment. Analytic Hierarchy Process suffers from the problem of imprecision and subjectivity. This paper proposes to use Fuzzy AHP (FAHP) instead of traditional AHP method. The usage of FAHP helps decision makers to make better choices both in relation to tangible criteria and intangible criteria. The paper provides a clear guide on how FAHP can be applied, particularly in the software engineering area in specific situations. The conclusion of this study would help and motivate practitioners and researchers to use multi-criteria decision making approaches in the area of software engineering.**

***Keywords*—Multi-criteria Decision Making; Analytic Hierarchy Process; Fuzzy AHP; Software Engineering**

I. Introduction

Multi-Criteria decision making (MCDM) approaches take decisions in the presence of multiple, usually conflicting, criterion. MCDM approach handles both quantitative and qualitative choices and is able to combine the historical data and expert opinion by quantifying subjective judgement [1]. There are many MCDM models which include Analytic Hierarchy Process (AHP), PROMOTHEE, ELECTRE, TOPSIS, VIKOR each having different algorithm [2]. Most widely used MCDM technique is AHP, developed by Saaty and inspired by the intelligent behaviour of human beings. Since judgments given by decision makers are relative, any change in the relative values of the choices may significantly change the weights of affected choices, resulting in a problem known as Rank Reversal [3]. The problem of imprecision and subjectivity in the weight calculation process is not handled in AHP and these problems can be overcome by using Fuzzy AHP. Software Engineering has always been an area of concern for researchers because of its real time applications in the era of computer science. In most of the applications the final decision is dependent on the outcome ranking of alternatives in respective to criterion [4]. Software development and evolution is characterized by multiple objectives and constraints [5]. Nowadays the problems have been become more and more complex and depend upon multiple factors. So applying multi-criteria decision making (MCDM) approaches for solving complex problems dependent on multiple aspects is required than simple linear algorithmic approaches. This paper focuses on AHP, FAHP and their comparison by taking a working example and how FAHP is widely accepted approach in the field of software engineering. The next section discusses about AHP, Fuzzy AHP process in detail. The further section summarizes the different application areas in which Fuzzy AHP can be used. Then an example illustrates the use of Fuzzy AHP in selecting the quality model. The last section concludes and gives the future scope of the paper.

II. Multi-criteria Decision Making Approaches

A. Analytic Hierarchy Process

Analytic Hierarchy Process (AHP) is an MCDM approach, proposed by Saaty [6], for handling multi objective problems. This approach selects best alternatives based on criterion [7]. AHP is well structured mathematical approach uses consistent matrices and their associated eigenvectors to produce relative weights[8]. AHP combines historical data and expert opinion by quantifying subjective judgement [9]. It structures the given problem as a hierarchy, with required goal as parent node and criteria for assessing it are placed in levels below it. Weights are assigned to each node and many pairwise comparisons and matrix multiplications are made assessing the relative importance of these criteria. The end result of this method is to provide a formal, systematic means of extracting, combining, and capturing expert judgements and their relationship to analogous reference data [10].

The steps followed by AHP for concluding the relative rankings of alternatives are as follows [3]:

1) Decomposition of problem to required goal, criterion, alternatives.
2) Read the decision values/variable.
3) Creating the reciprocal matrix for the pairwise comparisons of criterion.
4) Find Eigen values and calculate the Eigen vector for computing weights.
5) Find the consistency index of the weight.
6) Repeat the steps from 1 to 5 for each value criterion.
7) Calculate the overall weight vector of the hierarchy.
8) Infer the alternative based on the overall weight vector.

TABLE I: Saaty's scale for pairwise comparison [9]

Saaty's scale	The relative importance of the two sub-elements
1	Equally important
3	Moderately important with one over another
5	Strongly important
7	Very strongly important
9	Extremely important
2,4,6,8	Intermediate values

Based on the complexity of the problem the number of levels in the hierarchy may increase. Saaty 9 point scale in tableI is used by expert to describe the relative ranking of one alternative over other alternative.

Based on Saaty's scale, experts develop reciprocal matrix A, in which values are representing the dominance of ith element on jth element as shown in equation 1.

1) $a_{ij} = 1/a_{ji}$, for $a_{ij} \neq 0$
2) $a_{ij} = 1$, for $i = j$ and $i, j = 1, 2, ..., n$.

The reciprocal matrix is as given below

$$A' = \begin{bmatrix} 1 & a12 & ... & a1n \\ 1/a12 & 1 & ... & a2n \\ . & . & . & . \\ 1/a1n & 1/a2n & . & 1 \end{bmatrix} \quad (1)$$

Finally, vector of weights is generated as normalized eigenvector using equation 2.

$$Aw{=}\lambda_{\max} w \quad (2)$$

Where λmax represents eigenvalues and the resultant vector of relative weights is given as in equation 3.

$$w = [w_1 w_2 ..... w_n]^T \quad (3)$$

In a situation where many pairwise comparisons are performed, inconsistencies may typically arise. The AHP has an effective method for identifying the inconsistency in the comparison made by the decision maker, called Consistency Index. It is calculated by equation 4.

$$CI = \frac{\lambda_{\max}}{n-1} \quad (4)$$

AHP offers advantages in comparison to other approached as being flexible, intuitive and simplistic approach for experts. Despite being popular approach, there are certain issue needs addressing. Firstly, as judgements given by experts are relative, so any arbitrary change in value of alternatives may affect the weights of other alternatives resulting in a problem known as Rank Reversal [3]. Another issue with AHP is its subjectivity and imprecision due to Saaty's 9 point scale [9]. Also, the AHP method cannot represent some comparison in which the alternative are bigger than 9 point.

B. Fuzzy AHP

Due to shortcomings incurred by the subjectivity of human judgements and vagueness of the data, the fuzzy logic, introduced by Zadeh, can be utilized. The combination of fuzzy logic and analytic hierarchy process is a hybrid approach for both qualitative and quantitative criteria comparison using expert judgements to find weights and relative rankings. Since most of multi-criteria methods suffers from vagueness, FAHP approach can better tolerate this vagueness [11], [12]. The combination of both generates credible results than conventional AHP [13], [14]. In FAHP, expert judgement is represented as a range of values instead of single crisp values. The range values can be given as optimistic, pessimistic or moderate [10]. For finding the weights the extent analysis method given by Wang *et al* [15] is utilized. The membership function used for creating the fuzzy set is given in equation 5, where x is the weight of relative importance of one criterion over other criterion.

$$\mu_A(x) = \begin{Bmatrix} 0 \quad for \quad x \leq l \\ \frac{x-l}{m-l} \quad for \quad l \leq x \leq m \\ \frac{u-x}{u-m} \quad for \quad m \leq x \leq u \\ 0 \quad for \quad x \geq n \end{Bmatrix} \quad (5)$$

Triangular fuzzy numbers (TFN) provide an opportunity in deciding the weight of one alternative over the other. TFN is represented by equation 6.

$$aij = (lij, mij, uij) \quad (6)$$

Where l,m,u are pessimistic, moderate and optimistic values respectively.

The modified Saaty scale using TFN is given in table II. In FAHP table 2 is used for construction comparison matrix A= (aij) nxn where i, j = 1, 2, 3n. The next step is to use extent analysis method to calculate the relative ranking of alternatives, the synthetic extent values are obtained by equation 7.

$$S_i = \sum_{j=1}^{m} N_{ci}^j \otimes \left[\sum_{i=1}^{n} \sum_{j=1}^{m} N_{ci}^j \right]^{-1} \quad (7)$$

The degree of possibility of $M_1 \geq M_2$ is defined in equation 8.

$$V(N_1 \geq N_2) = sup[min(\mu_{N1}(x)), \mu_{N2}(y))] \quad (8)$$

$$V(N2 \geq N1) = hgt(N_1 \cap N_2) = \mu_{N1}(d) = \begin{pmatrix} 1 \quad if \quad m2 > m1 \\ 0 \quad if \quad l1 \geq u2 \\ \frac{l_1 - u_2}{(m_2 - l_2) - (m_1 - l_1)}, otherwise \end{pmatrix} \quad (9)$$

In equation 9, d is representing ordinate of the highest intersection point between between μ_{N1} and μ_{N2}.

The degree of possibility for a convex fuzzy number, is defined by equation 10.

$$V(N \geq N_1, N_2, ..., N_k) = [(N \geq N_1), ..., (N \geq N_k)] = \min V(N \geq N_i) \quad (10)$$

TABLE II: Linguistic Scale for Fuzzy AHP

Linguistic scale for importance	Fuzzy numbers for FAHP	Membership function	Domain	Triangular fuzzy scale (l, m, u)
Just Equal	1			(1.0, 1.0, 1.0)
Equally important	1	$\mu M(x) = (3-x)/(3-1)$	$1 \leq x \leq 3$	(1.0, 1.0, 3.0)
Weak importance over each another	3	$\mu M(x) = (x-1)/(3-1)$	$1 \leq x \leq 3$	(1.0, 3.0, 5.0)
		$\mu M(x) = (5-x)/(5-3)$	$3 \leq x \leq 5$	
Essential importance over each other	5	$\mu M(x) = (x-3)/(5-3)$	$3 \leq x \leq 5$	(3.0, 5.0, 7.0)
		$\mu M(x) = (7-x)/(7-5)$	$5 \leq x \leq 7$	
Very strong importance over other	7	$\mu M(x) = (x-5)/(7-5)$	$5 \leq x \leq 7$	(5.0, 7.0, 9.0)
		$\mu M(x) = (9-x)/(9-7)$	$7 \leq x \leq 9$	
Extreme importance over other	9	$\mu M(x) = (x-7)/(9-7)$	$7 \leq x \leq 9$	(7.0, 9.0, 9.0)
The value of second element in comparison to first would be by reciprocal of TFN given as (1/u1,1/m1,1/l1)				

In order to normalize the weight vector,equation 11 is used.

$$W_A = \frac{W^T}{\sum(W^T)} \tag{11}$$

After calculating the weights of criteria, the scores of alternatives with respect to each criterion is evaluated and composite weights of the decision alternatives are determined by aggregating the weights through hierarchy.

III. FUZZY AHP IN SOFTWARE ENGINEERING

Many application areas in the field of software engineering have been identified as the problems, where MCDM is utilized for solving multi-objective problems. Some application areas are discussed below:

A. Evaluation and Assessment

Sarfaraj *et al* [16] have successfully used Fuzzy AHP technique for identifying the appropriate web development platform. The proposed model took into account four criteria, namely security (C1), compatibility (C2), performance (C3) and licensing cost (C4) for choosing the best platform. Three alternatives namely, (Linux/Apache/ MySQL/PHP (LAMP) (A1), Microsoft's ASP.NET (A2) and Sun's Java 2 Enterprise Edition (J2EE) (A3) are evaluated at the level of problem hierarchy. The conclusion of the work is that criteria 'security' is most significant of all others and LAMP is chosen as web development platform. Fuzzy AHP approach is proposed by Vatansever and Akgul [17] for assessing the quality of service delivery of websites. In the study, the quality of four eCommerce company web sites which operate in Turkey having the highest sales volume have been analysed with the fuzzy AHP approach. The criteria used for evaluating the web site quality were 4 main and 22 sub-criteria. The most significant criteria affecting the quality of the Web site were determined as the information quality, system quality, service quality, and vendor specific quality. The most significant factor affecting the quality of the website is the vendor specific quality as per the proposed model using fuzzy AHP. A study has been conducted by Kong and Liu [18], about ranking of the factors behind the success of E-commerce. They have considered Trust, System Quality, Content Quality, Use and Online Service as 5 main criteria and 17 sub-criteria. They have concluded that Trust is the most critical factor and Security is the most critical sub-factor of Trust. All other factors have also been ranked by using FAHP.

B. Risk Analysis and Ranking

Risk analysis is procedure of finding, analysing and handling identified risk factors throughout the life cycle of a software project. The Fuzzy AHP provides the flexible and easily understood way to analyse project risks. Kahraman and Tuysuz [19] have suggested that MCDM can be used evaluation and assessment of project risks. They have measured the risk level of an information technology product by considering six different risk groups which can further be divided into 28 sub-risk factors by using FAHP. Lee [20] has used FAHP for information security risk assessment. He considered four criteria namely, assets, threats, vulnerability and safety measures for pairwise comparisons. Risk factors and the Customer-to-Customer E-commerce transaction system's security risk level can be identified by incorporating Fuzzy AHP as suggested by Wei et al [21]. Ranking of risks has been achieved by Askari et al [22] by identifying the project objectives and alternatives i.e. risks of creating a FAHP model. By considering the different alternatives, the weights are calculated and then a ranking is assigned to the risks.

C. Quality Evaluation

It is hard to measure and quantify the quality of software product, due to that the approach followed is to evaluate development software quality of vendors. The selection of vendors has been shown in numerically by taking the data of a company designing and manufacturing smartphone. The criterion namely functionality, reliability, usability, efficiency, maintainability and portability of software quality model ISO/IEC 9126 are chosen as evaluation criteria for selecting an alternative i.e. vendor in the case. Challa et al [23] have developed a tool based upon the algorithm using Fuzzy AHP as the base for selecting the quality parameters. They considered the developer's perspective, the user's perspective, and the project manager's perspective. They also added several new sub characteristics to the base model i.e. ISO/IEC 9126.

D. Software Project Selection

Selection of software projects can be done by using MCDM methods. For this selection three phases are proposed by Bakshi et al [24]. In first phase set of alternatives are identified, second phase uses quality function to find best alternatives. The sensitivity analysis is performed in the last phase for checking the robustness of selection methodology. Jusoh *et al*

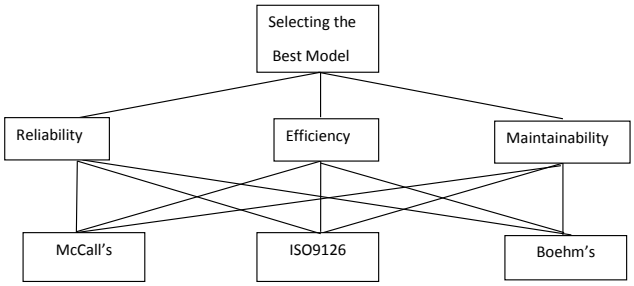


Fig. 1: Hierarchy of problem

TABLE III: Normalized weights of Criterion Using FAHP

Goal	Reliability	Efficiency	Maintainability	Normalized weight
Reliability	[1,1,1]	[1/7,1/5,1/3]	[3,5,7]	0.306
Efficiency	[3,5,7]	[1,1,1]	[1/9,1/7,1/5]	0.302
Maintainability	[1/7,1/5,1/3]	[5,7,9]	[1,1,1]	0.391

[25] have proposed FAHP for the selection process of open source software (OSS) products based on selection criteria for adopting the OSS are reliability, usability, performance efficiency, functionality and competence. According to Khan *et al* [26], the success of a software system developed depends on the software development life cycle (SDLC) models during the development process. A method for project selection is suggested by Mahmoodzadeh et al [27] using fuzzy AHP and TOPSIS technique. The four methods namely, net present value, rate of return, benefit cost analysis and payback period, used for comparing investment alternatives are taken as criteria in FAHP.

E. Testing Adequacy Criteria for UML Design Models

A model is presented by Srivastava and Ray [28] for comparing an automated functional and regression testing tool using the FAHP. Triangular Fuzzy numbers are used by decision makers from CMM level five organizations. They discussed the use of FAHP approach for incorporating the three aspects namely, DCD criteria, Interaction diagram criteria and deriving test objectives in testing adequacy criteria for UML design models. They stated that FAHP has the ability to cater to uncertain and imprecise data. Upon evaluating the framework it was identified that DCD criteria is the preferred decision testing adequacy criteria for UML Design Models. Belton and Stewart [29] have assessed four aspect oriented programs qualitatively based on the five factors of software testability i.e. controllability, observability, built in test capability, understandability and complexity. Different testing environments and software change characteristics can affect the choice of regression testing techniques.

IV. NUMERICAL EXAMPLE FOR SELECTION OF SOFTWARE QUALITY MODEL

The essential part of software product is its quality, different quality models are preferred for quantifying the software product quality. The quality model evaluates the quality of model based on certain parameters. These software quality models define parameters that are related to quality of system or software product. For selecting the best quality model to quantify the software product depends upon many criteria [30]. The identified criterion used are Reliability, Efficiency, Maintainability and alternatives are McCAll, Boehm and ISO 9126 software quality models as represented in figure 1.

The equation 11 has been used to calculate the normalized weights of the criteria comparison matrix and yielded the result w= (0.306, 0.302, 0.39) as shown in table III. After calculating the final weights for criterion same methodology is applied to find the weights for alternatives for each criterion. Table IV depicts the normalized weights for alternatives using FAHP. Table V depicts the pairwise comparison for an individual criteria with respect to different alternatives. Table VI presents the comparison of AHP and FAHP. A weight factor of 1.39 in case of AHP shows that Boehm's model has clear dominance. But when uncertainty has been considered by decision maker, results displayed that a weight factor of 0.38 in ISO 9126 has clear dominance over other software quality models.

V. CONCLUSION

AHP is an intuitive approach for solving decision making problems by breaking the it into alternative, assessment criterion and overall goal at the top of the hierarchy. Due to vagueness in human judgement and Saatys scale, sometimes it is hard to correctly making the relative ranking, which generates rank reversal problem. FAHP is modified version of traditional AHP as it uses fuzzy logic. FAHP is able to tolerate the human vagueness to greater extent. In this paper, different application areas of software engineering have been identified in which MCDM methods can be applied. FAHP approach is successful in solving MCDM problems such as objective of assessing and finding the right alternative for different applications like finding the web development platform, assessing the quality of websites and evaluating the success factors of e-commerce. Since project risks are multidimensional in nature, ranking and assessment of risks has been realized by using Fuzzy AHP. The application areas and numerical example discussed justify that Fuzzy AHP can be effectively implemented in software engineering application domains.

REFERENCES

[1] Pohekar, SD and Ramachandran, M., "Application of multi-criteria decision making to sustainable energy planningA review," *RENEW SUSTAIN ENERGY REV - Renewable & Sustainable Energy Reviews*, vol. 8, no. 4, pp. 365–381, 2004.

[2] W.-S. Lee and W.-S. Tu, "Combined {MCDM} techniques for exploring company value based on ModiglianiMiller theorem," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8037 – 8044, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410015010

[3] Y.-M. Wang and Y. Luo, "On rank reversal in decision analysis," *Mathematical and Computer Modelling*, vol. 49, no. 5-6, pp. 1221–1229, Mar. 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0895717708002860

[4] E. Kornyshova, R. Deneckre, and C. Salinesi, "Improving Software Development Processes with Multicriteria Methods," *arXiv preprint arXiv:0911.1496*, 2009. [Online]. Available: http://arxiv.org/abs/0911.1496

TABLE IV: Weights generated from given comparison matrix

	McCall's Model	ISO 9126 Model	Boehm's Model	Normalized weights
Evaluation of alternatives with respect to Reliability (C1)				
McCall's Model	[1, 1, 1]	[1/7, 1/5, 1/3]	[3, 5, 7]	0.43
ISO 9126 Model	[3, 5, 7]	[1, 1, 1]	[1, 3, 5]	0.55
Boehm's Model	[1/7, 1/5, 1/3]	[1/5, 1/3, 1]	[1, 1, 1]	0.013
Evaluation of alternatives with respect to Efficiency (C2)				
McCall's Model	[1, 1, 1]	[1, 1, 1]	[1, 1, 1]	0.23
ISO 9126 Model	[1, 1, 1]	[1, 1, 1]	[1/5, 1/3, I]	0.19
Boehm's Model	[1, 1, 1]	[1, 3, 5]	[1, 1, 1]	0.56
Evaluation of alternatives with respect to Maintainability (C3)				
McCall's Model	[1, 1, 1]	[3, 5, 7]	[1/7, 1/5, 1/3]	0.3
ISO 9126 Model	[1/7, 1/5, 1/3]	[1, 1, 1]	[5, 7, 9]	0.39
Boehm's Model	[3, 5, 7]	[1/9, 1/5, 1/7]	[1, 1, 1]	0.3

TABLE V: Results obtained by fuzzy AHP method

	Criterion Weights	McCalls Model	ISO 9126 Model	Boehms Model
Reliability	0.27	0.43	0.55	0.013
Efficiency	0.35	0.23	0.19	0.56
Maintainability	0.36	0.3	0.39	0.3
Global Weights		0.32	0.38	0.29

TABLE VI: Rankings computed using (AHP and FAHP)

Software Quality Models	Normalized weights (AHP)	Normalized Weights(FAHP)
McCalls Model	0.26	0.32
ISO 9126 Model	0.067	0.38
Boehm's Model	1.39	0.29

[5] G. Ruhe, "Software Engineering Decision Support A New Paradigm for Learning Software Organizations," in *Advances in Learning Software Organizations*, ser. Lecture Notes in Computer Science, S. Henninger and F. Maurer, Eds. Springer Berlin Heidelberg, 2003, vol. 2640, pp. 104–113. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-40052-3_10

[6] T. Saaty, "Decision making the Analytic Hierarchy and Network Processes (AHP/ANP)," *Journal of Systems Science and Systems Engineering*, vol. 13, no. 1, pp. 1–35, Mar. 2004. [Online]. Available: http://dx.doi.org/10.1007/s11518-006-0151-5

[7] R. W. Saaty, "The analytic hierarchy processwhat it is and how it is used," *Mathematical Modelling*, vol. 9, no. 3, pp. 161–176, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0270025587904738

[8] W. Sheng, L. Zhang, W. Tang, J. Wang, and H. Fang, "Optimal multi-distributed generators planning under uncertainty using ahp and ga," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no. 4, pp. 2582–2591, 2014.

[9] T. L. Saaty, "Decision making with the analytic hierarchy process," *International Journal of Services Sciences*, vol. 1, no. 1, pp. 83–98, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1504/IJSSci.2008.01759

[10] S. K. Sehra, Y. S. Brar, and N. Kaur, "Multi Criteria Decision Making Approach for Selecting Effort Estimation Model," *International Journal of Computer Applications*, vol. 39, no. 1, pp. 10–17, 2012.

[11] L. Mikhailov and P. Tsvetinov, "Evaluation of services using a fuzzy analytic hierarchy process," *Applied Soft Computing*, vol. 5, no. 1, pp. 23–33, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494604000535

[12] L. Liu, H. Chen, and R. Zhang, "Comprehensive evaluation of examination quality based on fuzzy ahp," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 9, pp. 5384–5394, 2013.

[13] C.-N. Liao, "Fuzzy analytical hierarchy process and multi-segment goal programming applied to new product segmented under price strategy," *Computers & Industrial Engineering*, vol. 61, no. 3, pp. 831–841, Oct. 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0360835211001410

[14] Y.-C. Tang, M. J. Beynon, and others, "Application and development of a fuzzy analytic hierarchy process within a capital investment study," *Journal of Economics and Management*, vol. 1, no. 2, pp. 207–230, 2005. [Online]. Available: http://www.aiecon.org/conference/efmaci2004/pdf/yuchen_tang_paper.pdf

[15] Y.-M. Wang, Y. Luo, and Z. Hua, "On the extent analysis method for fuzzy AHP and its applications," *European Journal of Operational Research*, vol. 186, no. 2, pp. 735–747, Apr. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0377221707002251

[16] A. Sarfaraz, P. Mukerjee, and K. Jenab, "Using fuzzy analytical hierarchy process (AHP) to evaluate web development platform," *Management Science Letters*, vol. 2, no. 1, pp. 253–262, Jan. 2012. [Online]. Available: http://www.growingscience.com/msl/Vol2/msl_2011_67.pdf

[17] K. Vatansever and Y. Akgul, "Applying Fuzzy Analytic Hierarchy Process for Evaluating Service Quality of Private Shopping Website Quality: A Case Study in Turkey," *Journal of Business Economics and Finance*, vol. 3, no. 3, pp. 283–301, 2014. [Online]. Available: http://dergipark.ulakbim.gov.tr/jbef/article/download/5000075891/5000070192

[18] F. Kong and H. Liu, "Applying fuzzy analytic hierarchy process to evaluate success factors of e-commerce," *International Journal of Information and Systems Sciences*, vol. 1, no. 3-4, pp. 406–412, 2005. [Online]. Available: http://www.math.ualberta.ca/ijiss/SS-Volume-1-2005/No-3-05/SS-05-03-22.pdf

[19] F. Tysz and C. Kahraman, "Project risk evaluation using a fuzzy analytic hierarchy process: An application to information technology projects," *International Journal of Intelligent Systems*, vol. 21, no. 6, pp. 559–584, Jun. 2006. [Online]. Available: http://doi.wiley.com/10.1002/int.20148

[20] M. Chang Lee, "Information Security Risk Analysis Methods and Research Trends: AHP and Fuzzy Comprehensive Method," *International Journal of Computer Science and Information Technology*, vol. 6, no. 1, pp. 29–45, Feb. 2014. [Online]. Available: http://www.airccse.org/journal/jcsit/6114ijcsit03.pdf

[21] b. Wei, F. Dai, and j. Liu, "C2c E-commerce Risk Assessment Based on AHP and Fuzzy Comprehensive Evaluation," *International Journal of Engineering and Manufacturing*, vol. 1, no. 1, pp. 34–39, Feb. 2011. [Online]. Available: http://www.mecs-press.org/ijem/ijem-v1-n1/v1n1-6.html

[22] M. Askari, H. R. Shokrizadeh, and N. Ghane, "A Fuzzy AHP Model in Risk Ranking," *European Journal of Business and*

Management, vol. 6, no. 14, pp. 194–202, 2014. [Online]. Available: http://iiste.org/Journals/index.php/EJBM/article/view/13347

[23] J. S. Challa, A. Paul, Y. Dada, V. Nerella, P. R. Srivastava, and A. P. Singh, "Integrated Software Quality Evaluation: A Fuzzy Multi-Criteria Approach," *Journal of Information Processing Systems*, vol. 7, no. 3, pp. 473–518, Sep. 2011. [Online]. Available: http://koreascience.or.kr/journal/view.jsp?kj=E1JBB0&py=2011&vnc=v7n3&sp=473

[24] T. Bakshi, B. Sarkar, and S. K. Sanyal, "A Novel Integrated AHP-QFD Model for Software Project Selection under Fuzziness," *International Journal of Computer Applications (09758887), c*, 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.7185&rep=rep1&type=pdf

[25] Y. Y. Jusoh, K. Chamili, N. C. Pa, and J. H. Yahaya, "Open source software selection using an analytical hierarchy process (AHP)," *American Journal of Software Engineering and Applications*, vol. 3, no. 6, pp. 83–89, 2014. [Online]. Available: http://article.sciencepublishinggroup.com/pdf/10.11648.j.ajsea.20140306.13.pdf

[26] M. Khan, A. Parveen, and M. Sadiq, "A method for the selection of software development life cycle models using analytic hierarchy process," in *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*, Feb. 2014, pp. 534–540.

[27] S. Mahmoodzadeh, J. Shahrabi, M. Pariazar, and M. S. Zaeri, "Project selection by using fuzzy AHP and TOPSIS technique," *International Journal of Human and social sciences*, vol. 1, no. 3, pp. 135–140, 2007. [Online]. Available: http://www.waset.org/publications/128

[28] P. R. Srivastava and M. P. Ray, "Multi-attribute Comparison of Automated Functional and Regression Testing Tools using Fuzzy AHP." in *IICAI*, 2009, pp. 1030–1043.

[29] V. Belton and T. Stewart, *Multiple criteria decision analysis: an integrated approach*. Springer Science & Business Media, 2002. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=mxNsRnNkL1AC&oi=fnd&pg=PR11&dq=%22defines+it+as+%E2%80%9Can+activity+in+which+a%22+%22in+terms+of+efforts+needed,+money+as+well+as%22+%22Multicriterion+Decision+Making%22+%22Stewart+%5B5%5D,+analyzed+the+MCDM+process%22+&ots=DKGvIODvBF&sig=CFV70qyKGLWESmABEcQ_kPuDM-Y

[30] R. Kohli and S. K. Sehra, "Fuzzy Multi Criteria Approach for Selecting Software Quality Model," *International Journal of Computer Applications*, vol. 98, no. 11, pp. 11–15, 2014.

Arabic Text Question Answering from an Answer Retrieval Point of View: a survey

Bodor A. B. Sati[1], Mohammed A. S. Ali[2], Sherif M. Abdou[2]
[1] Najran University, Saudi Arabia
[2] Information Technology Department,
Faculty of Computers and Information,Cairo University
Cairo, Egypt

***Abstract*—Arabic Question Answering (QA) is gaining more importance due to the importance of the language and the dramatic increase in online Arabic content. The goal of this article is to review the state-of-the-art of Arabic QA methods, to classify them into different categories from an answer retrieval viewpoint and to present their applications, issues and new trends. The main components of question answering systems are also presented. Finally, this survey provides a comparative study of systems of each type of QA based on several criteria.**

***Keywords*—Question answering; Information retrieval; Answer retrieval; Arabic NLP**

I. Introduction

Nowadays, the Internet has become the main source of information and search is a daily activity for many people throughout the world. The need to retrieve related information for request became increasingly important. It became necessary to find useful and accurate information from large amounts of information. The current techniques in Information Retrieval (IR) allow a user to retrieve only the relevant documents which match a given query. Then, the users look for the information they need within the relevant documents. Therefore, a new need emerged: the possibility of obtaining a brief and accurate answer has motivated the interest in question answering (QA) systems.

QA is a special and sophisticated form of information retrieval. It has been created to automatically satisfy a specific need of information which requested by users who are looking for answer using a natural language question [1]. QA systems are composed of three main subtasks: question analysis, passage retrieval, and answer extraction. Most QA systems follow these tasks but it may differ in the way of implementation of each sub-task.

Arabic language is the 6th important language in the world with more than 300 million speakers [2]. Moreover, it is the language of the Quran, so it has a great attention from Muslims all over the world which means about 1 billion people around the world may be interested in it. Arabic QA gains more and more attention due to the dramatic increase of the Arabic content on the Internet.

QA systems are classified into two main categories. Open-domain QA which deals with questions about nearly everything. The second category is closed-domain QA which deals with questions in a specific domain (Fatwa, weather forecasting, medical applications etc.) [3].

QA from answer retrieving viewpoint can be divided into two subcategories: the first one is QA by extracting an answer from unstructured documents such as web pages or via generating it; where the answer is drawn from multiple sentences or multiple documents [4]. The second type is QA based on Frequently Asked Questions (FAQ). In this type, a user query is matched with already existed questions which are associated with their answers in a database to retrieve the closest possible answer to a given question.

Actually, few surveys have been published that investigate Arabic QA such as [1], [5]. Unlike the previous surveys, in this survey a new taxonomy for QA is presented. In such a taxonomy, both types of QA (QA where the answer is generated and QA based on FAQ) have been investigated. Moreover, our survey has shown a comparative study of the systems of each type of QA based on several criteria as shown in the upcoming section.

A. Challenges of Arabic QA

Design of QA system for Arabic language has become a greater challenge because of the nature of the language. There are some characteristics in Arabic language which slow down the progress in Arabic Natural Language Processing (NLP) [2], [6], [7]:

- Arabic has a very complex morphology (inflectional and derivational characteristic).
- The absence of diacritics in the written text creates many ambiguity problems to question analysis and answer extraction.
- The absence of capitalization which makes problems in Named Entity Recognition (NER).

The major challenges that are faced by Arabic QA are:

- The lack of accessibility to Arabic linguistic resources such as WordNet.
- The lack of technologies like basic NLP tools (tokenizers, morphological analyzers, information extraction tools)

The rest of this paper is organized as follows: Section 2 introduces the-state-of-the-art of both types of Arabic QA. Section 3 presents the applications of Arabic AQ. In Section 4 some issues and expected future trends of Arabic QA are exposed. And finally, The conclusion are introduced in Section 5.

II. THE-STATE-OF-THE-ART

As it has been aforementioned, QA form answer retrieving perspective can be divided into two subcategories: QA based on FAQ and QA where answer is generated from raw text. The-state-of-the-art of the two types are investigated in the following sections:

A. QA where answer is generated from raw text

In this type, the answer is generated and formulated from unstructured documents such as web pages. The answer is drawn from single or multiple documents. In this section the general architecture of QA is discussed. Moreover, the section discusses the approaches used to implement each task in this architecture.

1) The general architectures of a typical question answering systems: A typical QA system consists of three distinct modules: question analysis, passage retrieval, and answer extraction. Figure 1 shows the general architecture of typical QA system.

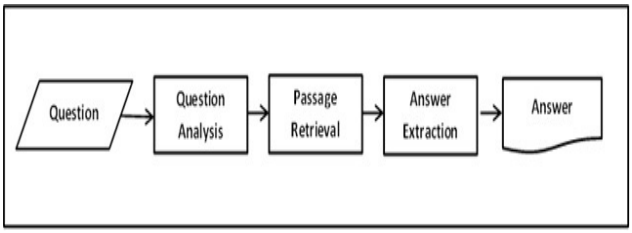


Fig. 1: Generic architecture of the Typical question answering system

Question Analysis: Question analysis is the first module in QA which identifies the aim of the question, classifies the question types, derives the expected answer types and performs the query expansion. Moreover, the question analysis module determines the named entities appeared in the question [8]. Question analysis could be considered as the most important step in a QA system. Since a right classification of the question will allow to limit the candidate answers to be considered.

In AQuASys [9] , question analysis module identifies the type of the expected answer by defining a number of rules for question types identification and their expected answers recognition. Additionally, it classifies the question words into three groups: interrogative noun, questions verb and questions keywords. Then, the additional keywords are generated and added to the questions keywords. Finally, the system performs stemming on question's and documents words.

Kanaan et. al. In [10] use a set of Natural Language Processing (NLP) tools in this module which tokenize and tag the text, identify some features of the tokens and identify proper names in the question. In DefArabicQA [11], the module of question analysis is performed by identifying the topic of question and determining the expected answer type. The question topic is identified by using two lexical question patterns: (Who+be+¡topic¿) and (What+be+¡topic¿). The expected answer type is deduced from the interrogative pronoun of the question.

Question analysis In QARAB [6] is achieved by performing tokenization and stop-words removing. Then, The remaining words are tagged for part-of-speech. Also, this module includes the following task: identifying proper names, identifying the type of the expected answer, applying query expansion to achieve better results, classifying the question and tagging the question keywords for part of speech.

In ArabicQA [12], question analysis module determines the type of the given question, the question keywords and the named entities in the question. for example: if the question is: متى استقلت السودان ؟ (When Sudan became independent?), the question type is: Time, the question keyword is: استقلت (Become independent: Verb), the name entity is: السودان (Sudan) . In this work the authors have used the NER system which has been built by the same authors.

In JAWEB [13], The question analysis module identifies the type of the question. For example (if the question begin with "أين", "where"; the question type will be location). In addition, it has five sub-modules: tokenizer, answer-type detector, question keyword extractor, extra keywords generator and question words stemmer.

The question analysis module in QASAL [14] starts by using a set of linguistic resources in NooJ [1] that is applied to the given question to analyze and annotate it. Then, NooJs graph editor was used to carry out some local grammars. These grammars translate each question into one or more regular expressions and that helps to represent the pattern of the answer corresponding to this question. The question analysis module allows the generation of all the candidate answer pattern regular expressions.

In Yes/No Arabic QA system [15], The question analysis module is performed by applying the following tasks: removing a question mark, removing an interrogative particle, tokenizing, normalize the (Alef) letter, removing the stopwords, removing the negation particles. In addition, apply tagging to determine the type of a word and obtain its root. Moreover, this module apply parsing and query expansion. The query expansion retrieves a list of synonyms and antonyms. Finally, the system represents a question using logical representation . The authors create 12 Logical representations for the Nominal and verbal sentences for both affirmative and negated questions. Figure 2 shows an example of logical representation of verbal sentence.

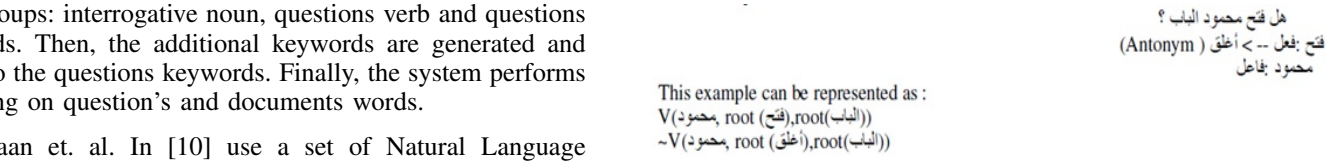


Fig. 2: An example of logical representation by [15]

Passage Retrieval: Passage retrieval is considered as the core of QA system. In this module the passages of documents that are relevant to the answer is retrieved. Usually, a

[1] NooJ is an NLP environment. For more information http://www.nooj4nlp.net

distance between question and documents is calculated. based on this distance and weighting schemes such as tf-idf [16], document retrieval systems supply a set of ranked documents. Most QA systems are based on IR methods that have been adapted to work on passages instead of the whole document [8].

Kanaan et. al. In [10] performs this module by applying information retrieving system which uses Saltons vector space model to measure the similarity between the query and the document. They use a database to store the information that is related to a word, query weight and the similarity of the query. then, a list of ranked documents that may contain the answer is generated. While in AQuASys [9] ,the passage retrieval module filters the sentences based on the number of question keywords they contain.

In DefArabicQA [11], this module returns the top-n snippets which are retrieved by the Web search engine. Then, this system performs definition extraction task. In this task, it identifies and extracts candidate definitions from the snippets by using lexical patterns. Finally, the candidate definitions will be Filtered by using heuristic rules.

The passage retrieval system in QARAB [6] is based on Saltons vector space model. First, an inverted file system is constructed from the collection of text. Then use relational database management system (RDBMS) to hold and retrieve the passages.

In ArabicQA [12], the passage retrieval module (JIRS) [2] retrieves the passages which are most probable to contain the answer. JIRS relies on using an n-gram model. This module apply three steps: First, it searches the relevant passages and assigns a weight to each of them. Second, It extracts the necessary n-grams from each passage. Finally, it compares between the question and the passage n-grams using the density distance model.

The passage retrieval module In JAWEB [13], retrieves the candidates answers from the corpus by retrieving set of passages. These passages are characterized by that they contain sentences which contain a pattern that matches words from the question keywords and extra keywords.

In QASAL [14], the passage retrieval module starts to select one or more regular expressions from those which are generated in the previous module. Then, applies those expressions to the answer text in order to identify the potential answers. And finally, the detected answers are displayed in a concordance table to be used later.

In Yes/No Arabic QA system [15], The passage retrieval module is performed by applying two levels of IR techniques: the first is on documents and the second on paragraphs. The paragraphs technique splits the documents into paragraphs and retrieve the top 5 paragraphs regardless of from which document they are, according to some indexing scheme. The document technique, in turn, retrieves the top 5 documents after they are ranked, then use the first indexing scheme to retrieve the top 5 paragraphs.

[2] JIRS is a QA-oriented passage retrieval tool. For further information http://jirs.dsic.upv.es

Answer Extraction: Answer extraction is the final module in most QA systems. This module distinguishes between QA systems and the usual sense of text retrieval systems [3].

In answer extraction module, the final answer to the question is extracted from the passages or documents that retrieved in the previous module. The search operation in this module is performed using information retrieved from the first module (question analysis). This information includes the focus and the target of the question. The answer extraction module extracts a list of ranked relevant answers, and finally returns the most probable one(s) [8].

In AQuASys [9] ,the answer extraction module uses a sophisticated scoring formulas are based on a large number of parameters. The aim of these formulas is to calculate the similarity between the given question and the candidate sentences. Finally, the answer that gives the highest score will be selected.

Kanaan et. al. In [10] choose the most appropriate document according to the similarity values, which are calculated by the IR system. Then,the system generates the answer. While Trigui et. al In DefArabicQA [11], strat to rank the definitions. Such ranking is achieved by applying a statistical approach. They use a global score to rank the candidate definitions. The global score is a combination of three criteria: pattern weight ,snippet position, and word frequency criterion. Finally, The first top-5 candidate definitions will ranked according to their global scores.

In QARAB [6], the input to the answer extraction module is the question words and the top ranked relevant documents. The authors assume that the best answer usually includes most of the words which appear in the query. Therefore, the system retrieves the answer that contains most words appear in the question in addition to the proper nouns that should appear in the final answer.

To extract a list of candidate answers from the relevant passage in ArabicQA [12], This module takes into consideration the type of expected answer and performs the following steps. First, It tags all named entities within the relevant passage. Second, It performs pre-selection of the candidate answers. Finally, It decides the final list of candidate answers by means of a set of patterns. Additional module is applied in this system, answer validation module , which estimates the probability of correctness for each of the candidate answers and ranks them.

In JAWEB [13], the answer extraction module consists of three sub-modules: answer keywords stemmer, answer similarity checker and answers ranker. The answer keywords stemmer returns the roots of the keywords in the retrieved answer. The answer similarity checker measured the similarity by counting the number of matching keywords between the question and retrieved answer. The answers ranker sorts answers according to their relevance. Then, return the top relevant answer.

To extract the answer in QASAL [14], in the previous module a concordance table of the potential answers is generated. So, in this module use such a concordance table to automatically extract the answer to the given question .

In Yes/No Arabic QA system [15], the answer extraction module applies the following steps: Split the paragraphs into

their sentences. The module focuses in topic when it deals with normal sentences and it focuses in subject when dealing with verbal sentence. Then, the module looks for the remaining terms that derived from the question in logical representation , assigns those indexes according to their position in the sentence. So each sentence will have its own rank. After that, look for negation particles in the selected answer. Finally, use the selected answer and the logical representation of the question to generate yes or no as follows : return "Yes" if the question and the answer are affirmative . The question and the answer are negated. Or return "No" if The question is affirmative and the answer is negated. The question is negated and the answer is affirmative.

A comparative study of these eight systems is presented in table I

B. QA based on FAQ

The previous section has introduced an overview of those approaches that based on retrieving the documents that contain the answer and then extract and craft this answer. However, in this section we will shed some light on those approaches that depend on FAQ. Indeed, what we mean by this is that there is a bank of questions and associated answers and the systems are going to receive a user query to be answered and looks for the most appropriate answer(s) and retrieves them.

According to the prediction of Boris Katz et al. in [17], the next generation of search engines would be based on question answering in which users would receive explicit answers extracted from documents to their natural language queries instead of the current ones which retrieve only the relevant pages.

A tremendous amount of questions and associated answers are available online and one can get almost any answer to their questions based on previous answered ones. Therefore, this type of question answering systems are one of the important areas of research in IR. Several research studies have been conducted and reported in the literature to facilitate this type of IR. Unfortunately, a majority of these studies were not oriented to Arabic Language.

However, in the past few years several researchers have tried to find clues to this problem in Arabic. Actually, we can differentiate between two type of researches that have been conducted to do such a task. The first and most common one is called answer selection and the second is answer extraction based on FAQ.

The former task is interested in identifying pertinent answers from a pool of user-generated comments related to a question. This means that all answers in the pool are related to the question with different degrees. Moreover, the pool of answers which are going to be nominees are very small. Though this type receives significant interest in several research [18] [19] in CLIF 2012 and [20] in CLIF 2013 [21]–[23], the most real-life QA-based-on-FAQ applications is not working like that. Instead, it looks for an answer from a flat huge collection, thousands, tens of thousands or maybe millions, of question-answer pairs. Several studies have been conducted to resolve such a problem using different technologies will be demonstrated in the following:

- *Using Textual Case-based Reasoning in Intelligent Fatawa QA System* [24]
  In this work Elhelwany et.al. have proposed an Arabic Fatwa Intelligent system based on textual case based rezoning which was firstly used in [25]. In their system, they started by extracting a representative term for each cluster which were later called clusters attractors. Then, the cases clustered around these attractors. Eventually, they used Jensen-Shannon divergence to assign a newly posed question to its appropriate cluster and, subsequently, to find the closest possible question among questions in such a cluster.

- *Enhancements to knowledge discovery framework of SOPHIA textual case-based reasoning* [26]
  In [26] The same authors of the previous system have enhanced their previous study by adding one tier in the middle. This tier was created by manually clustering the dataset into several groups so the SOPHIA will be applied to each group separately and that what make an enhancement as they reported.

- *A Case Based Tool As Intelligent Assistance To Mufti* [27]
  Nabila Nouaouria et.al in [27] have designed El Bayane. In their system, the authors started by representing cases manually in the following structure ¡product features, exceptions, product type¿ and in hierarchal order (most general to most specific). Their system is closed on the field of drinking and smoking in the Islamic legislation. To answer a new question, the system requests selection for specific predefined parameters as a representation to the question:¡question type, action, product name, product type, features, exceptions¿. Finally, they look for the most similar cases and return associated answer. This system is very domain-dependent because it is limited to a subdomain of drinking legislation in Islamic fatwa, limited on factoid questions, require exhausting manual work and usually designed for the situation where all cases have similarly structured content.

- *Intelligent Tool for Mufti Assistance* [28]
  In [28] Amari et. al start to organize already answered cases in a cases memory which lately will be known as case-base. They represent a case using two dimensions: case discretion ¡action type, product name, product type, features exceptions¿ and case solutions. As it was reported, they use a new way to represent cases based on constructing a problem neighborhood to ease the retrieval of the cases later. In the test phase they introduce a system of five modules to retrieve the similar cases to a query: neighborhood computation, associative access, adaptation, validation and storage. Like the previous system, this system constraint a user to formulate predefined questions and require exhausted manual work to build the case-base.
  Unfortunately, none of the previously mentioned works have reported the evaluation approach they used or reported results.

- *Answer Extraction System Based on Latent Dirichlet Allocation* [29] In [29] Ali et. al have proposed a new system that based on Latent Dirichlet Allocation

TABLE I: Comparative study of eight QA systems when the answer is generated

System	Aim	Domain	Dataset	Results	Shortcomings
A Question Answering System For Arabic (AQuASys) [9]	answer unformatted fact-based questions written in an Arabic natural language.	Close domain	ANERcorp: 150,000 tagged tokens) as well as few gazetteers (ANERgazet) available online	recall rate of 97.5% and 66.25% as a precision rate	The system focused only on factoid questions
A New Question Answering System for the Arabic Language [10]	provide short answers for Arabic natural language questions	Close domain	collection of Arabic text documents	not presented	Does not include the other types of question (How and Why)
Arabic Definition Question Answering System (DefArabicQA) [11]	Arabic definitional Question Answering system based on a pattern approach to identify accurate definitions about organization using Web resources	Close domain	Google search engine and Wikipedia Arabic version	90% of the questions used have complete definitions in the top-five answers and 64% of them have complete definitions in the top answer. MRR was (0.81).	limited to a specific type (definition)
A Question Answering System to Support the Arabic Language (QARAB) [6]	provide short answers for Arabic natural language questions	Close domain	collection of Arabic newspaper text extracted from Al-Raya, a newspaper published in Qatar	not presented	Limited dataset
Arabic Question Answering System (ArabicQA) [12]	provide short answers for Arabic natural language questions	Close domin	ANERcorp which contain two corpora for training and testing build by the developer	The precision is 83.3%	The Answer Extraction focused only on factoid questions
A Web based Question Answering system for Arabic Language (JAWEB) [13]	provide short answers for Arabic natural language questions	Close domain	an extended version of the Arabic corpus developed by [9]	The system provided 15-20% higher recall	The system focused only on factoid questions
An Arabic Question-Answering system for factoid questions [14]	provide short answers for Arabic natural language questions	Close domain	a collection of Arabic text documents containing factoid questions as well as their different answers	not presented	the system focused only on factoid questions
Development of Yes/No Arabic Question Answering System [15]	design a formal model for a semantic based yes/no Arabic question answering system based on paragraph retrieval	Open domain	20 documents which used to test the system and a collection of 100 different yes/no question	The results of using documents technique:85% when 20 documents are used. The result of using paragraphs technique: 88% when 20 documents are used	The system focused only on yes/no questions. and the corpus size is small (20 documents)

(LDA) [30] and word tow vector space word representation [31]. In their work the authors started to cluster the cases (documents contain questions and associated answer) into similar thematic groups. Then, to reply to a new query they started to assign this query into appropriate cluster and subsequently retrieve the most suitable answer to this question. As they reported, they achieved accuracy of 83.6 %.

A comparative study of these five systems have been presented in table II

III. Applications of Arabic question answering

Question answering has many applications. Arabic language is the 6th important language in the world with more than 300 million speakers [2]. due to the dramatically increase of the Arabic content on the internet, the increase of Arabian internet users, and increase demand for information that traditional information retrieval methods can not satisfy, an inevitable need for an effective information retrieval system is required.

Distance education also is gaining a lot of attention and has become a popular research topic. No matter where or when the teacher or student is, the communication between students and teachers is a very important. However, face to face communication is not possible. Question answering system based on FAQ is the solution in this case. Where the student asks a question, the answer is retrieved, if the question-answer pair is already in the database or it would be answered by the teacher later and saved in the corpus as well. Arabic Question-Answer pair also available in many and many Arabic websites such as Islamic Fatwa websites, Arabic medical websites, distance educational systems etc.

IV. The future and issues of Arabic question answering

According to the prediction of Boris Katz et al. in [17], the next generation of search engines would be based on question answering in which users would receive explicit answers extracted from documents to their natural language queries instead of the current ones which retrieve only the relevant pages. Moreover, we will have to deal with that queries that will posed by voice. That means, we will need more sophisticated speech recognition algorithms and techniques to deal with different Arabic dialect. Accordingly, more challenges will be posed to deal with detected users queries with different accents. We recommend [32] and [33] for further information in a speech based question answering.

TABLE II: Comparative study of five QA- based-on-FAQ systems

System	Aim	Domain	Dataset	Results	Shortcomings
Using Textual Case-based Reasoning in Intelligent Fatawa QA System [24]	Answering new questions based on pre-answered ones	Open domain	5k QA pairs (Fatwa)	No reported results	• No result was reported and No evaluation was performed. • Negation and numbers were not taken into account
"Enhancements to knowledge discovery framework of SOPHIA textual case-based reasoning [26]"	Answering new questions based on pre-answered ones	Open domain	5k QA pairs (Fatwa)	The only result was reported is clustering results 80% accuracy	• Requires exhaustive manual work in a new added tier. • The results of QA were not presented. • Negation and numbers were not taken into account
El Bayane [27]	Answering new questions based on pre-answered ones	limited to Islamic legislation on drinking	No reported dataset	No reported results	• Predefined template. • The results were not shown • The dataset was not described. • limited to specific domain Requires manual work
Intelligent Tool for Mufti Assistance [28]	Answering new questions based on pre-answered ones	Limited to Islamic legislation on drinking and as reported it can be expanded to new domains.	No reported dataset	No reported results	• Predefined template. • The results were not shown • The dataset was not described. • Limited to specific domain Requires manual work
Answer Extraction System Based on Latent Dirichlet Allocation [29]	Answering new questions based on pre-answered ones	Open domain	11k pairs of questions and answers (Fatwa)	Accuracy 83.5 /%	• Negation was not handled. • Numbers were not taken into account.

Additionally, a tremendous amount of information is available in the form of video, images, maps and sounds. We need new tools in the next generation of question answering to deal with this multimedia. These tools have to organize, search, understand and extract the answers out of this diverse representation of information.

In spite of this promising future of QA, it still has several limitations and issues that could be improved in the future. Some of these issues is due to the challenges of Arabic language that aforementioned in section 4. However, the others are associated with the system itself. Information retrieval models, for instance, still fails to return an appropriate set of answers at an acceptable level of precision and recall. The current systems cope with this problem by query performing expansion for those queries which got too many candidates to answer or by removing and eliminating some words from those queries which got too few candidates to answer [34]. Another problem with question answering systems is that they still has limitations to capture semantic content for queries and answers.

V. Conclusion

In this article, an extensive survey of Arabic QA is presented and a new taxonomy for QA is introduced. This taxonomy is based on an answer retrieval perspective. QA is categorized into two groups: QA where the answer is generated and QA based on frequently asked questions. Note that all these classes involve retrieving the answer to a newly posed question using natural language. From the analysis of the relevant literature, we have organized and compared the studies of the first group based on the three tasks of the general structure of a typical QA (question analysis, passage retrieval and answer extraction). In this survey, the studies have been investigated and compared based on several criteria such as, aim, domain, datasets and performance results of the systems.

Despite the efforts that have been made in the field of Arabic QA, there are still some issues and limitations that have not yet been addressed. Some of these important issues have been presented above. Finally, applications and the expected future directions of Arabic QA have been discussed.

References

[1] M. Shaheen and A. M. Ezzeldin, "Arabic question answering: Systems, resources, tools, and future trends," *Arabian Journal for Science and Engineering*, vol. 39, no. 6, pp. 4541–4564, 2014.

[2] A. Ezzeldin and M. Shaheen, "A survey of arabic question answering: Challenges, tasks, approaches, tools, and future trends," in *Proceedings of The 13th International Arab Conference on Information Technology (ACIT 2012)*, 2012, pp. 1–8.

[3] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, vol. 2, no. 3, 2012.

[4] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, vol. 7, no. 04, pp. 275–300, 2001.

[5] W. Bakari, P. Bellot, and M. Neji, "Literature review of arabic question-answering: Modeling, generation, experimentation and performance analysis," in *Flexible Query Answering Systems 2015*. Springer, 2016, pp. 321–334.

[6] B. Hammo, H. Abu-Salem, and S. Lytinen, "Qarab: A question answering system to support the arabic language," in *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*. Association for Computational Linguistics, 2002, pp. 1–11.

[7] W. Brini, M. Ellouze, O. Trigui, S. Mesfar, H. Belguith, and P. Rosso, "Factoid and definitional arabic question answering system," *Post-Proc. NOOJ-2009, Tozeur, Tunisia, June*, pp. 8–10, 2009.

[8] Y. Benajiba, P. Rosso, L. Abouenour, O. Trigui, K. Bouzoubaa, and L. Belguith, "Question answering," in *Natural Language Processing of Semitic Languages*. Springer, 2014, pp. 335–370.

[9] S. BEKHTI and M. AL-HARBI, "Aquasys: A question-answering system for arabic," in *WSEAS International Conference. Proceedings. Recent Advances in Computer Engineering Series*, no. 12. WSEAS, 2013.

[10] G. Kanaan, A. Hammouri, R. Al-Shalabi, and M. Swalha, "A new question answering system for the arabic language," *American Journal of Applied Sciences*, vol. 6, no. 4, p. 797, 2009.

[11] O. Trigui, H. Belguith, and P. Rosso, "Defarabicqa: Arabic definition question answering system," in *Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta*, 2010, pp. 40–45.

[12] Y. Benajiba and P. Rosso, "Arabic question answering," *Diploma of advanced studies. Technical University of Valencia, Spain*, 2007.

[13] H. Kurdi, S. Alkhaider, and N. Alfaifi, "Development and evaluation of a web based question answering system for arabic language," *Computer Science & Information Technology (CS & IT)*, vol. 4, no. 2, pp. 187–202, 2014.

[14] W. Brini, M. Ellouze, S. Mesfar, and L. H. Belguith, "An arabic question-answering system for factoid questions," in *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*. IEEE, 2009, pp. 1–7.

[15] W. N. Bdour and N. K. Gharaibeh, "Development of yes/no arabic question answering system," *arXiv preprint arXiv:1302.5675*, 2013.

[16] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 13, 2008.

[17] B. Katz, J. J. Lin, D. Loreto, W. Hildebrandt, M. W. Bilotti, S. Felshin, A. Fernandes, G. Marton, and F. Mora, "Integrating web-based and corpus-based techniques for question answering." in *TREC*, 2003, pp. 426–435.

[18] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Idraaq: New arabic question answering system based on query expansion and passage retrieval," 2012.

[19] O. Trigui, L. H. Belguith, P. Rosso, H. B. Amor, and B. Gafsaoui, "Arabic qa4mre at clef 2012: Arabic question answering for machine reading evaluation." in *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[20] A. M. Ezzeldin, M. H. Kholief, and Y. El-Sonbaty, "Alqasim: Arabic language question answer selection in machines," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2013, pp. 100–103.

[21] H. Abdelnasser, R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky, and M. Torki, "Al-bayan: an arabic question answering system for the holy quran," *ANLP 2014*, p. 57, 2014.

[22] R. Mohamed, M. Ragab, H. Abdelnasser, N. M. El-Makky, and M. Torki, "Al-bayan: A knowledge-based system for arabic answer selection," *SemEval-2015*, p. 226, 2015.

[23] Y. Belinkov, "Answer selection in arabic community question answering: A feature-rich approach," in *ANLP Workshop 2015*, 2015, p. 183.

[24] I. Elhalwany, A. Mohammed, K. Wassif, and H. Hefny, "Using textual case-based reasoning in intelligent fatawa qa system," *The International Arab Journal of Information Technology*, vol. 12, no. 5, 2015.

[25] D. Patterson, N. Rooney, M. Galushka, V. Dobrynin, and E. Smirnova, "Sophia-tcbr: A knowledge discovery framework for textual case-based reasoning," *Knowledge-Based Systems*, vol. 21, no. 5, pp. 404–414, 2008.

[26] I. Elhalwany, A. Mohammed, K. T. Wassif, and H. A. Hefny, "Enhancements to knowledge discovery framework of sophia textual case-based reasoning," *Egyptian Informatics Journal*, vol. 15, no. 3, pp. 211–220, 2014.

[27] N. Nouaouria, F. Atil, M. Laskri, D. Bouyaya, and A. H. Amari, "A cased based tool as intelligent assistance to mufti," *Arabian Journal for Science and Engineering*, vol. 31, no. 1, pp. 75–90, 2006.

[28] H. Amari, F. Atil, N. Bounour, and N. Nouaouria, "Intelligent tool for mufti assistance," *International Journal on Islamic Applications in Computer Science And Technology*, vol. 3, no. 2, 2015.

[29] M. A. Ali and S. M. Abdou, "Answer extraction system based on latent dirichlet allocation," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 462–465, 2015.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[31] M. A. Zahran, A. Magooda, A. Y. Mahgoub, H. Raafat, M. Rashwan, and A. Atyia, "Word representations in vector space and their applications for arabic," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 430–443.

[32] T. Mishra and S. Bangalore, "Qme!: A speech-based question-answering system on mobile devices," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 55–63.

[33] E. Schofield and Z. Zheng, "A speech interface for open-domain question-answering," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*. Association for Computational Linguistics, 2003, pp. 177–180.

[34] M. A. Pasca and S. M. Harabagiu, "High performance question/answering," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 366–374.

Comparative Analysis of ALU Implementation with RCA and Sklansky Adders In ASIC Design Flow

Abdul Rehman Buzdar, Liguo Sun, Abdullah Buzdar
Department of Electronic Engineering and Information Science
University of Science and Technology of China (USTC)
Hefei, People's Republic of China

***Abstract*—An Arithmetic Logic Unit (ALU) is the heart of every central processing unit (CPU) which performs basic operations like addition, subtraction, multiplication, division and bitwise logic operations on binary numbers. This paper deals with implementation of a basic ALU unit using two different types of adder circuits, a ripple carry adder and a sklansky type adder. The ALU is designed using application specific integrated circuit (ASIC) platform where VHDL hardware description language and standard cells are used. The target process technology is 130nm CMOS from the foundry ST Microelectronics. The Cadence EDA tools are used for the ASIC implementation. A comparative analysis is provided for the two ALU circuits designed in terms of area, power and timing requirements.**

***Keywords*—Arithmetic Logic Unit; Ripple Carry Adder; Sklansky Adder; ASIC Design, EDA Tools**

I. Introduction

An Arithmetic Logic Unit (ALU) is key the element of a processor which performs arithmetic and logical operations on binary numbers [1-7]. In this work we have designed an ALU for a 32-bit processor using VHDL hardware description language. The ALU designed can perform four major tasks addition, subtraction, logical and shift operations. Fig. 1 shows the block diagram of the ALU designed, as can be seen that it is made edge-triggered by having flip-flops on its inputs and outputs. Two types of ALUs were designed and the difference in both the ALUs was of adder type. In the first one a ripple carry adder [8-17] was used while in the second one a pre-fix tree of sklansky type adder [18-25] was used. The adder is a very important component in digital systems, so lot of research has been done in past on various types of adders to improve the speed and area requirements [26-41]. Apart from adder both the ALUs consisted of a Logical and Shifter blocks.

In case of ALU-RCA two different designs were used to compare the performance of both the designs in terms of area and power usage. The first design contains a demux for selecting which block of ALU would be used based on the opcode and the blocks which are not used are shutdown to save power. While the second ALU-RCA design does not contain this demux and all the three units of ALU perform their respective operations and only one result out of three units is sent to the output based on the opcode which is selected by the mux which is placed before the output. In both ALUs designs a mux is placed before the output for selecting the result from only one of the three blocks based on the opcode.

The rest of the paper is organized as follow: In the next section we describe the design and verification of ALU circuits, followed by ALU circuit basic Synthesis. Later we describe the process of Design Respin, Power analysis, Place and Route of ALU circuits. Finally, we summarize our conclusions.

II. ALU Design- Verification

Initial verification of both the ALUs i.e. ALU-RCA and ALU-SKL were performed based on the waveform approach using ModelSim software tool [47] as we made both the ALUs generic so we reduced the ALU size to 8-bit just to make initial verification simple. The waveform based approach of verification is only useful during the initial phases of small designs and always requires more comprehensive verification in the later design stages. There were some small bugs found in the code which were later corrected.

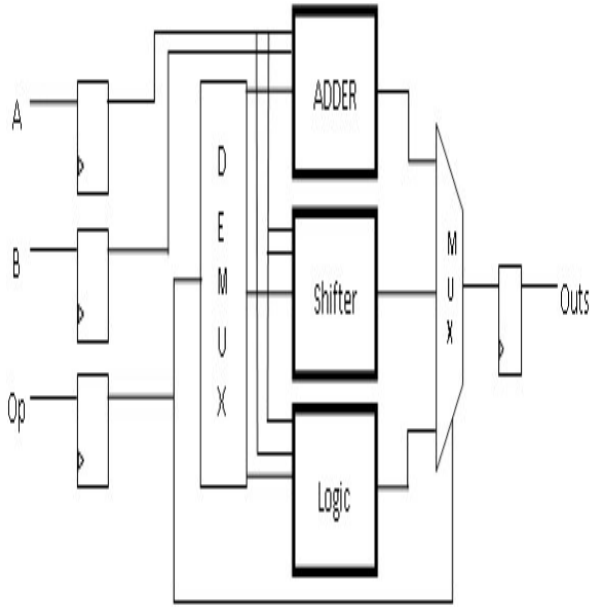


Figure 1: ALU Block Diagram

The next task was to define a VHDL TestBench for more comprehensive verification of ALU, than waveform approach. The TestBench written takes the required input values of A, B and opcode from stimuli file and compares the result to an expected output which is also taken from a stimuli file. If the result is incorrect against any of the testvectors then false is written in a dumped file else true is written. If the results are correct against all the test vectors then a message is displayed that simulation was successful else an error message

is displayed in the end of simulation. The first task in the verification was to functionally verify our 32-bit ALU meaning that whether our design conforms to specification and does this proposed design do what is intended. We can view the test vectors as a functional specification. The NCSIM logic simulator was used to logically verify our designs. The logic simulation is the use of some computer program to stimulate the operation of digital circuit before it is actually built. The designs were simulated against the 1000 randomized reference test vectors and found no errors. To check the correctness of our TestBench we changed some of the vectors in the testvectors stimuli files and again simulated our design this time it gave errors against those vectors which we had changed as expected. We cannot say that this verification is complete as we checked our design using only 1000 randomized testvectors and the possible combinations in our case is 268, but it is a good starting point of verification process and as they say verification is never complete.

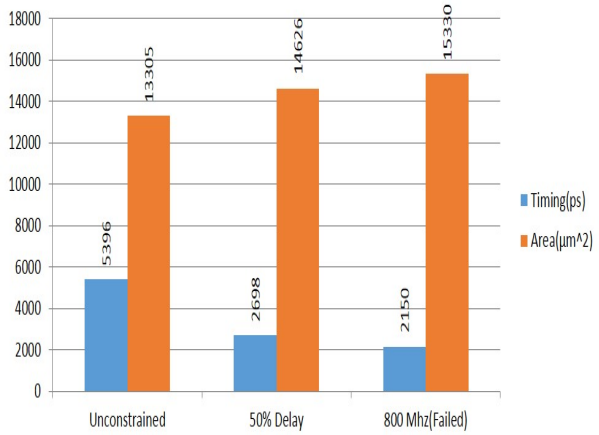


Figure 2: Timing and Area results of ALU-RCA

III. ALU Design- Basic Synthesis

The next task was to synthesize the VHDL code of ALUs using Cadence RTL compiler [42-45]. The VHDL descriptions of ALU will be mapped to certain process technology which in this case is 130-nm technology provided by STMicroelectronics [46]. After starting the RTL compiler the technology files were added by giving path and file name of library files. The next step in synthesis process was to read the VHDL files of ALU-RCA. The important thing to be noted here is that the TestBench was not included here since it is not synthesizable rather just behavioral description. The RTL compiler was instructed to assemble the VHDL descriptions of ALU-RCA into an internal representation i.e. network of logic gates by typing the elaboration command. The VHDL code of ALU-RCA was found synthesizable as it was kept in mind while writing it. The VHDL code of ALU-RCA can be called RTL since it is found to be synthesizable. Some time was spent in studying the gate level netlist produced during elaboration and later using GUI. Up till now we have done initial logic synthesis in which no process technology is used rather the RTL compiler makes use of a virtual gate library. The next step was to assign our hardware descriptions to real standard cells, which is commonly known as technology mapping. The initial synthesis was done without any timing constraint and using low effort to study the intrinsic behavior of implementation. The reason for using low effort here is that we do not want to optimize the timing at all to study the actual behavior of our design which is also known as Static Timing Analysis (STA). The timing and area of our design was documented by giving the appropriate commands. The worst-case delay value and estimated area of the implementation was found to be $5396ps$ and $13305\mu m^2$ respectively.

The worst-case signal propagation path of ALU-RCA was found to be between input and output registers through the chain of adders starting from index 0 to 31, this was observed by looking at the GUI window. The design was re-synthesized using the new timing goal of 50% of the delay we obtained in the last task which comes to be $2689ps$ using medium effort. Here we are using medium effort because we want the Compiler to put some more effort to meet this timing constraint. The worst-case delay value and estimated area of the implementation was found to be $2698ps$ and $14626\mu m^2$ respectively.

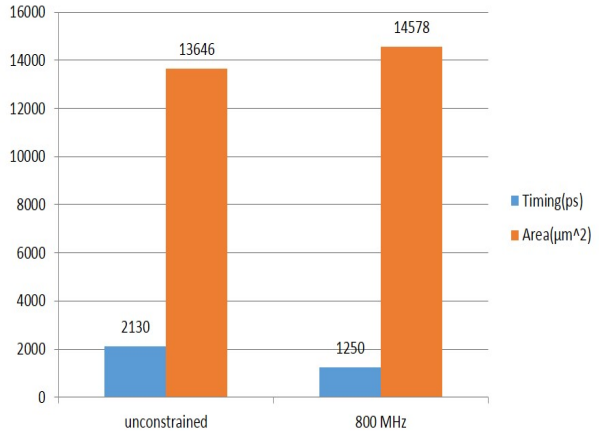


Figure 3: Timing and Area results of ALU-SKL

The RTL Compiler was able to meet this timing constraint at the expense of increased area because it has to put more effort to meet this timing constraint. The worst-case timing path was still passing through the chain of adders. The ten longest signal propagation paths were documented and they all were passing through the chain of full adders. The data books were studied to see what standard cells are being used. In the last task when we synthesized the design without timing constraint it was using one bit fuller adder cells but with timing constraint of $2698ps$ it is now using half adders those having less area than full adders. But as the half adders are used in more numbers than the full adders, so the total area of design has increased.

Later it was found out that the original intension of our ALU design is actually to put in inside a 800-MHz processor. So the ALU-RCA was re-synthesized using $1250ps$ timing constraint and with medium effort. The ALU-RCA was unable to meet this timing constraint as there was negative slack time, so we cannot use ALU-RCA for 800-MHz processor. The worst-case path was still passing through the chain of adders.

This time the standard cells used in implementation were different from the previous task as Compiler tried its best to meet the timing constraint using fast standard cells. The worst-case delay value and estimated area of the implementation was found to be $2150ps$ and $15330\mu m^2$ respectively.

The compiler was only able to reach the timing of $2150ps$ with increased area as it has to put extra effort in trying to meet this stricter timing constraint but was unsuccessful. The next step in the ASIC design flow is the verification of synthesized netlist of ALU-RCA. The VHDL description of ALU-RCA was synthesized again with new timing constraint of $2698ps$ for which function is guaranteed using medium effort. The TestBench was used with test vectors for the verification of synthesized netlist. The clk period used in the TestBench was 50% of timing constraint i.e. $1349ps$ meaning that clock used in the TestBench would be high for $1349ps$ and low for $1349ps$. The netlist was successfully verified without any errors, which proves the idea that netlist has the same functionality as that of VHDL description of design. Fig. 2 shows the timing and area results of ALU-RCA at different settings.

IV. ALU Design- Design Respin and Power analysis

The ALU-SKL was synthesized without timing constraint and using low effort to study the intrinsic implementation. The worst-case delay value and estimated area of the implementation was found to be $2130ps$ and $13646\mu m^2$ respectively. The ALU-SKL was re-synthesized using stricter timing constraint corresponding to 800-MHz using medium effort. The worst-case delay value and estimated area of the implementation was found to be $1250ps$ and $14578\mu m^2$ respectively. The ALU-SKL was able to meet this timing constraint at the expense of increased area which means that we can use ALU-SKL for 800-MHz processor. The worst-case timing path of ALU-SKL was found to be passing through shifter block. The main reason of using Sklansky adder is that it is faster than RCA based adder. The worst-case path also proves this as in the case of ALU-RCA the worst case was passing through the chain of adders and now it is through the shifter block because of high performance of Sklansky adder in terms of speed. The data books were studied again to see which standard cells Sklansky adder is using and it was found out that it was using completely different cells than RCA e.g. 4 Input NOR gate etc and that's why it has more speed at the expense of larger area. The ten longest paths were documented and it found that first nine paths were passing through the shifter while the tenth path was passing through the Sklansky adder. Fig. 3 shows the timing and area results of ALU-SKL at different settings.

The Fig. 4 and 5 shows that the area of ALU-RCA changes more rapidly than ALU-SKL as the ALU-RCA has to put more effort to meet the stricter timing constraint and its area increases, while ALU-SKL is fast adder easily meets the stricter timing constraint without increasing the area. The netlist of ALU-SKL was also successfully verified.

The next task was to perform power analysis of both designs for a timing constraint that both ALUs satisfy. So Both the ALUs were synthesized with the timing constraint of $2500ps$ using medium effort. The estimated area of the implementation was found to be $14828\mu m^2$ for ALU-RCA and $13825\mu m^2$ for ALU-SKL. The ALU-RCA with DeMux was also synthesized using this timing constraint to compare its area and timing with the ALU-RCA without the DeMux.

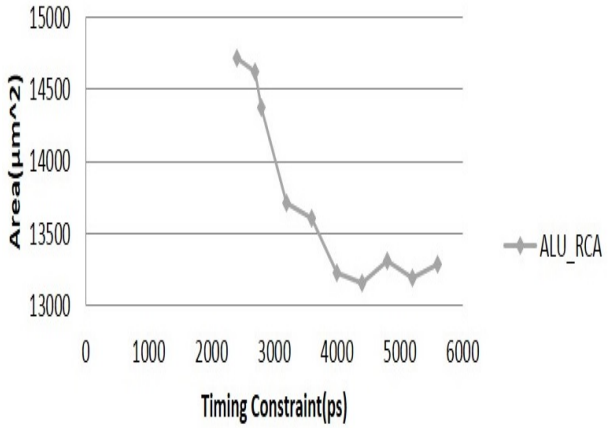


Figure 4: Scaling of Area of ALU-RCA with Timing Constraint

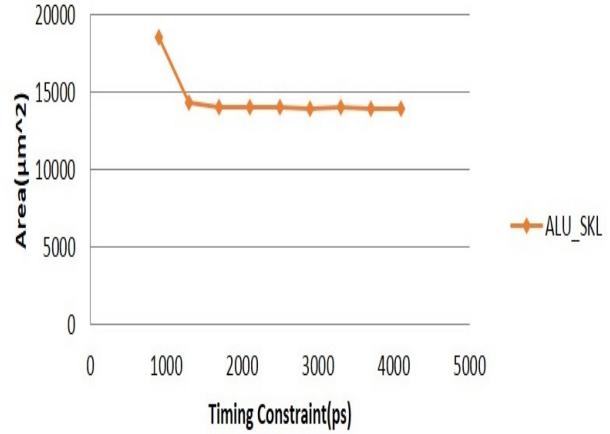


Figure 5: Scaling of Area of ALU-SKL with Timing Constraint

The initial power analysis was performed by assigning some switching probabilities on the primary inputs. This is common practice as initially the the test vectors are not available but the correctness of the power analysis highly depends on the test vectors used for power analysis. The Table I and II clearly show that increasing the toggling probability increases the switching power where as the leakage power almost remains the same. Table III shows the clk power and Capacitance of ALU-RCA and ALU-SKL. The ALU-RCA is consuming more power and has more area compared to ALU-SKL for stricter timing constraint of $2500ps$. This is because ALU-RCA has to put more effort to meet this timing constraint which result in more area and high power, while ALU-SKL has no problem in meeting this timing constraint which result in more area and power efficient design. The individual power for the three blocks was also compared and it was found out that adder was consuming the most power then logical block and shifter was consuming the least power. The power

dissipated in the clock net for both the ALU-RCA and ALU-SKL was documented and was found to be in agreement with the common expression $f * V_{DD}^2 * C$, it should be noted here that the RTL compiler shows wrong unit for capacitance.

Table I: Power results of ALU-RCA and ALU-SKL

	Toggling Probability(ns)	Dynamic Power(nW)	Leakage Power(nW)
ALU-RCA	0.02	2005282	377251
ALU-SKL	0.02	1806029	325786

Table II: Power results of ALU-RCA and ALU-SKL

	Toggling Probability(ns)	Dynamic Power(nW)	Leakage Power(nW)
ALU-RCA	0.1	5039712	377858
ALU-SKL	0.1	4443770	326528

Table III: Synthesis results

	CLK Power(nW)	Capacitance(fF)
ALU-RCA	219974	381
ALU-SKL	178560	310

The next task was to relax the timing constraint for both the ALUs to see the impact of this on area and power. The timing constraint was set to $4500ps$ and both the designs were re-synthesized using medium effort. The estimated area of the implementation was found to be $13468\mu m^2$ for ALU-RCA and $13819\mu m^2$ for ALU-SKL. Now it can be seen that the ALU-RCA using less area and power as compared to ALU-SKL. It means that it is better to use ALU-RCA if the timing constraint is not high and we require area and power efficient ALU. Table IV and V shows the power results for timing constraint of $4500ps$.

Table IV: Power results of ALU-RCA and ALU-SKL

	Toggling Probability(ns)	Dynamic Power(nW)	Leakage Power(nW)
ALU-RCA	0.02	1315083	318043
ALU-SKL	0.02	1332294	325224

Table V: Power results of ALU-RCA and ALU-SKL

	Toggling Probability(ns)	Dynamic Power(nW)	Leakage Power(nW)
ALU-RCA	0.1	3877682	317800
ALU-SKL	0.1	3974884	325085

Table VI shows the power and area results for two different ALU-RCA designs. The design with DeMux is more power efficient as the switching is only taking place in the block that is needed at that time depending on the opcode but has little area overhead compared to the other design which is without DeMux. The power analysis was performed using toggling probability of 0.1 ns on the primary inputs.

The next task of power analysis was to do power analysis of ALU-RCA using three different set of test vectors, as discussed earlier that the correctness of power analysis highly depends on the test vectors used for power analysis. The TestBench

Table VI: Synthesis results of ALU-RCA

ALU-RCA Type	Dynamic Power(nW)	Area(μm^2)
With DeMux	4887000	14931
Without DeMux	5039712	14828

was used to generate the VCD file for each set of test vectors using the synthesized netlist generated from the design with correct timing constraint, here timing constraint was set to $2500ps$ with medium effort. One important thing to note is that the clock period used in the TestBench should be 50% of timing constraint. Table VII shows the power results using test vectors. The result of power analysis using test vectors shows that the toggling probability of Random test vectors is higher as compared to Regular and Real trace because random test vectors has highest switching power where as the toggling probability of regular and real trace seems to be almost same and seems to be close to 0.1 ns. As instructed the TCF files were checked to compare the signals A[16] and B[15] in all the three TCF files. It was found that in the regular trace test vectors the signal A[16] was changing state from 0 to 1 and vice versa all the time and has the high state toggling probability of almost 0.5 ns, where as the signal B[15] was all the time zero and has the high state toggling probability of almost 0.0 ns.

Table VII: Power results of ALU-RCA

	Random	Regular	Real Trace
Dynamic Power(nW)	9524859	4578158	4509525
Leakage Power(nW)	378506	379258	387047

V. ALU Design- Place and Route

The final task in the ASIC design flow is place and route step which takes considerable amount of experience to make good place-and-route. The first step was to generate netlist file of our own design using timing constraint of 3.2 ns also need to produce the constraint file as we decided to work with our ALU-RCA design. Then these files were placed in the proper directories as directed. The partitioning step was performed followed by the Floorplanning, we placed the input registers on the left and the output registers on the right side of core. Then pin placement and power routing was done. The standard cell placement was the next step, it was found out that the placement of cells was done according to our pre-placement constraint. Then Clock Tree Synthesis (CTS) step was performed, after Pre-CTS optimization timing was checked and the timing constraint was not met. The Pre-CTS step can be explained as mapping the design to logic gates, without mapping to actual cells i.e. buffers. The actual CTS step was performed which is like mapping the design to actual cells. The positions of clock buffers and clock tree were checked and it was found that our design has one level of buffers. The timing was checked again after this step and we were still unable to meet the constraint. The last step of CTS i.e. post-CTS optimization was performed to do the optimization based on existing clock tree. The timing was checked again and it was found that timing was improved a lot and the slack time was -0.466, much better than before when the slack time was -2.259. The routing and post-route

optimization was performed and the clock and reset signals should have highest priority because these signals have to be provided to every block in the design and therefore are critical. The Filler cells were used to fill the gaps and to connect them to the power rails. The layout verification was done, four MinCut violations were found which were later removed using the fixMinCutVia command. The final timing analysis was performed and the slack time was found to be -0.395.

VI. Conclusion

The aim of this research was to design a Arithmetic Logic Unit (ALU) for a 32-bit processor using two different adder circuits. The two ALU units were implemented in VHDL using Ripple Carry Adder and Sklansky Adder circuits. After the VHDL implementation synthesis was performed using Cadence RTL compiler to compare the performance of both the ALU units in terms of area, power and timing requirements. The VHDL descriptions of ALU were mapped to 130-nm process technology provided by STMicroelectronics. The synthesis results shows that the area of ALU-RCA changes more rapidly than ALU-SKL as the ALU-RCA has to put more effort to meet the stricter timing constraint at the expense of more area. While ALU-SKL which is a fast adder easily meets the stricter timing constraint without increasing the area and power consumption. It was also observed that the ALU-RCA uses less area and power as compared to ALU-SKL, so it is better to use ALU-RCA if the timing constraint was not high so in this way we can get more area and power efficient ALU Design.

Acknowledgment

This work is partially supported by the Chinese Academic of Sciences and The World Academy of Sciences CAS-TWAS President's Fellowship 2013-2017.

References

[1] Ravindran, N.; Lourde, R. Mary "An optimum VLSI design of a 16-BIT ALU", Information and Communication Technology Research (ICTRC), 2015 International Conference on, On page(s): 52-55

[2] Larsson-Edefors, P.; Jeppson, K. "Timing- and power-driven ALU design training using spreadsheet-based arithmetic exploration", Microelectronics Education (EWME), 10th European Workshop on, On page(s): 151 154

[3] Kaur, H.; Singh, H. "Advanced ALU with inbuilt selection modules for Genetic Algorithm processor", Signal Processing, Computing and Control (ISPCC), 2015 International Conference on, On page(s): 405 - 410

[4] Kaur, Harmeet; Kaur, Manpreet; Singh, Haramardeep "Genetic Algorithm processor with advanced ALU, memories and control unit", India Conference (INDICON), 2015 Annual IEEE, On page(s): 1 - 5

[5] H. Singh, "Design of enhanced arithmetic logical unit for hardware genetic processor", Advances in Computing, Communications and Informatics (ICACCI), 2014 International Conference on, pp. 549-553

[6] M. Suzuki, N. Ohkubo, T. Yamanaka, A. Shimizu, and K. Sasaki, "A 1.5 ns 32 b CMOS ALU in double pass-transistor logic", IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers, vol. 267, pp. 90-91, 1993

[7] Parameswar, A.; Hara, H.; Sakurai, T. "A swing restored pass-transistor logic-based multiply and accumulate circuit for multimedia applications", Solid-State Circuits, IEEE Journal of, On page(s): 804 - 809 Volume: 31, Issue: 6, Jun 1996

[8] T. -Y. Chang and M.-J. Hsiao, "Carry-select adder using single ripple-carry adder", Electronics Letters, vol. 34, no. 22, pp. 2101-2103, 1998

[9] Chung-Hsun Huang; Jinn-Shyan Wang; Chingwei Yeh; Chih-Jen Fang "The CMOS carry-forward adders", Solid-State Circuits, IEEE Journal of, On page(s): 327 - 336 Volume: 39, Issue: 2, Feb. 2004

[10] Senejani, M. Nadi; Ghadiry, M. Hossein "Low dynamic power high performance adder", CAD Systems in Microelectronics, 2009. CADSM 2009. 10th International Conference - The Experience of Designing and Application of, On page(s): 242 - 245

[11] Senejani, M. Nadi; Hosseinghadiry, M.; Miryahyaei, M. "Low Dynamic Power High Performance Adder", Future Computer and Communication, 2009. ICFCC 2009. International Conference on, On page(s): 482 - 486

[12] C.-J. Fang, C.-H. Huang, J.-S. Wang, and C.-W. Yeh, "Fast and compact dynamic ripple carry adder design ", Proc. 3rd IEEE Asia-Pacific Conf. ASIC, pp. 25-28, 2002

[13] G. A. Ruiz, "Evaluation of three 32-bit CMOS adders in DCVS logic for self-timed circuits", IEEE J. Solid-State Circuits, vol. 33, pp. 604-613, 1998

[14] W. Hwang, G. Gristede, P. Sanda, S. Y. Wang, and D. F. Heidel, "Implementation of a self-resetting CMOS 64-bit parallel adder with enhanced testability", IEEE J. Solid-State Circuits, vol. 34, pp. 1108-1117, 1999

[15] A. Rothermel, "Realization of Transmission-Gate Conditional-Sum (TGCS) Adders with Low Latency Time", IEEE J. Solid-State Circuits, vol. 24, pp. 558-561, 1989

[16] N. Burgess, "Fast Ripple-Carry Adders in Standard-Cell CMOS VLSI", 20th IEEE Symposium on Computer Arithmetic, pp. 103-111

[17] Burgess, Neil "Fast Ripple-Carry Adders in Standard-Cell CMOS VLSI", Computer Arithmetic (ARITH), 2011 20th IEEE Symposium on, On page(s): 103 111

[18] M.Moghaddam and M. B. Ghaznavi-Ghoushchi ,"A New Low-Power, Low-area, Parallel Prefix Sklansky Adder with Reduced Inter-Stage Connections Complexity",IEEE Computer society,2011

[19] V. S. Veeravalli and A. Steininger, "Architecture for monitoring set propagation in 16-bit sklansky adder," in Quality Electronic Design (ISQED), 2014 15th International Symposium on, March 2014, pp. 412-419.

[20] N. Burgess, "New Models of Prefix Adder Topologies", J. VLSI Signal Processing Systems, Vol. 40, (June 2004), pp. 125 - 141.

[21] Roy, S.; Choudhury, M.; Puri, R.; Pan, D.Z. "Towards Optimal Performance-Area Trade-Off in Adders by Synthesis of Parallel Prefix Structures", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, On page(s): 1517 - 1530 Volume: 33, Issue: 10, Oct. 2014

[22] Choi, Y.; Swartzlander, E. E. "Speculative Carry Generation With Prefix Adder", Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, On page(s): 321 - 326 Volume: 16, Issue: 3, March 2008

[23] Choi, Y.; Swartzlander, E.E., Jr. "Parallel prefix adder design with matrix representation", Computer Arithmetic, 2005. ARITH-17 2005. 17th IEEE Symposium on, On page(s): 90 98

[24] Liu, Jianhua; Zhu, Yi; Zhu, Haikun; Cheng, Chung-Kuan; Lillis, John "Optimum Prefix Adders in a Comprehensive Area, Timing and Power Design Space", Design Automation Conference, 2007. ASP-DAC '07. Asia and South Pacific, On page(s): 609 615

[25] Chen, Jun; Stine, James E. "Enhancing parallel-prefix structures using carry-save notation", Circuits and Systems, 2008. MWSCAS 2008. 51st Midwest Symposium on, On page(s): 354 357

[26] Yingtao Jiang, Abdulkarim Al-Sheraidah, Yuke Wang, Edwin Sha and Jin-Gyun Chung, "A novel Multiplexer-based low power Full Adder", IEEE Transactions on Circuits and systems-II, vol. 51, 2004

[27] R. Zlatanovici, S. Kao and B. Nikolic, "Energy-delay optimization of 64-bit carry look-ahead adders with a 240 ps 90 nm CMOS design example", JSSC, vol. 44, no. 2, pp. 569-583, 2009

[28] H. Ling, "High-Speed Binary Adder", IBM J. Research and Dev., vol 25, p.156-166 (May 1981) J. Grad and J.E. Stine, "New algorithms for carry propagation", Proc. 15th ACM Great Lakes symposium on VLSI, Chicago, 2005, pp. 396 - 399

[29] Harris, D.; Sutherland, I. "Logical effort of carry propagate adders", Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on, On page(s): 873 - 878 Vol.1 Volume: 1, 9-12 Nov. 2003

[30] Patel, R. A.; Benaissa, M.; Powell, N.; Boussakta, S. "Novel Power-Delay-Area-Efficient Approach to Generic Modular Addition", Circuits and Systems I: Regular Papers, IEEE Transactions on, On page(s): 1279 - 1292 Volume: 54, Issue: 6, June 2007

[31] Roy, S.; Choudhury, M.; Puri, R.; Pan, D.Z. "Polynomial Time Algorithm for Area and Power Efficient Adder Synthesis in High-Performance Designs", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, On page(s): 820 - 831 Volume: 35, Issue: 5, May 2016

[32] Ge Yang; Seong-Ook Jung; Kwang-Hyun Baek; Soo Hwan Kim; Suki Kim; Sung-Mo Kang "A 32-bit carry lookahead adder using dual-path all-N logic", Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, On page(s): 992 - 996 Volume: 13, Issue: 8, Aug. 2005

[33] Esposito, D.; De Caro, D.; Napoli, E.; Petra, N.; Strollo, A. G. M. "Variable Latency Speculative Han-Carlson Adder", Circuits and Systems I: Regular Papers, IEEE Transactions on, On page(s): 1353 - 1361 Volume: 62, Issue: 5, May 2015

[34] Zlatanovici, R.; Kao, S.; Nikolic, B. "EnergyDelay Optimization of 64-Bit Carry-Lookahead Adders With a 240 ps 90 nm CMOS Design Example", Solid-State Circuits, IEEE Journal of, On page(s): 569 - 583 Volume: 44, Issue: 2, Feb. 2009

[35] Zamanlooy, Babak; Novak, Ashley; Mirhassani, Mitra "Complexity Study of the Continuous Valued Number System Adders", Multiple-Valued Logic (ISMVL), 2012 42nd IEEE International Symposium on, On page(s): 116 121

[36] Yuke Wang; Parhi, K.K. "A unified adder design", Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on, On page(s): 177 - 182 vol.1 Volume: 1, 4-7 Nov. 2001

[37] Kwak, Sanghoon; Har, Dongsoo; Lee, Jeong-Gun; Lee, Jeong-A "Design of Heterogeneous Adders Based on Power-Delay Tradeoffs", Embedded Computing, 2008. SEC '08. Fifth IEEE International Symposium on, On page(s): 223 226

[38] Roy, Subhendu; Choudhury, Mihir; Puri, Ruchir; Pan, David Z. "Polynomial time algorithm for area and power efficient adder synthesis in high-performance designs", Design Automation Conference (ASP-DAC), 2015 20th Asia and South Pacific, On page(s): 249 254

[39] Baliga, Akansha; Yagain, Deepa "Design of High Speed Adders Using CMOS and Transmission Gates in Submicron Technology: A Comparative Study", Emerging Trends in Engineering and Technology (ICETET), 2011 4th International Conference on, On page(s): 284 289

[40] Vergos, H.T.; Efstathiou, C.; Nikolos, D. "Diminished-one modulo 2n+1 adder design", Computers, IEEE Transactions on, On page(s): 1389 - 1399 Volume: 51, Issue: 12, Dec 2002

[41] Sun Xu-guang; Mao Zhi-gang; Lai Feng-chang "A 64 bit parallel CMOS adder for high performance processors", ASIC, 2002. Proceedings. 2002 IEEE Asia-Pacific Conference on, On page(s): 205 208

[42] Cadence EDA Tools. [Online]. Available: http://www.cadence.com

[43] Abdul Rehman Buzdar, Liguo Sun, Azhar Latif and Abdullah Buzdar, "Distance and Speed Measurements using FPGA and ASIC on a high data rate system" International Journal of Advanced Computer Science and Applications(IJACSA), 6(10), 2015, 273-282.

[44] Abdul Rehman Buzdar, Liguo Sun, Azhar Latif and Abdullah Buzdar, "Instruction Decompressor Design for a VLIW Processor", Informacije MIDEM-Journal of Microelectronics, Electronic Components and Materials Vol. 45, No. 4 (2015), 225-236.

[45] Abdul Rehman Buzdar, Azhar Latif, Liguo Sun and Abdullah Buzdar, "FPGA Prototype Implementation of Digital Hearing Aid from Software to Complete Hardware Design" International Journal of Advanced Computer Science and Applications(IJACSA), 7(1), 2016, 649-658.

[46] STMicroelectronics. [Online]. Available: http://www.st.com

[47] ModelSim. [Online]. Available: https://www.mentor.com

Computational Modeling of Proteins based on Cellular Automata

Alia Madain, Abdel Latif Abu Dalhoum, Azzam Sleit
Department of Computer Science
King Abdulla II School for Information Technology
The University of Jordan
Amman, Jordan

***Abstract*—The literature of building computational and mathematical models of proteins is rich and diverse, since its practical applications are of a vital importance in the development of many fields. Modeling proteins is not a straightforward process and in some modeling strategies, it requires to combine concepts from different fields including physics, chemistry, thermodynamics, and computer science. The focus here will be on models that are based on the concept of cellular automata and equivalent systems. Cellular automata are discrete computational models that are capable of universal computation, in other words, they are capable of doing any computation that a normal computer can do. What is special about cellular automata is its ability to produce complex and chaotic global behavior from local interactions. The paper discusses the effort done so far by the researchers community in this direction and proposes a computational model of protein folding that is based on 3D cellular automata. Unlike common models, the proposed model maintains the basic properties of cellular automata and keeps a realistic view of proteins operations. As in any cellular automata model, the dimension, neighborhood, boundary, and rules were specified. In addition, a discussion is given to clarify why these parameters are in place and what possible alternatives can be used in the protein folding context.**

Keywords*—*Proteins 3D Folding; Bioinformatics; Computational Modeling; Cellular Automata; Theoretical Computer Science;

I. Introduction

Modeling any complex biological phenomenon is essentially a form of abstraction. The game of building a meaningful model usually falls back to making choices of what to keep and what to eliminate from available information. Nevertheless, Models have many advantages, as they tend to be more accessible and convenient for understanding the subject of study. Additionally, models can act as objects of further experimentation [1]. This perfectly applies to modeling proteins, because not only they are diverse, but also the simplest protein endures a huge amount of details.

Artificial Intelligence and image processing concepts are heavily used in the domain of modeling proteins such as neural networks [2], optimized evidence-theoretic K-nearest neighbor classifier [3], complexity measure factor [4], moments [5], in addition to fusing multiple classifiers [6].

A cellular automaton (CA) is a discrete model of computation that is studied in computability theory. CAs are simple since they are based on local interactions only but they are capable of exhibiting complex behavior [7].

Simply a CA has a collection of identical cells that are distributed spatially in one dimension, two dimensions or higher. Every cell in the CA has a finite number of possible internal states, the CA evolves from one iteration to the other based on transition rules that are applied simultaneously to all CA cells. The rules depend mainly on the cell neighborhood and may or may not consider the cell state itself.

There are many options for almost all aspect of CAs. CAs differ in their spatial distribution, cell neighborhood, transition rules, cell possible states, boundary, number of generations (iterations), cells shape, and the initial configuration from where the CA starts.

Although, all proteins composition is based on twenty amino acids, proteins are diverse and cover multiple functions in nature. Some proteins contain a surfeit of one amino acid whereas others may have one or two members of the twenty amino acids missing entirely [8]. Since there are many details in real proteins, simplified models called simple exact models (SEMs) were proposed. The most common one is the HP model, which consists of only hydrophobic (H) and polar (P) Monomers [9].

This paper discusses the CA potential in the domain of protein modeling and shed light on the possibilities offered by the CA concept. In addition, it focuses on the process involved in protein modeling when CA is used which is quite different from other computational paradigms. Finally, a 3D CA model is proposed and the challenges of protein modeling in terms of CA are discussed.

The remaining of this paper is organized as follows: Section II gives the related work; Section III includes background information about proteins; Section IV discusses the CA potential in the context of protein modeling; Section V presents the proposed model; Finally, section VI concludes the work done and gives direction to future work.

II. Related Work

The CA concept is related to many disciplines including mathematics, physics, biology, and computer science [10]. The idea of employing CA to the central dogma of molecular biology is not new and many attempts were made to model the central dogma in terms of structure, function, and evolution. CA models were used in modeling DNA sequences [11], evolution [12], mutation prediction [13], and gene networks [14].

One of the most attractive properties of CA is its ability to represent global behavior, and this is truly important in modeling the central dogma of molecular biology since the initial state of the protein synthesis process does not help in understanding the system behavior as a whole.

In this section, the discussion covers the work that depends on elementary cellular automata combined with pseudo-amino acid composition. In addition to the methods that combine CA with evolutionary algorithms, and finally work done in L-Systems is covered, since L-Systems were proved to be equivalent to CAs.

One work that is used in predicting multiple protein attributes is that based on elementary Rule 84 and pseudo-amino acid composition. This line of research depends on amino acid coding language proposed in [15] to act as the initial configuration of the elementary CA. This model is used to predict protein subcellular location [16], the G-protein-coupled receptor functional classes [17], and protein structural classes [18] [19].

The process starts with converting the protein amino acid sequence to the binary encoding and assumes the binary representation of each protein sequence as the initial configuration of the CA, after the CA runs for 100 generations, the resulting image parameters are extracted as given in Figure 1. These CA image parameters are then considered along with 20 more attributes to calculate the PseAA representation of each protein and each group of proteins as given in Figure 2.

In fact, PseAA proposed in [20] is widely used in protein modeling, which differs from traditional AAC in that it adds the protein sequence order effect in a set of discrete numbers. Surveys tailored to methods depending on the pseudo amino acid composition are given in [21] and [22].

In addition, CA was combined with evolutionary algorithms. An interesting work is the one that proposes a CA-like structure or a neural CA, where the cellular automaton is implemented by means of a simple feed forward neural model. The artificial neural network output correspond to the possible relative movements.

The idea was implemented in two-dimensions (2D) [23] and three-dimensions (3D) [24]. In 2D case the possible movements are forward, left and right while in the 3D case, the possible movements are forward, up, down, left, and right.

The work done in [25] combines CA with genetic algorithms to predict the protein secondary structure, where the genetic algorithm is used to optimize the parameters (Rules) involved. The authors summarizes what effects the prediction to three factors: the neighborhood, weights assigned to the neighborhood and the number of generations.

Specially designed cellular automata were proposed to model the chemical reactions of DNA replication, mRNA transcription, and splicing process in [26], where the protein synthesis process was left for future work.

Moreover, Systems proven to be equivalent to cellular automata such as L-systems [27] [28] were used to model proteins in [29] [30] and [31].

In this paper, the design of a 3D CA proteins model is discussed. The model is meant to be as simple as possible and within the known CA configurations. It does not require any combined representation nor does it require searching capabilities and evolutionary algorithms. The discussion also covers challenges in such design and present alternative solutions to each.

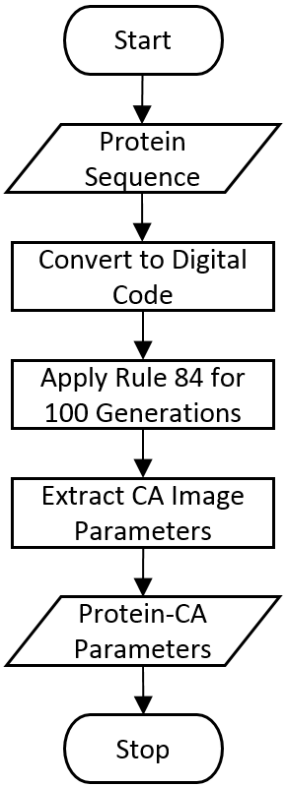


Fig. 1. Workflow of Protein Parameters Extracted from CA Images

III. Proteins

Proteins are the end product of the DNA decoding process. The central dogma of molecular biology states that DNA is transcribed into messenger RNA (mRNA), which is translated into proteins. This way of viewing the process is quite simplified, in reality this biological process is a rich and complex set of events [32].

In a cell, proteins are the workhorses and lead performers of cellular functions [33], they can be considered as specialized machines, each of which fulfills its own task. All the complex molecules of the cell are proteins except DNA and RNA which are not proteins and considered complex as well [32].

To simplify things, proteins are all united through their reliance on the same group of twenty amino acids, they consist of a linear arrangement of amino acid residues assembled together into a polypeptide chain and the order of linking the residues together is ultimately derived from the genes information.

Amino acids contain amine (-NH2) and carboxylic acid (-COOH) functional groups, usually along with a side-chain usually referred to as an R group that is specific to each amino acid. The key elements of an amino acid are carbon, hydrogen,

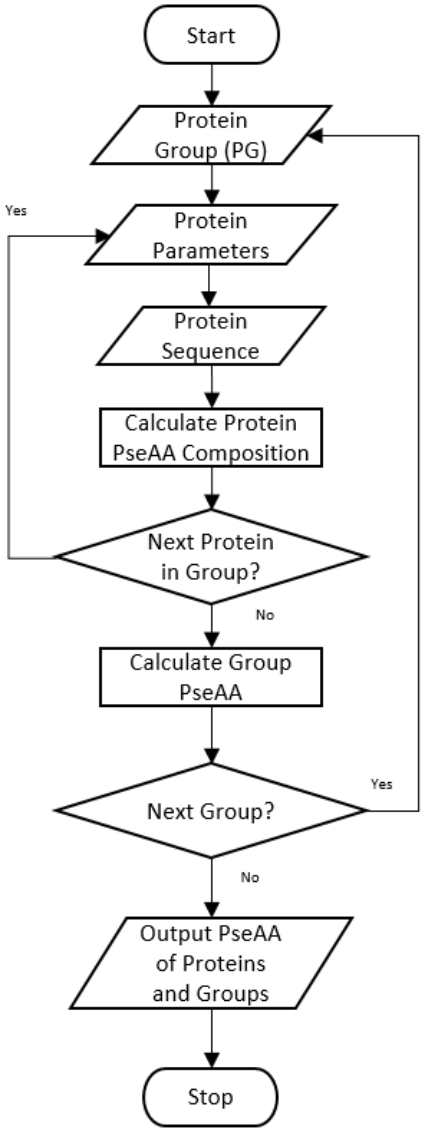


Fig. 2. Workflow of PseAA Composition of Proteins and Their Groups

oxygen, and nitrogen, though other elements are found in the side-chains of certain amino acids [34].

The translation to proteins starts by the ribosomes which proceed along the mRNA one codon at a time incorporating one amino acid at each step and finally leaving the mRNA from the last codon.

This process results in a chain of amino acids called the primary structure of a protein. Mathematical and computational modelling of ribosomal movement along with a discussion of the impact of modeling studies on experimentalists is summarized in [35].

Protein folding is the process that folds the protein primary chain to its native three-dimensional structure, which is a specific and stable structure. The three-dimensional structure of a protein defines its function [33].

IV. Cellular Automata

A. Cellular Automata Potential

Starting this section with some historical information may sound redundant but it is essential to get a feel of why CA has potential to model proteins properly. The use of CA as a model is justified by its roots in biology and that it is especially relevant to the problem of protein modeling.

CAs were originally proposed as formal models of self-reproducing organisms. In the forties, John Von Neumann wanted to design a machine that can reproduce itself. He suggested a programmable assembly machine that can build a copy of itself, and he defined two phases in the machine blueprint, which are translation and transcription. The problem of this machine is in its components, which are sophisticated logical units. This is when Stanislaw Ulam suggested that Von Neumann use cellular automata, which Ulam used to study the growth of crystals at the time.

Also, the theory of Konrad Zuse is very relevant in this context, which suggests that physics is just computation. Zuse tried to apply an information and automata theory approach to certain problems of physics [36] in his article written in German (Rechnender Raum) which literally means space that is computing. In 1969, he published the book Rechnender Raum [37] which was translated into English as ”Calculating Space”, Zuse proposes that the universe is computed by some sort of CA or other discrete computing machinery.

In addition to being a suggested framework for researching connections between biology and automata theory, CAs design is open and flexible, there is no restrictions or mathematical formulas that restricts the construction of CAs. Another advantage of using CAs is the different behavior dynamics resulting from different rules namely, stable, periodic, chaotic and complex ones.

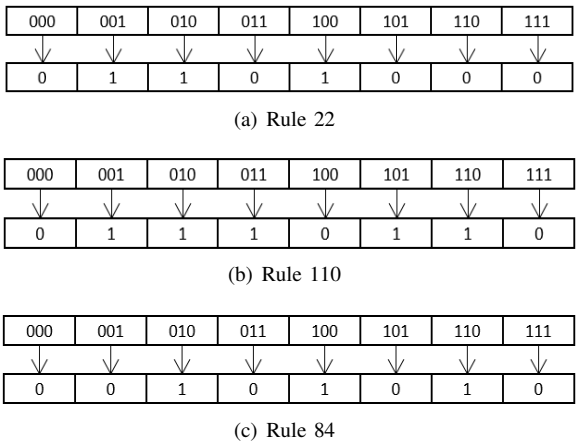
Finally, CAs are parallel in their nature which can be applied in many different ways using commercially-available parallel computers where the state of cells can be updated simultaneously, or using specialized CA machines.

B. Cellular Automata Technical Details

CA can be described as a set of cells arranged in any dimension, for example, cells can be arranged in a two dimensional grid or a one dimensional array. These cells can take a finite number of states and the states can be of any type for example the set of states can be binary (0,1) or an integer number or any other finite set of states.

The state of each cell may change or stay the same at every generation, the stability or change of the states depends on predefined rules (a transition function). The rules use the state of a cell neighbors as input and may or may not use the cell state itself to determine the cell state in the next generation.

According to Wolfram, it seems that the patterns which arise from different types of cellular automata can almost always be assigned to one of just four basic classes [38] [39] [10]. In class 1, patterns evolve into a stable, homogeneous state; in class 2, patterns evolve to a periodic state; in class 3 a chaotic behavior appears; and in class 4, configurations contain structures that interact in complex ways.



(a) Rule 22

(b) Rule 110

(c) Rule 84

Fig. 3. Different CA Rules

Langton [40] defines the Lambda parameter, which is a way to quantify the qualitative approach of Wolfram. In ideal cases, transition rules with the same Lambda evolve to a similar behaviour [41].

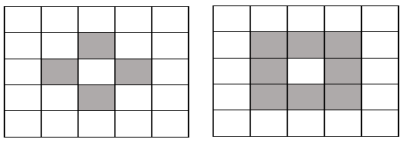
Although it is widely accepted that the power of CA is its ability to exhibit fascinatingly complex behavior from local rules, there is no superior behavior in CAs. In biological modeling, the appropriate behavior is the one that represents reality. In practical applications, the best behavior is the one that achieves the goals of the application, for example, work done in [42] uses chaotic behavior in image security and chaotic elementary CAs had an equivalent effectiveness as complex game of life in a multimedia related application [43] [44].

The simplest cellular automata (elementary CA) is one dimensional and the rules depend on the cell state and the state of its left and right neighbors (values of the nearest neighbor). So the combinations of each cell and its neighbors have 8 possibilities only. There are only 256 elementary cellular automata, each of which can be indexed with an 8-bit binary number. All the behavioral cases defined by Wolfram are covered within the 256 rules of the elementary CA.

The CA is referenced by its rule number, which can be easily computed in the case of elementary CA. The rule number is simply the decimal number representing the rule output, so for every combination of the three cells (core cell and its neighbors) the rule give an output that is either zero or one, this output is then concatenated to a binary string and converted to a decimal number representing the rule number. Figure 3 shows the 8 states of rules 22, 110, and 84.

In a two-dimensional context (2D) some parameters are different. In 2D CA Moore and von Neumann are two widely used neighborhood configurations. In von Neumanns neighborhood, every cell has four neighbors: the cells at its North, South, East, and West, whereas in Moores neighborhood the cells at the four diagonals are also considered, as given in Figure 4.

One famous two-dimensional CA is the Game of life proposed by John Conway [45]. The rules of Conways game of life are simple and assumes a Moore Neighborhood.



(a) von Neumann (b) Moore

Fig. 4. Different CA Neighborhood

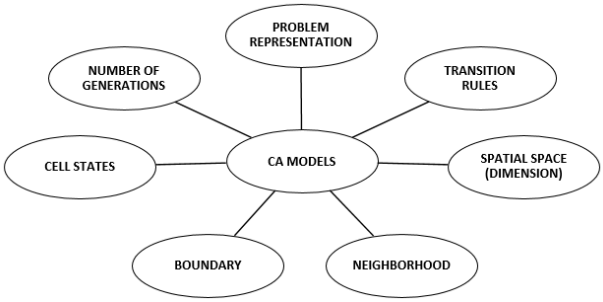


Fig. 5. CA based Modeling

V. Proposed Model

The problem of modeling proteins using cellular automata requires representing the problem in a way that maps to reality and that is suitable for cellular automata at the same time, it also requires setting multiple parameters such as the dimension, neighborhood, boundary, number of generations, and most importantly the transition rule as given in figure 5. In addition, there must be a way to check the validity of the CA used and to make sure it represents a realistic behavior, maybe by finding certain attributes specified by this process.

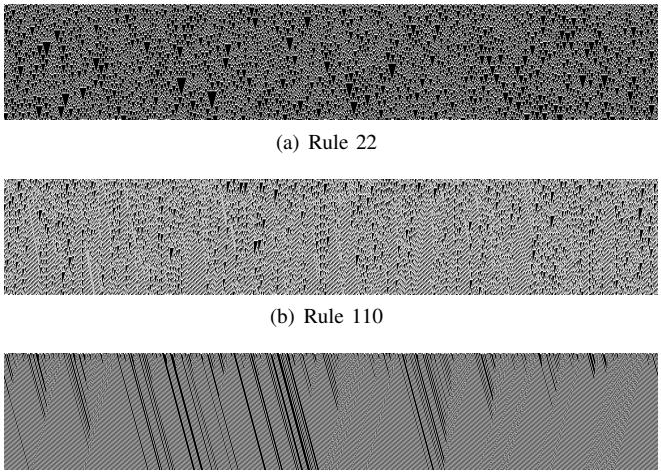
A. Problem Representation

The process of modeling proteins starts with the challenge of representing the problem in a set of finite, discrete values. Proteins in reality are full of details. Until today, the functional motions of proteins usually operate at timescales and conditions that are beyond the limits of current technology [46].

One way to specify the values that each cell in the CA uses, is to convert the twenty amino acids to a five digits binary representation. The binary representation has many advantages in the context of CA since the properties of CA are mostly studied in the binary domain.

The conversion between amino acids and binary representation is not random. Authors in [15] and [47] makes use of similarity rule, complementarity rule, molecular recognition theory, and information theory to give the digital coding of amino acids. Figure 6 shows the use of this coding in protein modeling using rules 22, 110, and 84. The figure shows the initial configuration and 100 generations of the same protein. The reason why these rules were chosen is that rule 22 is known for the chaotic behavior and rule 110 proved to being capable of complex behavior and rule 84 was used before in the context of protein modeling.

A comparison of four coding methods is given in [48], the binary codes presented are either based on biochemical properties or generated by artificial intelligence (AI) methods.



(a) Rule 22

(b) Rule 110

(c) Rule 84

Fig. 6. Different CA Rules representing the same protein

Another approach is the use of the HP model, which classifies the amino acids into two classes, the hydrophobic ones and the polar ones. Although this is a very useful abstraction, keeping the 20 amino acids with their one letter representation might allow for adding more rules of biochemistry.

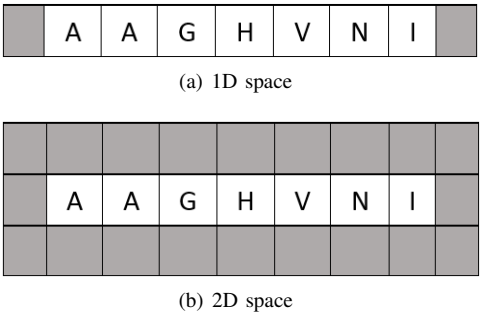
Notice that some approaches combines two methods together. In [16] [18], [17] and [19], the binary representation is followed by the Pseudo-amino acid composition (PseAA). In those Papers, CA is applied to the binary representation, then the CA image parameters are extracted by methods such as the geometric moments of Hu [49] and the GLCM texture features [50]. As explained before in the related works section, the discrete numbers identifying the protein are added to 20 attributes to form the PseAA composition of proteins.

B. Environment and CA parameters

The environment effects the protein folding, for example, the hydrophobic and hydrophilic properties of the amino acids forming the proteins are important in the context of protein folding since the environment surrounding the protein contains water.

The work done in [29] [30] and [31] models the folding of protein-like structures using local rewriting rules with environmental interaction. In the context of cellular automata, the dynamic environmental factors may be modeled by the CA rules whereas the static environmental factor such as the existence of water may be assumed as a part of the initial configuration.

It is assumed that CAs have an infinite grid then every cell has neighbors. Nevertheless, the actual implementation of space is usually finite and therefore there must be a way to handle the neighborhood over the edges. The proposed model uses water as the boundary of the initial configuration. Therefore, if the CA is implemented in a one-dimensional space the neighbors of the first cell and the last cell are two cells of water and in the case of two-dimensional space, the extreme cells in the four dimensions are assumed water cells, as shown in Figure 7.



(a) 1D space

(b) 2D space

Fig. 7. Suggested CA initial configuration in 1D and 2D spaces

The proposed model uses 3D CA space, so now the CA used is more like a cube. The reason why the 3D model was chosen is that it maps to proteins reality. At first, the finite spatial space and the water boundary are assumed, then the flexibility of expanding the spatial space is given.

One main issue here is that the amino acids move and since CA is highly parallel, one might expect that these amino acids will end up in the same cell, which makes the possible state of a cell more complicated. Based on the basic principles of CA, there must be a finite set of states for the CA cells. In order to overcome this problem there are many possible solutions. The first thing that comes to mind is to restrict the possible movement of each cell, but that might effect the model accuracy.

Another work around is to change the cell shape to become hexagonal (still a homogeneous grid). What is gained from this conversion is that it is possible to map the hexagonal cell to a group of square cells with adding the radius of neighborhood and without effecting the assumption of a finite set of states in each cell. This work around adds to the complications of the design.

Another alternative solution is the use of a timed CA where neighboring cells or competing cells to the same position does not calculate the transition function at the same time. Each cell has its turn based on its position. So the cells in the second round can check the availability of the position. A similar solution is to use block CA or a partitioning CA, where groups of cells are divided into non-overlapping blocks and instead of applying the transition rule to each cell individually, it is applied to a whole block at each time step.

An interesting challenge in modeling proteins in terms of CA is keeping the amino acid connected to its neighbors in the primary structure, one way to do this is to number the primary chain of the protein, and check that there is no separation between originally connected amino acids. This problem can be partially fixed after the final generation or it can be ignored. A more complicated solution is to backtrack illegal moves after each generation.

The actual neighborhood chosen is usually crucial for the global behavior of a CA. most CA studies restrict the neighborhood to Moore or von Neumann [51]. The details of the proposed 3D CA is given in Table I. The use of Moore neighborhood has an advantage in modeling protein folding, since the presence of certain amino acids and the connections between them effect the folding. In addition,

TABLE I. CA PARAMETERS

No.	Parameter	Value
1	Dimension	3D
2	Boundary	Water
3	Neighborhood	3D Moore Neighborhood
4	Spatial Space	Infinite (finite and can be extended)
5	Number of Generations	Specified based on the protein stabilization
6	Possible States	Amino acids of the primary structure and water

Moore neighborhood might serve the heuristic rules described in section V-C and gives a meaningful abstraction of each cell environment.

C. CA Rules

Section V-B discusses the CA configuration before moving from one generation to the other. The rules or the transition function is what cause the global behavior to occur; they are essential to understand the behavior of proteins.

According to Chou [52], the three main strategies developed in structural bioinformatics, are pure energetic approach, heuristic approach, and homology modeling approach. Pure energetic approach depends on the thermodynamics principle. The heuristic approach on the other hand collects the physical, chemical, and biological principles as much as possible.

Finally, the homology modeling approach, which is a well-known method of modeling proteins, compares the protein in hand (target protein) with related proteins stored in a database (template proteins). When the target and template proteins are closely related, homology modeling can produce accurate structural models with more reliable results than other methods. Nevertheless, the quality of the homology model depends on the data used and the quality of the sequence alignment and template structure.

In the CA context, the heuristic approach seems to be the most relevant. the priority is given to the chemical properties of hydrophobic and hydrophilic amino acids. Therefore, if the amino acid is hydrophobic and is surrounded by water, it must change its position preferably towards other hydrophobic amino acids.

The following subsections discuss the possible use of simple rules and principles of chemistry and thermodynamics.

1) Chemistry: Usually the abstraction of the chemistry behind protein folding depends on the hydrophobic and hydrophilic amino acid properties. In the living cell, Ribosomes read the mRNA to produce the amino acid chain. After that, proteins are in an environment full of water (around 70% of the living cell), so it will spontaneously fold.

In the protein folding process, one can imagine the hydrophobic amino acids cluster in the core of the protein since those amino acids move away from the water in the environment. On the other hand, hydrophilic amino acids fold around this core as if they are trying to protect the hydrophobic amino acids.

Moreover, the interactions that stabilizes the protein can be added such as the salt bridge where positively charged side chains likes to be close to negatively charged side chains. The salt bridge is a combination of two non-covalent interactions, namely, hydrogen bonding and electrostatic interactions.

One more rule that can be added is the contribution of cysteine in folding. Similar to the case of salt bridges, cysteine plays an important role in stabilizing the protein because of the disulfide bridges.

2) Thermodynamics: Thermodynamics is the study of energy. The second law of thermodynamics states that in an isolated system, the total entropy always increases or remains the same but never decreases.

The measure of a molecule energy is Gibbs free energy (G), let the change in free energy be ΔG and the energy of the final state be G_f and the energy of the initial state be G_i, then ΔG is calculated as follows:

$$\Delta G = G_f - G_i \tag{1}$$

When $\Delta G<0$ the process goes from a high free energy state to a low free energy state which implies that the process is spontaneous and releases energy, so the process is a favored reaction and would happen if it could. On the other hand, if the $\Delta G>0$ the process is not spontaneous.

Gibbs energy takes into account the total energy or enthalpy (H), the total disorder or entropy (S), and the temperature (T), as shown in the following equation:

$$\Delta G = \Delta H - T\Delta S \tag{2}$$

Temperature plays a role in how much the entropy effects the change in (ΔG), if a process occurs in a high temperature environment then the entropy has a higher role in determining (ΔG) or how spontaneous the process occurs.

The role of thermodynamics laws in protein folding and stabilization is explained in [53]. In protein folding enthalpy changes from a high value in primary structures to a lower value in the 3D structures which makes ΔH negative, but entropy is also negative since it is higher in primary structures, so the temperature need to be low in order for the protein to fold. Thermodynamics are important since it is usually assumed that the protein's native state corresponds to its free energy minimum. This is tricky since it needs a global view and CA models work on the level of local rules. The point is that one should make the CA go towards this global energy model at each step in each cell.

D. General Steps

In this subsection, the proposed process of modeling proteins is summarized in some general steps. The input of the modeling process is the amino acid sequence of the target protein and the CA number of generations and the output of the algorithm is the final 3D CA representing the input protein. The process can be summarized as follows:

- Initialize the CA cells with the amino acid sequence and initialize the boundary with the water state
- Run CA rules for the number of generations
- at each step, if the spatial space needs extension then extend it
- Return the folded protein in the form of the input amino acids in their new positions in the 3D space

From an implementation point of view, it can be easier to define a 3D space that is double the size of the input chain so that the extensions are not needed often. The number of generations is assumed to be given in the input. Nevertheless, there are multiple methods that can be used to find an average number of generations that is suitable for the problem.

The output of the process shows the position of each Amino acid and how the hydrophobic core of the protein is created. The accurate results of the model means that CA is capable of modeling proteins without knowing the global view beforehand. It also means that proteins depend to some extent on some internal rules that results in its global behavior.

E. Yet More things to consider

Building a model that is based on CA starts with building a dataset, although protein information is generously available online, choosing the proteins or the benchmark needs to be accounted for. Many researchers depend on more than one benchmark to test their work. After applying the CA model, there must be a meaningful way to evaluate the model.

Although not using the CA model, some useful Literature compares between two or more distance matrices in terms of predicting protein attributes. For example, work done in [54] reports detailed results of protein structural classes prediction using Hamming, Euclidean, and Mahalanobis distances. Comparisons also exists in predicting protein subcellular Location [55].

VI. Conclusion and Future Work

The aim of many who work in the field of modeling natural phenomena is to add a step in discovering new things. Usually models try to capture the main properties and factors of the phenomenon to get results that are more meaningful in the sense of modeling one or more realistic attributes.

In this paper, protein modeling using cellular automata was discussed. Work in this area was analyzed and a suggested 3D model with heuristic rules was given. In addition, the general process of modeling proteins using cellular automata was discussed and alternative solutions to possible design issues were given.

The actual implementation of the design proposed in this paper is left for future work. In general, measuring the model effectiveness includes comparisons against data from laboratory experiments, but still the similarity between proteins is an interesting issue for further investigation.

References

[1] J. C. Wooley and H. S. Lin., *Catalyzing Inquiry at the Interface of Computing and Biology*. Washington (DC): National Academies Press (US), 2005, ch. Computational modeling and simulation as enablers for biological discovery, pp. 117–202.

[2] Y. Cai, J. Hu, Y. Li, and K. Chou, "Prediction of protein structural classes by a neural network method," *Internet Electronic Journal of Molecular Design*, vol. 1, no. 7, pp. 332–338, July 2002.

[3] H. Shen and K.-C. Chou, "Using optimized evidence-theoretic k-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 288–292, 2005.

[4] X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang, and K.-C. Chou, "Using complexity measure factor to predict protein subcellular location," *Amino Acids*, vol. 28, no. 1, pp. 57–61, 2005.

[5] X. Zhou, X. Li, M. Li, and X. Lu, "Predicting protein functional class with the weighted segmented pseudo-amino acid composition moment vector," *MATCH Communications in Mathematical and in Computer Chemistry*, vol. 66, no. 1, pp. 445–462, 2011.

[6] K.-C. Chou and H.-B. Shen, "Predicting protein subcellular location by fusing multiple classifiers," *Journal of Cellular Biochemistry*, vol. 99, no. 2, pp. 517–527, Oct 2006.

[7] P. Sarkar, "A brief history of cellular automata," *ACM Comput. Surv.*, vol. 32, no. 1, pp. 80–107, 2000.

[8] D. Whitford, *Proteins: Structure and Function*. Wiley, 2005.

[9] E. Ferrada, "The amino acid alphabet and the architecture of the protein sequence-structure map. i. binary alphabets," *PLOS Computational Biology*, vol. 10, no. 12, pp. 1–20, December 2014.

[10] S. Wolfram, *A New Kind of Science*. Wolfram Media Inc., 2002.

[11] C. Burks and D. Farmer, "Towards modeling dna sequences as automata," *Physica 10D*, vol. 10, no. 1-2, pp. 157–167, 1984.

[12] G. Sirakoulis, I. Karafyllidis, C. Mizas, V. Mardiris, A. Thanailakis, and P. Tsalides, "A cellular automaton model for the study of dna sequence evolution," *Computers in Biology and Medicine*, vol. 33, no. 5, pp. 439–453, 2003.

[13] C. Mizas, G. Sirakoulis, V. Mardiris, I. Karafyllidis, N. Glykos, and R. Sandaltzopoulos, "Reconstruction of dna sequences using genetic algorithms and cellular automata: Towards mutation prediction?" *Biosystems*, vol. 92, no. 1, pp. 61–68, 2008.

[14] J. A. de Sales, M. L. Martins, and D. A. Stariolo, "Cellular automata model for gene networks," *Phys. Rev. E*, vol. 55, pp. 3262–3270, Mar 1997.

[15] X. Xiao, S. Shao, Y. Ding, and X. Chen, "Digital coding for amino acid based on cellular automata," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 5, Oct 2004, pp. 4593–4598.

[16] X. Xiao, S. Shao, Y. Ding, Z. Huang, and K.-C. Chou, "Using cellular automata images and pseudo amino acid composition to predict protein subcellular location," *Amino Acids*, vol. 30, no. 1, pp. 49–54, 2006.

[17] X. Xiao, P. Wang, and K.-C. Chou, "Gpcr-ca: A cellular automaton image approach for predicting g-protein-coupled receptor functional classes." *Journal of Computational Chemistry*, vol. 30, no. 9, pp. 1414–1423, 2008.

[18] X. Xiao and W. Ling, "Using cellular automata images to predict protein structural classes," in *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*, July 2007, pp. 346–349.

[19] X. Xiao, P. Wang, and K.-C. Chou, "Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image," *Journal of Theoretical Biology*, vol. 254, no. 3, pp. 691–696, 2008.

[20] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *PROTEINS: Structure, Function, and Genetics*, vol. 43, pp. 246–255, 2001.

[21] ——, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[22] X. Xiao, P. Wang, and K.-C. Chou, "Cellular automata and its applications in protein bioinformatics," *Current Protein and Peptide Science*, vol. 12, no. 6, pp. 508–519, 2011.

[23] J. Santos, P. Villot, and M. Dieguez, "Cellular automata for modeling protein folding using the hp model," in *Evolutionary Computation (CEC), 2013 IEEE Congress on*, June 2013, pp. 1586–1593.

[24] J. Santos, P. Villot, and M. Diéguez, "Emergent protein folding modeled with evolved neural cellular automata using the 3d HP model," *Journal of Computational Biology*, vol. 21, no. 11, pp. 823–845, 2014.

[25] P. Chopra and A. Bender, "Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature," *In Silico Biology*, vol. 7, no. 7, pp. 87–93, 2006.

[26] D. Takata, T. Isokawa, N. Matsui, and F. Peper, "Modeling chemical reactions in protein synthesis by a brownian cellular automaton," in

2013 First International Symposium on Computing and Networking, Dec 2013, pp. 527–532.

[27] A. L. A. Dalhoum, A. Ortega, and M. Alfonseca, "Cellular automata equivalent to d0l systems," in *3rd WSEAS International Conference on Systems Theory and Scientific Computation, Special Session on Cellular Automata and Applications*, 2003, pp. 15–17.

[28] A. Ortega, A. A. Dalhoum, and M. Alfonseca, "Grammatical evolution to design fractal curves with a given dimension," *IBM Journal of Research and Development*, vol. 47, no. 4, pp. 483–493, 2003.

[29] G. B. Danks, S. Stepney, and L. S. D. Caves, *Folding Protein-Like Structures with Open L-Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1100–1109.

[30] G. Danks, S. Stepney, and L. Caves, "Protein folding with stochastic l-systems," *Artificial Life XI*, pp. 150–157, 2008.

[31] G. B. Danks, S. Stepney, and L. S. D. Caves, *Cotranslational Protein Folding with L-systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 289–296.

[32] M. Gibson and E. Mjolsness, *Computational modeling of genetic and biochemical networks*. Cambridge MA: MIT Press, 2004, vol. 8, no. 1, ch. Modeling the Activity of Single Genes, pp. 3–48.

[33] P. Koehl, *Protein Structure Classification*. John Wiley & Sons, Inc., 2006, pp. 1–55.

[34] D. E. Krane and M. L. Raymer, *Fundamental concepts of bioinformatics*. San Francisco, London, Paris: B. Cummings, 2003.

[35] T. von der Haar, "Mathematical and computational modelling of ribosomal movement and protein synthesis: an overview," *Computational and Structural Biotechnology Journal*, vol. 1, no. 1, pp. 1–7, 2012.

[36] K. Zuse, "Rechnender raum," *Elektronische Datenverarbeitung*, vol. 8, pp. 336–344, 1967.

[37] ——, *Rechnender Raum*. Friedrich Vieweg & Sohn, Braunschweig, 1969.

[38] S. Wolfram, "Statistical mechanics of cellular automata," *Rev. Mod. Phys.*, vol. 55, pp. 601–644, Jul 1983.

[39] ——, "Universality and complexity in cellular automata," *Physica D: Nonlinear Phenomena*, vol. 10, no. 12, pp. 1–35, 1984.

[40] C. G. Langton, "Computation at the edge of chaos: Phase transitions and emergent computation," *Phys. D*, vol. 42, no. 1-3, pp. 12–37, 1990.

[41] Z. Aleksic, "Artificial life: growing complex systems," in *Complex Systems*, T. R. J. Bossomaier and D. G. Green, Eds. Cambridge University Press, 2000, pp. 91–126, cambridge Books Online.

[42] R. Ye and H. Li, "A novel image scrambling and watermarking scheme based on cellular automata," in *Electronic Commerce and Security, 2008 International Symposium on*, Aug 2008, pp. 938–941.

[43] A. Madain, A. Abu Dalhoum, H. Hiary, A. Ortega, and M. Alfonseca, "Audio scrambling technique based on cellular automata," *Multimedia Tools and Applications*, vol. 71, no. 3, pp. 1803–1822, 2014.

[44] A. L. Abu Dalhoum, A. Madain, and H. Hiary, "Digital image scrambling based on elementary cellular automata," *Multimedia Tools and Applications*, pp. 1–16, 2015.

[45] M. Gardner, "Mathematical Games: The fantastic combinations of John Conway's new solitaire game "life"," *Scientific American*, vol. 223, pp. 120–123, 1970.

[46] L. Orellana, "Protein dynamics studied by coarse-grained and atomistic theoretical approaches," Ph.D. dissertation, University of Barcelona, Department of Fundamental Physics, 2014.

[47] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, and K.-C. Chou, "Using cellular automata to generate image representation for biological sequences," *Amino Acids*, vol. 28, no. 1, pp. 29–35, 2005.

[48] H. Fu and E. Mephu Nguifo, *Clustering Binary Codes to Express the Biochemical Properties of Amino Acids*. Boston, MA: Springer US, 2005, pp. 279–282.

[49] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, February 1962.

[50] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems Man and Cybernetics SMC3*, vol. 3, no. 6, pp. 610–621, 1973.

[51] H. Nishio, "How does the neighborhood affect the global behavior of cellular automata?" in *Cellular Automata*, ser. Lecture Notes in Computer Science, S. El Yacoubi, B. Chopard, and S. Bandini, Eds. Springer Berlin Heidelberg, 2006, vol. 4173, pp. 122–130.

[52] K.-C. Chou, "Structural bioinformatics and its impact to biomedical science," *Current Medicinal Chemistry*, vol. 11, no. 16, pp. 2105–2134, Aug 2004.

[53] A. Cooper, *Protein: A Comprehensive Treatise*. JAI Press Inc., 1999, vol. 2, ch. Thermodynamics of Protein Folding and Stability, pp. 217–270.

[54] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[55] K.-C. Chou and Y.-D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *The Journal of Biological Chemistry*, vol. 277, no. 48, pp. 45 765–45 769, November 2002.

Conditions Facilitating the Aversion of Unpopular Norms: An Agent-Based Simulation Study

Zoofishan Zareen
Bahria University
Islamabad, Pakistan

Muzna Zafar
Bahria University
Islamabad, Pakistan

Kashif Zia
Bahria University
Islamabad, Pakistan

***Abstract*—People mostly facilitate and manage their social lives adhering to the prevalent norms. There are some norms which are unpopular, yet people adhere to them. Ironically, people at individual level do not agree to these norms, but, they still follow and even facilitate them. Irrespective of the social and psychological reasons behind their persistence, sometimes, for societal good, it is necessary to oppose and possibly avert the unpopular norms. In this paper, we model theory-driven computational specifications of Emperor's Dilemma into an agent-based simulation, to understand the the conditions that result in emergence of unpopular norms. The reciprocal nature of persistence and aversion of norms, thus, is utilized to define situations under which these norms can be changed and averted. Simulation is performed under many interesting "what-if" questions. The simulation results reveal that under high density conditions of agent population with a high percentage of norm aversion activists, the aversion of unpopular norms can be achieved.**

***Keywords*—Agent-based Modeling and Simulation; Emperor's Dilemma; Complex Adaptive Systems.**

I. Introduction

Norms are regular principles of conduct that organize our communications with others. Norms plays an important role in the development of the individual behavior and the social order [1]. Since, norms affect individual and group behavior, emergence of behavior at societal level reform the norms. Hence, norms and their manifestation on the society are related to each other in a reciprocal manner, being influenced and influencing each other [2].

Intuitively, norms are supposed to be the conventions and rules which prevail in a society due to their goodness at the societal level. However, the factors influencing the emergence of such a collective thought are not homogeneous and fair in nature. Sometimes, the human factors of fear, greed, honor, influence, and belief etc., collectively manifest into a norm which is not really popular, but still prevails. While, an unpopular norm emerges, majority of the people at individual level do not agree with that, however, due to social pressure they have no option but to adhere to it. Surprisingly, they even become advocate of the norm due to potential embarrassment as a consequence of revealing their private thought to a group of people which they think, think otherwise. Some interesting examples quoted in the literature are foot binding [3], female genital mutilation [4], Johannes town massacre [5] and acceptability of corruption [6].

To reason about emergence of norms in a society, two predominant approaches are: economic and social approach. The economy-based approach resides on the basics of cost-benefit analysis, where each individual or group is required to be "somewhat" rational. For many, this cannot be true for a human society. Hence, the social approach focus on societal factors which effect the decision-making of individuals and groups, such as preferences, networking and interactions, and cooperation and externalities [7]. Naturally, the social approach is better to reason with unpopular norms.

A general description of using Agent-Based Modeling (ABM) as a bottom-up approach for modeling social interactions is presented in [8]. A more generic study with focus on cultural differentiation based on social factors of homophily, influence and network structure is presented in [9]. Similar studies exists throughout the research literature [10], [11], [12], [13], [14], [15].

The seminal work by Damon Centola, Robb Willer and Michael Macy [16] explains the social factors influencing the emergence of an unpopular norm. In this work, they have proposed a computational model of self-enforcing norms, stating the conditions necessary to enforce an unpopular norm. Additional work also exists modeling the compliance and/or enforcement of a an unpopular norm, covered in next section.

In this work, we argue that for societal good, it is necessary to oppose and possibly avert the unpopular norms. Hence, we attempted to realize the conditions that result in emergence of unpopular norms and define situations under which these norms can be changed and averted. To achieve it, we used and extended the social interaction model proposed by Centola, et.al. [16].

II. Related Work

There are two cognitive aspects that describe the human reaction to a situation related to norms. First is the *compliance*; the person confronted with such a situation may decide to comply to the norm or not. The second is the *enforcement*; the person confronted with such a situation may decide to enforce the norm or not. Related work exists in both directions. Additionally, the methodologies to do research on the topic (and many other related social phenomena) can broadly be categorized into four types: (i) purely theoretical, (ii) experimental (performing experiments on human subjects), (iii) mathematical and (iv) model-based (with agent-based models being most popular).

Irrespective of whether the norms are popular or not, Helbing, et. al. [17] presented an excellent description of conditions for the emergence of shared norms in populations with incompatible preferences. They have presented an agent-based model and simulated interesting scenarios to reach to following conclusions. First, effects of punishments on emergence of norms by defining behavior-based and preference-based punishment. Due to punishment there is probability of emergence of norms that is sometimes called unpopular norms. Individuals publicly follow the norms due to fear of punishment but privately oppose these norms. Second, is externalities needed for norm emergence. Externalities help in cooperation. Third, norms emerge due to social networks where everyone interacts. Agents in social networks communicate with each other and try to convince to comply or deviate from current norm. Fourth, path-dependent is important feature of norms. The paper concludes that the agents are influenced by neighbors and they set their preferences according to their neighbors preferences and neighbors current behavior to set a shared norms in a society.

In [18], authors argue that dependency and ease of access of information in the current technological age have given birth to propagation of false beliefs, named by them as "Infostorms". They blame three factors responsible for such untrue storms: *informational cascades*, *bystander effect / pluralistic ignorance* and *group polarization*. This paper gives a theoretical understanding of phenomena assisted by real-life examples. The authors conclude that technological systems developed with epistemic approaches with emphasis on truth tracking can prove to be beneficial to avert the infostorms.

Informational cascades are related with social interactions. These connections can be with ones own social network or some random network. Studying the evolution and emergence of norms influenced by such interactions is an interesting domain. For this purpose, the authors in [19] defined a weighted selection algorithm that determines the probability of a person to meet a stranger on the basis of individuals path distance. Using this algorithm, the paper elaborates four cases. In the first case, the agents take rational decision to opt for a norm based on highest utility. Whereas, in second case the agents use Markov decision process to select a norm by assigning weights. Third and fourth cases focus on examining the effect of social interaction on evolution of norms as it spreads throughout the masses. The simulation results showed that in first case people converge to a single norm but the second case takes lesser time in convergence of the norm. Defining the norm as an n-bit sequence indicated that increasing the random interaction had some adverse effects.

A logic-based approach to pluralistic ignorance is proposed in [20], [21]. Being one of the main reason of propagation of unpopular norms, it is important to define the pluralistic ignorance. It can be defined as a situation where "no one believes, but everyone believes that everyone else believes." The logic behind diffusion of unpopular norms is derived from theoretical understanding and intuition. However, the authors argue that the pluralistic ignorance is a fragile phenomenon such that a simple act of public announcement can suspend it. They further present the conditions of dissolution of pluralistic ignorance by stating that either all agents need to announce or an information from a trusted sources would help.

Beyond compliance, the enforcement of an unpopular norm is also evidenced. It has been observed that people enforce unpopular norm to which they privately disapprove. This paper [22] is based on discovering the reason of false enforcement. The authors are of the opinion that people enforce norms to create an illusion of sincerity rather than conviction. The study has been tested in two experiments of wine tasting and text evaluation. Both experiments reveal that the people who enforced the norm, against their actual belief, under social pressure criticized the deviants of the norm. These outcomes indicate how social pressure can lead to false enforcement of an unpopular norm.

On aversion of unpopular norms, the literature available is quite thin. But, there is a need to work in this direction as it is evidenced [23] that people often are not gratified with the norms that already persist in the society and they want to change it or terminate it. The authors relate the mechanism of change with the reasons of why people obey unpopular norms. First is the lack of accurate information about others' behaviors. Second is the herding effect (do as others do). And the third, is the panic punishment. In this paper researchers discussed that having accurate information through communication can avert the unpopular norms and can counter the first two reasons. Third reason is countered through time interval. Lower the time interval between discussion and outcome, the lower the probability of uncertainty leading to more chances to change unpopular norm. However, the study is experimental without any formal model.

We, in this paper, propose a model of aversion of unpopular norm. It is an agent-based model, facilitating the analysis of interesting scenarios in a systematic manner. The model's motivation comes from [16], in which, an agent-based computational model of Emperor's Dilemma is presented. The model resolves the conflict between compliance and enforcement of a unpopular norm, supported by few true believers and privately disapproved by a majority of disbelievers. The authors quantify the influence of networking and population distribution onto the diffusion of unpopular norms or otherwise. In our model presented in the next section, followed by Centola's original specifications, we focus on possibility of aversion of unpopular norms, introducing the reciprocal of behavior of true believers enforcing the unpopular norm unconditionally.

III. Models

A. Agent-Based Computational Model of Emperor's Dilemma [16]

In [16], authors state the Emperor's Dilemma as:

"Hans Christian Andersen ... tells the story of three rogues who sell a foolish monarch a nonexistent robe that they claim cannot be seen by those who are "unfit for office" or "incorrigibly stupid." Fear of exposure leads the emperor, and in turn, each of the citizens, to express admiration for the new clothes, which then reinforces the illusion of widespread support for the norm. The spell is broken when a child, innocent of the norm, laughs at the naked old man."

The agent-based computational model of Emperor's Dilemma, proposed in [16] formalizes the phenomena of spreading of unpopular norms through equations of *compliance*

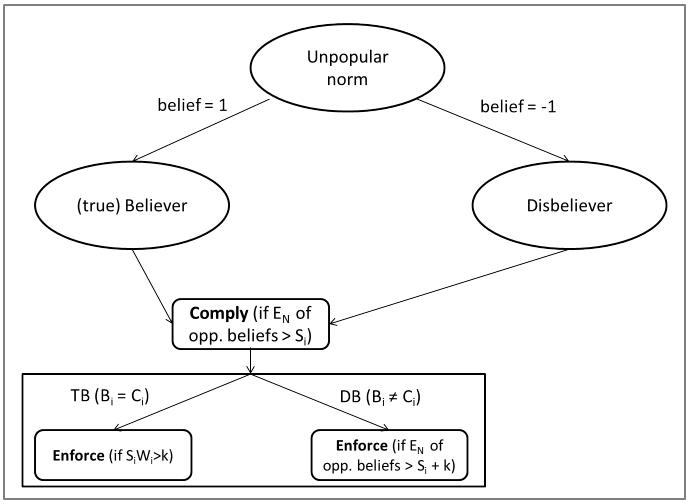


Fig. 1: Flow Diagram of the model based on Emperor's Dilemma [16].

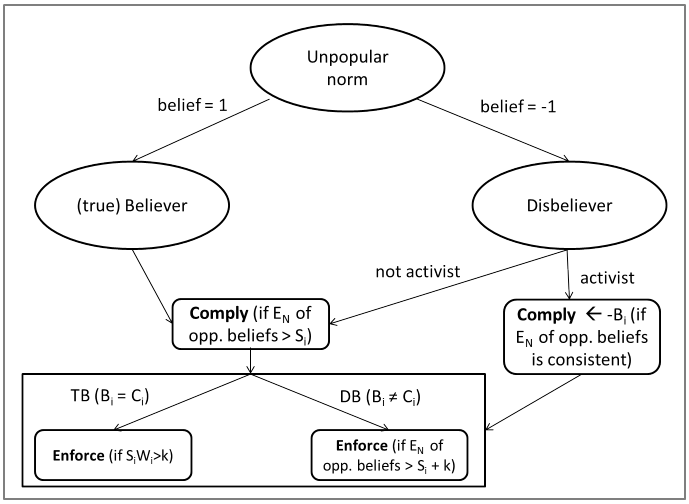


Fig. 2: Flow Diagram of the Proposed Model.

and *enforcement*. They model two behaviors for the agent; *compliance*, when agents comply to the norm; and *enforcement*, when agents, in addition to compliance, also enforce the norm in their region of influence. The factor of compliance transforms an agent either to comply with the norm or choose to deviate from it based on its personal belief, strength of the belief and the neighborhood dynamics. Once complying, an agent may choose to enforce the norm as well.

The model is based on two type of agents; true *believers* (TB) and *disbelievers* (DB). Believers are those agents who truly believe in the sanctity of the unpopular norm, while disbelievers are those agents who disbelieve (at least privately) in the sanctity of the the unpopular norm. Hence the *belief* of an agent on the truthfulness of an unpopular norm corresponds to its type; 1 for TB, and -1 for DB. The initial value of *compliance* of the norm is also set to 1 for TB, and -1 for DB. The *strength* of the belief corresponds to how strongly an agent believes what it believes; hence, the value is equal to 1 for TB and a low random fraction for DB.

In discrete time simulation, each agent *i* in the simulation space, performs the following actions:

- **Interact** with the neighbors and calculate the value of *Enforcement Need* (W_i) as:

$$Wi = \frac{1 - \left(\frac{Bi}{Ni}\right) \sum_{j=1}^{Ni} Cj}{2} \tag{1}$$

  where B_i is agent's belief, N_i is neighbors count and $\sum_{j=1}^{Ni} Cj$ represents the neighbors count whose compliance is not equal to i's belief.

- **Comply** with the norm if the value calculated below is equal to 1:

$$Ci = \begin{cases} -Bi & if \ \frac{-Bi}{Ni} \sum_{j=1}^{Ni} Ej > Si \\ Bi & otherwise \end{cases} \tag{2}$$

  i.e. compliance (C_i) is set to opposite of the belief, if strength of enforcement of opposite belief by the neighbors is greater than i's own strength; otherwise the C_i remains equivalent to agent's belief.

- **Enforce** the norm; whether it is true enforcement or false enforcement:

$$Ei = \begin{cases} -Bi \ if \ (\frac{-Bi}{Ni} \sum_{j=1}^{Ni} E\ j > (Si + k)) \bigwedge (Bi \neq Ci) \\ +Bi \qquad\qquad if \ \ (SiWi > k) \bigwedge (Bi = Ci) \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases} \tag{3}$$

  i.e. enforcement (E_i) is set to opposite of the belief, if strength of enforcement of opposite belief by the neighbors is greater than i's own strength plus a factor k and current belief is not equal to compliance; otherwise if belief is equal to compliance, and and product of S_i and W_i is greater than k, the E_i remains equivalent to agent's belief.

Figure. 1 depicts the transition diagram of this model. An agent whether TB or DB complies with the norm if the enforcements of agents with opposite belief in the neighborhood exceeds its strength. Since, the enforcement values range from -1 to 1, and we take an average (see equation (2)), for a TB, it can never be more than its strength (= 1). Hence, a TB would never be affected by the influence of the neighborhood even when all the neighbors are DBs and are not complying. For a DB, as a result of compliance, the value C_i changes to negation of its belief (= 1).

All agents who are complying would enforce the norm. For a TB, the value E_i would always be equal to B_i, i.e. 1, if the value of W_i is greater than k (i.e. in the neighborhood, there are sufficient non-complying DBs). The constant k determines the sensitivity of a TB for a need of assertion of norm. A TB would always be applying *true enforcement*. For a DB, the value E_i is equal to opposite of its belief (= 1), if there is sufficient enforcement pressure from the surrounding. A TB would always be applying *false enforcement*.

B. The Proposed Model Extension

As it is evidenced in the model presented above that a TB is not a normal agent; i.e., it would never be affected by whats happening in the surrounding. Our model is based on reciprocity of this behavior. It means that a minority of DBs are assumed to be more enthusiastic about averting the unpopular

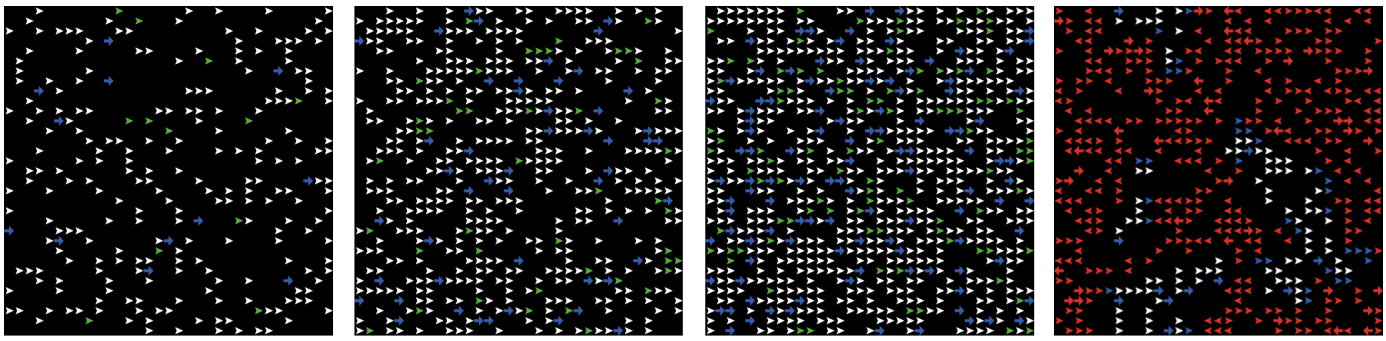


Fig. 3: Simulation Setup of 3 of 27 cases: agents in white are disbelievers, agents in blue are (true) believers, and agents in green are activists, in a (a) sparse population (250 agents) with *TBPerc* = 5% and *ACTPerc* = 5%, (b) medium dense population (500 agents) with *TBPerc* = 10% and *ACTPerc* = 10%, and (c) dense population (750 agents) with *TBPerc* = 15% and *ACTPerc* = 15%. The *ACTPerc* is a percentage out of disbelievers. The orientations of all agents are towards "right" initially. (d) Agents' state after end of the simulation of case 14: all agents in red are applying enforcements; even some true believers are applying enforcement to avert the norm.

norm. We name such an agent as an *activist* (ACT). Like a TB, an activist would never be affected by whats happening in the surrounding. Like TBs, they would be ambitious about fulfilling their role which is acting to avert the norm. The role with be triggered by presence of TBs in the surrounding, particularly who are enforcing.

A transition diagram of the extended model is shown in Figure. 2. An activist would change its belief from -1 to 1 after being encountered by enforcements of norms from the neighborhood. This is achieved by progressive increment of the value of S_i by a constant k. If this value reaches to 1 or greater, the belief of the agent is changed from -1 to 1, which means that now the agent believes in *aversion* of the unpopular norm and acts to avert it. The exact formulation is given in equation (4).

$$Ci = \begin{cases} -Bi & if\ Si > 1 \\ Bi & otherwise \end{cases} \tag{4}$$

Now an activist acts like a true believer and enforces the norm unconditionally (because the value of S_i is sufficiently hight already). Since this enforcement will be in the opposite direction, it would be aversion of unpopular norm. Even the TBs are susceptible to be enforced as we will see in the analysis section.

IV. Simulation and Results Discussion

A. Simulation Setup

The simulation is performed in Netlogo [24], a popular agent-based simulation tool with grid space support. Netlogo provides an automatic mechanism of applying a function on a set of agents in a random order in successive iterations. Hence a fairness between agents is ensured even though they use the current state of required agents (e.g. the neighborhood of an agent) while making a decision. The agents reside on cells of a spatial grid. We have used Moore's neighborhood to represent the surrounding of an agent which has been a popular strategy in many cell-based spatial configurations [25].

A specific simulation setting is represented as a simulation *case*, which corresponds to *density* of the population of agents, percentage of TBs in population *TBPerc*, and percentage of ACTs in population *ACTPerc*. For visual assistance, we represent the compliance of unpopular norm with a directional clue. Hence, the agents averting the unpopular norm have a completely opposite direction. The agent type and state is also visually differentiated. This is further explained in Figure. 3. A complete list of simulation cases is given in Table. I. Each simulation case was run for 10 times and the results were averaged.

B. Simulation Results

The simulation results are analyzed based on four quantities:

- DBAvertPerc: The percentage of disbelievers who ended up averting the unpopular norm.
- DBComplPerc: The percentage of disbelievers who ended up complying the unpopular norm.
- TBAvertPerc: The percentage of true believers who ended up averting the unpopular norm.
- TBComplPerc: The percentage of true believers who ended up complying the unpopular norm.

The Table. I shows the 27 simulation cases and the corresponding values for above measures. Qualitatively the norms is considered as been averted if majority of the population starts refuting it (acting against it). The result verify a trend that can be stated as:

- The percentage of agents averting the unpopular norm increases with increase in population.
- The percentage of agents averting the unpopular norm increases with increase in ACTPerc.
- The percentage of agents averting the unpopular norm increases with increase in relative positive difference between ACTPerc and TBPerc.

TABLE I: Complete list of all 27 simulation cases.

Case No.	Population	*TBPerc*	*ACTPerc*	DBAvertPerc	DBComplPerc	TBAvertPerc	TBComplPerc
1	250	5	5	11.39	88.61	8.33	91.67
2	250	5	10	24.05	75.95	16.67	83.33
3	250	5	15	32.49	67.51	16.67	83.33
4	250	10	5	9.33	90.67	8	92
5	250	10	10	21.33	78.67	12	88
6	250	10	15	28.89	71.11	24	76
7	250	15	5	11.79	88.21	5.41	94.59
8	250	15	10	20.28	79.72	13.51	86.49
9	250	15	15	33.96	66.04	29.73	70.27
10	500	5	5	17.47	82.53	8	92
11	500	5	10	31.79	68.21	28	72
12	500	5	15	48	52	32	68
13	500	10	5	14.89	85.11	4	96
14	500	10	10	31.78	68.22	24	76
15	500	10	15	48.67	51.33	42	58
16	500	15	5	14.35	85.65	13.33	86.67
17	500	15	10	33.41	66.59	24	76
18	500	15	15	42.12	57.88	33.33	66.67
19	750	5	5	23.88	76.12	21.62	78.38
20	750	5	10	47.33	52.67	43.24	56.76
21	750	5	15	63.48	36.52	51.35	48.65
22	750	10	5	26.37	73.63	28	72
23	750	10	10	46.07	53.93	38.67	61.33
24	750	10	15	64	36	58.67	41.33
25	750	15	5	26.37	73.63	18.75	81.25
26	750	15	10	49.92	50.08	46.43	53.57
27	750	15	15	61.07	38.93	45.54	54.46

- The aversion is not only experienced by disbelievers but also believers.

A visual representation of simulation of case 14 is given in Figure. 3 (d). Next, we analyze three interesting cases, resulting in norm aversion qualitatively. It has been observed that case 21 & 24 display aversion of unpopular norm whereas we can see partial aversion in case 27.

In case 21, we observe that the unpopular norm is averted at approximately 7.37 ticks. The graph in Figure. 4 shows that the disbelievers enforcing alternative norm (other than the unpopular norm), increasing afterwards, averting the unpopular norm. In addition to this, we also observe that the number of true believers enforcing the unpopular norm gradually decline up to tick 9.55, and start averting the norm afterwards.

In case 24, we observe that the unpopular norm is averted more quickly, at 6.75 ticks. The graph in Figure. 5 shows that the disbelievers enforcing alternative norm (other than the unpopular norm) increase afterwards, averting the unpopular norm. In addition to this we also observe that the number of true believers enforcing the unpopular norm gradually decline up to tick 6, and start averting the norm afterwards.

In case 27, we observe that the unpopular norm is averted more quickly, at 6.74 ticks. The graph in Figure. 6 shows that the disbelievers enforcing alternative norm (other than the unpopular norm) increase afterwards, averting the unpopular norm. Contrary to this, we observe that the number of true believers enforcing the unpopular norm gradually decline but does not fall below the percentage of false believers thereby indicating that unpopular norm is only partially averted as shown in Figure. 6.

V. Conclusion

In this paper, an agent-based simulation model of unpopular norm aversion is presented. We have modeled a theory-driven computational specifications of Emperor's Dilemma into an agent-based simulation, to understand the conditions that result in emergence of unpopular norms. The reciprocal nature of persistence and aversion of norms is utilized to define situations under which these norms can be changed and averted. Following is concluded from the analyzes of the simulation results. The percentage of agents averting the unpopular norm increases with increase in population. Further, the percentage of agents averting the unpopular norm increases with increase in agents actively participating in averting the unpopular norm.

References

[1] H. P. Young *et al.*, *Social norms*. Department of Economics, University of Oxford, 2007.

[2] D. Kübler, "On the regulation of social norms," *Journal of Law, Economics, and Organization*, vol. 17, no. 2, pp. 449–476, 2001.

[3] C. F. Blake, "Foot-binding in neo-confucian china and the appropriation of female labor," *Signs*, vol. 19, no. 3, pp. 676–712, 1994.

[4] B. Essén and S. Johnsdotter, "Female genital mutilation in the west: traditional circumcision versus genital cosmetic surgery," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 83, no. 7, pp. 611–613, 2004.

[5] M. M. Maaga, *Hearing the voices of Jonestown*. Syracuse University Press, 1998.

[6] C. Bicchieri and Y. Fukui, "The great illusion: Ignorance, informational cascades, and the persistence of unpopular norms," *Business Ethics Quarterly*, vol. 9, no. 01, pp. 127–155, 1999.

[7] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[8] M. Macy and A. Flache, "Social dynamics from the bottom up: Agent-based models of social interaction," *The Oxford handbook of analytical sociology*, pp. 245–268, 2009.

[9] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. San Miguel, "Homophily, cultural drift, and the co-evolution of cultural groups," *Journal of Conflict Resolution*, vol. 51, no. 6, pp. 905–929, 2007.

[10] D. Centola and M. Macy, "Complex contagions and the weakness of long ties1," *American journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007.

[11] M. Michaeli and D. Spiro, "Inverted preferences and skewed norm," *Journal of Philosophical Logic*, 2013.

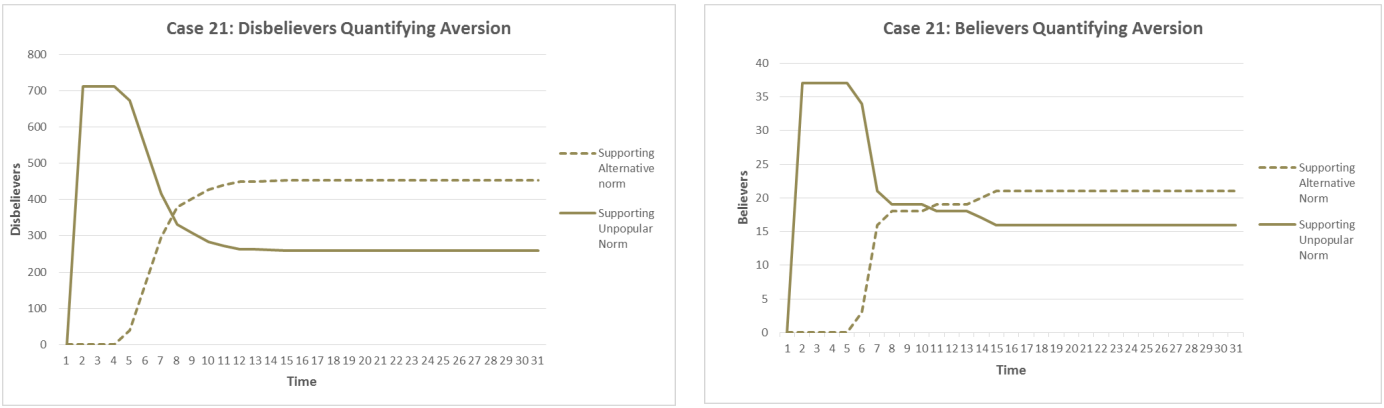


Fig. 4: Time-line of Case 21.

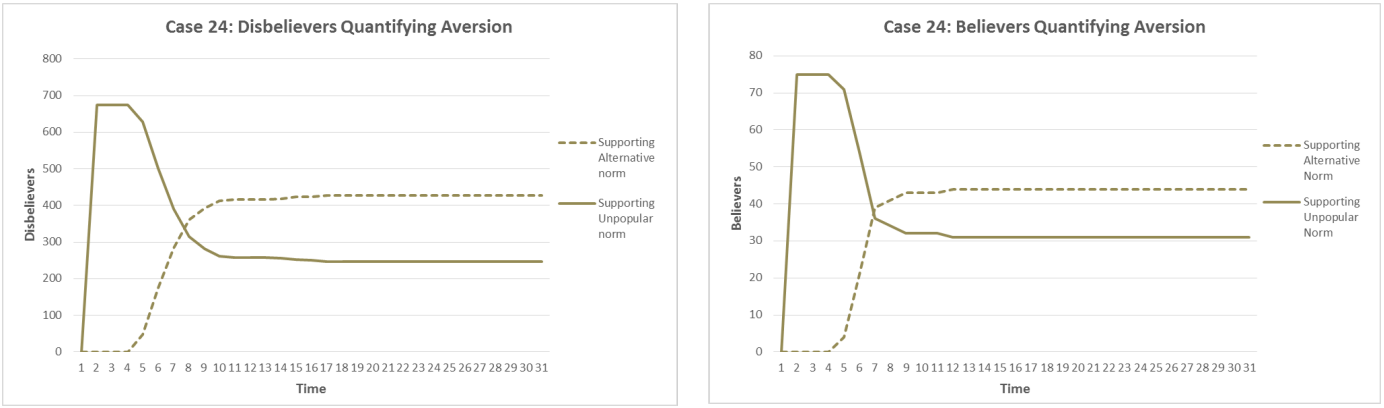


Fig. 5: Time-line of Case 24.

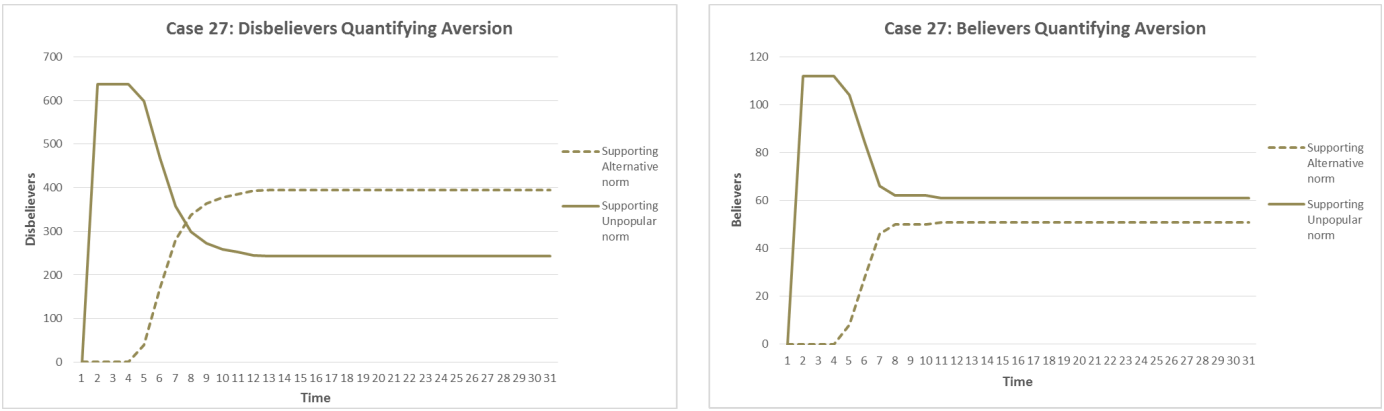


Fig. 6: Time-line of Case 27.

[12] M. Granovetter, "Threshold models of collective behavior," *The American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[13] A. C. Martins, "Trust in the coda model: Opinion dynamics and the reliability of other agents," *Physics Letters A*, vol. 377, no. 37, p. 23332339, 2013.

[14] L. Muchnik, S. Aral, and S. J. Taylor, "Social influence bias: A randomized experiment," *Science*, vol. 341, no. 6146, pp. 647–651, 2013.

[15] M. A. Pachucki and R. L. Breiger, "Cultural holes: Beyond relationality in social networks and culture," *Annual review of sociology*, vol. 36, pp. 205–224, 2010.

[16] D. Centola, R. Willer, and M. Macy, "The emperors dilemma: A computational model of self-enforcing norms1," *American Journal of Sociology*, vol. 110, no. 4, pp. 1009–1040, 2005.

[17] D. Helbing, W. Yu, K.-D. Opp, and H. Rauhut, "Conditions for the emergence of shared norms in populations with incompatible preferences," *Plos One*, vol. 9, no. 8, pp. 1–14, 2014.

[18] P. G. Hansen, V. F. Hendricks, and R. K. Rendsvig, "Infostorms," *Metaphilosophy*, vol. 44, no. 3, pp. 301–326, 2013.

[19] D. Mungovan, E. Howley, and J. Duggan, "The influence of random in-

teractions and decision heuristics on norm evolution in social networks," *Computational and Mathematical Organization Theory*, vol. 17, no. 2, pp. 152–178, 2011.

[20] J. U. Hansen, "A logic-based approach to pluralistic ignorance," in *Logic and Interactive RAtionality Yearbook*, vol. 2. The Institute for Logic, Lanuage and Computation, 2014, pp. 226–245.

[21] C. Proietti and E. J. Olsson, "A ddl approach to pluralistic ignorance and collective belief," *Journal of Philosophical Logic*, vol. 43, no. 2, pp. 499–515, 2014.

[22] R. Willer, K. Kuwabara, and M. W. Macy, "The false enforcement of unpopular norms," *American Journal of Sociology*, vol. 115, no. 2, pp. 451–490, 2009.

[23] K. Gerxhani and J. Bruggeman, "Time lag and communication in changing unpopular norms," *Plos One*, pp. 1–17, 2015.

[24] U. Wilensky, "{NetLogo}," 1999.

[25] A. Ferscha and K. Zia, "Lifebelt: Silent directional guidance for crowd evacuation," in *Wearable Computers, 2009. ISWC'09. International Symposium on*. IEEE, 2009, pp. 19–26.

Developing a Real-Time Web Questionnaire System for Interactive Presentations

Yusuke Niwa, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani
Department of Computer Science,
Graduate School of Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555 Japan

***Abstract*—Conducting presentations with bi-directional communication requires extended presentation systems, e.g., having sophisticated expressions and gathering real-time feedback. We aim to develop an interactive presentation system to enhance presentations with bi-directional communication during presentations. We developed a hybrid interactive presentation system that is a collaboration between the traditional presentation supporting system, e.g. PowerPoint, and a web application. To gather feedback from audiences at presentations, the web application delivers presentation slides to audiences. The client system provides a feature of creating annotations and answering the questions on delivered presentation slides for making feedback. Specifically, the system provides a real-time questionnaire function where the result is displayed on a shared screen in real time while gathering answers. Since users can make their questionnaire on PowerPoint, the task becomes quite easy. This paper explains the development of the system and demonstrates that the real-time questionnaire system realizes high performance scalability.**

Keywords*—*Interactive Presentation; Real-time Web questionnaire; Collaborative tools; communication aids; information sharing; Web services

I. Introduction

In an interactive presentation, a lecturer gives a presentation for the audience and the audience provides feedback about the presentation slide by slide. We proposed a hybird interaction presentation system [1], [2], [3] that consists of three subsystems: a presenter system, audience system and back-end system. The presenter system utilizes an existing presentation software PowerPoint to display presentation slide content and extends the presentation software to interactive presentation functions. The audience system is a web application for collecting comments from the audience and shows the same slide image with the presenter's display. In this paper, we describe the implementation of the real-time questionnaire function and evaluate the performance.

We implemented new functions for an easy-to-use and quick feedback system on web browsers from the audience to realize the interactive presentation. The existing presentation softwares (e.g. Microsoft PowerPoint and Apple Keynote) and the presentation web applications (e.g. Google Drive, iCloud, Office 365) do not support the functions for the interactive presentation. We attempt to implement the functionalities as a co-application with PowerPoint for easy deployment. The real-time questionnaire function uses existing presentation software functions to display the questionnaire result as a graph, pie chart or bar graph.

The hybrid interactive presentation system provides the real-time questionnaire function to count votes and show the result in the presentation software slide show mode and the synchronizing slide content function to display same slide with the presenter screen. The questionnaire function has a crucial part in the interactive presentations to attract the audience and increase the depth of understanding of presentations. In a decision-making scene, the questionnaire function is utilized as a supporting tool for making a decision.

The paper is organized as follows. In Section II, we discuss the interactive presentation. In Section III, we explain a design goal of the real-time questionnaire function. In Section IV, we explain a system for interactive presentation system. In Section V, we describe the implementation of the real-time questionnaire. In Section VI, we show the evaluation of the real-time questionnaire function. In Section VII, we discuss the evaluation and the performance. In Section VIII, we conclude the paper and summarize the effects of real-time voting.

II. Interactive Presentation

The interactive presentation plays a key role in education. Clicker Assessment and Feedback (CAF) is an instructional assessment and feedback strategy that is incorporated with interactive technologies for higher education, often referred to as clickers. CAF is important as one use for the interactive presentation. Wireless systems that enable professors to ask questions and have students respond using hand-held devices (clickers) are proposed in existing studies [4], [5]. Existing studies of CAF confirm the efficacy of the interactive functionality at presentations in education, but the hand-held devices for gathering feedback from students are a hindrance in a large class.

In recent years, the software clickers have been proposed to avoid the trouble of having to install specific physical devices. Hauswirth et al. proposed a software clicker to teach Java programming for higher education, which allows for much richer types of problems than the traditional multiple-choice questions [6]. The burden of installing devices is certainly removed with the use of the software clickers, but the approach requires setting applications for collecting opinions of audiences before each presentation. The approach Triglianos et al. proposed resolves the problem via an interactive presentation system such as a web application [7]. The proposed method lacks flexibility of designing presentations provided in the conventional applications for supporting presentations.

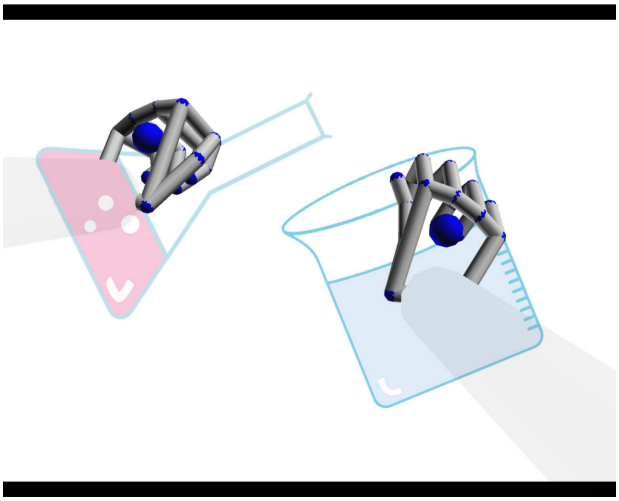


Fig. 1: Presenter can manipulate slide objects with motion controllers during an ongoing presentation.

When lecturers deliver presentations in an unexpected context, i.e., the situation such as the place, knowledge of audiences and questions is unexpected, designing presentation slides and giving a presentation are challenging. To achieve flexible presentations for solving issues in presentations, tangible interfaces based on direct interaction with digitally encoded printed handouts are proposed in previous studies [8], [9].

Our ongoing efforts at Silhouette Effects are dedicated to scaling conventional presentation tools to instant manipulations of slide objects at the presentation mode [10]. Fig. 1 shows that a presenter using Silhouette Effects can manipulate slide objects with motion controllers duraing an ongoing presentation. While Silhouette Effects supports only the features for delivering presentations, in our work we target the hybrid interactive presentation system that is intended to improve the presentations by having interaction between lecturers and audiences without loss of the flexible presentation functionality. We demonstrate a real-time questionnaire system that is based on the hybrid interactive presentation system and gathers feedback from audiences in presentations with the HTML5 web technologies.

III. Real-Time Questionnaire for Interactive Presentations

Our system has a real-time questionnaire function that enables a lecturer to collect what audiences are thinking about the presentation without questionnaire papers. In the section, we define the requirements and the required functions to realize the real-time questionnaire.

A. Requirements

We aim to offer a simple method to have the real-time questionnaire, i.e., the type of the questionnaire in which the result is displayed on a shared screen in real time during the implementation of presentations. The real-time questionnaire in presentations helps to maintain focus on presentations and to comprehend and increase the depth of understanding of presentations. In conventional research, the specific devices are needed to gather the answers from audiences. Our approach requires that the answers are collected using the web application and then simplifies the operation of performing the real-time questionnaire.

Our objective is to have our system provide seamless questionnaires in the interactive presentation easily. Therefore, the real-time questionnaire system should be independent of troublesome specific coded communication channels and authentication mechanisms for ensuring safety and anonymity in questionnaires. We take no thought of challenges of the web questionnaire system, such as cheating via multiple responses and assurance of anonymity. Even if the problems are out of consideration, our approach has the efficacy of decision-making at a certain scale of meetings in enterprises and organizations. As you perform the questionnaires you want to consider problems of safety; thus, we implement a new questionnaire mode for tackling the problems with existing approaches [11].

We suppose that the real-time questionnaire is used in a situation wherein a lecturer and audiences are in the same location and there are dozens or hundreds of audience members. Hence, we consider the scalability of the real-time questionnaire because an increase in audience members leads to network congestion and high processing time, that is, reason for a delay in displaying the results of the questionnaires in real time. Moreover, because of the requirements of the specific devices expressed in the approaches of the previous studies, the scale of the effort to carry out questionnaires is vast in cases where the number of audience members is large, wherein the system does not scale well.

B. Design Goal

The critical goal of our study is to make it an open possibility to provide the real-time questionnaire at the interactive presentation easily when a lecturer and audiences are in the same location. We set the following concrete objectives:

- Interoperability: The approach should be independent of special devices and applications for collecting opinions of audiences, i.e., audiences should answer the questionnaires using the personal devices without reference to the type and OS of devices.
- Interactive Functionality: A lecturer and audiences should share presentation slides and the results of the questionnaires in real time when delivering a presentation. Audiences should send feedback for the questionnaires at presentations to improve the communication among participants during the real-time questionnaire.
- Scalability: The approach should absolutely perform summary of a questionnaire and display the results in real time in cases where there is a large audience at a presentation.

To attain interoperability, we implement a user interface for audiences such as the web application. The specific devices and applications for collecting comments from audiences are not required, and personal devices such as notebooks, smart phones and tablet computers are used in our system.

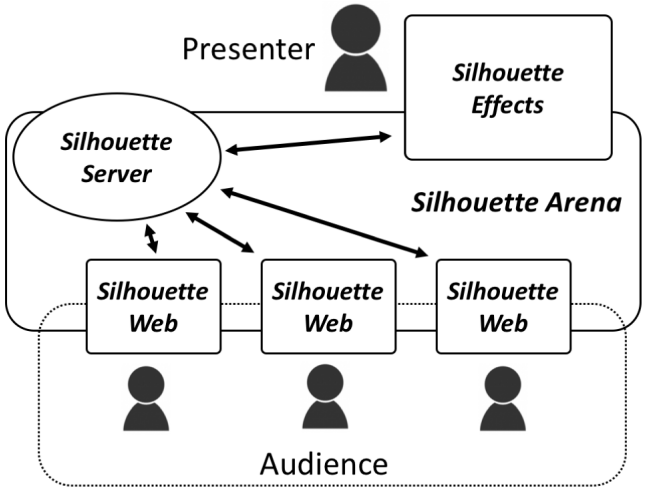


Fig. 2: Silhouette Arena: Hybrid Interactive Presentation System.

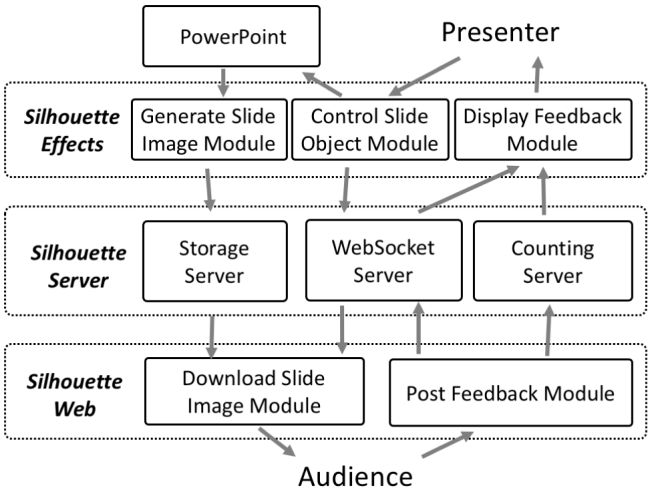


Fig. 3: Architecture of Silhouette Arena.

Our system supports interactive functionality that ensures communication between a lecturer and audiences during the real-time questionnaire. The results of the questionnaires are displayed on a shared screen while submitting answers from audiences using the graph provided by Microsoft Excel. When updating the result on the presentation slide in accordance with the received answers, audiences share presentation slides in real time using the web application operating with general web browsers. Concretely, audiences submit handwritten memos and text comments as feedback for presentations. Submitted handwritten memo and comments are sent to the device of a lecturer and reflected on presentation slides. The communication mechanism can be crucial to improvement of the consensus of opinions. Our system supports real-time information presented for lecturers and the features of feedback posting from audience members.

The approach of the web application contributes to the scalability of our system by reason of independence from the specific devices for gathering feedback. Besides, to avoid the network congestion and the high processing time, we design the mechanism of the interactive presentation that is based on the tallying server to handle the crush of demand. To confirm the efficacy of the mechanism, we performed an experiment of the time delay of sharing the results of the questionnaires among a lecturer and audience members.

IV. Silhouette Arena: System for Hybrid Interactive Presentation

A. Architecture

The architecture of the hybrid interactive presentation system, called Silhouette Arena, is shown in Fig. 2. The Silhouette Arena consists of three main parts: the Silhouette Effects, the Silhouette Web, and the Silhouette Server in Fig. 2.

The Silhouette Effects controls a PowerPoint to share presentation slides with clients called Silhouette Web. The Silhouette Effects provides the presentation with supporting features, e.g., sending the information of shared presentation slides. Moreover, the Silhouette Effects enables users to manipulate slide objects. The slide objects, such as text and figures, can be moved, zoomed, modified and removed using the presentation controller at the presentation mode of PowerPoint. Furthermore, lecturers can start a questionnaire at any time in presentations. While audiences submit the answers of the questionnaire, the result of the questionnaire is visualized using Excel graphs in real time on a presentation slide.

The Silhouette Web provides the user interface for displaying presentation slides and adding annotations and handwritten memos on shared slides in order to share presentation slides and submit opinions of the audience. The web application consists of two layers, i.e., a layer to recreate presentation slides and to share feedback from audiences in real-time. The first layer recreates presentations made of presentation slides that are converted into images. The approach is known to perform reliably and accurately since converting presentation slides to images was commonly achieved by web conferences in previous research [12]. On the second layer on the front of the web application, audiences generate annotations and handwritten memos that are composed of DOM elements for sending opinions and feedback to a lecturer.

Fig. 3 shows the architecture of the system. The Silhouette Effects communicates with the Silhouette Web via the Silhouette Server.

The Silhouette Server of managing the interactive presentation is divided into three internal subsystems. The first subsystem is a storage server keeping presentation slides that have been converted into images. When lecturers start presentations or manipulate the slide objects in presentations, the presentation controller converts the presentation slides into images and uploads the images to the storage server. The function of the storage server is to issue the URL of each uploaded image.

The Silhouette Effects sends the URLs of the images to the second subsystem, which is a WebSocket server for real-time communication among lecturers and audiences. The

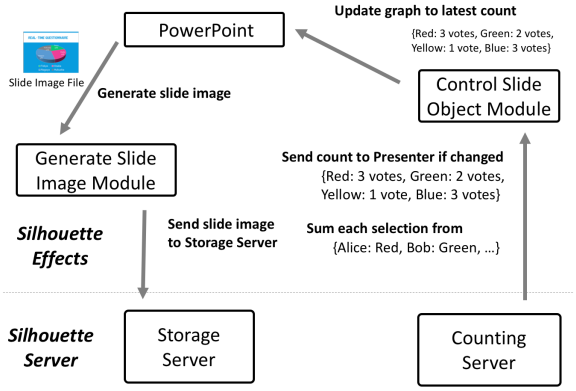


Fig. 4: Control Flow of Updating Pie Chart Graph.

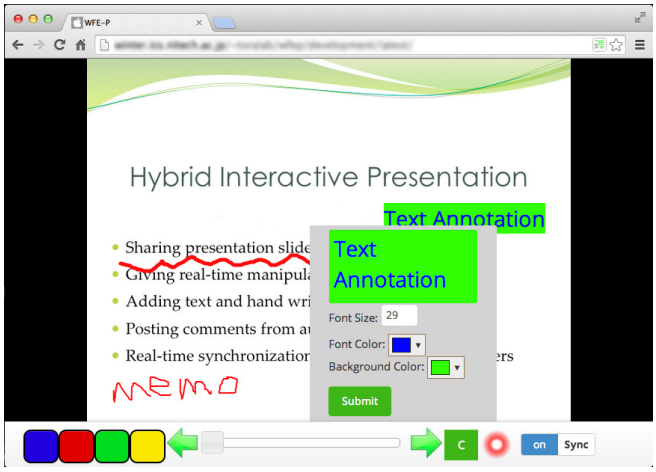


Fig. 5: Silhouette Web: Audience can make feedbacks, text annotations and freehand drawings on the Web application.

WebSocket server pushes received data in real time. The presentation controller and the web application communicate through the WebSocket server.

The Silhouette Web serializes and sends the created annotations and handwritten memo to the WebSocket server of the Silhouette Server. The Silhouette Effects deserializes the received data and generates the slide objects. Then, the feedback from audiences is displayed on the presentation slides in real time.

The third subsystem is the counting server that tallies answers of the real-time questionnaire. The count process consists of the following three steps. First, the system notifies an issue of a questionnaire by the lecturer to the client web application via the WebSocket server, then the web applications launches a questionnaire mode and presents voting answers. Second, the clients send the answer selected by their users to the counting server, and then the server tallies the answers and saves the results to the storage. Finally, the counting server sends the results of the questionnaires to the presentation controller at regular intervals.

Fig. 4 presents how to update a pie chart graph with PowerPoint material in the Silhouette Arena in order to prevent unexpected crushes of the PowerPoint system. The most important part is the counting server in Fig. 4. The counting server keeps pace by updating the pie chart without crashes of the PowerPoint. Then the counting server receives voting data from Silhouette Web clients, collects the votes and sends the result to the Control Slide Object Module periodically. When the Control Slide Object Module receives the voting result, it operates the PowerPoint to update the graph, and it sends an image file of the slide to the Silhouette Server to distribute the image to the clients.

B. Silhouette Effects: System for Enhancing Prsentation Expressions

Our approach supports the lecturers to give a presentation effectively with the proposed method of Silhouette Effects. There is a designing mode that modifies slide objects and a presentation mode that gives a presentation in Microsoft PowerPoint, and normally the users cannot edit slide objects at the presentation mode. Silhouette Effects expands the methods for utilization of the presentation mode. When delivering a presentation, running our system together with PowerPoint allows the users to modify slide objects at the presentation mode.

The proposed method for manipulating objects on slides in real time realizes the reactive presentation. Manipulating slide objects such as moving, zooming and adding some visual effects allows the lectures to guide and keep the audience's attention, as well as to communicate information in visual space through representing the animations and effects of slide objects reactively. Moreover, Silhouette Effects provides presentation-supporting functions except for manipulations of slide objects, that is, the handwritten memo and the pointing feature are available in the presentation mode.

The lecturers perform the presentation-supporting functions with input operations from a mouse and a keyboard or gesture inputs using Leap Motion and Kinect. Since Leap Motion and Kinect have an advantage over the typical input devices such as a mouse and a keyboard, our system corresponds with the gesture input interface. Leap Motion supports very precise finger tracking, and then we construct a highly intuitive interface by using Leap Motion. Additionally, Kinect has high user mobility that supports skeleton tracking in the wide recognizable area. Therefore, by using Kinect, we construct the interface with high user mobility and interact through the movement of users' whole body. The gesture input interface to manipulate slide objects supports the gestures bound to mundane actions such as the action of grasping something a person would hold, and the gestures are widely known such as Swipe and Pinch-in/out. Thus, the gesture input interface using Leap Motion and Kinect achieves seamless input operations for the presentation-supporting functions.

C. Silhouette Web: Feedback System for Audience

Fig. 5 shows the interface of the web application for submitting feedback from audiences, called Silhouette Web. To participate in the interactive presentation, audiences open

the web application. The web application runs on common web browsers and devices to achieve the interoperability necessary for audiences to send feedback with respective devices. Since the web application assumes a crucial role in the interactive presentation, we describe the user interface in detail.

Audiences need to enter an ID in the window when opening the web application to get into the interactive presentation. The counting server in the real-time questionnaire that gathers answers from audiences and tallies the real-time questionnaire manages answers from audiences with the IDs. The web application associates the answers of the questionnaires with the entered ID and sends them to the counting server. Then, the counting server keeps a correct tally in cases of receiving multiple answers from the same web application. The ID is also sent to the presentation controller through the WebSocket server when entering the ID. The presentation controller enumerates the number of audience members that submits the ID. The number of audience members is displayed on the screen in presentations for the duration of the questionnaire. If the number of entrants of the questionnaire is apparent, the lecturer compares the actual number of entrants with the displayed number, thus helping the lecturer to determine whether there is a repeater that votes answers with multiple web applications or not.

The lower bar of the screen has the four-color buttons for answering questionnaires: the controller of page transitions, the comment-posting button, the mode switch and the synchronization switch. As audiences give a reply to the questionnaire using the four-color buttons, the web application sends the answer to the counting server. The dialogue for inputting comments is displayed when clicking the comment-posting button. The comments are sent to the presentation controller on a device of a lecturer and saved in presentation slides. After presentations, the lecturers check submitted comments to assess the audience's feedback. The users define the behavior of pressing and holding the mouse button and dragging by the mode switcher, i.e., users select the pointer mode or the handwritten mode. The synchronization on/off button configures whether other users' manipulations are synchronized in real time or not. When the synchronization is off, the manipulations of a lecturer and other audiences are not synchronized. The button allows the users to genuinely browse the presentation slides.

When audiences press-and-hold the mouse button and drag, a pointer or handwritten memo is displayed on a presentation slide, and audiences indicate a slide object when delivering feedback. Right clicking on presentation slides displays the dialogue for adding an annotation like Fig. 5. The annotation consists of DOM elements that have a handler of the user's operation, so the user edits and drags the created annotation.

V. Implementation of Real-Time Questionnaire

A. Mechanism of The Real-Time Questionnaire

The hybrid interactive presentation shares presentation slides, manipulations in presentations and feedback from audiences in real time. The presentation controller on the device of the lecturer and the web application for audiences communicate through the intermediately of the counting server in the real-time questionnaire.

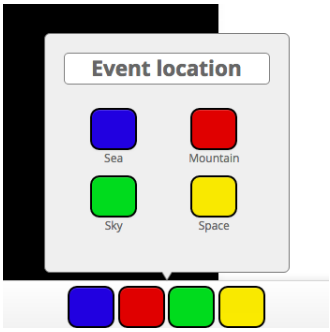


Fig. 6: Vote interface.

The web application sends the answers of questionnaires as JSON data that contain the color selected using the four-color button, the ID for identifying the users and a token for identifying the questionnaires to the counting server. The counting server deserializes the received data and tallies the selected answers with respect to the token.

The result is sent to the presentation controller at an interval of two seconds. Although the counting is accomplished in the way that the web application sends answers to the presentation controller directly and the presentation controller compiles the total amount, the device of the lecturer and the network have a heavy processing load. The heavy processing load can be a hindrance to the interactive presentation. Our approach reduces the load by counting results of questionnaires in the external server and sending the results to the lecturer's device at intervals of two seconds. We consider the chances of continuing to change an answer at intervals of less than two seconds are low. Thus, our approach provides real-time information about the questionnaires.

Using the synchronization mechanism of presentation slides as follows, the results of the questionnaires on presentation slides are shared in real time on the web application for audiences. When receiving a questionnaire result, the presentation controller updates a graph for displaying the result using Excel. Then, the presentation controller uploads the image of the updated presentation slide and broadcasts the image URL as JSON data to the web application through the WebSocket server. After the web application updates the image on the first layer that displays presentation slides with the received image, the result of the questionnaire is shared in real-time.

B. Real-Time Web Questionnaire Protocol

The interactive presentation provides the real-time questionnaire system that enables a lecturer to publish an electrical questionnaire to the audience members and collect the answer during the presentation. In this section, we describe a messaging protocol for the real-time questionnaire system.

A questionnaire session management in the real-time questionnaire system consists of six steps: announcement, opening, assignment, waiting, closing and cleanup. At the first step, the system broadcasts a title and options on a vote session to clients, and then the clients display the title and the options. The title is 'Event location' in Fig. 6. The buttons on the client

display the options. The buttons display 'Sea, Mountain, Sky and Space' in Fig. 6. The system does not accept any vote at this time. At the second step, the system begins to accept votes from the audience devices. At the third step, the system assigns unique numbers to each client to identify voters. At the fourth step, the system waits for votes. The clients send selected options with assigned numbers to the system. At the fifth step, the system closes the vote session. The system finishes accepting votes, and the clients disable the buttons on audience devices. At the last step, the system releases the unique numbers to prevent the tracking of voters. Finally, the system reveals the results.

We explain protocol messages for each sessions. At the first step, the system uses **Prepare Vote Notification Message**. At the second step, the system uses **Open Vote Notification Message**. At the third step, the system uses **Vote Request Message** and **Vote Response Message**. At the fourth step, the system uses **Close Vote Notification Message**. At the last step, the system uses **Result Notification Message**. We explain the details of the messages.

- **Prepare Vote Notification Message:** The message has three fields and notifies the title of the vote and button labels of the selection. The first is a vote session ID. The second is a vote session title. The third is pairs of selection item IDs and labels. The audience system updates the UI selection button labels of UI when receiving the message.
- **Open Vote Notification Message:** The message has three fields and notifies when the vote has begun. The first is a vote session ID. The second is a timestamp of when the message sent from the presenter system. The third is a temporary unique number. The client should contain the number into a Vote Request Message described below. The message should be one in a vote session. The audience system enables UI selection buttons of UI when the message is received.
- **Close Vote Notification Message:** The message has two fields and notifies when the vote is ended. The first is a vote session ID. The second is a timestamp of when the message sent from the presenter system. The message should be one in a vote session. The audience system disables UI selection buttons of UI when the message is received.
- **Vote Request Message:** The message is sent from the client system to post a selection. The message has four fields. The first is a vote session ID. The second is a timestamp of when the message is sent from the audience system. The third is a temporary unique number. The fourth is a selection id. The message should be one or more in a vote session. The message is available if the open vote notification message is received or the close vote notification message is not received with the same vote session ID.
- **Vote Response Message:** The message is sent from server system in response to the vote request. The message has four fields. The first is a vote session ID. The second is a timestamp of when the message sent from the server system. The third is a temporary unique number. The fourth is a response code that indicates whether the request is accepted or not.
- **Result Notification Message:** The message is sent from the presenter system periodically to display the current selection count on audience devices. The message has three fields. The first is a vote session ID. The second is a timestamp of when the message is sent from a presenter system. The third is an array of selection counts in the same order with the pairs of selection item IDs and labels of Prepare Vote Notification Message. The messages broadcast the current selection status in the questionnaire session. The vote session ID is valid between when the vote session begins and when the session is closed.

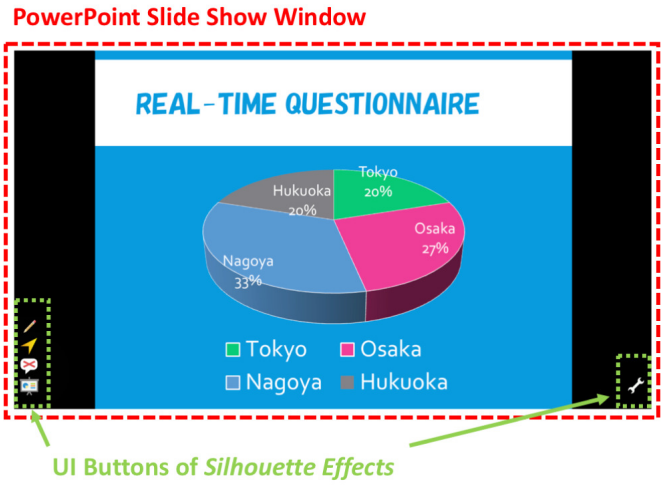


Fig. 7: The Silhouette Effects system controls a PowerPoint system to display questionnaire results.

C. Displaying Results of Real-Time Questionnaire

Generally, changing the shape object attributes (coordinating position in a slide, changing background color of the shape) on a slide in high frequency requires high CPU loads. To suppress the high loads, the graph update module has a minimum interval for updating the graph properties. In the experimental case, we fixed the minimum interval to 1 second.

Fig. 7 indicates the screenshot displaying the questionnaire results on the device of the lecturer. The icons located in the left and right lower angles are a presentation controller for supporting the lecturers in giving a presentation effectively with the proposed method Silhouette Effects. Silhouette Effects expands the methods for utilization of Microsoft PowerPoint to modify slide objects, such as text, graphs and figures, at the presentation mode. The presentation controller based on Silhouette Effects is a PowerPoint supporting tool that provides the presentation-supporting features and sends the information of sharing presentation slides. The lecturers run the presentation controller together with PowerPoint.

The presentation controller allows lecturers to begin a questionnaire at any time during presentations. Since our real-time questionnaire supports a type of survey of which there are

up to four alternatives, the lecturer configures the alternatives using the presentation controller. Then, a graph like Fig. 7 is generated for displaying the result of the questionnaire in real-time. The presentation controller inputs the lecture's configuration of the alternatives to a Microsoft Excel spreadsheet and visualizes the configuration as a graph object. The graph object in Excel gets copied to the PowerPoint presentation. When receiving the result of questionnaires from the counting server that tallies answers from audiences, the presentation controller passes the results to the Excel spreadsheet that runs in the presentation background, and the graph object visualizes the result. When receiving the second and subsequent results, the presentation controller rewrites the data in the Excel, spreadsheet and then the graph object is updated in real time.

The presentation controller saves the results of the questionnaires to a presentation file as PowerPoint graphs. In general conference, the results need to be recorded in the conference note for future reference. The real-time questionnaire system avoids the trouble of recording by automatic saving.

We propose the hybrid interactive presentation system that connects the traditional presentation-supporting tool and the web application for gathering feedback from audiences. The hybrid interactive presentation system occupies an important role as the basis of the real-time questionnaire system.

D. Preparation before Presentation

To use real-time questionnaire function in the presentation slideshow, the presenter prepares a slide and graph shape object with the following steps: First, you add a new slide if needed. Second, you add a graph shape object (pie chart, bar graph) on the slide. Finally, you set the graph label title using the Excel cell editor.

In the presentation slide show, the presenter system checks whether a slide has a graph shape object when the slide is shown the first time. If the slide has a graph shape object, the presenter system sends a "Prepare Vote Notification" message and "Open Vote Notification" message to the audience system via the server system. The presenter system automatically sets the graph label color corresponding to the client web application's vote buttons.

VI. Evaluation

In this section, we describe the evaluation procedure of the real-time questionnaires function and the evaluation result. We measured the execution time of updating graph parameters.

A. Experimental Procedure and Environment

We made an evaluation script to measure the execution time of updating graph parameters. First, the script opens a PowerPoint presentation file that has a slide, and the slide contains a pie chart graph for the real-time questionnaire. Next, the script performs the following steps 100 times after pre-execution and 10 times for warm-up. For the first step, the script records the start time (t_0). For the second step, the script generates four random numbers and updates four graph parameters on the slide graph. For the third step, the script records the end time (t_1). Finally, the script closes the file. The execution time is calculated by $t_1 - t_0$. We used a MacBook Pro as a presenter's device that is composed of Intel Core i7 2.3 GHz, 16 GB 1,600MHz DDR3 of RAM, Windows 10, and the version of PowerPoint is 16.0. We performed the script 10 times; therefore, we sample 1,000 execution times.

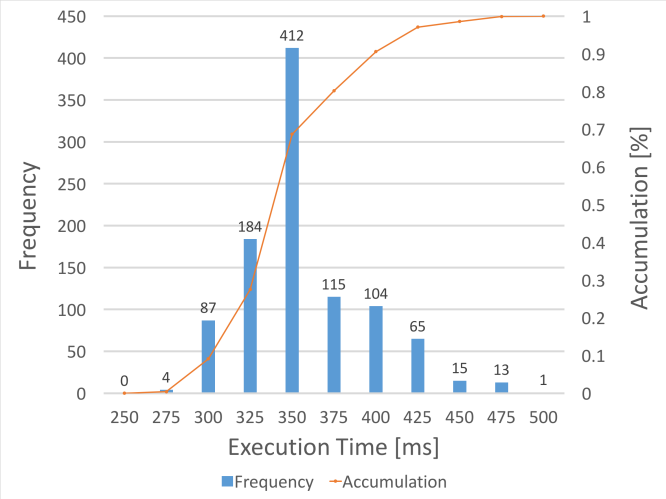


Fig. 8: Execution Time Histogram of Updating Pie Chart Graph.

B. Experimental Result

Fig. 8 shows the execution time histogram of updating graph parameters. We sampled 1,000 times of the execution time. In Fig. 8, the horizontal axis indicates the sampling period of execution time, the left-side vertical axis indicates the frequency and the right-side vertical axis indicates accumulation of the frequency. The peek appeared in 350ms to less than 375ms sampling period, 90% of the execution times are below 425ms and all execution times are below 525ms. Therefore, the slide object controller can update graph parameters about 2 times in a second.

VII. Discussion

In this section, we discuss the usability and the application of our approach. We confirm the usability of the real-time questionnaire, and we entertain the progressive approach for use from the decision-making methods point of view.

A. Usability

The hybrid interactive presentation system allows the lecturer and audiences to communicate using the presentation-supporting system that extends the traditional system (i.e., PowerPoint) and the simple web application. Our approach reduces the time and expense associated with learning and running the interactive presentation system. The lecturers only have to know the use of commonly-used tools, and audiences submit their feedback with the user-friendly interface that is illustrated in Fig. 5.

The result of the questionnaire is visualized using the Excel spreadsheet. The type of the graph object for displaying the results is only the pie chart, but the type of visual effects is

expanded easily by the simple expansion of the presentation controller. When the extended presentation controller allows the users to draw a variety of graphs, the expressive faculty of the results is improved. Our approach can handle the complex data via the functions of Excel. The users are able to perform more complex questionnaires, i.e., not only the type where the answer has a certain number of choices, but the type where the answerer makes multiple selections and selects answers from a number of choices. The approach that incorporates presentations and the existing data-handling system contributes to non-trivial extensions of functions of presentations.

B. Meeting Support for Decision-Making

We discuss the effect of applying the real-time questionnaire system to the ordinary conferences in enterprises and organizations. In previous researches, the approaches to implementing real-time questionnaires in education are properly validated. The experimental result ensures that the performance of the real-time questionnaire has efficacy as the feedback mechanism at a large conference. The system can be applicable to larger-scale presentations, e.g., an ordinary conferences in enterprises and organizations.

Let us discuss the system from the application as the decision-making system point of view. The decision-making system supports various methods of decision-making such as Brainstorming and KJ Technique so that the participants offer opinions on the same plane and organize the opinions using each procedure. The posting feedback feature of our system allows for collecting the opinions of the users and other data as a key step, and then the submitted opinions and data are arranged on the shared plane that is the screen of the presentation. Using the real-time questionnaire system, furthermore, the users perform the decision-making techniques and forge a convergence of opinion effectively.

The function of adding annotations allows the users to offer opinions and compiled data on the shared presentation slide. Dragging annotations and handwritten memos are synchronized in real time, and the features are supportive of the method of decision-making such as grouping opinions in KJ Technique. The operations are performed on the shared screen, and users deal with the operations cooperatively.

In the step of gathering and organizing opinions, the real-time questionnaire is available to summarize options and reach an agreement. The real-time visualization of the questionnaire results has the potential to have an impact on the process of decision-making, i.e., the users can be influenced by other people's opinion. The real-time questionnaire can prompt the users to defer to or to deprecate the answers voted by others. The approach that integrates the real-time questionnaire into the traditional decision-making methods can develop a new methodology of decision-making.

Compared with the traditional system of supporting decision-making, our system widens the scope of participants, i.e., the method based on the interactive presentation functions successfully in a large group. Then, since functions for submitting opinions give anonymity, the users submit opinions at ease, and our approach simulates the discussions. The real-time questionnaire that is lacked in the existing applications improves to form an opinion about the discussions.

VIII. Conclusion

We proposed the interactive presentation system, the Silhouette Arena, which provides the efficient real-time questionnaire function to collect feedback from attendants of presentations. The system helps presenters to have a sophisticated presentation with the slide object-manipulation function on the Silhouette Effects. Moreover, the Silhouette Web supports audiences to make real-time feedbacks in presentations. We explained the implementation of the system to improve the efficiency of the real-time questionnaire function. We demonstrated the efficiency of the system, finding that the system can be used at a large conference in organizations and in an educational class. The approach to the real-time questionnaire based on the hybrid interactive presentation system compensates for the lack of current systems and contributes to establishing a new type of presentation.

References

[1] Ryota Inoue, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, *Visualizing Real-Time Questionnaire Results to Promote Participation in Interactive Presentations*, Proc. of the 2014 IIAI 3rd International Conference on Advanced Applied Informatics (IIAI-AAI 2014), pp.64-69, 2014.

[2] Ryota Inoue, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, *An Interactive Presentation System Based on Feedback to a Presentation Material of an Ongoing Presentation*, IPSJ Journal, Vol.56, No.10, pp.2011-2021, 2015.

[3] Yusuke Niwa, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, *An Efficient Method for Distributing Animated Slides of Web Presentations*, International Journal of Advanced Computer Science and Applications(IJACSA), Vol.7, Issue 1, pp.612-620, 2016.

[4] S. M. Keough, *Clickers in the Classroom A Review and a Replication*, Journal of Management Education, Vol.36, No.6, pp.822-847, 2012.

[5] J. H. Hana and A. Finkelsteinb, *Understanding the effects of professors' pedagogical development with Clicker Assessment and Feedback technologies and the impact on students' engagement and learning in higher education*, Journal of Computers & Education, Vol.65, pp.6476, 2013.

[6] M. Hauswirth and A. Adamoli, *Teaching Java programming with the Informa clicker system*, Journal Science of Computer Programming, Vol.78, pp.499-520, 2013.

[7] V. Triglianos and C. Pautasso, *ASQ: Interactive Web Presentations for Hybrid MOOCs*, Proceedings of the 22nd international conference on World Wide Web companion, pp.209-210, 2013.

[8] K. Kanev, *Tangible interfaces for interactive multimedia presentations*, Journal of Mobile Information Systems, Vol.4, No.3, pp.183-193, 2008.

[9] B. Signer and M. C. Norrie, *PaperPoint: a paper-based presentation and interactive paper prototyping tool*, Proceedings of the 1st international conference on Tangible and embedded interaction, pp.57-64, 2007.

[10] Hiroyuki Yamada, Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, *A Reactive Presentation Support System based on a Slide Object Manipulation Method*, Proceedings of The 2014 International Conference on Computational Science and Computational Intelligence, Vol.2, pp.46-51, 2014.

[11] Y. Goto, *Information Assurance, Privacy, and Security in Ubiquitous Questionnaire*, Proceedings of the fourth international conference on Frontier of Computer Science and Technology, pp.619-624, 2009.

[12] H. Louafi, S. Coulombe and U. Chandra, *Efficient Near-Optimal Dynamic Content Adaptation Applied to JPEG Slides Presentations in Mobile Web Conferencing*, Proceedings of the 27th international conference on Advanced Information Networking and Applications, pp.724-731, 2013.

FPGA implementation of filtered image using 2D Gaussian filter

Leila kabbai *, Anissa Sghaier†, Ali Douik* and Mohsen Machhout† *
National Engineering School of Monastir, University of Monastir Tunisia
† Faculty of sciences Monastir, University of Monastir-Tunisia
‡ National Engineering School of Sousse, University of Sousse-Tunisia

***Abstract*—Image filtering is one of the very useful techniques in image processing and computer vision. It is used to eliminate useless details and noise from an image. In this paper, a hardware implementation of image filtered using 2D Gaussian Filter will be present. The Gaussian filter architecture will be described using a different way to implement convolution module. Thus, multiplication is in the heart of convolution module, for this reason, three different ways to implement multiplication operations will be presented. The first way is done using the standard method. The second way uses Field Programmable Gate Array (FPGA) features Digital Signal Processor (DSP) to ensure and make fast the scalability of the effective FPGA resource and then to speed up calculation. The third way uses real multiplier for more precision and a the maximum uses of FPGA resources. In this paper, we compare the image quality of hardware (VHDL) and software (MATLAB) implementation using the Peak Signal-to-Noise Ratio (PSNR). Also, the FPGA resource usage for different sizes of Gaussian kernel will be presented in order to provide a comparison between fixed-point and floating point implementations.**

***Keywords*—Gaussian Filter; convolution;fixed point arithmetic;Floating point arithmetic;FPGA**

I. Introduction

Convolution has been widely used in computer vision and image processing, including object recognition [2] and image matching [3], However, convolution operation typically requires a significant amount of computing resources [4]. Image filtering is applied as pre-processing to eliminate useless details and noise from an image. It is produced by convolution between an image and 2D Gaussian mask. In the literature, several efficient FPGA implementations of the 2D convolution operation have been proposed [5]–[9].
Hanumantharaju et al. [10] proposed a hardware architecture suitable for FPGA/ASIC implementation of a 2D Gaussian surround function for image processing application which offers a savings of memory. Barbole et al. [11] implemented steerable Gaussian smoothing filters on an FPGA platform based on a VirtexV ML506 using the pipelined approach and DSP which reduces memory requirements. Talbi et al. [5] developed architecture for separable and two-dimensional Gaussian smoothing filters, which was implemented in the VirtexV FPGA platform. They prove that the first approach is significantly faster than the second one. In the same year, Cabello et al. [2] implemented a 2D Gaussian Filter in FPGA using fixed-point arithmetic and floating point arithmetic, they found that increasing the kernel sizes, they reduced the computational costs using floating point arithmetic.
In this paper, a Gaussian filter on an Field Programmable Gate Array (FPGA) platform will be implemented. We will focus in the main bloc which is the convolution module based on the multiplication operation. Thus, the multiplier is in the heart of the proposed design. For this, the standard multiplier will be firstly implemented. Then, in order to accelerate calculus and to minimize resource use, FPGA features will be used which are DSP (Digital Signal Processor) and RAMs. Finally, in order to have more precision in image output, a real multiplier proposed in [13] will be used to implement the entire architecture. It is a new way to do a multiplication between two real numbers. Our application is implemented by two tools such as MATLAB and VHDL, and simulated on the ISE simulator.

The remainder of this paper is as follows. Section 2 introduces the image filtering algorithm. The hardware implementation of image filtering is presented in section 3. In section 4, the hardware optimization of convolution module based on changing the multiplier will be discussed. Experimental results are given in section 5. Finally, a conclusion will be done in section 6.

II. Image filtering algorithm

Smoothing filters are widely used in many applications such as object recognition, matching, classification, etc. They are applied as pre-processing for removing useless details and noise [14]. We will focus on image filtering based on Gaussian filter.

A. Gaussian mask

Gaussian filter is one of the most important and widely used filtering algorithms in image processing [5]. Gaussian filter (G) is defined in equation 1.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (1)$$

where G is the Gaussian mask at the location with coordinates x and y, σ is the parameter which defines the standard deviation of the Gaussian. If the value of σ is large, the image smoothing effect will be higher.

B. Convolution operation

In general, smoothing can be effected by convolve the original image I(x,y) of the size h x w with a Gaussian mask G(x,y) as illustrated in equation 2. It is obtained by computing the sum of products among the input image and a smaller Gaussian matrix of the size (3 × 3). A 2D convolution using a 3 × 3 mask and 3 × 3 input image is illustrated in Figure reffig1.

$$f(x,y) = \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} G(i,j)I(x-i,y-j) \qquad (2)$$

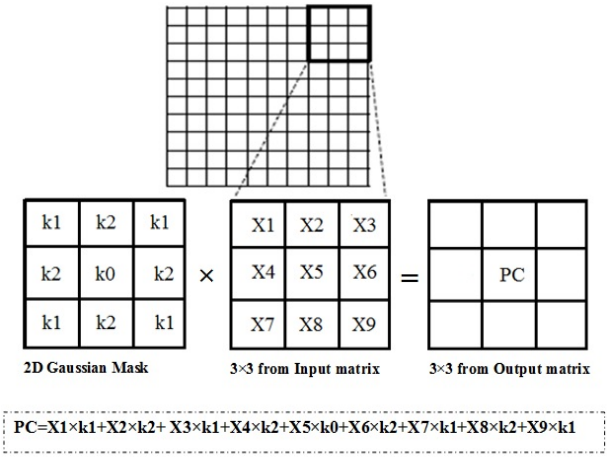


Fig. 1: Convolution operation

III. Hardware implementation of image filtering

In this section, the proposed architecture design of the Gaussian filter will be presented.

A. Block diagram of image filtering

Figure 2 illustrates the block diagram of image filtering. First, the input image and the Gaussian mask are read and saved by MATLAB. Next, These values are converted into a vector in a text file extension *.coe using the MATLAB tool and loaded the text file in block RAM (BRAM). The text file of Gaussian mask and image is stored respectively in BRAM1 and BRAM2. After that, the convolution operation is effected between these pixel values of two BRAM (1 and 2) using VHDL tool and saving the obtain results in another block (BRAM3). Finally, the text file of BRAM3 is converted by MATLAB tool in order to display the results form an image. The next step, we defined each block of diagram in Figure 2.

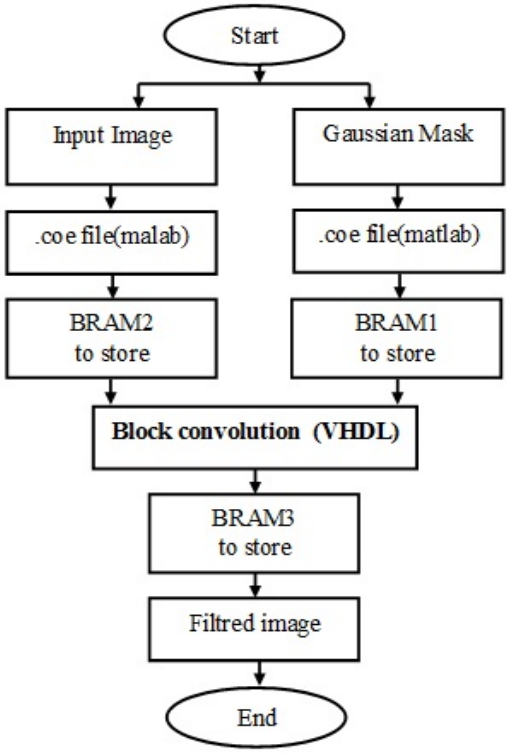


Fig. 2: Block diagram of image filtering

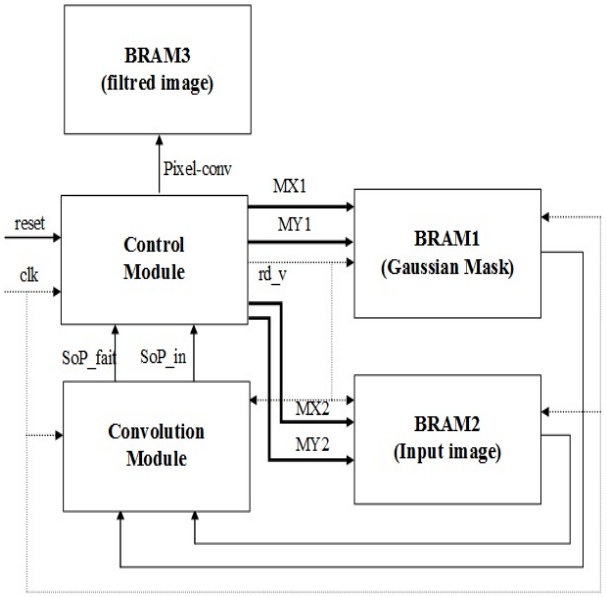


Fig. 3: Synchronous architecture of image filtering

B. Synchronous architecture hardware of image filtering

Figure 3 depicts the block diagram of synchronous image filtering which contains a set of modules: Control Module, 3 BRAMs (matrix of input image, matrix of Gaussian mask, matrix of filtered image) and convolution Module.

1) **Gaussian Filter**
The convolution of an image with a Gaussian mask involves floating point multiplications, which consumes considerable hardware resources. The Gaussian mask size (3 × 3) is presented by the matrix below by choosing the standard deviation equal to 0.5.

$$\begin{bmatrix} 0.0113 & 0.0838 & 0.0113 \\ 0.0838 & 0.6193 & 0.0838 \\ 0.0113 & 0.0838 & 0.0113 \end{bmatrix}$$

Then, it is necessary to convert the floating point coefficients to fixed integer point coefficients for hardware implementation of the Gaussian filter. In the convolution process, each mask values has to be multiplied with each element of the image and then divided by a power of 2 [15], [16]. The approximation of the Gaussian mask is presented by equation below.

$$\begin{cases} G(x,y) = \frac{1}{2^8}\begin{bmatrix} 3 & 21 & 3 \\ 21 & 158 & 21 \\ 3 & 21 & 3 \end{bmatrix} \\ = \begin{bmatrix} 0.0117 & 0.082 & 0.0117 \\ 0.082 & 0.6172 & 0.082 \\ 0.0117 & 0.082 & 0.0117 \end{bmatrix} \end{cases}$$

2) **Block RAM**
In Xilinx FPGAs, a Block RAM (BRAM) is a dedicated two-port memory that stores up to 36Kb of data. The FPGA contains many of these blocks. Inside of each, small logic block is a configurable lookup table. It is normally used for logic functions, and it can be also reconfigured as a few bits of RAM. Several of them can be combined into a larger RAM which is denoted by a distributed RAM. BRAM is synchronous, this means that the read and write operations from and to the memory are based on the clock input signal. The read and write operations are also dependent on the read/write enable ports. In our case, BRAM2 is used to store the data test image using .coe file which is generated with Matlab tool, and a BRAM1 is used to store the .coe file of Gaussian mask, which are then read by the control module. BRAM3 will save the data filtered.

3) **Control module**
The control unit is an important step of the proposed synchronous architecture. It allows to generate the address to BRAMs (1 and 2) and transfers the data from each BRAM to the corresponding convolution module for computing the Sum of Products (SoP) between these values, after that the convoluted value is stored in BRAM3. The control module is designed as a Finite State Machine (FSM) simulated in VHDL. Figure 4 illustrates the Finite State Machine (FSM) of the control module.
In the first state, initialization parameter will be affected. Then in state 1, the signal rd-v will be putted to 1 to access both memories. FSM increments the counter MY1 and MY2 when the MX1 and MX2 counter are finished addressing a line of image pixel block (3 by 3) and the same Gaussian block. This process is repeated the addressing of the blocks, if it is completed then goes to state 2 if not it returns to state 1. States 2 and 3 represent two late cycles to synchronize system signal. After that, it goes to state 4 where the machine puts the rd-v signal to zero in order to stop the addressing of the two memories and goes to state 5. In the state 5, the machine tests the SoP-fait signal, if it is equal to zero then it returns to the same state, if not it stored the value of SoP-in a table. After that, it increments the counter one " i " or " j " in order to read a new block, if " i " is different to the (length of size image -1) and " j " is different (width of size image -1) then returns to state 1. If not goes to state 6 (end process). Where, X is the length of size image -1 and Y is the width of size image -1.

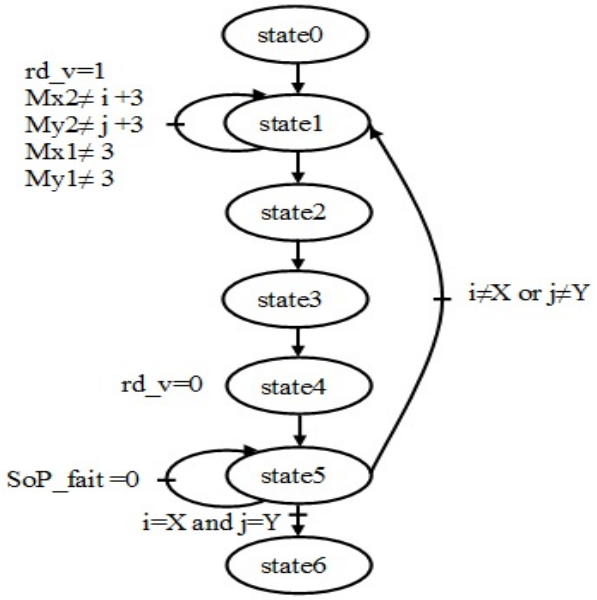


Fig. 4: FSM of the control unit

4) **Convolution Module**
Convolution module focuses on the calculation of the sum of products (SoP) between pixels in BRAM1 and BRAM2 for a window of 3 by 3. Equation 5 depicts an example of the convolution module between Gaussian mask integer and matrix 3×3 from input image.

IV. Hardware optimization of convolution module

The main operation in the convolution module is the multiplication.

A. Convolution module using standard method

Signal multiplication of 3×3 image by 3×3 mask will be done. Multiplier input values are loaded into the RAM block addresses port registers (the outputs of the RAM blocks (BRAM1 and BRAM2) are the inputs of the multiplier). One multiplication is completed in one clock cycle, thus the other 9 multiplications will take 9 clock cycles. The partial product of each multiplication will be summed to obtain the final result. The final result of the multipliers will be stored also in RAM blocks (BRAM 3).

B. Convolution module using FPGA multiplier

FPGA devices have dedicated architectural features that make it easy to implement high performance multipliers. FPGA devices feature embedded high-performance multiplier-accumulators (MACs) in dedicated Digital Signal Processor (DSP) blocks. For high performance applications, DSP blocks can speedup different operations. Embedded multiplier blocks using DSP will be used in our Gaussian filter for low cost and

speedup smoothing image.
These multipliers are implemented in a combination of DSP blocks or embedded multipliers and logic resources. DSP is a multiplication-intensive technology and to achieve high speeds, these multiplication operations must be accelerated. The base of many DSP algorithms is multiplication. In this operation, each element of the multiplier is multiplied by each bit of the multiplicand. Then, the partial product of each multiplication is accumulated according to the weight of the partial product, where the weight indicates the location of a bit corresponding to other bits.

- Multiplication: Multiple memory blocks produce one multiplication result every clock cycle. This mode is useful for high-speed data scaling.
- Sum of multiplication: One memory block or group of memory blocks produces the sum of multiplication results.

C. Convolution module using real multiplier

From a practical point of view, using floating point multiplications to calculate the convolution can consumes considerable hardware resources but it offers more precision and a good smoothing of the input image. In this section, we will present the used multiplier proposed in [13]. Hence, if we want to multiply real coefficients like this example: $a = 3,14$ and $b = 4,15$. We may first define α_a and α_b which are number of terms after comma. In our example: $\alpha_a = 2$ and $\alpha_b = 2$. And, we divide a and b into two parts:

$$\begin{cases} X_a = 3 \quad and \quad Y_a = 14 \\ X_b = 4 \quad and \quad Y_b = 15 \end{cases}$$

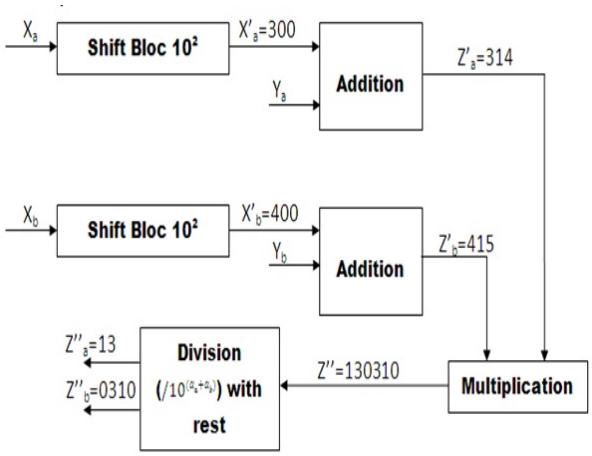


Fig. 5: Proposed Multiplication architecture

In Figure 5, the first part of a should be multiplied by 10^{α_a} and b by 10^{α_b}. To ensure the multiplication we can use the left shifting function. Then we add the result to the correspondent second part of a and b, so we will have Z'_a and Z'_b. Then after multiplying these two terms we have to divide the result by $10^{(\alpha_a+\alpha_b)}$. To implement this division, we can use right shifting function. The final result is: $R = a \times b = 13,0310$.

V. Experimental results

In this section, simulations and implementation results will be discussed.

A. Performance Measures

The Peak Signal to Noise Ratio (PSNR) is the most used parameter to evaluate image quality in the literature [11], [17]–[20], [22]. PSNR value can be computed by comparing two images which are original image and filtered image. The PSNR was used to measure the image quality. A higher PSNR value indicates that the filtered image contains better image quality. The PSNR has been calculated as follows;

$$PSNR = 10\log_{10}(\frac{255^2}{MSE}) \quad (3)$$

Where, MSE is the Mean Square Error (equation 4) between the original image (I1(m,n)) and the filtered image (I2(m,n)), with, m and n are pixels of image M N.

$$MSE = \frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N}(I_1(m,n) - I_2(m,n))^2 \quad (4)$$

B. Simulation results in MATLAB and VHDL

In this section, simulation and implementation results will be done. Figure 6 presents the filtered image by two tools which are MATLAB and ModelSim-SE (VHDL).‘

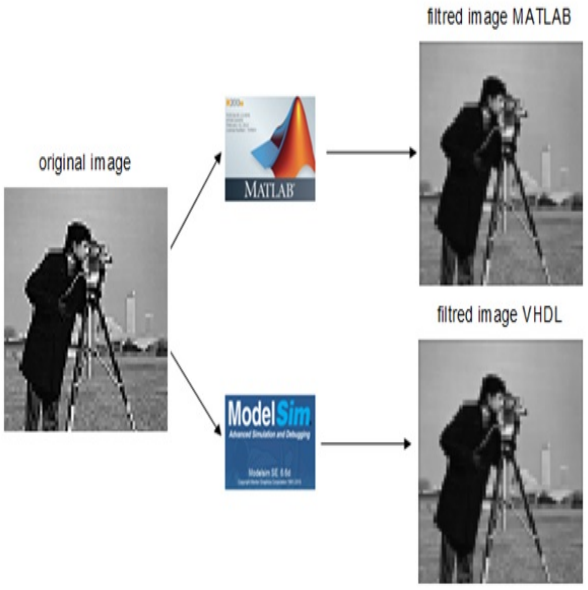


Fig. 6: Resulting filtered image in both MATLAB and VHDL

The kernel size 3×3 will be conserved and sigma values will be changed in order to see their impact in the filtered image. Figure 7, 8 and 9 illustrate the filtered image by the software (MATLAB) and hardware (VHDL) implementations. We can deduce that the blurring effect increases proportional to the sigma value (respectively 0.5, 1 and 1.5).

For different sigma values, Table I resumes the corresponding PSNR of images (in both VHDL and MATLAB).

Fig. 7: Filtered image with sigma=0.5

Fig. 8: Filtered image with sigma=1

Fig. 9: Filtered image with sigma=1.5

For sigma equal 0.5, we observe that the PSNR (VHDL) obtains better result compared to PSNR (MATLAB). So, when increase sigma, the PSNR value of MATLAB and VHDL are decreased. Figure 10 shows the comparison between PSNR values both resulting image in MATLAB and VHDL.

Normally, if PSNR value is more than 40 dB, this is an indication that the quality of the image is good. But, if the image is mean quality, the PSNR value is less than 30 db which is the case of our selected image. We note that when we vary the sigma value the effect of smoothing increase and the PSNR decrease.

TABLE I: PSNR values for different output images in VHDL and MATLAB

	PSNR (MATLAB)	PSNR (VHDL)
Sigma = 0.5	25. 2236	27.3294
Sigma = 1	19.8879	20.3760
Sigma = 1.5	18. 0441	19.6098

C. Simulation results and resources utilization

Modern FPGA families integrate many features into the silicon. These features reduce the area required and increase

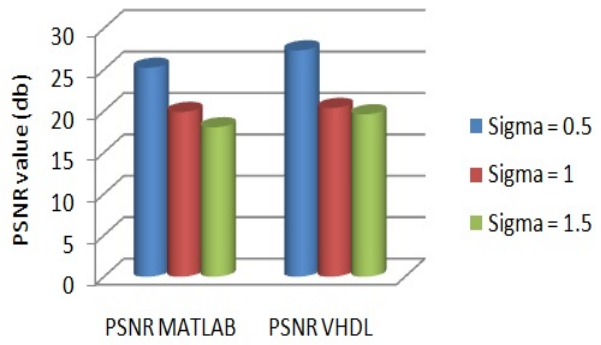


Fig. 10: Comparison between PSNR values to both resulting image in MATLAB and VHDL

speed compared to building them from primitives. For examples: multipliers, generic DSP blocks, embedded processors, high speed I/O logic and embedded memories. To ensure the correctness of the proposed architecture, the algorithm of Gaussian filter has been firstly coded and tested in MATLAB (Version 12.1), then an FPGA implementation was coded in RTL compliant VHDL and the hardware simulation results have been obtained using ModelSim (Version SE 6.4) and synthesized using Xilinx ISE 12.4.
The proposed design has been implemented on Xilinx VirtexV device. In the proposed work, Gaussian design is generic, thus it can be upgraded to any size without an appreciable increase in the hardware. Hence, the functional modules are control module, BRAM, multiplier and adder. Here, experiments are performed on image of size 8×8 using 2D Gaussian mask of 3×3. The simulation results of convolution are shown in Figure 11 and device utilization summary of the implementation is given in Table II, III, and IV.

Figure 11 presents the addressing of the two blocks (BRAM1 et BRAM2) and the result obtained by calculation the Sum of Products (SoP) between pixels in BRAM1 and BRAM2 using VHDL tool.

TABLE II: Performance comparison with the state of the art implementations

	Slices Registers	Slices LUTs	DSP48Es
Ours	127	176	9
2D [5]	228	2089	6
[23]	369	480	-

Comparing our results to those in [5], we note that we decrease the number of slice registers by 44.3% and by 65.5% compared to [23]. Table III compares the results of both fixed point arithmetic and floating point arithmetic. As we can see, we decrease the number of slice registers and slice LUTs comparing to [12].

In addition, we note that floating arithmetic uses more than fixed one but still little compared to the state of the art and we should not forget that it gives more precision to filtered image.

TABLE III: Comparing fixed arithmetic results to floating arithmetic one

	Slices Registers	Slices LUTs
Fixed Arithmetic (ours)	127	176
Float Arithmetic (ours)	138	3080
Fixed Arithmetic [12]	135	209
Float Arithmetic [12]	151	5052

If we increase kernel size we obtain results in Table IV. As we can see, kernel size have a big influence in design performances so that area occupation increase by the increase of kernel size.
Our results outperform those in [12] in term of slices registers and LUTs by 6% and 15% for fixed arithmetic using kernel size $[3 \times 3]$. For floating arithmetic and $[3 \times 3]$ kernel size, the area use decrease by 8% and 39% in term of slices registers and LUTs. It is the same to the other kernel sizes.

VI. Conclusion

Hardware implementation of the Gaussian filter is faster than software one. Thus, using FPGA we are able to process the filtering at the same time of reading the image. In this paper, we have presented the implementation of two-dimensional convolution on a Xilinx VirtexV FPGA platform based on a state machine. We implemented Gaussian filters with different sigma values. Then we optimized the proposed architecture using different multipliers. At the first, we used the standard multiplication "$\times$" used in VHDL language. Then we explored FPGA features and DSP blocks. Finally, we introduced floating point arithmetic. Performances and results show that area and resources utilization decrease specially when using DSP and BRAM of FPGA. Also, speed increase comparing to the other solutions. By using floating point arithmetic the image has more precision and result seems to be is better.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] DG. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision; 60(2), pp. 91-110, 2004.

[3] L. Kabbai, M. Abdellaoui, A. Douik, *New robust descriptor for image matching*, Journal of Theoretical and Applied Information Technology, 87(3), pp. 451- 460, 2016.

[4] L. Rao, B. Zhang, J. Zhao, *Hardware Implementation of Reconfigurable 1D Convolution*, Journal of Signal Processing Systems, 82(1), pp. 1-16, 2016.

[5] F.Talbi, F.Alim, S. Seddiki, I. Mezzah, B. Hachemi , *Separable Convolution Gaussian Smoothing Filters on a Xilinx FPGA platform*, International conference on innovative computing technology (INTECH), Galcia, pp.112-117, May 2015.

[6] M.Neggazi, M.Bengherabi, A.Amira, Z.Boulkenafet, *An Efficient FPGA Implementation of Gaussian Mixture Models Based Classifier*, IEEE.International Workshop on Systems, Signal Processing and their Applications (WoSSPA), Algiers , pp. 367-371.May 2013.

[7] H. Zhang, M. Xia, and G. Hu, *A Multiwindow partial buffering scheme for FPGA based 2-D convolvers*, IEEE Transactions on Circuits and Systems II: Express Briefs, 54(2), pp. 200 - 204, February 2007.

[8] L. Chang, J. Hernndez Palancar, L.E. Sucar, M. Arias-Estrada, *FPGA-based detection of SIFT interest key points*, Machine vision and applications, 24(2), pp.371-392, 2013.

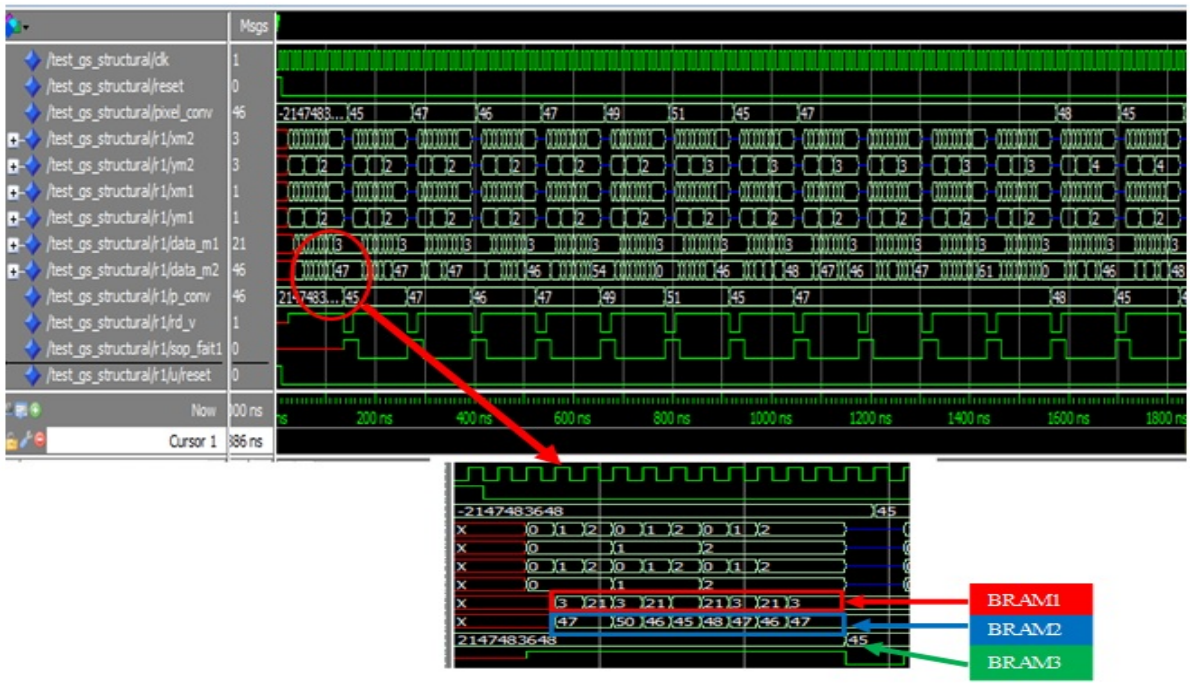


Fig. 11: Simulation results of two-dimensional convolution method

TABLE IV: Implementation results for different kernel sizes

	kernel size					
	[3 × 3]		[5 × 5]		[7 × 7]	
	Unsigned	Float	Unsigned	Float	Unsigned	Float
Slices Registers(ours)	127	138	378	579	546	774
LUTs	176	3080	1270	14559	20380	
Slices Registers [12]	135	151	1181	583	1687	883
LUTs	209	5052	2296	32557	2626	54988

[9] V. Bonato, E. Marques, G. A Constantinides, *A parallel hardware architecture for scale and rotation invariant feature detection*, Circuits and Systems for Video Technology, IEEE Transactions on, 18(12), pp.1703-1712. 2008.

[10] M. C Hanumantharaju, M. Ravishankar, D. R Rameshbabu, *Design and FPGA Implementation of an 2D Gaussian Surround Function with Reduced On-Chip Memory Utilization*, IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, pp. 604 - 609, August 2013.

[11] S. Barbole and S. Shah, *Efficient Pipelined FPGA Implementation of Steerable Gaussian Smoothing Filter*, International Journal of Science and Research, 3(8), pp.1753-1758, 2014.

[12] Frank Cabello, Julio Leon, Yuzo lana, Rangel Arthur: *Implementation of a Fixed-Point 2D Gaussian Filter for Image Processing based on FPGA*, Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, pp.28 - 33, September, 2015.

[13] A. Sghaier, M. Zeghid, M. Machhout, *Proposed efficient arithmetic operations architectures for Hyperelliptic Curves Cryptosystems (HECC)*, The International Multi-Conference on Systems, Signals and Devices, Mahdia, pp.1-5, March 2015.

[14] S. Rashid, S. R. Dixit, A. Y. Deshmukh, *VHDL Based Canny Edge Detection Algorithm*, International Journal of Current Engineering and Technology, 4(2), pp.2277, 4106, 2014.

[15] M. N. Alsharif, *Real Time Image Processing for Lane Following*, Master Thesis, 2014.

[16] N. S. Tahiyah, P. Vikramkumar, K. Sridharan, T.Vineetha, J. Arthi, *Very large-scale integration architecture for video stabilisation and implementation on a field programmable gate array-based autonomous vehicle*, IET Computer Vision, 9(4), pp. 559-569,2015.

[17] B. Sankur, K. Sayood, I. Avcibas, *Statistical evaluation of Image quality measure*, Journal of Electronic Imaging, 11(2), pp.206-223, 2002.

[18] M. Carnec, *Critre de qualit d'images couleur avec rfrence rduite perceptuelle gnrique*, Polytechnique de Nantes, These, 2004.

[19] C. Delgeorge, C. Rosenberger G. Poisson, P. Vieyres, *Towards a new tool for the evaluation of the quality of Ultrasound compression Images*, IEEE transactions on Mdical Imaging, 25(11), pp.1502-1509, 2006.

[20] P. Marziliano, F. Dufaux, S. Winkler, T.Ebrahimi, *Perceptual blur and ringing metrics : application to jpeg 2000*, Signal Processing : Image communication, 19(2), pp. 163-172, 2004.

[21] J.L. Olives, *Optimisation globale d'un systme imageur l'aide de critres de qualit visuelle*. Ecole nationnale suprieur de l'aronautique et de l'espace, 1998.

[22] B. Rajan, S. Ravi, *FPGA based hardware implementation of image filter with dynamic reconfiguration architecture*, IJCSNS International Journal of Computer Science and Network Security, 6(12), pp. 121-127, 2006.

[23] S. Eswar, *Noise reduction and image smoothing using gaussian blur*, Masters of Science in Electrical engineering, California State University, Northridge, 2015.

From Emotion Recognition to Website Customizations

O.B. Efremides
School of Web Media
Bahrain Polytechnic
Isa Town, Kingdom of Bahrain

***Abstract*—A computer vision system that recognizes the emotions of a website's user and customizes the context and the presentation of this website accordingly is presented herein. A logistic regression classifiers is trained over the Extended Cohn-Kanade dataset in order to recognize the emotions. The Scale-Invariant Feature Transform algorithm over two different part of an image, the face and the eyes without any special pixel intensities preprocessing, is used to describe each emotion. The testing phase shows a significant improvement in the classification results. A toy web site, as a proof of concept, is also developed.**

***Keywords*—Emotion recognition; classification; computer vision; web interfaces**

I. Introduction

The development of emotion aware systems is the next step in creating effective, trustworthy and persuasive web applications and websites. Websites capable of *sensing* and reacting to user's emotional state by adjusting their context and their *look & feel (L&F)* can be used from e-companies as powerful recommendation and advertisement tools or by website designers as a medium to increase user satisfaction.

Automatic detection of human emotions from digital images and videos is an active research area attracting a lot of attention in recent years. Interdisciplinary in nature it combines image processing, computer vision and machine learning and can be applied in large number of application especially in the area of human-computer-interaction (HCI).

Recognizing emotions with a high accuracy is a difficult task though. The emotions being communicated by human facial expressions are complex and vary constantly. Variations related to camera/face pose, occlusions of main facial components (e.g., eyes and nose), features as glasses and beards along with illumination conditions and camera technical characteristics make the problem even harder.

In this paper a system that automatically recognizes user emotions and customizes the context and the L&F of a website is presented. The representation of emotions by Scale-Invariant Feature Transform (SIFT) [4] descriptor is investigated. The descriptor applied on a dense grid of keypoints on two images; the face and the eyes of the user as they captured by a web camera. A logistic regression model is used to label the emotion and the context of the website along with its presentation are customized accordingly.

This works is organized as follows. Section II briefly presents the related work. The overview of the system and the description of its parts are given in section III. Settings, experiments and their acquired results are presented in section IV. Finally conclusions are drawn and further work is given in section V.

II. Related Work

Automatic emotion recognition consists of two key factors: the emotion representation and the development of a classifier. A set of features which effectively represents the emotion must be derived and from these features a model must be learned.

Pantic et al [7] proposed a ruled-based classifier in order to recognize facial actions based on contours changes. A multistage model is used to extract and encode features. Initially face, face profile and facial components detectors are used to locate contours. Then a number of fiducial points are extracted and selected defining a mid-level feature parameters used for the final encoding of the actions. They reported 86% accuracy.

Recognition accuracy 81.4% on Cohn-Kanade (CK) [5] dataset reported by Buciu et al. [1]. They used Principal Component Analysis (PCA) as baseline and proposed a non-negative matrix factorization and a local non-negative matrix factorization technique for recognizing six facial expressions.

Shan et al. [8] worked with Local Binary Patterns (LBP). To extract the features and recognize the emotions they proposed a Support Vector Machine (SVM - RBF kernel) classifier with Boosted-LBP features. They used a fixed distance between the two eyes to normalized the faces and they manually labeled the eyes location, to evaluate LBP features in conditions without face registration errors. They reported a 91.4% accuracy.

Combining PCA with an SVM based classifier Vretos et al. [10] achieved 90% accuracy on the Cohn-Kanade dataset. They worked using analysis of vertices on Candide model (a parameterized face mask developed for model-based coding of human faces).

Kalita et al. [3] used an eigenvector based method. Their images are cropped to produce five different eigen-spaces and Euclidean distance is used for classification. They achieved 95% recognition rate.

Donia et al. [2] used histogram of oriented gradients (HOG) to extract features and trained an SVM classifier. The face is cropped for five regions to be created and HOG is calculated

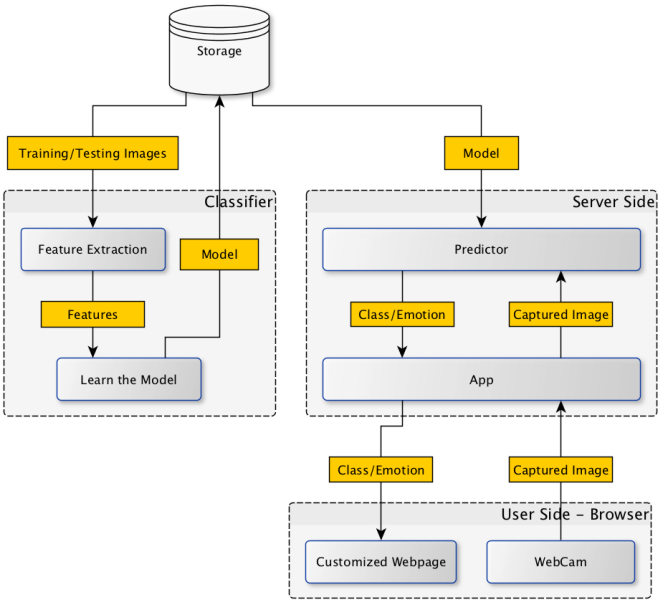


Fig. 1: System Overview

for each region. Using a linear kernel for their SVM they achieved 95% recognition accuracy on static images.

Scale-Invariant Feature Transform (SIFT) [4] has been extensively used for face but not for emotion recognition. In a recent study Neeru [6] reported 89.67% accuracy on the JAFEE dataset using the algorithm and also proposed an modified version of SIFT achieving 97.65% on the same dataset.

III. The System

The system is divided into two different modules as shown in Figure 1. The first one deals with the training of a classifier and results to a model. This model is used from the web part of the system (second module) for predictions. Based on these prediction the website is customized.

A. Features Extraction and Classification

Initially the original images of the dataset are loaded and for each one of them the face is detected using Viola-Jones [9] algorithm. The detected face is cropped, resized and saved to the disk (this is not an essential part of the process but it can significantly improve the speed later (e.g., during cross-validation phase); the entire process can be done on-the-fly (as it happens to the web part of the system)). A second Haar cascade detector [9] is used then to detect the eyes on this new image of the face. The detected eyes are also cropped and resized. Since the entire area of the face and the eyes must be clearly described a dense grid of keypoints is applied to both of them.

Every keypoint on the two grids is described by a SIFT descriptor. SIFT features are characterized by a high distinctiveness power and they are invariant to minor affine distortions, noise and illumination changes. In this work we will use only the description (and not the detection) part of the algorithm. For each keypoint a set of orientation histograms is created ($4x4$ pixel neighborhoods, 8 bins each). These histograms are computed by sampling the gradient magnitude and orientation values around the keypoint ($16x16$ region). A weight is assigned to the magnitude of each sample drawn by a weighted Gaussian function with σ equal to one half the width of the descriptor window. The values of the histograms forms the vector of the descriptors which is normalized to unit length, thresholded (less than 0.2 threshold value is given on the original paper) and normalized again. The method produces a feature vector with 128 elements for each keypoint.

The resulted feature vectors for the face and the eyes are concatenated to form the final feature vector and the process repeats for the next image. The results is a NxD-dimensional array (where N is number of examples and D is the number of features) (see Table II) which is used for the training of the classifier.

A logistic regression classifier (softmax) is trained in this work. In this model, the probabilities of the possible outcome label for a single example are modeled using a logistic function (Eq. 2). The implementation used herein can fit a multi-class logistic regression with L2 regularization (Eq. 1) by following the *one-vs-rest* technique. That mean that a single binary classifier is trained per class.

$$L = -\sum_{i=}^{n} log\, g(y_i z_i) + \frac{C}{2}\sum_{k}^{l} w_k^2 \tag{1}$$

where g is the logistic fiction:

$$g(z) = \frac{1}{1+e^{-z}} \tag{2}$$

and $z_i = \sum_k w_k x_{ik}$, with $w_k, k \in \{0, \ldots, l\}$ the weight for the k^{th} feature and l the number of the features, w_0 the bias weight and C balances the tradeoff between the two terms.

The learnt model is store to the disk in order to be available to the web part of the system.

The proposed approach requires no preprocessing corresponding to pixel intensities. The images are used as they captured by the camera (just cropped and slightly resized). Both the eyes together are detected and handled as a separated single image. Previous approaches (see section II) depended on preprocessing or on facial landmarks in order to work with different parts of the face (e.g, the eyes). In some cases these landmarks should be manual registered. This preprocessing time can negatively affect the total elapsed time of the system when it is finally deployed. It should be noted that even though this is not a critical-time system it remains a real time system.

B. Website

Concerning the client-end of the system the functionality is simple enough. At predefined intervals (and after user's permission) a web camera captures frames from the users while they are visiting the website. Each captured image is sent to the server for processing. As soon as the result (class number - emotion) comes back from the server the interface is changed. On the server-end the learnt model is loaded from the disk and the web application is ready to process the images.

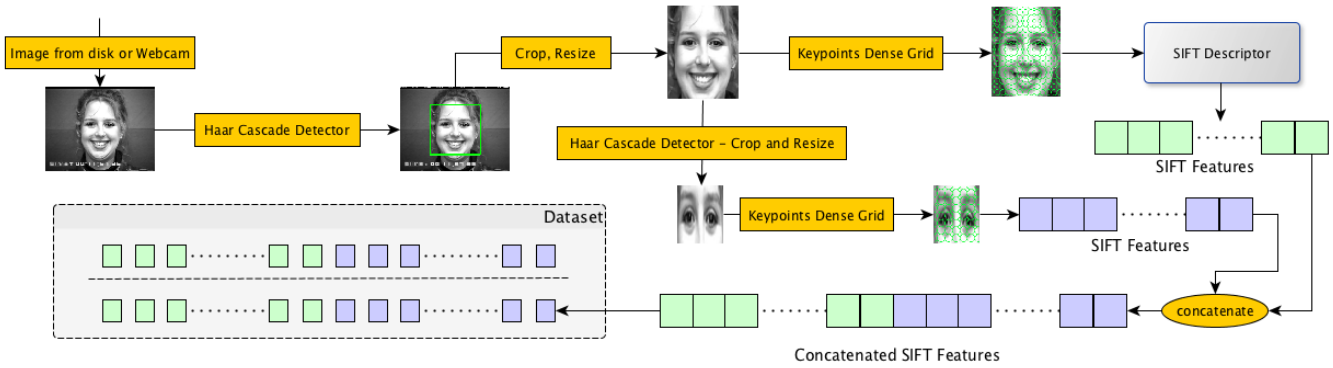


Fig. 2: Feature Extraction from a Single Image

When an incoming image arrives a process similar to the one used for the training of the classifier is applied. The face and the eyes are detected resized and cropped, a dense grid of keypoints is applied, the SIFT descriptors are produced and concatenated (see Figure 2), and the prediction is made. The resulted emotion is sent back to the user-end of the application.

A toy website is built as a proof of concept. This site is only capable of recognizing 3 different emotions (neutral, happy, sadness). These three have been chosen as their facial expressions can last more and they are highly likely to be presented while a user is surfing on the internet. Of course there might be moments that the user is surprised or feels disgust but these emotions change the facial characteristics for a very small period of time which is not justify the change of the environment. A model is also trained to recognize only these three emotions.

IV. Experiments and Results

The system developed using python as programming language and the experiments were conducted on a 2.3 GHz Intel Core i5 mini-Mac system with 8GB main memory.

The Extended Cohn-Kanade (CK+) dataset [5] is used in this work. Currently, the set is one of the most commonly used datasets for facial emotion recognition. Facial behavior of 210 adults from 18 to 50 years of age belonging to different gender and racial groups is shown. 23 facial displays (began and ended in a neutral face) are performed by each participant, Image sequences are digitized into 640x490 pixel arrays with 8- bit gray-scale or 24-bit color values. The emotions included in this dataset are *neutral*, *anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*.

A. Classification

Four different models of the haar cascade detectors for the face and two for eyes where checked. A class creating the dense grid is developed and the SIFT descriptor is used to provide the feature vector for each image. The number of the feature produced are 33792 per image.

A number of linear, non-linear classifiers and ensembles are initially checked to find those who might perform well on the data under investigation. For this phase 80% of the data were used as training set and the rest 20% for testing. Since the dataset is small a 10-fold cross-validation resampling process was applied for the hyper-parameters tuning.

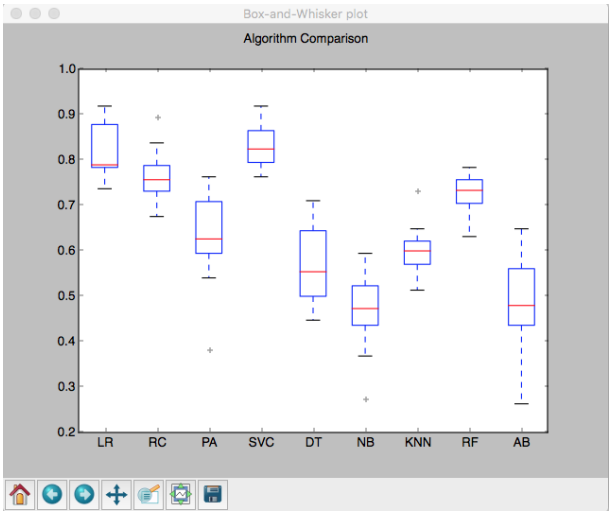


Fig. 3: Comparing the Classifiers (LR: Logistic Regression, PA: Passive Aggressive, SVM: Support Vector Machine, DT: Decision Tree, NB: Gaussian Naive Bayes, KNN: k-Nearest Neighbors, RF: Random Forest, AB: Adaboost)

Numerical results (Figure 3, Table I) suggested that logistic regression and linear SVM classifiers were promising for good results and further investigation. Their performance was more or less the same as expected. To improve the results further an exhausted grid search with cross-validation approach were taken. For almost all trials the logistic regression performed slightly better compared to SVM with linear kernel for this specific dataset. To boost even more the performance two ensembles were tested. A voting ensemble combining both classifiers did not improve the results. Then a boosting ensemble for the logistic regression was built and tested but again without any improvement of the results. The logistic regression was finally selected as the appropriate classifier for the dataset.

TABLE I: Initial Result - Comparing the Classifiers

	SIFT Features	
	Mean (%)	St. Dev (%)
Logistic Regression	81.81	5.98
Passive Aggressive	62.93	11.02
SVM (linear kernel)	83.12	4.56
Dessision Trees	57.08	8.76
Gaussian Naive Bayes	46.66	9.08
k-Nearest Neighbors	60.34	5.59
Random Trees	72.68	4.57
Adaboost	48.32	11.29

TABLE II: Classification Results

	Full Emotions Set	Reduced Emotions Set
Parameters		
Total Images	467	237
Features per Example (D)	33792	33792
Training Examples (N)	373	189
Testing Examples (N)	94	48
Validation k-fold	10	10
Classification Report (average)		
Precision	0.90	0.94
Recall	0.89	0.94
F1-score	0.89	0.94
Support	94 (total)	48 (total)
Accuracy		
Training/Validation	87.10 (+/- 4.35)	93.65 (+/- 3.93)
Testing	89.36	93.75

After tuning the parameters, the classifier was trained and evaluated using the test set. The numerical results are presented in Table II and the corresponding confusion matrices are depicted in Figures 4, 5. As it is mentioned before, two different models are trained. The first is trained to recognize the full range of emotions in our dataset while the second a reduced set of emotions (in order to be used as an example in the toy website). Thus, two different set of results are presented. As it is shown the method improves the initial accuracy results from 81.81% too 87.10% while reduces the standard deviation from 5.98% to 4.35%. The testing time on the experimentaion system is $0.119sec$ for 94 images.

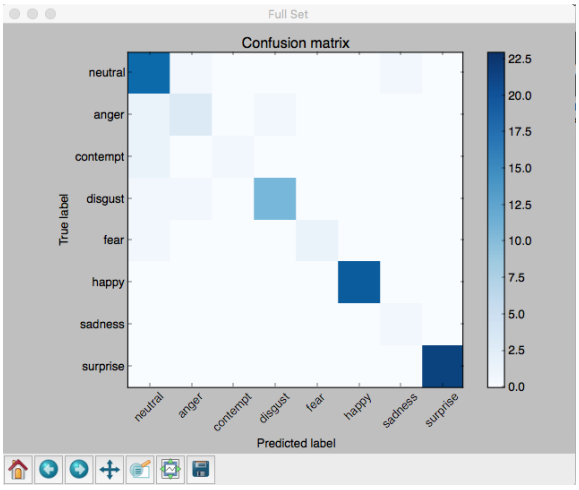


Fig. 4: Confusion Matrix - Full Emotions Set - Testing Phase

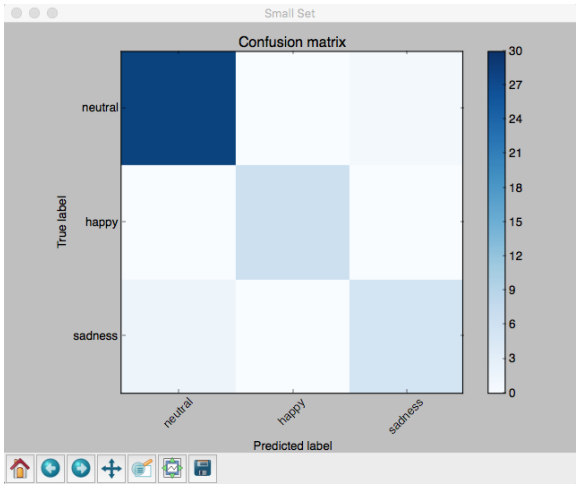


Fig. 5: Confusion Matrix - Reduced Emotions Set - Testing Phase

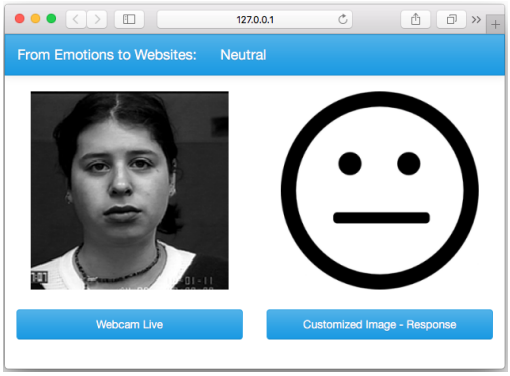


Fig. 6: Toy Website Customized according to User Emotion - Neutral

B. Website

The client-side of the website is developed using HTML for the context part of the page and the Bootstrap framework for CSS styling. The most important component is the WebcamJS javascript library (and open source MIT licensed library) which provides all the necessary functions for an image to be captured and sent to the server. It is an AJAX based communication and a callback function accepts the server response. The JQuery library is used for accessing and altering the Document Object Model (DOM) of the page. Customizations are i) the bootstrap theme is changed and b) a different images are shown to the user (each wiht the recognized emotion). The server-side part is developed using the Flask framework (BSD licensed) and python as programming language. Figures 6, 7 and 8 show three different snapshots of the site. It presents the image captured and the adjustment made to the context and to the look & feel of the site.

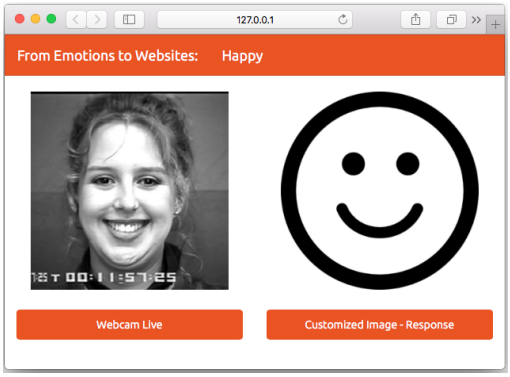


Fig. 7: Toy Website Customized according to User Emotion - Happy

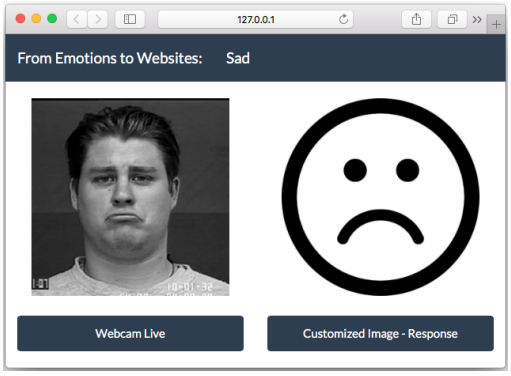


Fig. 8: Toy Website Customized according to User Emotion - Sad

V. Conclusions

In this work a system capable of customizing a web site according to its user emotions is presented. Initially, the SIFT algorithm was used to extract descriptors from a dense grid applied on two different images: the face and the eyes. Using these features a logistic regression (softmax) classifier was trained to recognize emotions. No preprocessing to pixels intensities is applied. The images are used as they are captured by the camera in term of illumination, subject pose, etc. After that, the learnt model is loaded on a website application which accepts images of a website's user through a web camera. Three different emotions are used as an example. The toy site reacts to the emotion recognized by changing its L&F as well as the images presented to the user.

The classifier is the key factor in this application. For the Cohn-Kanade Extented (CK+) dataset which used herein a logistic regression classifier has been chosen and tuned. Cross-validation shows an estimated training/validation accuracy of 87.10% and a testing accuracy of 89.36%. The recognition accuracy goes up to 93% concerning the three emotions the website reacts to.

Further work can be done to improve the classification results for the entire range of emotions included in this dataset. Techniques as Convolutional Neural Networks (CNNs) for this and other bigger datasets can used for models to be trained and tested. Concerning the website, proper usability tests and experiments must be contacted in order determine parameters like the time intervals between the changes of the website interface without frustrating the user or the set of the emotions that must be recognized along with the system's reactions them.

References

[1] I. Buciu and I. Pitas. Application of Non-Negative and Local Non Negative Matrix Factorization to Facial Expression Recognition. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 1, pages 288–291, 2004.

[2] M.M.F. Donia, A.A.A. Youssif, and A. Hashad. Spontaneous Facial Expression Recognition Based on Histogram of Oriented Gradients Descriptor. *Computer and Information Science*, 7(3):31–37, 2014.

[3] J. Kalita and K. Das. Recognition of Facial Expression Using Eigenvector Based Distributed Features and Euclidean Distance Based Decision Making Technique. 4(2):196–202, 2013.

[4] D. G Lowe. Distinctive image features from scale invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004.

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, (July):94–101, 2010.

[6] N. Neeru. Modified SIFT Descriptors for Face Recognition under Different Emotions. 2016:1–19, 2016.

[7] M. Pantic and L.J.M. Rothkrantz. Facial Action Recognition for Facial Expression Analysis From Static Face Images. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 34(3):1449–1461, 2004.

[8] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[9] P Viola and M Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition (CVPR)*, 1:511–518, 2001.

[10] N. Vretos, N. Nikolaidis, and I. Pitas. A model-based facial expression recognition algorithm using Principal Components Analysis. *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3301–3304, 2009.

New mechanism for Cloud Computing Storage Security

Fragmentation-redundancy-scattering as security mechanism for Data Cloud Computing

Almokhtar Ait El Mrabti, Najim Ammari,
Anas Abou El Kalam, Abdellah Ait Ouahman
OSCARS Laboratory, National School of Applied Sciences,
Cadi Ayyad University
Marrakesh, Morocco

Mina De Montfort
ARTIMIA,
75 Street Guy Mˆoquet, 92240
Malakoff, France

***Abstract*—Cloud computing, often referred to as simply the cloud, appears as an emerging computing paradigm which promises to radically change the way computer applications and services are constructed, delivered, managed and finally guaranteed as dynamic computing environments for end users. The cloud is the delivery of on-demand computing resources - everything from applications to data centers - over the Internet on a pay-for-use basis. The revolution of cloud computing has provided opportunities for research in all aspects of cloud computing. Despite the big progress in cloud computing technologies, funding concerns in cloud, security may limit a broader adoption. This paper presents a technique to tolerate both accidental and intentional faults, which is fragmentation-redundancy-scattering (FRS). The possibility to use the FRS technique as an intrusion tolerance one is investigated for providing secure and dependable storage in the cloud environment. Also a cloud computing security (CCS) based on the FRS technique is proposed to explore how this proposal can then be used via several scenarios. To demonstrate the robustness of the proposal, we formalize our design and we carry out a security as well as performance evaluations of the approach and we compare it with the classical model. The paper concludes by strongly suggesting future research proposals for the CCS framework.**

***Keywords*—Cloud Computing; Data Security; Data Encryption; Fragmentation-Redundancy-Scattering;**

I. Introduction

The cloud is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. While many organizations are looking to take advantage of the cloud, data security remains a top matter. Nevertheless, effective data protection and strong encryption in the cloud is possible and available through several of cloud solutions. For a small and medium size business (SMB), the benefits of cloud is currently driving adoption. In the SMB sector, there is often a lack of time and financial resources to purchase, deploy and maintain an infrastructure (e.g. the software, server and storage). Then, SMBs can easily add or remove services and typically will only pay for what its do use.

For cloud computing, there are numerous security issues. Indeed, open systems and shared resources increased security challenges, and made security one of the barriers that face cloud computing technology adoption. Cloud computing is now the hot spot of computer business and research, and its use has grown rapidly in many businesses, especially SMB because it provides many benefits in terms of low cost and data accessibility. Cloud computing adoption leads to gain efficiency development, effectiveness deployment and cost saving in purchasing and maintaining infrastructure. This indicates that the cloud industry is promising, except that existing vulnerabilities in this technology will increase the threat of pirates particularly for data security in the cloud computing.

First, two important terms are defined as indicated in [1] and that can merge on cloud:

- Cloud computing: An information technology (IT) model or computing environment composed of IT components (hardware, software, networking, and services) as well as the processes around the deployment of these elements that, together enable us to develop and deliver cloud services via the Internet or a private network.
- Cloud services: Those are expressed by a cloud and delivered over the Internet or a private network. Services range from infrastructure-as-a-service (IaaS), to platform-as-a-service (PaaS), and software-as-a-service (SaaS), and include other services that are layered on these basic service models.

Data security is a common concern for IT, but it becomes a major challenge when users must rely on their suppliers for adequate security [2]. Indeed, the data present the head of computer networking and all the responsible parts in IT try to protect it from certain attacks. In general, the data are treated and stored clearly in the cloud. However, when data flow in the network from their source to their destination through a series of routers, and across multiple networks, they could be intercepted and falsified. Furthermore, the SaaS provider, for example, is solely responsible for data security (storage, transmission, and processing). Moreover, data backup is a critical aspect to facilitate recovery in the event of a disaster, but it has some security problems [2].

Cloud Service Provider (CSP), who is responsible for

providing a secure service, must address issues related to data and network security in terms of data locality, data integrity, web applications security, data segregation, data access, authentication, authorization, data privacy, as well as issues of data breaches, and various other factors [2].

CSP must also ensure data security, and that customers are able to run queries on the data and the results must be protected and not visible to the provider [2][3]. Data encryption, Secret Sharing algorithms and Private Information Retrieval (PIR) are the techniques widely used for securing outsourcing data [3].

CSPs should be able to manage their infrastructure without exposing internal details to their customers or partners. The goal is allowing customers to run their everyday IT infrastructure in the cloud. In fact, many questions that have been raised, on the client side, in terms of:

- Trust on the CSP,
- Capabilities and limitations of the centers administration to access client data,
- Data isolation achieved between Cloud Computing Customers (CCC).

Finally, the most important and strategic question to ask, especially with the PRISM event of the National Security Agency (NSA) [4][5], is the following: can the administrative authorities request a full or partial access to customer data without his knowledge?

Several studies have addressed the problems and challenges of cloud computing [6]. This proves that the provision and search for solutions and improvements of other security practices are an area of active research.

Security in cloud computing is a shared responsibility between the IT department of an enterprise and the cloud service provider. Therefore, even when IT infrastructure can be moved into the cloud, the responsibility for information security cannot be entirely outsourced to the CSP [7].

Today, a static storage system is unreliable because the data will not be available if the storage location is not available for any reason. To avoid such problems, distributed storage networks are used, which consist of several different locations in computers interconnected via the Internet or a private network. However, in such systems, there is no forced data replication. Thus, if one machine is disconnected, data will not be then available [8].

This paper presents the hypothesis that the FRS technique adoption in the cloud computing is more beneficial than the classic model of data storage. In fact, our proposed framework considers these security challenges and seeks to improve data security in the three known aspects of security (confidentiality, integrity and availability). It presents the solution design for handling communications between entities in a single one that will be mainly installed, for example, at the client's terminal. The experiments show more robustness in our proposal, especially in terms of time to recover data.

Our contribution and targets for CCS are :

- To use FRS technique as a principal security mechanism for data storage in cloud computing.
- To create two communication channels between user and cloud computing; each channel transfers a part of data. Consequently, it is difficult for attackers to understand the relations between the two channels and then to reconstruct the original data.
- To propose new scenarios for data storage in cloud computing.
- To protect data not just from the extern hacker but also from CSP.
- To obtain a good security level without contradicting the quality of service (QoS).
- To possibly expand our proposal application to multi-cloud.

The other part of the paper is organized as follows: Section 2 gives a background about cloud computing architecture, discusses about few security issues, threats and challenges of CC, identifies a security technique to protect data from intrusion attacks and poor storage strategies, which is IDA, and discusses current state of data encryption in the cloud. In section 3, the overall design of the proposal is presented. In particular, an interesting security technique FRS is presented as core of the proposal, then a brief comparison is gived between FRS and IDA. Next, we demonstrate how the proposal can be applied via some scenarios. Section 4 presents details of the simulation and the results of our enquiry are then discussed. Finally, section 5 concludes the paper and proposes possible extensions of this work.

II. BACKGROUND

A. CC architecture

Cloud architecture refers to the various components in terms of databases, software capabilities, applications, etc. engineered to leverage the power of cloud resources and to solve business problems of companies.

According to the National Institute of Standards and Technology (NIST), cloud computing is a model for enabling convenient, on-demand network access to shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [9]. As specified in the NIST definition, cloud architectures can be analyzed from two different perspectives: an organizational (the deployment models) or a technical (the service models) point of view [9]. It specifies five essential characteristics of cloud computing, three different service models, and four different deployment models. Figure 1 illustrates this architecture. Cloud computing has some advantages like scalability, resilience, flexibility, efficiency and outsourcing non-core activities. Likewise, the cloud model cannot work for the client without reliable network connectivity and the right bandwidth. Cloud computing helps a company with an advanced business model to accept the IT service without any investment [10].

This is why an existence of SLAs (Service Level Agreements), which include QoS requirements, must be ideally set up between customers and CSP to act as warranty [11]. The CSP offers storage and treatment services in one sever

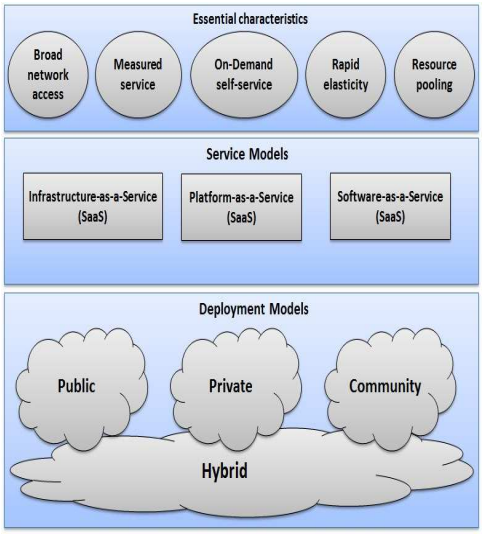


Fig. 1: NIST cloud architecture

among others, and via redundancy, in other servers to keep the adequate availability level for the end user. The data, as one file, can be encrypted at rest and in transit, using encryption algorithms and are then placed on a storage cloud [12]. In these cases, the CSP's administrator has access to entire data in encrypted form, and this entails a real risk for a triad security. Cloud computing, then, offers a multitude of opportunities and services, but also entails some risks.

B. CCS challenges and issues

Having sensitive applications and critical data in a company presents an impediment that avoids the migration to the cloud. Indeed, cloud service providers are more or less transparent about their practices and customer support for incident resolutions. Security threats present an obstacle to the rapid adoption of cloud computing paradigm [13]. The more and more information that are placed in the cloud by users, the more and more they become vulnerable to attacks and threats via the Internet. Therefore, the users need to understand the risk of security in cloud computing. The paper [14] proposes a survey of threats that present various cloud security risks, and presents the service delivery of cloud computing system and security issues. Important efforts have been devoted by research organizations and universities to build a secure cloud computing environment. Despite these efforts, there are a number of gaps and challenges that still exist in this environment [13][15].

In cloud computing, applications and data users are stored in some specific providers' platforms called data centers. This will make users concerned about the security of their data, and, in particular, their privacy. In addition, security threats can occur during deployment. The environment of cloud computing will preserve data integrity and privacy of users and will improve the interoperability between several providers of cloud computing [15]. Indeed, the active data security should be provided on several levels. At each level, it is necessary to satisfy the security requirements in order to preserve data security in the cloud (confidentiality, integrity, availability and non-repudiation). Also, at each level, there must be an insurance of the effectiveness of the measures, their strength, their resistance to attacks and their relevance to customer expectations and cloud administrators.

Resource virtualization is at the heart of most cloud architectures. The concept of virtualization allows an abstract, logical view on the physical resources and includes servers, data stores, networks, and software. This introduces a number of risks which are identified below [16]:

- Complex configuration: adding several layers of networks and systems, which increases the possibility of creating security vulnerabilities through improper configuration of virtual machines.
- Privilege escalation: It is possible for a hacker to access a virtual machine with a lower level of access rights and then attacks a machine, using a hypervisor.
- Inactive virtual machines: they store sensitive data, which creates security risks if these machines are incorrectly accessed.
- Poor access controls: A hypervisor facilitates access to all virtual machines and it may expose all the network systems.

Some of the important typical risks associated with cloud computing are [17][18][19]:

- Loss of governance: customers do not have security and administrative controls in cloud computing, which comprises transferring data to the cloud, and refers to losing control over location, redundancy and file system.
- Vendor lock-in problem: This process will require terming the requirements for the cloud providers to certify that they are able to assure that data migrate from the legacy provider.
- Data Loss or Leakage: It happens when data may be logically or physically detached from the organization or user either unintentionally or intentionally.
- Insecure or ineffective deletion of data: Deleting data from cloud storage does not entail data total removal from the storage or eventual backup media. The data might still be accessed at later time by another user.
- Malicious insider: Cloud architectures necessitate certain roles which are extremely high-risk. Examples include CP system administrators and managed security service providers. In fact, CSP personnel with privileged access can have access to customer data or even, dump the memory for extracting the bitlocker and/or the encryption/decryption keys.

Specifically, common safety issues around cloud computing are divided through four categories [13]:

- Access: it comprehends the concern over access to cloud control (authentication, authorization and access AAA), encrypting the communication (data), and the management of user identity.

- Cloud infrastructure: includes concerns about virtualization, storage and network vulnerabilities that may be inherent in the code and hosted in the cloud computing software. It can also include physical security aspects of the data center.
- Data: refer to the concerns about the integrity, conservation, availability, confidentiality and privacy of users.
- Compliance: because of its size and its disruptive influence, the cloud must address some issues related to the regulation like the safety audit, location data, non-repudiation and traceability.

Cloud computing is an outsourcing concept and a remote applications and data processing that is growing increasingly. Nevertheless, there are still challenges in terms of administration, interoperability and security tools. These challenges must be addressed before users can enjoy all the benefits of cloud computing and place their trust [7][13].

These CCS issues and challenges require some efficient mechanisms of data redundancy to protect them.

C. Related works

Security and reliability issues in distributed systems have been enquired for several years using some techniques. In this section, a technique is presented and that aims to tolerate both accidental and intentional faults, and also have others advantages to increase the system performance, especially the current statement of cloud computing security storage.

1) Information Dispersal Algorithm technique: The IDA is a technique applied to ensure reliable and secure storage and transmission of data in distributed systems [20]. Rabin describes the IDA as a tool for cutting a file into several parts based on some parameters according to the desired complexity [21]. Among the IDA applications:

- Secure and reliable storage of information in computer networks and simple hard drives,
- Fault tolerance,
- Transmission of information in computer networks,
- Communications between processors in computers working in parallel mode.

This allows the load balancing on the storage and transmission.

IDA presents a replication protocol or a theoretical coding technique, it allows reducing the cost of replication storage and the bandwidth size, but it is not able to update small portions of data files efficiently [22][23].

The IDA technique dispatches a file F of length $L = |F|$ into n pieces (segments or parts) F1, F2... Fn, each of size L/m with $(m < n)$. It is therefore a more efficient method compared to the traditional operation of transmitting or storing the data [21]. The Rabin's IDA is a technique ensuring high confidentiality [20]. To protect the file against the illegal modification, it is recommended to encrypt the file before the dispersing operation [21]. The algorithm IDA (n, m) is considered as a tool for converting a file into multiple files and any m files of n files are sufficient to recover the original file [24].

In [20], the author stated that there are two levels of confidentiality when applying IDA:

- Weak confidentiality: possibility of reconstructing the original file from fewer than m files; in the case of adoption of an arbitrary non-systematic erasure code,
- Strong confidentiality: it is necessary to have m files to form the original file.

The work presented in [24] proposes an efficient algorithm IDA (n, k) for the case $n/2k < n$ over Fermat field GF(2r + 1) for applications correction codes. IDA has proposed fewer operations than the algorithms based on FNT. In the context of better processing performance in a distributed system, the work [25] mentioned the interest of the IDA technique used in both iStore and FusionFS systems. Also, this technique was introduced into the fundamental management layer in the structural model of the integration information platform of high quality teaching resources in universities based on the cloud storage [26].

The concept of IDA is like Shamir's work who designed the first system of sharing in 1979. He published a regime based on polynomial interpolation. His goal with the plan is to take t points on the coordinate plane, and with these points, a polynomial $q(x)$ such that $y = q(x)$ for each of the given points. As an application, he showed how to divide data D into n pieces such that D was easily rebuild with the knowledge of k pieces, provided that knowledge of $(k - 1)$ pieces does not reveal any information about D. This technique allows the construction of robust key management schemes for cryptographic systems that can operate safely and reliably [27][28][29].

The work [3] implemented the IDA technique in an IaaS Cloud OpenStack. The environment of this experiment is composed of Linux machines as nodes (client and controller). The working file was a database of four million records. The IDA technique (n, k) is applied to this database file in the client side. The generated files were placed randomly on the storage server using SCP (Secure Copy Protocol) so that no server had k files. This implementation shows that Rabin's IDA is able to successfully rebuild the entire data even when $(n - k)$ files are unavailable. It was observed a considerable decrease in the dispersion time when the congestion decreases (k increases and therefore the file size decreases). The recovery time remains roughly constant with a maximum variation of 0.4 seconds with a best time of recovery at the threshold value of eight (expense at this time is 25%).

a) Example: Let F be a file, the IDA $(8; 4)$ approach is applied which involves the generation of 8 pieces of size $|F|/4$. The total size is $8/4|F|$ (see Figure 2). These parts are then distributed across three data centers so that such of them cannot receive more than three pieces $(< m-1 = 4-1 = 3)$. Even more, it allows preventing the rebuild the original file by the data center administrator. Figure 2 shows the distribution of different parts generated from the file F on three data centers.

Figure 3 shows the possibility of reconstructing the original

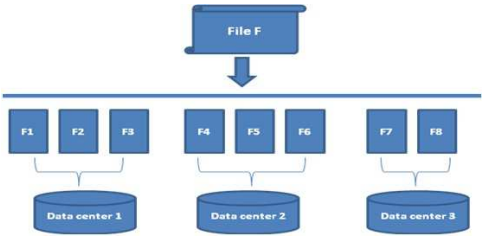


Fig. 2: IDA application in data centers

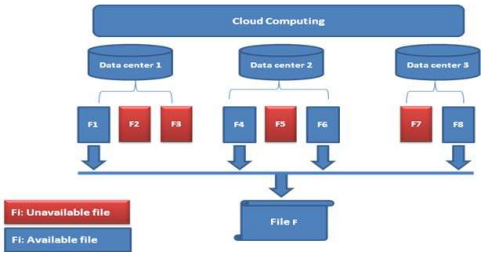


Fig. 3: Retrieving the original file F from the Cloud

file from four parts. It seems advantageous in the case where there is a loss of some parts in the data centers.

In this scheme, the four files are lost (or unavailable) for some reasons and there are four other files available and in good condition. In this case the user can download these four pieces from three data centers to build the original file. This shows that the IDA technique offers significant flexibility in the case of intrusion attacks or denial of service.

2) Data encryption in the cloud: Cloud computing is one of the enormous platform which provides storage of data in very low cost and with availability for all time over the internet. Encryption is considered the baseline technology that IT experts agree is the cornerstone of security. The risks to data security in the cloud are presented in two states: data that is at rest (or stored in the cloud) and the data is moving (or moving in or out of the cloud). Many clouds propose to use secure web connections, such as transport layer security (TLS) or HTTPS encryption, to transfer data from the user's terminal to the web application [30][31]. Some cloud storage applications, such as Barracuda's Copy.com, allow the user to create a secure link between their corporate network or mobile systems and the cloud storage application [32]. Once the data reaches the cloud providers' servers, the application provider generally encrypts it to secure the data at rest.

However, there is another challenge in this case. In the past, one of the most important tasks the IT manager was managing encryption keys [33]. In order to keep data secure, the recommendation is to separate the encryption key from the encrypted data.

For cloud computing provider, there is some issue about the management of encryption keys location. Normally, encryption keys should be kept on a separate server. A backup of all keys should also be kept in an offsite location in case of disaster. By the way, encryption keys also need to be refreshed regularly to keep a high level of data security [33].

In the beginning, many companies felt comfortable allowing the cloud provider to manage encryption keys, believing that security risks could be managed through contracts, controls and audits. Over time, it has become apparent, however, that cloud providers cannot honor such commitments when responding to government requests for information [34].

A lot of cloud providers do not just store client data, they do things with that data especially with the NSA Prism event [4]. By the way, the cloud does not allow storing the data encrpyted by user depending on the type of service. For example, Gmail as a SaaS, do not allow the mailing an encrpyted file as attachment for an unkown reason.

It is important for companies to create rules to identify what information rises to the need of encryption and what data can be stored safely in plain text [35]. User could have indeed role with attached key for accessing the confidential data.

In fact, protecting data at rest is essential. The best choice is to encrypt sensitive data when it is created so that when it is stored in a data center, be it locally or in the cloud, it will be protected. This is why, encryptions should be considered a standard business practice.

By the way, the Cloud Security Alliance, in its Security Guidance for Critical Areas of Focus in Cloud Computing, recommends that sensitive data should be [36]:

- Encrypted for data privacy with approved algorithms and long, random keys
- Encrypted before it passes from the enterprise to the cloud provider; should remain encrypted in transit, at rest, and in use
- The cloud provider and its staff should never have access to decryption keys.

Besides, another type of encryption technique exists: An homomorphic encryption scheme. It is a mathematical technique that allows operating over encrypted data, without ever needing to decrypt it. It is considered as the ultimate cryptographic tool to build more secure cloud computing services that respect the user's privacy. It allows to confidentialy share data, and the encrypted data can then be processed without ever needing to decrypt or reveal it [37][38].

The first fully homomorphic encryption system, built by Craig Gentry (now an IBM Research cryptographer), was incredibly slow, taking 100 trillion times as long to perform calculations of encrypted data than plaintext analysis [39].

The paper's abstract [40] explains how this technology could be used in the cloud to process encrypted data without needing the decryption keys: "The encryption ensures that the data remains confidential since the cloud does not have access to the keys needed to decrypt it. Nevertheless, we will show that the cloud service is capable of applying the neural network to the encrypted data to make encrypted predictions, and also return them in encrypted form."

As confirmed by Professor Kristin Lauter, principal research manager at Microsoft, there is still a lot of work to be done, but the initial results look very promising and could be used for a kind of secure machine learning-as-a-service

concept, or on specialist devices for medical or financial predictions: information sent over to the neural network remains encrypted all the way through the processing [39].

Therefore, the encryption technique is essential for data security but not sufficient. It needs to be modified, developped or to modify the way of how use it in the Cloud. The encryption assures that the data reside confidential since the cloud does not have the keys needed to decrypt it. Finally, the data are in a danger not from only the extern attack but also from the intern attack especially the data's confidentiality can easily be lost by the CCP administrators.

III. OUR PROPOSAL

A. Introduction

Based on a distributed system, our proposal defines a new approach in cloud computing. It mainly gives guarantees to the CSP and especially to users who require their data to be more secured.

The idea of creating a decentralized system is not new. Indeed, the integrity, confidentiality and availability of data with scalability in a distributed system probably requires a fragmentation and dispersion process in different nodes of the system [21][41]. The main idea of our approach is to dispatch the data into parts. Each of them is sent via a different link to gain in terms of processing speed (parallelism) and security.

The basic objective is to make the attack process very difficult when data transfered to cloud computing since it requires two steps:

- Knowing the communication parameters of the two channels (or more) between client and CSP.
- Establishing a relation between different fragments in both communication channels (and more), which requires a lot of treatments and time.

The cloud computing contains two types of servers: processing server and storage server. Therefore, our model's objective is to treat certain scenarios involving just storage servers. At the same time, the FRS technique will be adopted as the main key to the framework because it offers a minimum processing time compared with the IDA technique.

Three cases are differentiated in the framework when there is a need for the CCS:

- Backup data: the data will be sent to storage servers based on the FRS technique.
- Internal treatment: the treatment is local. For example, the middleware can easily collect data based on the FRS technique (data recovery), and then the user's terminal can handle the treatment.
- External treatment: the processing is delegated to the cloud because the cloud capacity is much higher than the terminal one.

However, in this paper, the target is to evaluate just our framework for the storage data in the cloud computing. The results of this study also highlight the crucial role of FRS in the CCS.

B. Fragmentation-Redundancy-Scattering technqiue

The problem of data availability in a traditional backup strategy was one of the motivations of the work discussed in [42]. The FRS technique consists in fragmenting confidential information in order to produce insignificant fragments and then scatter the fragments so obtained in a redundant fashion across a distributed system like data centers according to a particular algorithm [41][43]. The paper [8] described the principle of FRS and gave another name that best described the main steps of this technique: Encryption-Fragmentation-Replication- Scattering (EFRS).

First, the FRS was applied to the persistent file storage, and it used to implement distributed management system security in terms of authentication and authorization. It was then applied to processes that handled sensitive information, by using a design approach by object [43][44].

All fragments may never be on a single storage node. Thus, if some hackers manage to recover some of the fragments, the attack will probably be useless. Even if the intruder manages to get all the fragments, the task of fragments assembly in the correct order and decryption will be an almost impossible mission [44][45]. Thus, the system tolerates a passive intrusion. The problem of active intrusion was treated by using the hash verification. The system verifies the hash value of each fragment as it is recovered. If the hash value is incorrect, the fragment will be discarded and the system will try to find another replica in another site [45]. A forced replication fragment in multiple servers allows the continuation of the system even in case of the failure of some storage nodes [46].

The security level of data while they are processed, transferred, and stored depends on the service provider. Therefore, data leakage happens when they get into the wrong hands while they are being transferred, stored, audited or processed. The main benefits of FRS are [46]:

- When FRS is used without encryption, but the fragments are stored on n secure servers, the attacker must interfere into all n servers instead of one.
- FRS is more effective against the denial of service (DoS) or destruction of data.
- When the data are encrypted and then follow the process of FRS to generate n fragments, the intruder must cryptanalyze combinations of (n!) fragments.
- Redundancy in FRS provides a mechanism for fault tolerance, and thus a mechanism to tolerate intrusions.
- Data replication, while using FRS, introduces fault tolerance without the risk of further exposure.

In [47], the developed idea is based on the object fragmentation at design time to reduce data processing in confidential objects. Non-confidential objects can be produced at design time, and then be traded in untrusted shared computers. Classified material should be treated in positions of trust unshared ones.

The FRS technique aims to avoid successful intrusions in one or more non-reputable sites [48]. This approach does not presuppose a particular type of security policy.

Different types of policies can be implemented by security sites that control access to the servers for processing and storage. A distributed approach to managing security policy can be applied in this context [49][50].

The FRS allows strengthening the data confidentiality, integrity and availability [24][41][46]:

- Confidentiality: an intruder will have to collect all the fragments in the correct order before attempting cryptanalysis.
- Integrity and availability: an intruder should alter or destroy consistently all replicas of a fragment to be able to modify or destroy data.

The application of the FRS technique involves the following operations:

- Encryption before fragmentation;
- Secure management of the encryption key;
- Fragmentation of encrypted fragments of fixed size data;
- Secure naming of fragments;
- Fragments diffusion from the user site to all storage sites;
- Implementation of storage sites in a distributed algorithm to select which sites will actually store each fragment.

The fragmentation and scattering technique, when applied to file storage, involves cutting every sensitive file into several fragments in such a way that one or several fragments (but not all) are insufficient to reconstitute the original file [48]. The number of copies depends on the file criticality defined by the user [44][48]. The user site displays in random order all the fragments of each page to all storage sites. Then, the fragments are stored in several copies on different distributed sites, which can be viewed as fragment server machines [51]. Figure 4 illustrates this processing. The name of each fragment is generated by a hash function from the encryption key, file name, index of page and the index of the fragment. It prevents knowledge of the correct order of fragments of a given page by an intruder [44][48][52]. Figure 5 shows the cycle of file processing during the fragmentation operation. Thus, no information about a fragment can be derived from its name.

At this stage, the question that arises is how to build a map of effective dispersion of minimum computer resources used during the reconstruction of the file. In [53], the article proposed a scattering technique based on a tree structure. The dispersion map proposed the use of a Huffman encoding process based on the use of the frequency of the file. In the same direction, another study proposed two algorithms developed to maintain a constant number of replicated fragments: one based on the game of life, and the other based on roaming ants [54]. Each of them respects the following criteria:

- To maintain an acceptable number of copies of fragments;
- To resist the malicious attacks and multiple node failures;

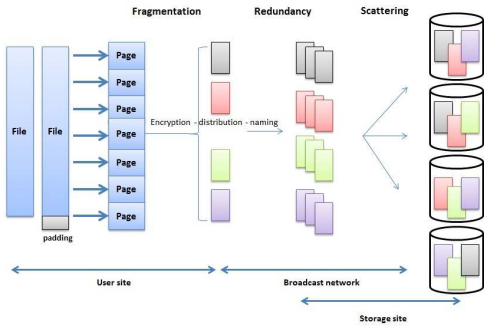


Fig. 4: FRS applied to persistent file storage

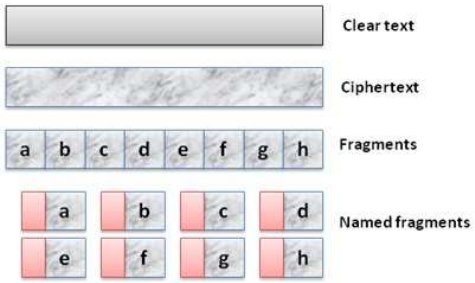


Fig. 5: Transformation cycle of data in the FRS

- To preserve the anonymity of data holders.

This study remarks that the number of replicas fragments generated by the game of life algorithm was higher than the ant swarm algorithm.

During the read operation, the user site reconstructs the names of all the fragments of pages that must be recovered (using a hash function), and diffuses in random read requests to all storage sites that possess a copy of the fragment. If all copies of each fragment are identical, the user site can easily restore encrypted pages, decrypt and verify the checksum. If different copies exist for a fragment, the user site recreates several encrypted pages and tries to decipher until it gets a correct checksum. Only the name of the fragments allows to find their location and this information is calculated during fragmentation operation (based on information as the key to fragmentation, file name, etc.) and dynamically recalculated using the same information at reassembly operation [41][43][44].

In [42], an approach was described, based on FRS in the context of a peer-to-peer architecture where each agent (client and/or server) has the ability to request data storage service from other agents to store elsewhere. The inconvenience here is that nodes (agents) constituting this system have a dynamic behavior (connectivity). The node can have all the fragments associated with a file without problem because the node cannot tell the difference between fragments belonging to a given file.

The effectiveness of FRS appears when the attacker is incompetent to differentiate between fragments of the same file across the transited network flow. This requires that the sending of fragments exchanged between archive sites and user sites should not be sequential in their normal order.

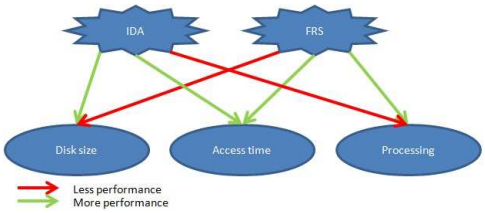


Fig. 6: FRS vs. IDA

The fragments should be transmitted at intervals sufficiently spaced. If necessary, this rate must be increased artificially [45]. The FRS performance depends on the fragmentation granularity. The FRS approach is a method to take advantage of the division file and distributed system to implement reliable applications that handle confidential data.

1) Comparison with Information Dispersal Algorithm technique: The IDA (n;m) algorithm is a tool for converting a file into multiple files and any m files from n files are sufficient to recover the original file. As a performance advantage, this technique provides an adequate level of security. Figure 6 illustrates in brief the comparison between the two techniques IDA and FRS. Based on the conclusions of various scientific research papers, it is argued that the IDA approach does not need the important disk space. With regard to the additional treatment, the FRS processing complexity is less than the IDA one. The fragmentation of a file (including encryption, distribution of bytes in fragments, naming) is faster than encrypting a file with conventional techniques [44]. The IDA technique necessitates high computation on large matrices especially if these calculations are made on traditional workstations. Finally, both techniques have the same access time to file by parallelization of access to different fragments on the storage sites [44][47][48].

C. Applying FRS for data storage and recovery

In cloud computing environment, the user needs some services related to data storage. Before using this service, the user needs some security mechanisms to data access (authentication and authorization). In fact, in this section, several required scenarios related to mono cloud are pesented in order to benefit from storage service.

1) User authentication and authorization: In order to enable security functions to tolerate faults and intrusions, despite the fact that these intrusions are made by security administrators, these functions are implemented as a distributed security system that contains several security sites managed by different administrators. The implementation is based on majority vote and threshold scheme algorithms. Therefore, any k of the n parts are sufficient to reconstruct the original secret [28][55][56][57]. As a majority of these sites are neither defective nor penetrated by an intruder, the security functions will be performed correctly, and no confidential data will be revealed.

The proposed framework is based on the use of two backup sites at least, belonging to a one or more of the clouds. It requires authentication and user authorization from cloud(s) contributing to serve the customer. This registration should be done separately in each cloud or each entity of the cloud. Authentication is perhaps the single, most common requirement of any application. First, the user must authenticate by the CSP to access the services. Then, he must be logged separately on the various security sites, and a user's authenticator is stored on each security site. This authenticator can be a secret shared by the user and site security (password) or biometric information characterizing the user (a fingerprint), or public information corresponding to a secret known only by the user. On the user side, the shared secrets should be stored in a specific device like smart cards or usb flash driver. Our framework requires a seperate registration for each security site, which significantly increases the level of security in our system; since a malicious security administrator cannot, without the help of other administrators, pass for a user, and he cannot create a new user (it would not be recognized by the other security sites). His authority is limited to his own security site and he has no power over other security sites. This approach provides some separation of powers. Figure 7 shows the process followed during authentication in our framework.

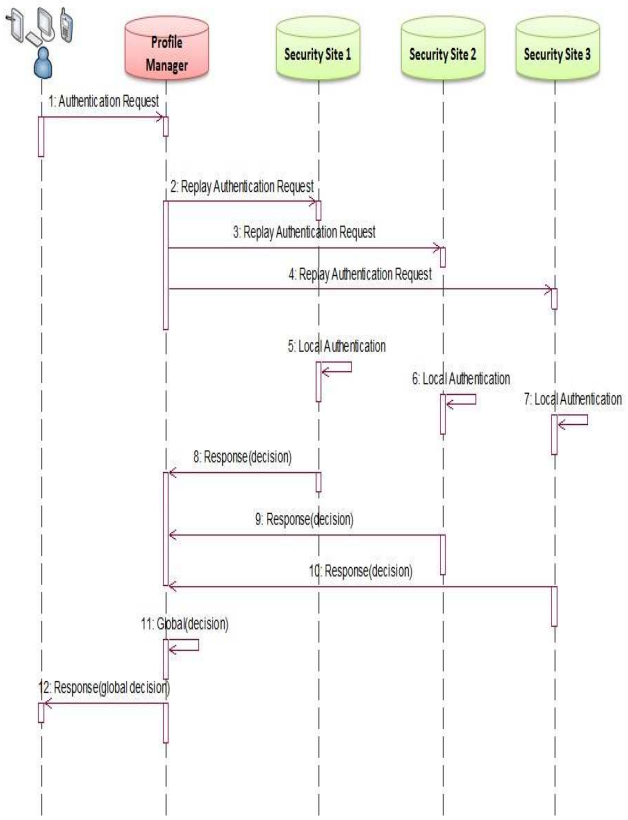


Fig. 7: Authentication scenario in our CCS framework

The user sends a request to an entity named Profile Manager to find the security sites to be able to authenticate (1). Certainly, there are many profile managers to ensure this security mission. Thereafter, the user sends authentication to multiple security sites (2, 3 and 4). Next, each site runs an independent security authentication protocol according to the local authentication scheme (5, 6 and 7) to verify the

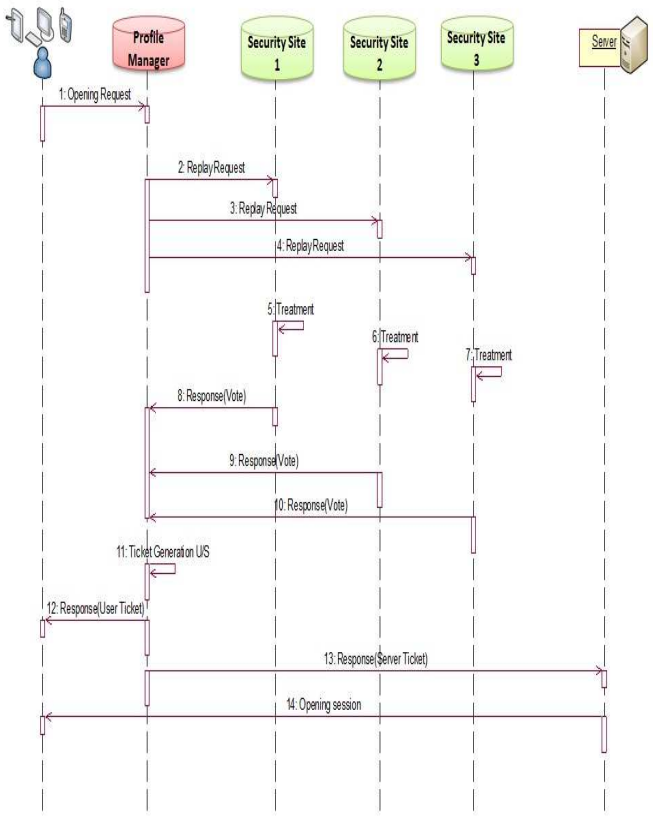


Fig. 8: Authorization scenario in our CCS framework

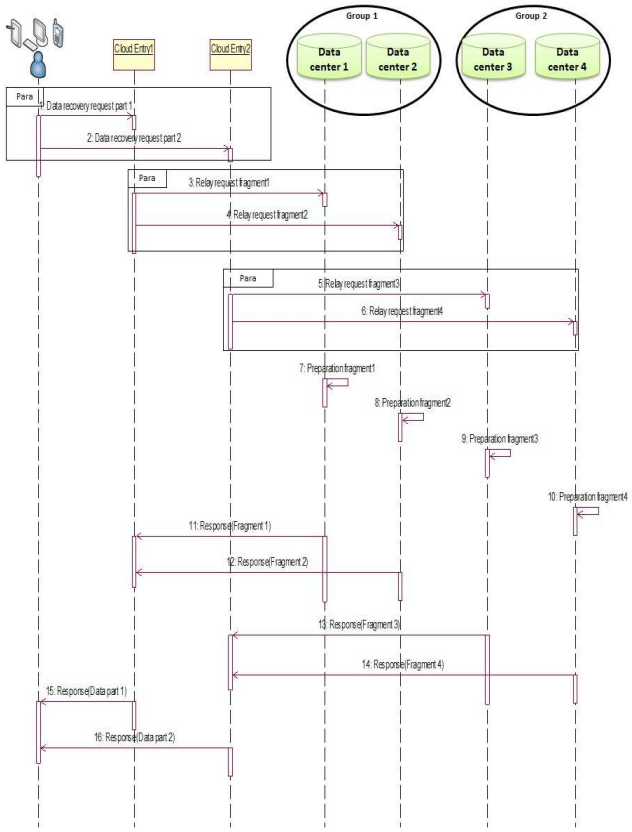


Fig. 9: Backup and data recovery scenario in our CCS framework

identity disclosed (password, biometrics, etc.). Then, these local decisions are shared to obtain an overall decision (8, 9, 10 and 11). Once the global decision is made, a response is sent to the user (12). Finally, the user can send queries to access other servers or objects of cloud computing. Thus, the user obtains session keys, one of them for site security for future accesses to every site security. All other requests will be encrypted by the session key. When there is a need to access an object, the authenticated user sends a request to the security server that allows or denies access. Figure 8 illustrates the protocol for authorization.

2) Storage and data recovery: In this scenario, two ingresses are used for access to cloud computing. The concept of the framework is to create two groups of data centers. Each group is built and assigned to ingress of the cloud (E1 or E2). To do this, during the authorization phase, the middleware (client side) must select two ingresses from the entry proposals of cloud (It is assumed that the cloud contains more than two entry points to its IT infrastructure). This can be done in a random or alternative manner in every time the user wants to establish a connection to a data center; he can choose a different entrance to his previous connection to a data center. Each data center is connected to the client via the two chosen entries (1 or 2) and each of them presents a group in our Framework. Figure 9 illustrates the proposed approach.

Therefore, any possible transaction (backup or data recovery) between the client and the cloud passes through the two communication channels between the client and the two chosen entries of cloud that serves as a bridge to the data center. These channels should be secured by establishing VPN (Virtual Private Network) connections. The figure 10 illustrates the process of data storage in our proposal.

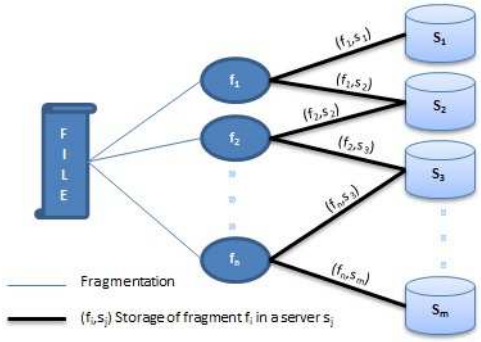


Fig. 10: Logical process of our proposal

Certainly, with many servers that present container of user's data, the confidentiality and availability are guaranteed and then security level rises relatively to actual strategy of clouds.

Let's suppose the probability $p(< 1)$ where an object is affected by an attack (integrity and/or confidentiality and/or availability) in a server. In fact, in the case when an entire file is

hosted in a server and is duplicated in another server, then the probability that this file is affected is $(p.p = p^2)$. However, in our proposal, a file is converted to n fragments that are stored in servers with duplication. Therefore, in this case, the probability that this file is affected is $(\prod_{i=1}^{n} p\ )(\prod_{i=1}^{n} p\ )$=$\prod_{i=1}^{n} p^2$. Consequently, our proposal offers more security than the traditional case because $\prod_{i=1}^{n} p^2 \ < p^2$.

D. The proposal in multi-clouds

This approach can be applied for the multi-clouds. Thus, several clouds can be used depending on the capacity and design of the overall architecture of the clouds. The multi-clouds architecture is the environment where there is cooperation between CSP (see Figure 11). The cooperation between clouds carries more benefits for the customer in terms of security and performance and this is achievable according to the SLA established between the client and the clouds.

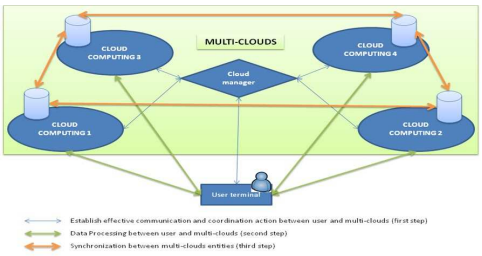


Fig. 11: Multi-clouds architecture in our CCS framework

Our projection of the Framework on the multi-cloud requires the addition of a new entity called cloud Manager (CM) that manages the communications between the client and the clouds. When constructing a communication context with clouds, the user must provide the list of clouds to the CM when he wants to communicate. Then, the scenario of creating a communication context with clouds begins between the user and the clouds selected, using the CM and then, it follows the same procedure as in the case of a single cloud. By the way, many possibilities of CM placement can be proposed:

- CM is implemented in user terminal. In this case, if the CM crashes, the user can restart, re-install or update the CM if there is a need.
- CM is a role that can be assigned in any cloud computing based on SLA (Service Level Agreement). In this case, if the CM crashes in one CC, another cloud computing can be chosen to operate as CM. The choice between cloud computing and CM can be based on priorities, service type oroperation's time, etc.
- CM can be implemented in a company's proxy. In this case, if the CM crashes, the IT department switches to another proxy as backup for the first one.

Independently of previous choices in CM's places, the CM should be available to orchestrate the communications between user and clouds. Certainly, policy manager for access to different clouds is required. The work in [49] treated a framework called Policy Management as a Service (PMaaS) that offers customers the ability to manage access policies to services running on a cloud infrastructure that can be accessed via user interfaces. The framework consists of four main components: cloud user, Policy Management Service Provider (PMSP), CSP and requester.

Other solutions can be integrated to increase the security level discussed in [58]. One consists of establishing a collaborative access control framework called PolyOrBAC. This approach provides each organization belonging to the CII (Critical Information Infrastructure) the ability to collaborate with others while maintaining control on its resources and its internal security policy. A contract is signed between the service provider and the service user to clarify certain degree of interaction and performance. The contract describes the parameters, functions of Web Service (WS), responsibility of each party and the security rules to control interactions. When running, respect for all interactions with these security rules is verified.

Another interesting point to discuss in the multi-cloud is about the impact of this framework on latency. This paper does not cover all QoS problem. However, a brief discuss about latency for this proposal in multi-clouds is necessary. The papers [59][60][61][62] analyzed some of the problems and challenges for achieving real-time cloud computing. QoS management in the cloud computing is linked to the problem of allocating resources to the application to guarantee a service level along dimensions such as performance and reliability [59].

QoS is considered in every side of the network - the user, the backbone network access, and the IP core network. In fact, the QoS depends on both cloud computing and the operator network. The paper [60] investigated if it is possible to use latency as an indicator for the other QoS parameters as throughput and jitter. It concluded that it was not possible to find a consistent relationship between latency and the other parameters. Also, the paper [61] presented PriorityMeister as a system that combines priorities and rate limits to provide tail latency QoS for shared networked storage in CC, even with bursty workloads. The paper [62] presented RT-VMs as a technology allowing virtualized applications to meet QoS constraints as stated in contractual agreements among customers and providers, formalized in proper SLAs.

In our framework, the requests for fragments increase compared to the case where the requests are for one file. However, the overall size of the received user will be the same for the case of one single file. The reconstitution of the original file is faster given the great performance of terminal processing. The problem is the amount of queries to request the fragments. In this case, a compromise is:

- File size,
- Fragments number,
- Number of storage servers,
- Location of servers.

By the way, the impact of the proposal can be decrease by reducing the fragments number and requests number. Normally, when the fragments number increases, the security level

raises. Certainly, this parameter depends on data security degree. Also, the requests number can be reduced by aggregation of fragments' names in few requests in each channel, and this entails a reduction of network traffic.

Finally, before the multi-cloud can offer a service to customer, it should verify everything about the QoS management via monitoring of some parameters that include latency, jitter, packet loss, and bandwidth.

IV. SIMULATIONS AND EVALUATIONS OF OUR PROPOSAL

A. Introduction

In this section, The simulation of the proposed CCS framework is presented for evaluating the performance of the EFRS technique as a dependable and secure solution for cloud storage. To evaluate the robustness, security (availability) and performance of the proposal during data exchange in cloud computing, Netlogo has been chosen as a modeling tool [63][64]. This allowed us to investigate the performance of our framework under various operational conditions.

Basically, during the simulation, some parameters are changed to evaluate the robustness of the framework and make a comparison with a classic model of storage data in the cloud computing. The Classic model is the case where the total file is transferred and saved in one place without cutting in fragments. Likewise, the other copies of this file are transferred and saved in other servers. In the proposal, the classic model is obtained if the fragments number is equal to 1. The figure 12 illustrates the general rule during simulation.

Subsequently, the target is to prove the strength of our framework in difficult conditions (increasing loss percentage of fragments) and have a performance's cartography of all simulations. Also, some conditions are made in automatic routine to evaluate the proposal.

In the simulation, there are:

- Five hundred files, each of them has a size of 150 Mo.
- One hundred servers, each of them has as maximum size of 1 To.
- The replicas number of fragments (or files for Classic model) is equal two.

The AES (Advanced Encryption Standard) is utilized as the symmetric cryptography algorithm before fragmentation. Because the length of plaintext and ciphertext are the same for AES, the advantages of the scheme are evaluated without adding any constraints about encryption operation complexity during the simulation.

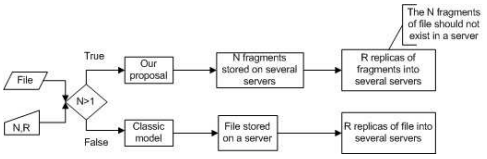


Fig. 12: The general schema of simulation

Some parameters are proposed and that can be changed in the simulation to have a result in output file. The parameters are:

- Number of servers: the number of storage server.
- Server size: the maximum size in the storage server.
- File size: the size file; to simplify, all created files have the same size (150 Mo).
- Fragments number: presents the number of fragments generated by the EFRS technique from the file can be imposed.
- Replicas requisite: the number of replicas fragments that should be normally kept in the cloud computing.
- Maximum buffer: the total size of the treatment data by second.
- Maximum buffer reference: the rest of the buffer during time unit of treatment (here, it is one second).
- FRS (OnOff): switches between our proposal and the classic model.

Here in the simulation, a tick is a time unit, so in this case, one tick is considered as one second.

Firstly, the tests adopted for the evaluation of the Framework are:

- One for classic storage strategy in the cloud computing.
- Second for new storage strategy in our CCS framework. In this case, five simulations are made. In each one, the fragments number is changed: 5, 10, 15, 20 and 25.

Next, the results and synthesis of these tests are presented to valorize the framework. Three options exist in all these simulation studies and that concern the number of servers failed in each trigger event of server failure.

B. Scenarios and results

Under the same experimental conditions, the global target of these scenarios is to show the difference between our proposal and the classic model by measuring some indicators for performance and security of cloud computing under some difficulties. Here, some events are made in automatic routine. Each 60 seconds, a failure server event is simulated. In fact, after each 60 seconds, a number of servers is shut down. The target is to evaluate the behavior of our framework and classic model. Thus, the Mean Time Between Failures (MTBF) for the simulation is 60 seconds.

There are three cases:

- Each 60 seconds, a server is shut down.
- Each 60 seconds, two servers are shut down.
- Each 60 seconds, three servers are shut down.

In each case, six simulations are made:

- Simulation 1: classic model (Sim1).

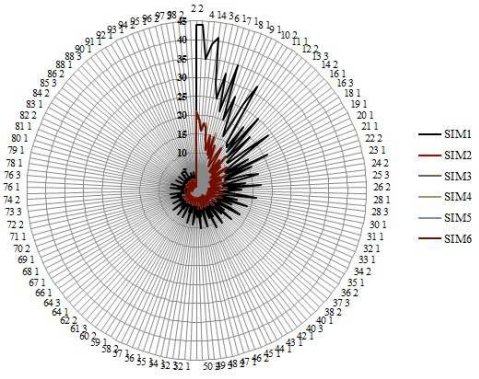


Fig. 13: MTTR comparaison between classic model and our proposal

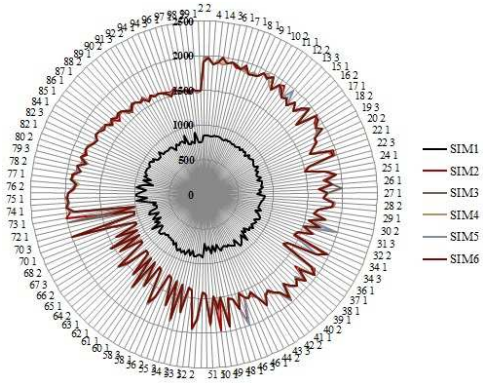


Fig. 14: RE comparaison between classic model and our proposal

- Simulation 2: our proposal with fragments number equal to 5 (Sim2).
- Simulation 3: our proposal with fragments number equal to 10 (Sim3).
- Simulation 4: our proposal with fragments number equal to 15 (Sim4).
- Simulation 5: our proposal with fragments number equal to 20 (Sim5).
- Simulation 6: our proposal with fragments number equal to 25 (Sim6).

Then, eighteen simulations are made in order to describe more the behavior of the model. So the simulations period is between thirty three and one hundred minutes. A radar chart is utilized to compare the aggregation values of multiple data of simulations for three cases.

When the failure event of a server is triggered, immediately the recovery processing of data tries to recover data from other servers that contain a copy of loss data. The target of this processing is to return to initial state before the server failure event. Also, Recovery Efficiency (RE) is defined as Maximum Data Size (MDS), that can be recovered, during Mean Time To Repair (MTTR) of the system. Then:

$$RE = \frac{MDS}{MTTR}$$

Figure 13 illustrates the variation of MTTR for all simulations. Here, the proposal reacts faster than the classic model to stabilize the global situation of the environment. Also, independently of fragments number, the proposal (sim2, sim3, sim4, sim5 and sim6), for three cases, has a stable behavior and approximately the same MTTR during all simulation. Finally, it is remarked that the maximum delay of reaction of the proposal is 20s whereas 45s for the classic model.

Figure 14 illustrates the variation of RE for all simulations. The RE of our proposal is higher than the classic model. Also, independently of fragments number, the proposal has a stable behavior and approximately the same RE during all simulation. As previous metric of MTTR, the RE of the proposal is always more than the classic model. Then, this proposal can be beneficial for the critical applications.

In the proposal, the confidentiality is guaranteed contrary to classic model. In fact, this proposal forbids storing all fragments of a file in one place. Indeed, the storage cloud security is based on encryption. The encryption and decryption keys are saved in memory in some places like USB flash drive. Some attacks, against the computer's memory, provide full access to protected (encrypted) information stored in the most popular types of crypto containers. The encryption keys can be derived from hibernation files or memory dump files. For example, while BitLocker may indeed protect against opportunistic stealing of a computer that is turned off at the time, there are several plausible scenarios for targeted attacks [65][66]. There are many ways available to acquire the original encryption keys. Then, the recovery of keys is easy and then the data are in danger of being lost or falsified for classic model.

In this respect, it is worth nothing that it seems difficult to break our proposal model, many obstacles exist:

- Knowing and access to different servers that contain the fragments.
- Search and collect fragments of file among existent fragments in the cloud; for example find ten fragments of a file among billions of fragments.
- Knowing the order of fragments; if n fragments are need it to generate the original file, so $(n!)$ operations are necessary to find the correct order of fragments.
- Knowing the encryption keys.

In conclusion, the proposal shows superiority in these comparisons. Also, it is remarked that the classic model needs more time than the proposal to stabilize the system situation. The synthesis is:

- Confidentiality in our proposal is higher than the classical model.
- Availability is approximately same in the two models.
- Consummation of memory is generally the same in the both models.
- Recovery efficiency of our proposal is more important than the classic model.

Furthermore, these results also indicate the important role of fragmentation operation, because when there are more fragments, the risk of loss data decreases.

V. Conclusion

In this paper, some issues related to data security have been discussed, also some concepts related to intrusion tolerance have been quoted. A technique IDA and the encryption data in the cloud have been discussed. Furthermore, the proposed CCS Framework based on the FRS technique have been presented in several situations to satisfy most user needs in terms of cloud computing operating, describing some scenarios (Authentication, Authorization, data backup and recovery). Furthermore, the robustness of the proposal have been evaluated by making a comparison with existing classical scheme. According to the results, our CCS framework presents an advantage over the classic model in terms of robustness. This study sheds some light on some advances in terms of data security and performance of cloud environment. The results demonstrate that this architecture can thus optimistically withstand a series of multiple failures.

Currently, other scenarios are being implemented in the real environment to assess the CCS framework in terms of security and performance compared to the current state of CCS. Consequently, as an extension of this work, a middleware will be developped and will be deployed in user terminals to handle all communications easily between the infrastructure provider and the final user. Future works should also examine other potential factors that might influence the global performance of cloud computing like the fragments number and dispersion algorithms. Also, It is planned to develop a dynamic approach in multi-cloud to increase performance especially QoS, based on the CCS framework.

References

[1] Vic (J.R) Winkler. *Securing the cloud - cloud Computer Security Techniques and Tactics*. Elsevier Inc, 2011.

[2] Abdul Nasir Khan, M.L. Mat Kiah, Samee U. Khan, Sajjad A. Madani. *Towards secure mobile cloud computing: A survey*. Future Generation Computer Systems (2012), doi:10.1016/j.future.2012.08.003.

[3] J. Sinduja, S. Prathiba. *Modified Security FrameWork for PIR cloud Computing Environment*. International Journal of Computer Science andMobile Computing-2013.

[4] Clara Leonard. *PRISM : la NSA argumente, le Guardian fait de nouvelles révélations*. From http://www.zdnet.fr/actualites/prism-la-nsaargumente-le-guardian-fait-de-nouvelles-revelations-39791924.htm. ZDNet, Jun 28,.2013. consulted Nov 20, 2013.

[5] Glenn Greenwald, Ewen MacAskill, Laura Poitras. *Edward Snowden: the whistleblower behind the NSA surveillance revelations*. http://www.theguardian.com/world/2013/jun/09/edwardsnowden-nsa-whistleblower-surveillance. The Guardian, jun 10,2013.

[6] Almokhtar Ait El Mrabti, Anas Abou El Kalam, Abdellah Ait Ouahman. *Data Security In The Multi-Cloud*. The International Conference On Networked Systems May 2-4, 2013, Marrakech, Morocco. The First International Workshop on Security Policies in cloud Environment (PoliCE2013)

[7] Keiko Hashizume, David G Rosado, Eduardo Fernndez-Medina, Eduardo B Fernandez. *An analysis of security issues for cloud computing*. Hashizume et al. Journal of Internet Services and Applications. SpringerOpen Journal. 2013, 4:5

[8] A.B. Chougule, G.A. Patil. *Implementation and Analysis of EFRS Technique for Intrusion Tolerance in Distributed Systems*. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.

[9] P. Mell and T. Grance. *The NIST definition of cloud computing*. Special Publication 800-145. Retrieved September 2011, from http://csrc.nist.gov/publications/PubsSPs.html.

[10] Joshna S, Manjula P. *Challenges and Security Issues in cloud Computing*. International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 558-563.

[11] Rajkumar Buyya, James Broberg, Andrzej M.Goscinski. *Cloud Computing: Principles and Paradigms*. John Wiley & Sons, 17 dc. 2010.

[12] K. Sudha, M.Tech. MISTE, B. Anusuya, P.Nivedha, A. Kokila. *A Survey on Encrypted Data Retrieval in cloud Computing*. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 1, January 2015.

[13] Almokhtar Ait El Mrabti, Anas Abou El Kalam, Abdellah Ait Ouahman. *Les défis de sécurité dans le cloud Computing - Problèmes et solutions de la sécurité en cloud Computing*. 2012 National Days of Network Security and Systems, IEEE Catalog Number CFP1239S-PRT

[14] Wenjun Luo, Guojing Bai, *Ensuring the data integrity in cloud data storage*. International Conference on cloud Computing and Intelligence Systems (CCIS), IEEE,240 243,15-17, 2011.

[15] Prashant Srivastava, Satyam Singh, Ashwin Alfred Pinto, Shvetank Verma, Vijay K. Chaurasiya, Rahul Gupta. *An architecture based on proactive model for security in cloud computing*. IEEE 2011.

[16] Christian Baun, Marcel Kunze, Jens Nimis, Stefan Tai. *Cloud Computing Web-Based Dynamic IT Services*. Springer, 2011.

[17] Thomas Haeberlen, Lionel Dupr. *Cloud Computing Benefits, risks and recommendations for information security*. The European Network and Information Security Agency (ENISA). Rev.B December 2012. from https://www.enisa.europa.eu.

[18] Cloud Security Alliance. *The Notorious Nine: cloud Computing Threats in 2013*. February 2013. from http://www.cloudsecurityalliance.org/topthreats/

[19] B. Rex Cyril, DR. S. Britto Ramesh Kumar. *Cloud Computing Data Security Issues, Challenges, Architecture and Methods- A Survey*. International Research Journal of Engineering and Technology (IRJET). Volume 02 Issue 04. July-2015.

[20] Mingqiang Li. *On the Confidentiality of Information Dispersal Algorithms and Their Erasure Codes*. the IBM China Research Laborato. 2013.

[21] Rabin, M.O. *Efficient dispersal of information for security, load balancing, and fault tolerance*. In: Journal of The ACM 36(2), pp. 335348 (1989)

[22] Abdelsalam A. Helal, Abdelsalam A. Heddaya, Bharat B. Bhargava. *Replication Techniques in Distributed Systems*. The Kluwer International Series on ADVANCES IN DATABASE SYSTEMS; KLUWER ACADEMIC PUBLISHERS 1996.

[23] M. TOUT Rabih. *Sauvegarde des données dans les réseaux P2P*. Thèse de l'université de Lyon - 2010.

[24] Sian-Jheng Lin and Wei-Ho Chung. *An Efficient (n; k) Information Dispersal Algorithm for High Code Rate System over Fermat Fields*. IEEE COMMUNICATIONS LETTERS, VOL. 16, NO. 12, DECEMBER 2012.

[25] Dongfang Zhao, Kent Burlingame, Corentin Debains, Pedro Alvarez-Tabio and Ioan Raicu. *Towards High-Performance and Cost-Effective Distributed Storage Systems with Information Dispersal Algorithms*. 2013 IEEE.

[26] Honglu Liu, Shugang Zhang, Xiaolan Guan. *Integration study of high quality teaching resources in universities*. Journal of Industrial Engineering and Management-2013.

[27] A. Shamir. *How to share a secret*. Communications of ACM, vol. 22, no. 11, pp. 612613 (1979).

[28] Martin Tompa. *How to Share a Secret with Cheaters*. 1998, Springer-Verlag.

[29] Hugo Krawczyk. *Secret Sharing Made Short*. 1998, Springer-Verlag.

[30] Chetan Bansal, Karthikeyan Bhargavan, Antoine Delignat-Lavaud, Sergio Maffeis. *Keys to the Cloud: Formal Analysis and Concrete Attacks on Encrypted Web Storage*. Second International Conference, POST 2013, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2013, Rome, Italy, March 16-24, 2013. Proceedings. Springer Berlin Heidelberg.

[31] *Guidance Implementing the Cloud Security Principles*. From https://www.gov.uk/government/publications/implementing-the-cloud-security-principles/implementing-the-cloud-security-principles. Consulted May 10, 2016.

[32] Claire Broadley. *Copy.com Cloud Storage: the Solution to Fair Sharing?*. From http://www.cloudwards.net/copy-com-cloud-storage-the-solution-to-fair-sharing/. Cloudwards.net (04 Feb 2016) consulted Jun 1, 2016.

[33] *Encryption key management is vital to securing enterprise data storage*. From http://www.computerweekly.com/feature/Encryption-key-management-is-vital-to-securing-enterprise-data-storage. Computer Weekly consulted Jun 1, 2016.

[34] Margaret Rouse. *Cloud encryption (cloud storage encryption) definition*. From http://searchcloudstorage.techtarget.com/definition/cloudstorage-encryption. consulted Dec 12,2015.

[35] Stephen Lawton. *Cloud Encryption: Using Data Encryption In The Cloud*. From http://www.tomsitpro.com/articles/cloud-data-encryption,2-913.html. Tom's IT Pro (APRIL 30, 2015) consulted Jun 01, 2016.

[36] *Security Guidance for Critical Areas of Focus in Cloud Computing V3.0*. From https://cloudsecurityalliance.org/download/security-guidance-for-critical-areas-of-focus-in-cloud-computing-v3/. Nov 14,2011

[37] Marten van Dijk, Craig Gentry, Shai Halevi, Vinod Vaikuntanathan. *Fully Homomorphic Encryption over the Integers*. International Association for Cryptologic Research. H. Gilbert (Ed.): EUROCRYPT 2010, LNCS 6110, pp. 2443, 2010.

[38] Craig Gentry, Dan Boneh. *A fully homomorphic encryption scheme*. Stanford University, Stanford, CA, 2009.

[39] Iain Thomson. *Microsoft researchers smash homomorphic encryption speed barrier*. From http://www.theregister.co.uk/2016/02/09/researchers_break_homomorphic_encryption. The Register (Feb 9,2016) consulted JUN 1, 2016.

[40] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*. Microsoft Research (February 24, 2016).

[41] J-C Fabre, Y. Deswarte and B. Randell. *A Framework for the Design of Secure and Reliable Applications by Fragmentation-Redundancy-Scattering*. Technical Report Report Series No. 410 February 1993.

[42] Rudi Ball, Vicki Spurrett and Rogério de Lemos. *Dependable and Secure Storage in Pervasive Peerto-Peer Systems*. Technical Report September 2006. University of Kent, Canterbury, Kent CT2 7NF, UK.

[43] Jean-Charles Fabre, Yves Deswarte and Brian Randell. *Designing Secure and Reliable Applications using Fragmentation-Redundancy-Scattering: an Object-Oriented Approach*. EDCC-1 Proceedings of the First European Dependable Computing Conference on Dependable Computing. Pages 21-38. Springer-Verlag London, UK 1994

[44] Jean-Charles Fabre, Yves Deswarte, Laurent Blain. *Tolérance aux fautes et sécurité par fragmentationredondance-dissémination*. Rapport de recherche - LAAS-CNRS& INRIA.

[45] Jean-Michel Fray, Yves Deswarte, David Powell. *Intrusion-tolerance using fine-grain fragmentation scattering*. 1986 IEEE.

[46] Rabih Zbib, Farooq Anjum, Abhrajit Ghosh, Amjad Umar. *Intrusion Tolerance in Distributed Middleware*. Information Systems Frontiers 6:1, 6775, 2004 Kluwer Academic Publishers. Manufactured in The Netherlands.

[47] J.-C. Fabre and T. Pérennou. *Processing of confidential information in distributed systems by fragmentation*. Computer Communications, vol. 20, pp.177 -188 1997.

[48] Yves Deswarte, Laurent Blain and Jean-Charles Fabre. *Intrusion Tolerance in Distributed Computing Systems*. 1991 IEEE.

[49] Hassan Takabi and James B. D. Joshi. *Policy Management as a Service: An Approach to Manage Policy Heterogeneity in cloud Computing Environment*. 2012 45th Hawaii International Conference on System Sciences. 2012 IEEE.

[50] Hassan Takabi, James B.D.Joshi and Gail-Joon Ahn. *Security and Privacy Challenges in cloud Computing Environments*. The IEEE Computer and Reliability Societies 2010.

[51] Yves Deswarte. *Fragmentation-Redundancy-Scattering: a means to tolerate accidental faults and intrusions in distributed systems*. Proceedings of the ERCIM Workshops, INESC, Lisbonne (Portugal), 14-15 novembre 1991, pp. 31-34.

[52] Bruno Martin. *Codage, cryptologie et applications*. Collection technique et scientifique des télécommunications 2004.

[53] Farooq Anjum and Amjad Umur. *Agent Based Intrusion Tolerance using Fragmentation-Redundancy-Scattering Technique*. 2000 IEEE.

[54] Rudi Ball, James Grant, Jonathan So, Victoria Spurrett, Rogério de Lemos. *Dependable and secure distributed storage system for ad hoc networks*. Springer-Verlag Berlin Heidelberg 2007. LNCS 4686, pp. 142-152.

[55] Jun Kurihara, Shinsaku Kiyomoto, Kazuhide Fukushima, Toshiaki Tanaka. *A New (k,n)-Threshold Secret Sharing Scheme and Its Extension*. Proceeding ISC 2008 Proceedings of the 11th international conference on Information Security, Taipei, Taiwan, September 15-18, 2008 Springer.

[56] A. Shamir. *How to share a secret*. Communications of ACM, vol. 22, no. 11, pp. 612613 (1979)

[57] Hugo Krawczyk. *Secret Sharing Made Short*. 1998, Springer-Verlag

[58] A.Abou El Kalam, Y.Deswart, A.Bana, M.Kaniche. *PolyOrBAC:A security framework for Critical Infrastructures*. International Journal Of Critical Infrastructure Protection 2(2009)154-69.

[59] Danilo Ardagna, Giuliano Casale, Michele Ciavotta, Juan F Prez and Weikun Wang. *Quality-of-service in cloud computing: modeling techniques and their applications*. Springer 2014.

[60] Jens Myrup Pedersen, M. Tahir Riaz, Bozydar Dubalski, Damian Ledzinski, Joaquim Celestino Jnior and Ahmed Patel. *Using latency as a QoS indicator for global cloud computing services*. John Wiley& Son 2013.

[61] Timothy Zhu, Alexey Tumanov, Michael A. Kozuch. *PriorityMeister: Tail Latency QoS for Shared Networked Storage*. ACM 2014.

[62] Marisol Garcia-Valls, Tommaso Cucinotta, Chenyang Lu. *Challenges in real-time virtualization and predictable cloud computing*. Elsevier 2014.

[63] S. Tisue and U. Wilensky,*Netlogo: A simple Environment for Modeling Complexity. International Conference on Complex System*. Boston, May 2004.

[64] Wilensky, U. (1999). *NetLogo*. From http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston.

[65] *Researcher Demonstrates Simple BitLocker Bypass*. By SecurityWeek News on November 18, 2015. from http://www.securityweek.com/researcherdemonstrates-simple-bitlocker-bypass. consulted MAR 10,2016.

[66] Sven Trpe, Andreas Poller, Jan Steffan, Jan-Peter Stotz and Jan Trukenmller. AAA. *Attacking the BitLocker Boot Process*. Second International Conference, Trust 2009, Oxford, UK, April 6-8, 2009, Proceedings. Pages 183 - 196. Springer Berlin Heidelberg.

Quality of Service Provisioning in Biosensor Networks

Yahya Osais
Department of Computer Engineering
King Fahd University of Petroleum and Minerals
Dhahran 31261

Muhammad Butt
Department of Computer Engineering
King Fahd University of Petroleum and Minerals
Dhahran 31261

***Abstract*—Biosensor networks are wireless networks consisting of tiny biological sensors (biosensors, for short) that can be implanted inside the body of human and animal subjects. Biosensors can measure various biological processes that occur inside the body of the subject under test. Applications of biosensor networks include automated drug delivery, heart beat rate monitoring, and temperature sensing. Since biosensor networks employ wireless transmission, heat is generated in the tissues surrounding the implanted biosensors. Human and animal tissues are very sensitive to temperature increase. Therefore, the generated heat is mitigated by the natural thermoregulatory system. However, excessive transmissions can cause a significant increase in temperature and thus tissue damage. Hence, there is a need for a mechanism to control the rate of wireless transmissions. Of course, controlling the rate of wireless transmissions will lead to Quality-of-Service (QoS) issues like the required minimum delay and throughput. In this paper, we are going to investigate the above issues using the framework of Markov Decision Processes (MDPs). We are going to develop several MDP models that will enable us to study the different trade-offs involved in QoS provisioning in biosensor networks. The optimal policies computed using the proposed MDP models are compared with greedy policies to show their vigilant behavior and viable performance.**

Keywords*—Biosensor networks; Quality of service; Markov decision processes

I. Introduction

Biosensors can be implanted inside the body of human and animal subjects to form a biosensor network that can be used for monitoring and observing various biological processes and detect anomalies. No processing is done on the biosensors. Therefore, measurements are transmitted to a Base Station (BS) for processing and recommendation of necessary actions. Biosensor networks can be used in daily medical tasks like sensing body temperature, calculating heart beat rate and automated drug delivery. Biosensor networks are powered by either rechargeable batteries or by continuously transmitting energy to them via electromagnetic waves.

Biosensor networks have the same technical challenges introduced by traditional wireless sensor networks. In addition, they introduce new challenges that are unique to them. For example, a major challenge to realizing the full potential of biosensor networks is the heat they generate as a result of power dissipation and wireless communication. Every wireless transmission generates heat. This heat increases the temperature of the tissues that surround the biosensor. The effect of the generated heat is balanced by the human thermoregulatory system. However, excessive transmissions may result in heat that is greater than what can be drained by the thermoregulatory system. If the temperature increase exceeds a certain threshold, the tissues may be damaged. In such a case, the biosensor should be shut down in order for the tissues to cool down and attain the normal body temperature.

As a consequence, the maximum safe temperature level that human tissues can withstand becomes an important factor while operating biosensor networks. Hence, there is a need for intelligent thermal management techniques to mitigate the thermal effect on human tissues. Such techniques, for example, would enable long-term monitoring and measurement to be performed. Furthermore, there is a need for a mechanism to optimize the transmission schedule of biosensors to prevent the potential damage to the human tissues and respect the required QoS. All these contradicting challenges need to be carefully and intelligently addressed.

Very little work has been done in the area of QoS provisioning in biosensor networks. The main focus has been to minimize the average temperature increase of the system with no consideration for QoS [1], [2], [3]. On the other hand, QoS issues such as data loss and late delivery are not studied in the context of temperature-sensitive environments like the ones in which biosesnor networks operate. These two specific QoS issues are studied in this paper using a new model that includes the state of the buffer inside a biosensor. In this way, a more accurate picture of the operation of biosensor networks can be painted.

The rest of the paper is organized as follows. Section II provides a survey of the relevant literature. Then, section III describes the newly proposed model. After that, section IV presents the numerical results and several insights. Finally, section V concludes the paper and provides directions for further research.

II. Related Work

The goal of this paper is to extend the models presented in [1], [2], [3] to include some QoS metrics. The current models consider only power and energy constraints with no regard for the effect of traffic and finite buffer size on the performance of biosensor networks. Hence, in this section, we are going to critique the current models and discuss their shortcomings. For more details about the problem and its context, the reader

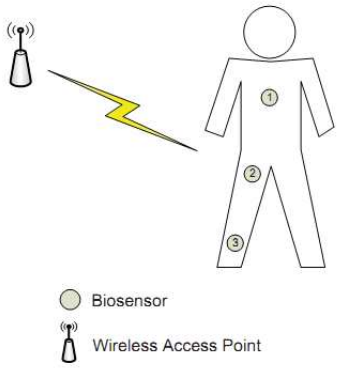


Fig. 1: Biosensors are implanted inside the body of a human to collect physiological measurements and transmit them over a wireless channel to an access point for further processing.

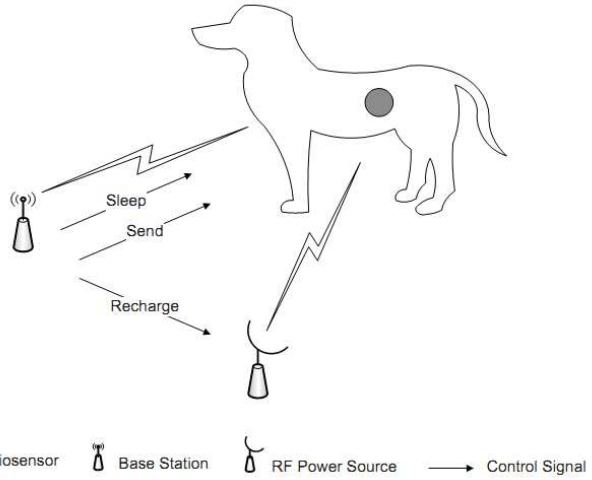


Fig. 2: A biosensor can be rechargeable. Recharging biosensors can increase their lifetime but it also increases the temperature of the tissues around them. A biosensor can be put to sleep to cool down.

is encouraged to read our previous papers and the references therein.

Support for various QoS requirements like low packet loss and delay is essential in the development of future wireless networks that employ tiny sensing devices. Several cross-layer optimization techniques have been proposed in the literature to tackle QoS-related issues. For example, the authors in [4] handle the issue of the time-varying nature of the wireless channel by constraining different system parameters like data rate, modulation schemes, and transmission power. The trade-offs between the average transmission power and average packet dropping probability and the average buffer delay are studied in [5]. The authors consider a system with a finite transmission buffer and a time-varying wireless channel. The system is formulated as both a constrained and unconstrained MDP with an average cost criterion.

The heating issue in biosensor networks is addressed in [1], [2]. The authors optimize the network lifetime under strict temperature constraints by considering different amounts of initial energy. The system consists of biosensor nodes whose wireless transmission affects the temperature level of the surrounding tissues. The system is modeled as a discrete time MDP that grows in discrete time steps. During each time slot, the scheduled sensor undergoes a change in its energy and temperature in accordance with its action. Temperature of the unaffected biosensors is assumed to decrease by a constant value. However, temperature of the affected biosensors increases according to a direct relationship with the biosensor scheduled for transmission and the state of its wireless channel with the base station. The system is solved to obtain an optimal operating policy that maximizes the network lifetime while keeping the system in a safe temperature zone to avoid tissue damages. The results obtained indicate that the optimal policy performs better when compared to several heuristic policies. Figure 1 shows the system used in the study.

Optimization of biosensor networks by increasing the number of transmitted samples is addressed in [3]. Three actions are considered as shown in Figure 2. The control signals are initiated by the base station which also controls the power source. The model is also formulated as a discrete time MDP whose state includes the current energy, transmission power, and temperature. The temperature is also used as a strict (i.e., global) constraint. The authors evaluate an optimal policy by solving the system using the value iteration algorithm with an average reward criterion. The obtained optimal policy maximizes the samples which can be transmitted by the biosensor network when compared with greedy and heuristic policies.

III. System Model

Figure 3 shows the layout of the system studied in this paper. Only one biosensor node is shown. Each biosensor has its own state. Multiple biosensor nodes share a common wireless channel that connect them to the base station. Each biosensor node contains a finite size buffer for storing the samples generated by the biosensing elements. These arriving samples may experience delay and loss while traveling to the base station . We assume that each biosensor node knows the state of the wireless channel and the size of its buffer. Hence, the state of the biosensor node is made up of three state variables: wireless channel, buffer size, and temperature. Based on the state of the biosensor, the controller should determine an efficient policy that optimizes certain QoS metrics. Basically, in each time slot, the controller decides whether to make a transmission or put the transmitter to sleep. Next, the details of the system model are given.

A. Wireless Channel Model

We consider a slotted Rayleigh fading channel with Additive White Gaussian Noise (AWGN) N_o and channel bandwidth W. The Rayleigh fading channel is assumed to be slowly varying so that the received Signal to Noise Ratio (SNR) remains constant during a single time slot. It is also assumed that transitions are only allowed to current or adjacent states. This slowly varying discrete time Rayleigh fading process can be represented by a Finite State Markov Chain (FSMC) which

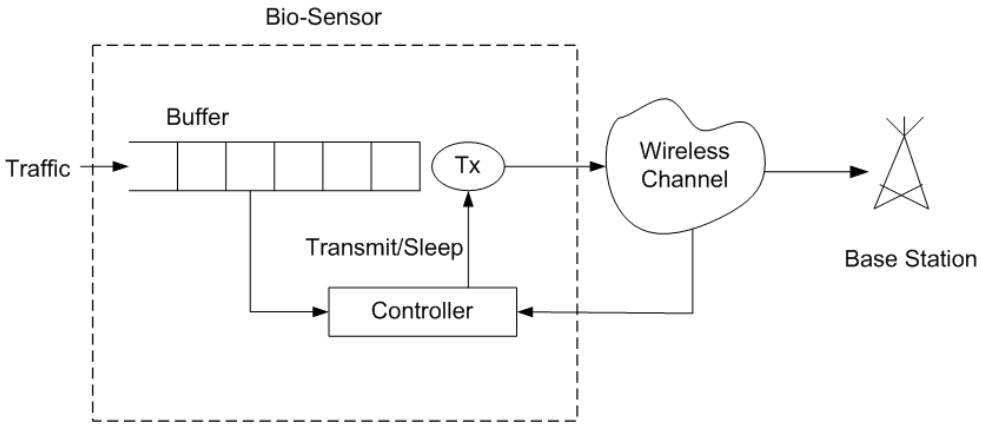


Fig. 3: System model for a biosensor node with a finite buffer and controller.

has K channel states [6]. The channel states are numbered from 0 to $K-1$. The channel gain for each state c, where $c \in (0, ..., K-1)$, is represented by θ_c. The probability distribution for the next channel state during a time slot n is given by

$$P_C(c, c') = P[C_n = c' | C_{n-1} = c] \tag{1}$$

$P_C(c, c')$ can be calculated by partitioning the range of channel gains into a finite number of intervals. The information about the fading process given in [7] is used. Further, we assume that the channel state transition probabilities for all channel states are available [8].

B. Buffer State Model

Samples generated by the on-board sensing elements are stored in a finite buffer of size β. Let σ_n indicate the number of arriving samples at the beginning of time slot n. Samples arriving in time slot n can only be transmitted in the next time slot $n+1$. Sample arrivals are Poisson distributed with an average arrival rate equal to λ. They are also independent of the channel fading process. A truncated Poisson process is considered since the number of on-board sensors is finite. This necessitates an upper bound, represented by Z, on the number of samples. It is assumed that the length of each time slot is equal to one time unit. Hence, the truncated Poisson process can be approximated as follows:

$$p(\sigma_n = i) = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = \{0, 1, ..., Z-1\} \tag{2}$$

$$p(\sigma_n = Z) = 1 - \sum_{i=0}^{Z-1} p(i) \tag{3}$$

It should be pointed out that $p(\sigma_n = Z)$ has a large probability due to truncation. In our model, this means that the likelihood that all sensors generate samples in one time slot is high.

Let B_n be a state variable indicating the number of samples in the buffer at the beginning of time slot n. Then, the number of samples in the buffer in time slot $n+1$ is given by

$$B_{n+1} = \min\{B_n - A_n + \sigma_{n+1}, \beta\} \tag{4}$$

where A_n is the number of samples transmitted in time slot n.

C. Transmission Model

The number of samples transmitted in a time slot n is equal to A_n which takes values from the set $\{0, 1, 2, ..., \alpha\}$. The transmitter is responsible for taking certain number of samples from the buffer and transmit them over the correlated faded channel. Let $A = a^0, a^1, a^2, ..., a^A$ indicate the set of actions performed by the transmitter where a^1 indicates one sample is transmitted , a^2 indicates two samples are transmitted and so on. a^0 represents the sleep action;i.e., no sample is transmitted by the biosensor node in this state.

Let $P(C_n, A_n)$ represent the power required to make action A_n in time slot n while the channel state is C_n. Power required to take a certain action in slot t must belong to $P_t(c, a) \in P_{op}$, where P_{op} indicates the set of power levels supported by the transmitter. Furthermore, we enforce a fixed Bit Error Rate (BER) constraint on all the transmissions done by the transmitter. Assuming an adaptive M-ary Quadrature Amplitude Modulation (MQAM) modulation scheme with ideal coherent phase detection, the power required to satisfy a particular BER can be evaluated by using the following equation from [8]:

$$P(c, a) \geq \frac{W \cdot N_o}{\theta_c} . (\frac{-(2^{a-1}) \log(5.E_b)}{1.5}) \tag{5}$$

In (5) N_o represents the channel noise, E_b represents the fixed BER constraint that is satisfied assuming coherent phase detection, θ_c represents the channel gain when the channel state is c and W represents the bandwidth of wireless transmission. If the required power is less than that described in (5) , it means that action is not feasible. Power calculated in (5) give a pessimistic estimate of the power required to achieve a certain BER for different channel states and actions.

In each time-slot the biosensor node's rate of transmission can be calculated by

$$Rate = \frac{G \cdot \Phi(A_n)}{F} \tag{6}$$

where Φ represents the number of bits per symbol used for transmission of A_n samples during F channel uses. G represents the size of incoming samples in terms of bits.

If we set $G = F$, the rate will be equal to Φ. We can transmit different number of samples by changing the number of bits per symbol. If we set number of bits per symbol equal to number of samples transmitted in a time slot, then $\Phi(A_n) = A_n$; i.e., the transmission rate becomes equal to the action suggested by the optimal policy.

IV. MDP Formulation

The global state of the system, denoted by S, consists of three variables and the state space is given by

$$S = C \times B \times T \tag{7}$$

where T is the temperature state variable. The size of the state space is thus the product of the number of channel states, number of buffer states, and number of temperature levels. In this section, two MDP formulations are given. They differ in whether the temperature is part of the global system state or a constraint.

An important element of any MDP formulation is the sytem state transition probability matrix. This matrix describes how the system transitions from one state to another. We assume that state variables are independent. Thus, the state transition probability matrix of the system can be calculated by simple multiplication of the transition probabilities of the channel and buffer state variables. The temperature state variable plays no role in the computation of the state transition probability matrix of the system. This is because it is not random.

Hence, the following equation gives the state transition probability matrix of the system.

$$P_S[s'|s,a] = P_C[c'|c] \times P_B[b'|b,a] \tag{8}$$

where s, c and b represent the current state of the system, wireless channel and buffer, respectively. On the other hand, s', c' and b' represent the next state of the system, wireless channel and buffer when action a is performed. The next state of the wireless channel is independent of the current action. The current action a determines the next state of the buffer only.

The solution of an MDP formulation is referred to as a policy which is a mapping from the system state space to action space. That is, a policy determines the best action that should be performed in each possible state of the system. An optimal policy guarantees an optimal behavior of the system.

Two objectives are considered. The first one is to minimize the expected long-term average transmission power.

$$P_{Avg}(\pi) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[P(s_i, \pi(s_i))] \tag{9}$$

where $\pi(s_i)$ represents the action suggested by policy π and $P(s_i, \pi(s_i))$ is the instantaneous transmission power.

The second objective, however, is to maximize the expected long-term average transmission rate.

$$R_{Avg}(\pi) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[R(s_i, \pi(s_i))] \tag{10}$$

where $R(s_i, \pi(s_i))$ represents the instantaneous transmission rate.

An important performance metric is the average loss rate which represents the expected number of samples that are dropped due to buffer overflow. The following equation shows how the number of samples lost in time slot n is computed for a specific state s and action a.

$$L_n(s,a) = \max\{b_n + \sigma_n - a_n - \beta, 0\} \tag{11}$$

The average number of lost samples can be computed using the first moment as follows.

$$L_{Avg}(s,a) = E(L_n(s,a)) \tag{12}$$

The instantaneous delay during a time slot n can be computed as follows.

$$D_n(b_n, a) = \frac{b_n}{\lambda} \tag{13}$$

where b_n is the instantaneous buffer size during time slot n. The expected long-term average delay is the following.

$$D_{Avg}(\pi) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[D(b_i, \pi(a_i, b_i))] \tag{14}$$

Finally, our thermal model is discussed. In this model, the increase in temperature is directly proportional to the magnitude of the action. For example, transmitting one sample during the best channel state ($C_n = 0$) will increase the temperature by one unit. The following equation is used for computing the instantaneous temperature increase.

$$T_{n+1}(s_n, a_n) = \begin{cases} -1 & a \in a^0 \\ a_t + K - c_t - 1 & a \in a^1, a^2, ...a^A \end{cases} \tag{15}$$

The long-term average temperature is mathematically expressed as follows.

$$T_{Avg}(\pi) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} T(s_i, \pi(s_i)) \tag{16}$$

Notice that the expectation operator is dropped since temperature is not a random variable.

Next, the details of the MDP models are given. First, in the average thermal increment model, the problem is formulated as a constrained MDP model where a particular objective function is optimized while putting various constraints on other QoS metrics . The first MDP formulation maximizes the system transmission rate while keeping the average power, delay, thermal increment and loss rate within given bounds. The second MDP model, on the other hand, optimizes the system power consumption while respecting a minimum transmission rate and keeping the biosensor network in a safe operating zone.

A. LP Formulation for The Thermal Increment Model

Let $x(s,a)$ indicate the decision variable in solving the MDP models obtained in previous section. $x(s,a)$ represents the steady state probability distribution when the system is in state s and action a is performed. Based on different rewards and depending on the QoS parameters, we want to optimize $x(s,a)$ to obtain an optimal policy which describes what action to take when the system is in state s. The MDP model proposed is solved using the LP algorithms in MATLAB [11] to obtain optimal operating policies for correlated wireless channel. The default mode for LP solver is to minimize the reward function.

Since the problem is formulated as an average cost constrained MDP, there are certain basic constraints that must applied for each implementation.

$$\sum_{s\in S}\sum_{a\in A} x(s,a) = 1 \tag{17}$$

$$\sum_{a\in A} x(j,a) - \sum_{i\in S}\sum_{a\in A} p_{ij}(a)\times x(i,a) = 0 \quad j\in S \tag{18}$$

$$x(s,a) \geq 0 \quad \forall\, s\in S, \quad \forall\, a\in A \tag{19}$$

The first constraint ensures that the $x(s,a)$ is a probability distribution with its sum over all pairs of system states and actions equal to one. The second constraint ensures that we are solving an average cost constrained MDP. The third constraint enforces that the decision variable $x(s,a)$ is always positive. These basic constraints are common to all the LP models given in this paper.

The first LP model is about the maximization of the transmission rate (i.e., throughput). The details of the model are as follows.

$$\max_x \sum_{s\in S}\sum_{a\in A} x(s,a)\times R(s,a) \tag{20}$$

subject to:

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times P(s,a) \leq P_O \tag{21}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times L(s,a) \leq L_O \tag{22}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times T(s,a) \leq T_h \tag{23}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times D(s,a) \leq D_O \tag{24}$$

The constraints in (21)-(25) makes sure that the average values of power consumption $P(s,a)$, loss rate $L(s,a)$, thermal increment $T(s,a)$ and delay $D(s,a)$ do not exceed their thresholds P_O, L_O,T_h and D_O, respectively.

In the next LP model, the objective is to minimize the average transmission power and use the other metrics as constraints. The following are the details of the model.

$$\min_x \sum_{s\in S}\sum_{a\in A} x(s,a)\times P(s,a) \tag{25}$$

subject to:

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times R(s,a) \geq R_O \tag{26}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times L(s,a) \leq L_O \tag{27}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times D(s,a) \leq D_O \tag{28}$$

The first constraint ensures that there is a minimum average throughput. The remaining two constrains put an upper limit on the loss rate and delay, respectively.

B. LP Formulation for the Strict Temperature Model

The LP formulation of the strict temperature model is similar to that of the thermal increment model discussed above. However, the reader is reminded that the system state now includes the temperature as a state variable. This represents a global constraint. Thus, there will be no explicit constraint on the temperature increase like in the previous LP models. The following are the details of the new LP model.

$$\max_x \sum_{s\in S}\sum_{a\in A} x(s,a)\times R(s,a) \tag{29}$$

subject to the following QoS constraints:

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times P(s,a) \leq P_O \tag{30}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times L(s,a) \leq L_O \tag{31}$$

$$\sum_{s\in S}\sum_{a\in A} x(s,a)\times D(s,a) \leq D_O \tag{32}$$

where $x(s,a)$ represents the decision variable for the optimization of average transmission rate. P_O, L_O, T_h and D_O represent the thresholds on the average transmission power, loss rate and delay, respectively.

C. Finding the Optimal Policy

After solving the above LP models, a probabilistic distribution over the state-action space is obtained. We would would like to find a policy that tells us what action should be performed in each system state with a probability of one. This can be achieved as follows.

$$\pi^*(s,a) = \frac{x^*(s,a)}{\sum_{i=1}^{A_s} x^*(s,a_i)} \quad \forall\, a\in A_s \quad and \quad s\in S \tag{33}$$

Here, A_s represents the set of feasible actions in each system state s.

TABLE I: Channel states and transition probabilities.

Channel states c	0	1	2	3	4	5	6	7
θ_c	0	0.1068	0.2301	0.3760	0.5545	0.7847	1.1090	1.6636
$P_{c,c}$	0.9359	0.8552	0.8334	0.8306	0.8420	0.8665	0.9048	0.9639
$P_{c,c+1}$	.0641	.0807	.0859	.0835	.0745	.0590	.0361	0
$P_{c,c-1}$	0	.0641	.0807	.0859	.0835	.0745	.0590	.0361

V. Results and Discussion

In this section we numerically solve the model proposed in the previous section to obtain the optimal policies and then simulate them. In our simulation, we are going to analyze the effect of various QoS constraints on the optimal policies. Then, we study the different optimal policies obtained by solving the average thermal increment and strict temperature models. The thermal behavior of the obtained policies is also discussed.

A. Configuration

TABLE II: Simulation parameters.

Parameter	Value
G	100 $bits$
B_{size}	8 $Samples$ = 800 $bits$
K	8
T_{size}	4
A	8
λ	3 $Samples$
W	100 MHz
N_O	10^{-12}
f_D	10 Hz
θ_{avg}	0.8

The following system parameters are used in the model formulation and simulation. They are also described in Table II. Arrivals at the buffer input are assumed to be Poisson with an average arrival rate of three. Buffer size is set to eight samples. Eight channel states are considered. The state zero is assumed to be the worst with a very small gain. There are eight possible actions in each state of the system;i.e., transmitting from one up to seven samples or no transmission. Based on these system parameters, the MDP model is formulated as a linear program and solved using MATLAB. The slowly varying Rayleigh model is described in Table I. It has an average power gain of 0.8 and a Doppler frequency of 10 Hz.

B. Analysis and Insights

For the purpose of analyzing the effect of various constraints on the optimization of average transmission rate and average power consumption, we vary the magnitude of the constraints on the average loss rate, delay and thermal increments to study their effects on the objective function. Values of the input parameters are also varied and their effects on both the constraints and objective function are studied.

First, the LP model expressed by equations (25)-28 is studied. Figure 4 shows the effect of varying L_O. It can be seen that the average transmission power decreases as the average loss rate increases. Since more samples are allowed to drop when the loss rate constraint is increased, the optimal policy will use the least amount of power possible for transmission. Also, increasing the arrival rate increases the average power consumption of the system. This is because there will be more samples in the buffer which need to be transmitted.

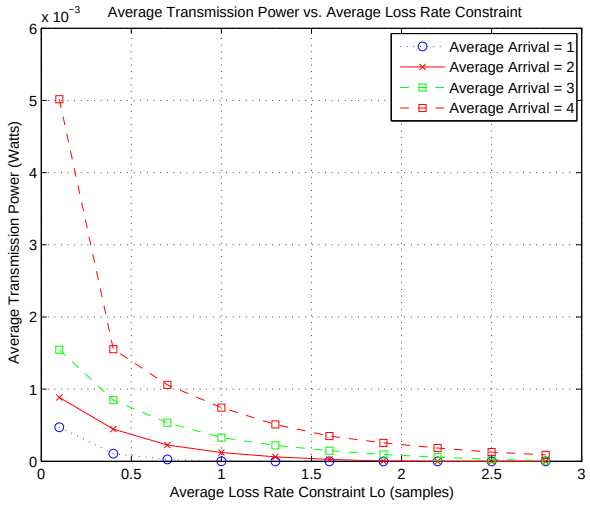


Fig. 4: Reduction in the optimal average transmission power as the average loss rate constraint (L_O) is varied.

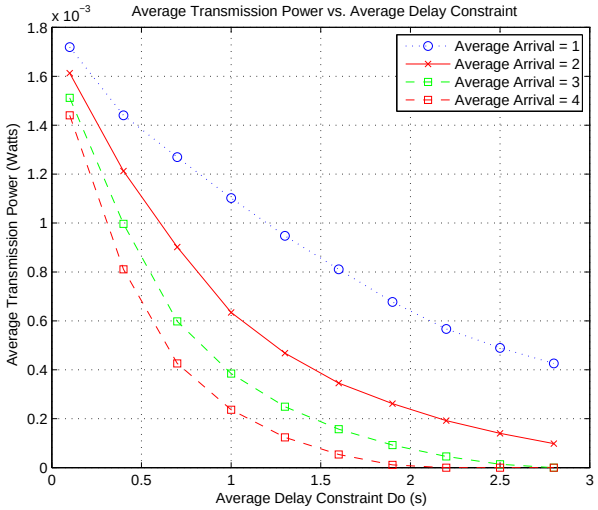


Fig. 5: The optimal average transmission power decreases as the average delay constraint (D_O) increases.

Figure 5 shows the effect of varying the delay (i.e., D_O). It can be seen that the value of the optimal average transmission power decreases as the average delay constraint is increased. This indicates that as the constraint on the average delay is increased, samples are allowed to experience more delays which results in a lesser average power consumption.

The effect of changing the average arrival rate λ on the

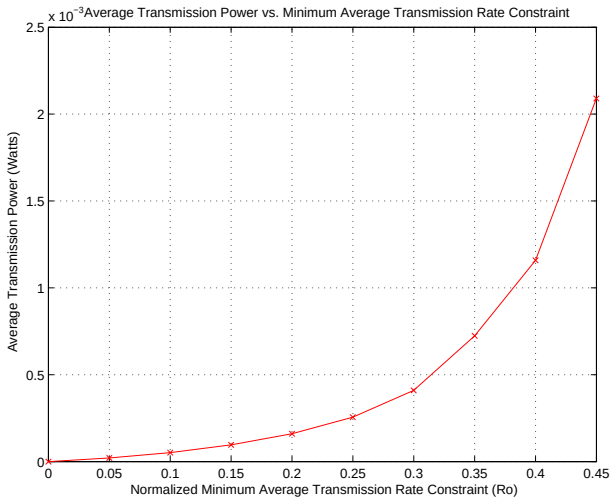


Fig. 6: The increase in the minimum average transmission rate constraint (R_O) causes an increase in the optimal average transmission power utilized.

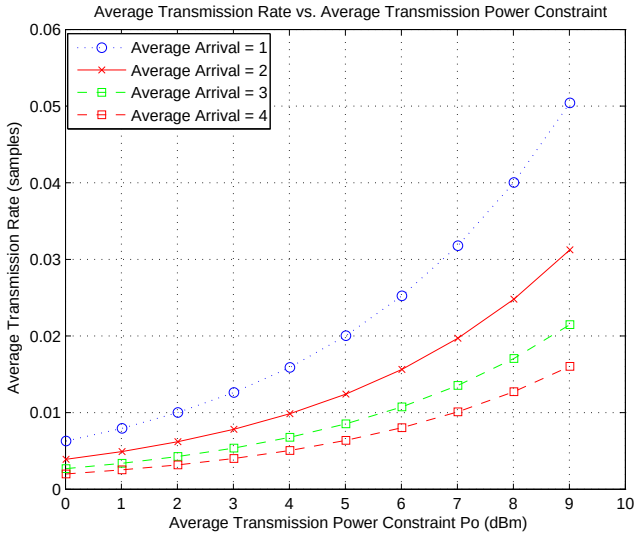


Fig. 7: The effect of increasing the average arrival rate (λ) on the optimal average transmission rate as the average power constraint (P_O) increases.

average delay constraint is studied next. Figure 5 shows the variations in the average delay and optimal average transmission power due to different arrival rates. The delay and average arrival rate have an inverse relationship. For example, for a fixed D_O, the left side of the delay constraint in equation 28 will be reduced if we increase the average arrival rate. This in turn should increase the optimal average power consumption in order to achieve the same delay constraint. By contrast, the behavior observed in Figure 5 is the opposite. This can be explained by the fact that the delay is directly proportional to the buffer occupancy while it is inversely proportional to the average arrival rate. So, based on the insights obtained from Figure 5, we can conclude that the effect of the increased delay dominates the reduction achieved by increasing the average arrival rate which in turns reduces the average power consumption.

We next study the effect of having a minimum average transmission rate requirement on the optimization of average power. The behavior obtained after applying the minimum average transmission rate constraint in equation (26) is shown in Figure 6. It can be seen that as the value of the constraint increases, the optimal average power consumption increases. This happens because the increase in the minimum average transmission rate constraint requires that the biosensor node transmits more samples. As a result, the optimal value of average power consumption increases.

Next, the LP model expressed by equations (20)-(24) is studied. In the same way, the value of P_O is varied. The results are then plotted in Figure 6. It can be seen that the optimal average transmission rate increases as the average transmission power P_O increases. This indicates that as the constraint on average power is increased, more power is available which can then be used to transmit a larger number of samples. Of course, this will result in higher transmission rates.

The effect of increasing the arrival rate on average transmission rate is depicted in Figure 7. It can be seen that as the average arrival rate increases the average transmission rate decreases. This is due to the fact that any increase in the average arrival rate causes an increase in the loss rate which in turns reduces the average transmission rate of the biosensor.

Maximization of the average transmission rate can cause the temperature of the system to increase by a large amount. The minimization of the average transmission power indirectly minimizes the system's thermal state increment by minimizing the power consumption. However, for the maximization of the average transmission rate, we need to explicitly include a constraint that controls the increase in the thermal state of the system at symbol level. In order to study the effect of the constraint in equation (23), the value of T_h is varied to obtain various optimal policies. The results are then used to calculate the optimal average transmission rates. Figure 8 shows that the average transmission rate increases as the average thermal increment increases. This is at the cost of damaging the tissues, of course. So, we should try to keep the thermal increase constraint as small.

It should be pointed out that a change in the average delay constraint does not affect the average transmission rate. The reason for such behavior is that the delay depends on the buffer state and the average arrival rate. If we keep the average arrival rate constant, the delay becomes directly related to the state of the buffer. But, changes in the buffer state also cause similar changes in the transmission rate. As a result, the optimal average transmission rate stays constant as the average delay constraint is varied. However, if we increase the arrival rate at the input of the buffer, the average loss rate and the delay both increase. This will cause a reduction in the optimal average transmission rate as shown in Figure 9.

C. Optimal Policies for the Thermal Increment Model

In this section, we study the thermal increment model and how the thermal increment constraint affects the optimal

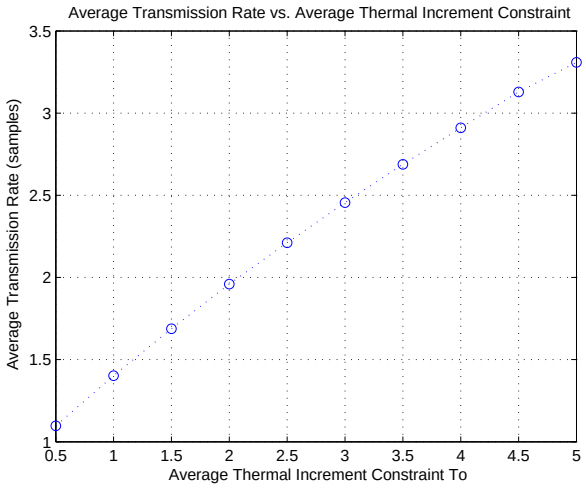


Fig. 8: The optimal average transmission rate increases as the average thermal increment (T_h) constraint increases.

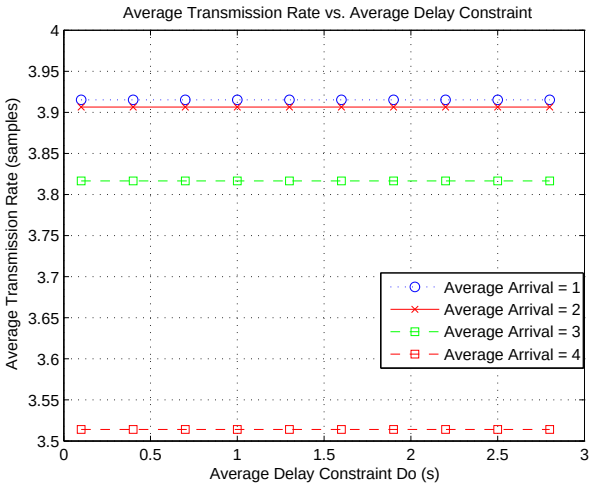


Fig. 9: Increasing the value of the average delay constraint does not have any effect on the average transmission rate. However, it decreases as the average arrival rate increases.

policies.

The optimal policy that results from solving the LP model in equations (25)-(28) is plotted in Figure 10. The minimum average transmission rate constraint R_O is set to 0.07, average delay constraint D_O is set to $10\,msec$ and average loss rate constraint L_O is set to $2\,Samples$. The 3D plot indicates that as the channel state improves, the policy suggests to make a transmission. Similarly, an increased number of samples in the buffer also indicates that the transmitter should start sending more samples to the base station. However, since the objective is to minimize the average power consumption and the minimum average transmission rate constraint is quite small, a maximum of one sample is transmitted even in the best channel state. This has the advantage of reducing the temperature increase of the biosensor node. However, if we increase the minimum average transmission constraint to 0.35,

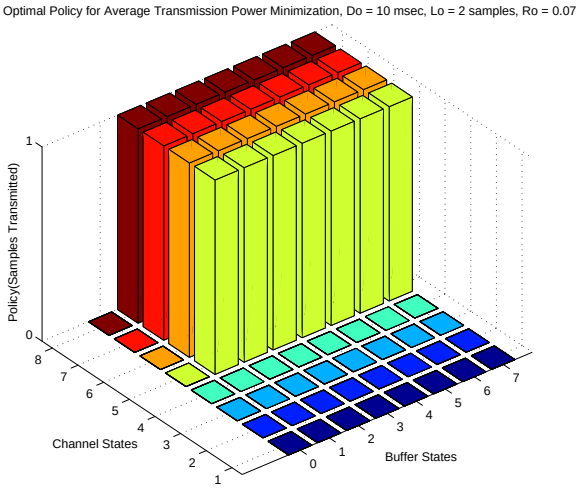


Fig. 10: Optimal policy for minimizing average power consumption with $R_O = 0.07$, $D_O = 10\,msec$ and $L_O = 2\,Samples$.

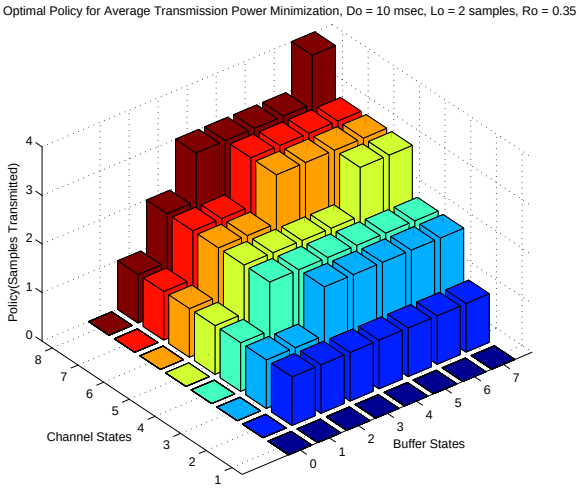


Fig. 11: Increase in the minimum average transmission rate constraint $(R_O = 0.35)$ results in an increased number of samples transmissions in the optimal policy

it can be seen in Figure 11 that the number of samples transmitted as the buffer state improves is increasing.

The optimal policies obtained from the different LP models have unique behaviors. They are observed to be monotonically increasing in the channel and buffer state of the system. This means that as the channel state improves or the buffer state increases, the optimal policy also increases monotonically. When embedding these policies into an actual hardware, we can define the actions in terms of increasing values of channel and buffer state information. The controller can make an easy decision based on these thresholds defined by the optimal policy. This behavior can thus help in the practical implementation

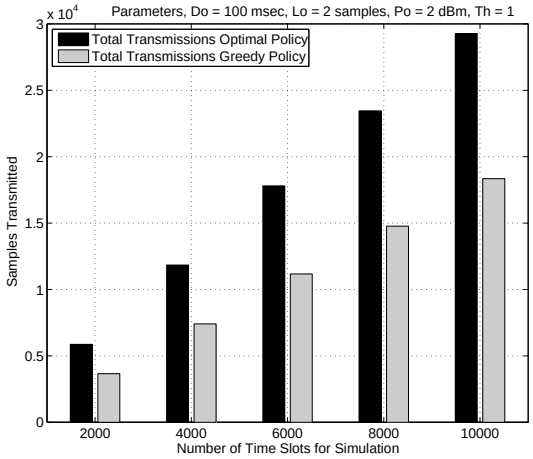


Fig. 12: Comparison of sample transmissions for different policies with a varying number of time slots.

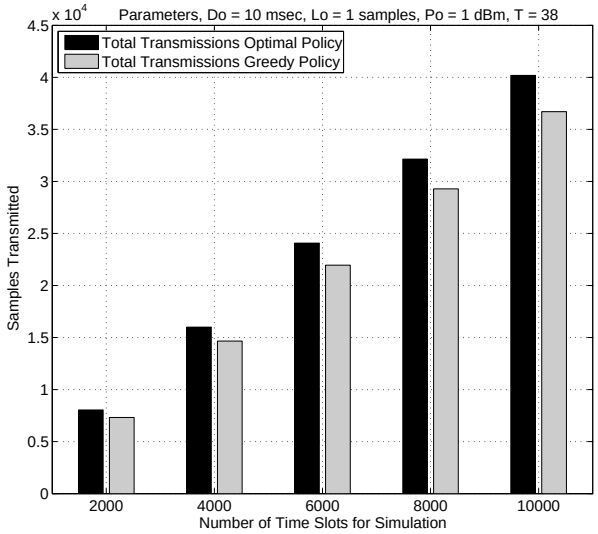


Fig. 13: Comparison of sample transmissions for different policies achieved by the average transmission rate maximization.

of these optimal policies on biosensor hardware.

The optimal policies computed in the previous section are simulated using MATLAB and the results are compared with a greedy policy. In the case of the transmission rate maximization, the greedy policy works on the principle that it always tries to transmit the maximum number samples that are allowed under the given system state without exceeding the constraints of the average loss rate, average thermal increment, average delay and average transmission power. As for the transmission power minimization, the greedy policy works by transmitting the least number of samples possible without violating the required average transmission rate. Each data point is the result of running the simulation five times. Both policies are simulated for a different number of time slots and their results are compared. The performance of the average transmission rate maximization policy against the greedy policy is shown in Figure 12. Clearly, this Figure indicates that the optimal policy outperforms the greedy policy in terms of the total number of transmitted samples.

D. Strict Temperature Model

In this section, we are going to study the LP model expressed by equations (29)-(32). The obtained optimal policy is simulated and the temperature variations are observed. Similar to the previous approach, a comparison is performed with a greedy policy. The conclusion is that the optimal policy provides better performance.

We choose four temperature levels to represent the temperature states in the model proposed for the strict temperature model. The lower and upper bounds on the temperature are set to 37^oC and 40^oC. The number of channel and buffer states are set to eight, respectively. The average arrival rate at the input of buffer is set to three. The optimal policy allows transmissions only when the temperature is in state one. For higher temperature states, the policy chooses the sleep action to keep the thermal state of the system within the provided constraints.

Again, the behavior of the optimal policy is observed to be monotonic in the channel and buffer states. The policy ensures that more samples are transmitted as the state of the wireless channel and buffer improves. The average temperature and power constraints are also kept within bounds. It is also observed that when the temperature is in its worst state, the policy suggests not to transmit any samples in order to save the biosensor from going into the highest thermal state. Therefore, the optimal policy is also monotonic in terms of the temperature states.

The optimal policy computed for the transmission rate maximization problem is also compared with a greedy policy that satisfies the constraints given in the model. The greedy policy always tries to transmit the maximum possible number of samples while respecting the QoS constraints. A running average for all the constraints is used to make the decision in each time slot. The simulation is run five times for each number of slots and the average results are calculated. Figure 13 shows the results obtained by running the simulation for up to 10000 time slots. The results indicate that the optimal policy again outperforms the greedy policy in terms of the total number of transmitted samples. However, the difference between the two is small as compared to the optimal policy for the previous average thermal increment model.

VI. Conclusion

In this paper, the problem of QoS provisioning in biosensor netoworks has been studied using the framework of MDPs. The newly proposed model captures the interaction between the wireless channel and buffer at a biosensor node. The obtained policies maximizes network throughput and lifetime under several QoS constraints. They are also monotonic which means that they can be easily realized. Further, the simulation of the thermal behavior of the optimal policies indicate that the strict temperature model provides a better control over temperature increase when compared to the average thermal

increment model. However, the strict temperature model has the disadvantage of requiring a high computation power which can be vital for battery-operated biosensor nodes that have limited energy. The average thermal increment model shows some promising results for average transmission power minimization since transmission power is indirectly related to thermal increase. However, in both cases, the optimal policies outperform the greedy policy in both network life time and transmission rate maximization. One possible direction for further research is to include the level of battery energy as part of the system state in the current model. The recharge action can also be taken into consideration for biosensor networks that have wireless recharging sources.

ACKNOWLEDGMENT

The authors would like to thank King Fahd University of Petroleum and Minerals (KFUPM) for support.

REFERENCES

[1] Y. Osais, F. R. Yu, and M. St-Hilaire, "Dynamic Sensor Scheduling for Thermal Management in Biological Wireless Sensor Networks," *International Journal of Distributed Sensor Networks*, vol. 2013, pp. 1–10, 2013.

[2] Y. Osais, F. Yu, and M. St-Hilaire, "Thermal management of biosensor networks," in *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*, 2010, pp. 1–5.

[3] F. R. Yu, Y. Osais, and M. St-Hilaire, "Optimal Management of Rechargeable Biosensors in Temperature-Sensitive Environments," *Vehicular Technology Conference Fall VTC 2010Fall 2010 IEEE 72nd*, 2010.

[4] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *Communications Magazine, IEEE*, vol. 38, no. 1, pp. 54–64, 2000.

[5] A. Karmokar, D. Djonin, and V. Bhargava, "Optimal and suboptimal packet scheduling over correlated time varying flat fading channels," *Wireless Communications, IEEE Transactions on*, vol. 5, no. 2, pp. 446–456, 2006.

[6] H. S. Wang and N. Moayeri, "Finite-state Markov channel- a useful model for radio communication channels," *IEEE Transactions on Vehicular Technologies*, vol. 44, no. 1, pp. 163–171, 1995.

[7] D. Zhang, "Analysis on Markov modeling of cellular packet transmission," 2002, pp. 876–880.

[8] A. Hoang and M. Motani, "Cross-layer Adaptive Transmission: Optimal Strategies in Fading Channels," *IEEE Transactions on Communications*, vol. 56, no. 5, pp. 799–807, May 2008.

[9] H. H. Pennes, "Applied physiology," *Journal of Applied Physiology*, vol. 1, no. 1, pp. 93–122, 1948.

[10] D. M. Sullivan, *Electromagnetic simulation using the FDTD method*. IEEE Press, 2000.

[11] "Matlab." [Online]. Available: http://www.mathworks.com

Software Architecture Quality Measurement Stability and Understandability

Mamdouh Alenezi
College of Computer & Information Sciences
Prince Sultan University
Riyadh 11586, Saudi Arabia

***Abstract*—Over the past years software architecture has become an important sub-field of software engineering. There has been substantial advancement in developing new technical approaches to start handling architectural design as an engineering discipline. Measurement is an essential part of any engineering discipline. Quantifying the quality attributes of the software architecture will reveal good insights about the architecture. It will also help architects and practioners to choose the best fit of alternative architectures that meets their needs. This work paves the way for researchers to start investigating ways to measure software architecture quality attributes. Measurement of these qualities is essential for this sub-field of software engineering. This work explores Stability and Understandability of software architecture, several metrics that affect them, and literature review of these qualities.**

***Keywords*—Software Engineering; Software Architecture; Quality Attributes; Stability; Understandability**

I. Introduction

Software systems are becoming complex, larger, more integrated, and are implemented by the use of several varieties of technologies. These various technologies need to be managed and organized to deliver a quality product. Quality attributes usually assessed and analyzed at the architecture level not at the code level. It is usually the case that when we decide on a appropriate architectural choice (i.e. the system will exhibit its required quality attributes) without the need to wait until the system is developed and deployed, since software architecture enables to predict system qualities.

The software architecture field has been inspired by other engineering domains. This inspiration led the movement to these well-known concepts such as stakeholders and concerns, analysis and validation, styles and views, standardization and reuse, best practices and certification. However, software is inherently different from all other engineering disciplines. Rather than delivering a final product, delivery of software means delivering blueprints for products. Computers can be seen as fully automatic factories that accept such blueprints and instantiate them.

In this work, we pave the way for researchers to start investigating ways to measure software architecture quality. The remainder of this paper is organized as follows. Section II introduces and defines software architecture and discusses its importance. Software metrics are discussed in Section III. Software architecture measurement is presented in Section IV. Tow samples of software architecture quality attributes are discusses is Section V. Section VI presents the software architecture measurement validation techniques. Conclusions are presented in Section VII.

II. Software Architecture

Over the past years software architecture has became an important sub-field of software engineering. There has been substantial advancement in developing new technical approaches to start handling architectural design as an engineering discipline. However, much research is yet to be carried to achieve that. Moreover, the changing nature of technology raises a number of challenges for software architecture.

Designing a software structure is the phase that comes immediately after gathering and analyzing the software requirements. During this phase the software is constructed in terms of components and relationships that link these components with each other [1]. These components and their relationships will illustrate the architecture for particular software. The software architecture of a system has many definitions in the field of software engineering. Software Architecture is defined in the IEEE standards [2] as "fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution". The authors of the book 'Software Architecture in Practice' [1] defined the software architecture as "the set of structures needed to reason about the system, which comprise software elements, relations among them, and properties of both." Software architecture represents the design decisions that are hardest to change and determine the overall system proprieties [3]. Those decisions have to be made before concurrent work on a system can be started. Architecture decisions will not be at a component level, but they span the overall system components and determine their interconnections and constrains. Once all architecture decisions are made, work in individual components can proceed independently [4].

Software architecture is the name of a particular form of abstraction, or model, of software systems. It is considered as abroad abstraction of the system, which contains information about both functional and non-functional requirements. The architecture lays the foundation of the shared communication platform for the various stakeholders. Software architecture typically bridges between requirements and implementation. Software architecture embodies the earliest decisions that shaped the impact on the success/failure of the software system. Software architecture serves as a reasoning, important

communication, analysis, and growth tool for software systems [1].

Several authors came to an agreement that software architecture is a skeleton produced in the early phase of design. Documenting the software architecture is useful both as a means of communication between stakeholders and in providing an overall picture of the system that is to be developed. Architecting the software system is a crucial task since it lays the groundwork for later activities in the software development process. The architecture plays an important factor in the software success or failure. Understanding the architecture is very important for both architects and developers to relate it to requirements, product, and process.

The software architecture influences greatly the system's quality as it can inhibit or enable product's quality attributes. The quality of a software system is largely attributed to its software architecture [5]. Thus, evaluation of this software architecture should be done on a regular basis. Such repeated evaluations ensure that the system remains sustainable and evolvable over a longer period of time [6].

The software architecture can be decomposed into more granular levels, namely packages, components, and modules. Package is used to represent a set of classes that might be hierarchically structured and to perform a series of related tasks [7]. A package is a group of classes that are related to each others or perform one higher purpose. Classes in the same package have special access privilege with respect to one another and may be designed to work together closely. Component in the context of object-oriented design is for organization purposes. Component contains a group of classes and other components as well. A component provides one or two similar system functionalities [8]. Module consists of a large number of classes and sometimes a module is referred to as a package. It provides information hiding for the module allowing a software engineer to see it as a black box [9].

The field of software architecture remains reasonably immature. Although it has an engineering foundation for software architecture, it is not clear yet, there are still several challenges. As a result, we anticipate major new advancements and developments in the software architecture field in the future.

A. Importance

The software architecture is very important in the software development life-cycle. It is considered as the blueprint of the system where important decisions are documented. It is a reference for the whole system in design, development, and maintenance. A poor software architecture may lead to a deficient software product that does not satisfy its customers and can not be adaptive to new changes. David Garlan [10] summarized the importance of software architecture in six aspects of software development:

1) Understanding: Software architecture can be seen as mechanism to simplify our ability to understand complex-large systems by presenting them at a higher level of abstraction [1]. Furthermore, the architecture exposes the high-level constraints on system design, as well as the rationale for making specific architectural choices [11].
2) Reuse: Software architecture supports reuse of components and frameworks. Platforms, frameworks, components, architectural patterns, libraries of plug-ins, add-ins, apps, and domain-specific software architectures are different promoters of reuse.
3) Construction: Software architecture provides a blueprint for development and implementation by showing the major components and dependencies between them. For instance, a layered architecture documents abstraction boundaries between parts of a system's implementation [11].
4) Evolution: Software architecture exposes the dimensions along which a system is expected to evolve. Software maintainers can easily understand the ramifications of changes, and accurately estimate costs of modifications [12].
5) Analysis: Software architecture can be seen as a way to analyze the whole system. These analyses can include satisfaction of quality attributes [1], system consistency checking [1], conformance to constraints forced by an architectural style, and domain-specific analyses for architectures built in specific styles.
6) Management: Software architecture can be seen as a viable milestone in any industrial software development process. Critical evaluation of an architecture leads to clear understanding of requirements, implementation plans, and possible risks, which will reduce the amount of rework required to address problems later in a systems life-time [1].

B. Quality Attributes

This section summarizes several important quality attributes across the software architecture domain. A quality attribute (QA) is a measurable feature of a system, which is utilized to stipulate how well the system satisfies stakeholders. You can consider a quality attribute as measuring the goodness of that property. ISO/IEC 9126 [2] classifies quality attributes of software as functionality, maintainability, usability, efficiency, reliability, and portability. These characteristics are attributes that can describe a software system. These quality attributes are further derives the sub-characteristics with more attributes. The quality characteristics are refined to sub-characteristics and these sub-characteristics are refined to attributes or measurable properties using several metrics. A metric is a defined measurement method that assigns a value to that attribute.

Quality attributes are strongly related to non-functional requirements of a system. One of the responsibilities of the software analyst to come up with a complete list of quality attributes before architecting and designing the system. Quality attributes commonly include efficiency (time, efficiency, resource economy), functionality (completeness, security, interoperability), maintainability (expandability, modifiability, testability), portability (hardware independence, software independence, installability, reusability), reliability (error tolerance, availability), and usability (understandability, user interface, learnability). Figure 1 shows the quality attributes in ISO/IEC 9126.

When software architects are able to measure and quantify these quality attributes, they will be able to enumerate feasible

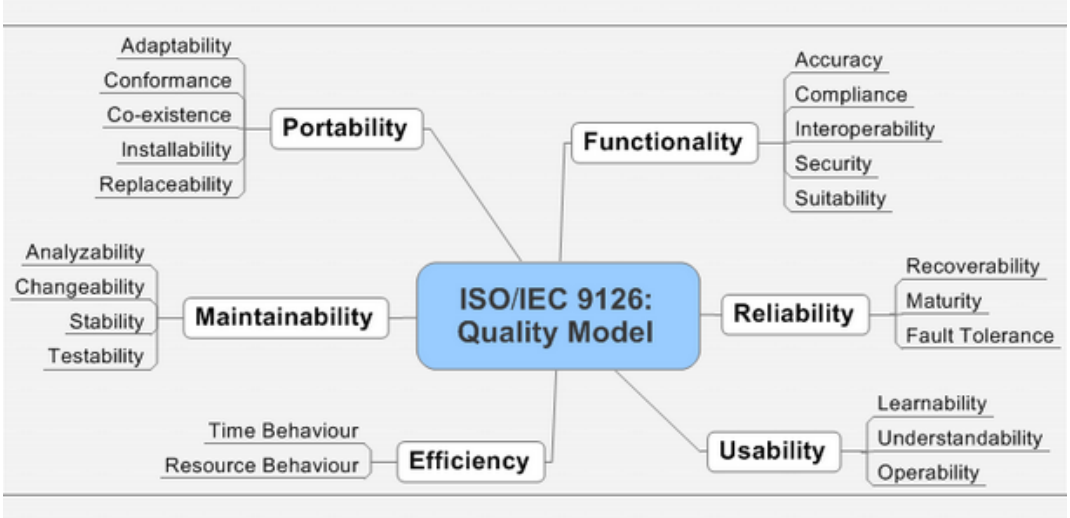


Fig. 1. Quality Attributes in SQuaRe

architecture design and evaluate all quality attributes. As each quality attribute is assigned a measure, a total score can be calculated in order to help the architect which design alternative to use. When the evaluation process is complete, the design with high score will be chosen.

III. Software Metrics

Measurement is crucial for any science or engineering field. Organizations strive to come up with meaningful measures that indicate progress or performance. Measurement in software engineering is considered a crucial factor to evaluate the software quality characteristics such as functionality, usability, reliability, efficiency, maintainability, and portability. In software engineering, there is still a lack in that discipline. We still need to work in consolidating terminology, principles and methods of software measurement [13]. Software measurement activities consist of direct and indirect assessments, as well as predictions [13], [14]. Measurement allows us to understand the current situation and to come up with clear benchmarks that are useful to set goals for the future behavior. Software measurement [15] is not limited only to evaluate a software product but it will be used to evaluate the software development process. Measurement is a crucial activity in all empirical studies.

Software metrics field is an interesting field in the software engineering community since more than 30 years. The interest in metrics by both academician and practitioners is growing rapidly. Software metrics are defined as [16] "standard of measurement, used to judge the attributes of something being measured, such as quality or complexity, in an objective manner". Software metrics are measures utilized to evaluate the process or product quality. These metrics helps project managers to know what is the progress of software and evaluate the quality of the various artifacts produced during development. The software requirements engineers can validate and verify requirements. Software metrics are required to capture various software attributes at different phases of the software development [17]. Software metrics are required to adequately measure various points in the software development process.

Software metrics constitute the main approach to software measurement [13], [18], [19]. Software metrics and quality are major players in measurement of software quality. Measuring software artifacts should focus on selecting the right metrics for each software and on how to apply them [13].

IV. Software Architecture Measurement

Software architecture measurement suffers from what people calls the tyranny of the dominant architectural principle. The assessments of certain principles are overstressed, other equally important design principles have been omitted in architecture measurement processes.

Each quality attribute can be measured using different characteristics of the software architecture. The characteristics can be of size, complexity, coupling, cohesion, or others. Furthermore, each quality attribute can be measured by combining several existing measures. Several quality attributes are very similar and can complement each other. Several software metrics can be combined together to measure a certain property using either composition or aggregation [20]. In composition, software metrics used to assess a property can be composed by (1) simple or weighted average of the metrics. This can be used only when the different metrics have similar range and semantic; (2) thresholding; (3) interpolating; or (4) a combination of these methods. In aggregation, several steps are required. (1) a weighting function is applied to each metric then (2) average the weighted values of the metrics then (3) we compute the inverse function of the average.

V. Architecture Quality Attributes

In this section we review several attempts to measure two quality attributes of the software architecture, namely stability and understandability.

A. Stability

The primary goal from the architecture evaluation is to assess and validate the software architecture using system-

atic methods and procedures [21]. This evaluation is accomplished to ensure the examined software architecture satisfy one or more of quality characteristics. One desired quality of the software architecture is stability. Stability is one of the maintainability characteristics of the ISO/IEC SQuaRe quality standard [2]. According to this standard, stability is defined as the degree to which the software product can avoid unexpected effects from modifications of the software [2]. Architectural stability reduces unnecessary architecture rework as the software system's functionality is expanding over multiple versions, thus reducing implementation costs.

As architectures have a profound effect on the operational life-time of the software and the quality of the service provision, architectural stability could be considered a primary criterion towards achieving the long-livety of the software. Architectural stability is envisioned as the next step in quality attributes, combining many inter-related qualities.

Several researchers proposed several metrics to measure the stability of software architecture. Stability is the ability of software to remain unchanged while facing new requirements or changing the environment. The software has to accommodate some of these changes and they should not affect the software stability, while other may harm the software stability. This section presents an overview of several attempts in the literature to measure the stability of particular software.

Ahmed et al. [22] proposed a new way to measure the architectural stability of an object oriented system by using similarity metrics. These metrics compare pair versions of a system. First metric is Shallow Semantic Similarity Metric (SSSM), and the purpose of this metric is to the measure the semantic similarity between components in a pair of systems. Second metric is Relationship-Based Similarity Metric (RBSM) and it is aiming to measure similarity between the relationships that exist in a pair of systems. A regression line is generated for the architecture changes with releases from these similarity values. A higher value indicates a stable architecture.

Ebad et al. [3] continued the work that is proposed in [22] by developing a new architecture stability metric (ASM) that measure cross-architecture components communications in term of inter-package connections (IPC). The idea behind IPC is when a pair of releases of a software system is compared; there are three types of changes that may happen (addition, deletion, and modification). ASM value will be between 0 and 1, where 1 means lowest possible amount of changes between two releases which means stable software architecture. ASM is validated by a set of mathematical properties which are: non-negativity, normalization, null value, maximum value, transitivity, package cohesion impact and change impact. Moreover, this metric is experimentally validated by using two open source projects: JHotDraw and abstract windowtoolkit. Measurements of the ASM are illustrated by lines of code for original IPCs and deleted IPCs, and added IPCs across releases in the two previously mentioned projects.

Aversano et al. [23] evaluated the software architecture for a set of open source software projects. Most of these projects are selected from sourceforge. Stability is the characteristic that is examined in order to evaluate the software core architecture. The evolution of certain software is considered when the software components are changed during the software releases. Two metrics are proposed to measure the stability of each release. These two metrics are Core Design Instability (CDI) and Core Call Instability (CCI). Both metrics provide a measure of how much the architecture of a software system changed passing from a release to another one. CDI metric finds the change in terms of number of packages and CCI finds the change in terms of number of the interactions among packages. Smaller values mean less change which means greater stability. All these metrics are based on calculating fan-in and self-call for software packages.

Alshayeb et al. [24] mentioned that none of the existing measures have included all class aspects such as class relationships, attributes, and methods. At the first of this study all properties that affect the class stability are identified; these properties are class access level, class interface level, inherited class name, class variable, class variable access- level, method signature, method access level , method body. Then from these properties the proposed metric is recognized. The name of the discovered metric is Class Stability Metric (CSM). Stability is calculated by counting the number of unchanged properties between two classes in version i+1 and version i divided by the maximum possible change value, then summation of all these properties is divided by the number of the properties which is eight. This metric is theoretically validated by some properties. Moreover, this metric is empirically validated through two Java systems. The result of this empirical study indicates that this metric is highly negatively correlated with maintenance effort.

Li et al [25] proposed new metrics to measure the stability for the software design. They highlighted that metrics that are discovered by Chidambe & Kemerer [26] cant measure all aspects of Object Oriented. Examples of these aspects are the change in the class name, class number, and class inheritance relations. From this imperfection of C&K metrics, authors proposed these three metrics: System Design Instability (SDI), Class Implementation Instability (CII), and System implementation Instability (SII). The main goal that pointed out in this study is to justify how the information that is gathered from theses metrics can help project manager to adjust the project plan. These metrics are experimentally examined against C&K metrics. They found out that SDI and CII measure Object Oriented aspects that are different from the aspects that are measured by C&K metrics.

Abdeen et al. [27] introduced a complementary set of coupling and cohesion metrics that assess packages organization in large legacy object-oriented software. These metrics are aiming to measure the modularization for an object-oriented system. Here are the metrics that are discovered by Abdeen: Index of Inter-Package Usage (IIPU), Index of Inter-Package Extending (IIPE), Package Focus (PF), Index of Package Service Cohesion (IPSC), and Index of Package Changing Impact (IPCI). These metrics are defined with respect to some modularity principles that are related to packages. Examples of these principles are information hiding, changeability and communality of goal. These metrics are defined with regard to two different types of object-oriented inter-class dependencies: method call and inheritance relationships. All metrics that are discovered in this work are validated against the mathematical properties that have to be existed in any cohesion or coupling metric.

Sethi et al. [28] mentioned that none of existing met-

rics, that are used to assess the modularity and stability of architecture decomposition, has considered environmental factors that drive software changes. From this point a suite of metrics is proposed: decision volatility, design volatility, impact scope, concern scope, concern overlap and independent level. Theses metrics consider environmental factors that drive software changes. Moreover they tried to measure how well a particular architecture produces independently substitutable modules. Furthermore, these metrics did not require having knowledge about the implementation details. In these metrics design dimensions and environmental conditions are modeled as variables and their relations are modeled as logical constraints. These metrics are evaluated using eight aspect-oriented and object-oriented releases of software product-line architecture.

It has been realized the work that is produced by Chidamber and Kemerer [26] has been cited in most of other studies. They developed a suite of metrics that is used for measuring a particular object- oriented design. The primary goal is to develop and validate theoretically and empirically a set of object-oriented metrics. These metrics are Weighted Method per Class (WMC), Depth of Inheritance of Tree (DIT), Number of Children (NOC), Coupling Between Object Class (CBO), Response for Class (RFC) and Lack of Cohesion of Metric (LOCM). Each one of this matric is evaluated theoretically against six properties: noun-coarseness, non-uniqueness, design details are important, Monotonicity, noun-equivalence of intersection and intersection increases complexity. Then, the implementation of these metrics is demonstrated through data collection from both C++and SMALLTALK implementations. From the obtained data, it has been shown how each one of this metric can help project managers and senior designers to obtain useful information about the entire evolution of a particular application.

Hassaine et al. [29] proposed a novel approach to investigate some metrics (code decay indicators) on software, that serve as symptoms, risk factors, and predictors of decay, in the context of an evolving architecture. The name of their approach is ADvISE and it aims for analyzing the evolution of certain software architecture at various levels (classes, triplets, and micro-architectures). The first step in observing architectural decay is to use a diagram matching technique to identify structural changes among versions of architectures. The second step is detecting the class renaming by using structure-based and text-based techniques. The third step is architecture diagram matching by using a bit vector algorithm to perform diagram matching between two programs versions in order to find the common triplets. The fourth step is architecture diagram clustering by applying the incremental clustering algorithm to find the sets of connected triplets. These sets will form the stable micro-architecture between two program versions. The fifth step is architecture evolution by performing a pairwise matching for programs architectures in order to identify sets of stable triplets and micro-architecture. The authors applied their approach on three open-source systems: JFreeChart, Rhino and Xerces-J to answering the following research questions: RQ1: What are signs of architectural decay and how can they be tracked down? The authors studied the graph of architectures evolution for each system, and then they showed these indicators to provide useful insights regarding the signs of software aging. RQ2: Do stable and unstable micro-architectures have the same risk to be fault prone? The authors showed stable micro-architectures, which are belonging to the original design, are significantly less bug-prone than unstable micro-architectures.

Jazayeri et al. [30] did retrospective analysis to evaluate and assess the architecture for telecommunication software. Twenty releases are selected to observe the evolution of this software. This kind of evaluation helps project managers to predict about how the future of the architecture will be look like. Metrics that are used in this work are likely to be the observation of the some simple measures between a pair of releases while the software is being evolved. Examples of these metrics are module size, number of modules changed, number of modules added, number of modules changed in the same sequence of release, number of programs in the same version of release.

Abreu and Melo. [31] evaluated the impact of object oriented design on software quality characteristics such as defect density, failure density, and normalized rework. There is a set of metrics for object oriented design MOOD. These metrics are empirically evaluated against the software quality characteristics by calculating the correlation coefficients where the quality characteristics are the dependents variables and the design metrics are the independent variables. Examples of these design metrics are 1) method hiding factor (MHF), attribute hiding factor (AHF), method inheritance factor (MIF), attribute inheritance factor (AIF), polymorphism factor (POC) and coupling factor (COF).To quantify the impact of OO design on software quality, a predictive model is developed. The results show that the design alternatives may have a strong influence on resulting quality. For the study validity multiple R, R square and adjusted R square are calculated for the software quality characteristics.

Olague et al. [32] utilized entropy to reduce spikes in the original SDI metric that is produced by Li.et al [25] and proposed the new SDIe metric. This study highlighted that the dynamic nature of the agile development process could obscure an analysis of software stability. Also, this study notified that the SDIe metric is easier than the SDI metric to compute the stability for a particular software system. The reason behind that is SDIe is able to be automated instead of requiring the close investigation of code by human judgment. The SDIe metric is calculated using the number classes added, deleted, changed and unchanged from the previous iteration. SDIe metric is theoretically investigated and validated using the Kitchenham criteria [33] and the Zuse requirements [34] for software measures. Moreover SDIe is empirically tested over two software projects by comparing SDIe metric with the original SDI, using SDIe to assess the software evolution, and comparing SDIe metric to the Chidamber and Kemerer [26].

Tonu et al. [21] proposed an approach that can helps developers in evaluating stability for a particular software architecture. Evaluating the software architecture is based on analyzing the changes in the softwares aspects form one release to another. Software aspects can be structural, behavior, or economical, in this research work the focus is only on the structural aspects. Growth rate, changes rate, coupling, cohesion are the measures that are applied in this approach to do retrospective analysis. Then, evolutionsensitive and evolution critical parts are identified by observing how the subsystems

are interconnected between each other. This approach is empirically evaluated on two spreadsheet applications by selecting nine releases for each application, and then the results of the architecture stability are discussed.

Roden et al. [35] performed empirical study on six different highly iterative projects with multiple iterations by observing the system packages. These projects are evaluated using Total Quality Index (TQI), System Design Instability (SDI) and System Design Instability using Entropy (SDIe) metrics. TQI is calculated by summation of quality factors. Each quality factor is calculated by a weighted formula of quality properties. This statistical analysis gives a string relation between TQI and SDI. Because of the similarity between TQI and stability metrics and since the stability metrics require human participation, the authors in this work suggest to use TQI instead of SDI.

Yu and Ramawamy [36] reintroduced an approach to represent and normalize the evolution stability of software modules. This approach is based on version differences of evolving software models by measuring the normalized distance of two versions for a module structure and module source code. For example by giving two versions V_i and V_j of a software component in a given release period let say m months, the component is said to be more stable in this period, if the measurement structure distance Di, j(source_code) or Di, j (structure) is considered to be small. A case study is applied on this model by evaluation the evolution of Linux and FreeBSD applications by selecting two versions.

Raemaekers et al. [37] highlighted backward compatibility as a very important concern to build an Application Programming Interface (API). API developers have to ensure the public interfaces are stabile because other systems are depending on theme. From this point, a way to measure interface and implementation stability of a library is introduced. Four metrics are proposed in this study to provide different insights in both implementation and interface stability. These four metrics are weighted number of removed methods, the change in metric values in existing units, the ratio between change in new and old methods and the percentage of new methods. Smaller value indicates greater stability. Moreover, they illustrate who these metrics can be used to help project managers or developers to make a decision regarding interface libraries by applying three scenarios. These metrics are theoretically evaluated by applying theme on the most frequently used Apache common libraries by selection a set of industrial systems which making use of Apache libraries. From the architectural perspective, the drawback of these metrics is the granularity level of the metric; they are not at package level (coarse-grain) but method level (fine-grain)

Ratiu et al. [38] started by defining two measurements that are used to identify which structure is considered a god class or data class. These measurements are based on object oriented design metrics and threshold for each metric. First measurement is used to identify god classes and it is based on these metrics: Access to Foreign Data (ATFD) and Weighted Method Count (WMC), Tight Class Cohesion (TCC) ,Number of Attributes (NOA). While the another measurement is used to identify data classes and it is based on these metrics: Weight of a Class (WOC), Number of Methods (NOM) ,Weighted Method Count (WMC) ,Number of Public Attributes (NOPA) and Number of Accessory Methods (NOAM). Then, they proposed two measurements that are applied on the history of a design structure. One of these measurements is used to measure the stability of a class (Stab) and another is used to measure the persistence of a design flaw (Pers). A class is considered stable with respect to measurement M version i and number of versions if there is no change in the measurement M. while a flaw is considered persistence in a class with respect to measurement M version i and number of versions if this flaw is exist in all versions of this class. Their approach is applied on three case studies: two in house projects, and one on a large open source framework. By observing the data while applying their approach, they discuss which classes, ether these classes are god classes or data classes, are considered to be harmless or harmful classes.

Bansiya et al. [39] introduced a methodology to evaluate framework architecture characteristics and stability that based on quantitative assessment on the change in framework versions using object oriented metrics. This approach consists of four steps that need to follow in order to calculate the extent-of-change measure. First step is identifying structure characteristics that evaluate the architecture of framework. There are two types of structure characteristics: static and dynamic. Example of static structure characteristics are number of classes, number of class hierarchies, number of single and multiple inheritances, and average depth and width of class inheritance hierarchies. Examples of dynamic structure characteristics are number of services a class provides, class coupling, and number of inheritance related classes. Second step is defining metrics for each one of these structure characteristics. Third step is collecting the data from the defined metrics by applying theme on a case study. Finally, for each release the extent-of-change is calculated by normalizing the values of these metrics. Once all values are normalized, the aggregate-change is calculated by summation of these values. Then the extent-of-change is calculated by taking the difference of the aggregate change value of a version i with the aggregate change value of the first version. The extent of change measure can be used as an indicator to identify the stability for a particular system structure, low number indicates high stability.

Alenezi and Khellah. [6] highlighted in their work that most of previous studies have not considered measuring the system instability at the system architecture level and most of them are focusing at the package level. From this point, they introduce a new approach to compute the instability changes for particular software architecture at a certain release while observing its evolution. Their approach is based on the instability I metric that is introduced by Martin [40] which tells how a flexible a package is able to change, the ration of the efferent coupling to the total the coupling for a package. The proposed approach is to reflect the instability change to evolution by expressing the aggregate system instability change for certain release as being composed of the average of two elements: ΔI: the amount of change in the system stability for all common packages and ASI: Aggregative System Instability for the current added packages. They illustrate how an improvement by just calculating ASI is incorrect and how the incorrectness would be solved by the proposed approach. This approach is empirically validated on two open source systems implemented in Java JEdit and PDFBox by selecting twelve releases.

Table I summarizes the discussed papers with their metrics

and granularity level. The literature reveals that the available architectural stability measures have some limitations. For instance, the metrics proposed by [39] and the methodology presented by [41] consider the method/class level which is fine-grain level. Bansiya work is suitable only when having cost and economic as his backdrop [39].

B. Understandability

One desired quality of the software architecture is understandability. Understandability in the context of software architecture simply means whether system architecture is understandable to average architects. Understandability is one of the usability characteristics of the ISO/IEC SQuaRe quality standard [2]. Understandability refers to the capability that to what extent users with different backgrounds can understand the architecture. Understandability is an essential characteristic of software quality since the difficulty of understanding the software architecture system inhibits its reuse and maintenance. Understandability is the capability of the software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use.

Several researchers have explored the relationships between several metrics and understandability. Table II summarizes their efforts, goals, and research methodology. Gupta and Chhabra [7] proposed a package coupling metric and empirically validated it against package understandability. Their study used one metric and they performed correlation analysis. They validated the package coupling metric with regard to the understandability of packages measured by assessing the effort required to fully understand the packages' functionalities. They concluded that there is a strong correlation between package coupling and the effort required to understand a package.

Elish [42] used several metrics (Size (NC), Coupling (Ca, Ce), and Stability (I, D)) and conducted a case study to correlate these metrics with package understandability of two open source software systems. The results of the study indicated statistically significant correlation between most of the metrics and understandability of a package.

Hwa et al. [9] have proposed hierarchical quality model (consist of 4 levels and 3 links to connect these levels) to assess the understandability of the modular design of an Object-Oriented software system. At the level 2 of their proposed model 6 design properties were identified that affect understandability of the modular design of a system. One of these properties is the coupling and they have found that the coupling property has a negative influence on understandability and that means the higher number of coupling the harder is to understand the system. in their proposed hierarchical quality model have found that there is a positive influence of the design size on understandability. The larger the size the harder is to understand.

Stevanetic and Zdun [8] carried a study to examine the relationships between the effort required to understand a component and component level metrics that describe component's size, complexity and coupling. Correlation, collinearity and multivariate regression analysis were performed. The results of the analysis show a statistically significant correlation between the metrics and the effort required to understand a component.

Stevanetic and Zdun [43] found that the architecture at the abstraction level that is sufficient to adequately map the systems relevant functionalities to the corresponding architectural components (i.e. , each component in the architecture corresponds to one systems relevant functionality) significantly improves the architecturelevel understanding of the software system, as compared to two other architectures that have a low and a high number o f elements . This means highly abstracted system; low in number of elements by merging systems relevant functionalities into one component would decrease the understandability of such systems. As well as, highly detailed systems; high number of elements by scattering the functionalities into several components would decrease the understandability of such systems. However, when each component in the software architecture corresponds to one of the systems functionalities would significantly improves the understandability at the architectural level.

Stevanetic et. al [44] have done a controlled experiment on 75 students of the Software Architecture lecture. The students were divided into 3 groups and each group had been given a different architectural representation of the same large system. The first architectural representation was hierarchical representation where all components at every abstraction level in the hierarchy are present. Second architectural representation concentrates on the lowest level no hierarchy used. Third architectural representation at concentrate on the highest level component in th hierarchy by does not use hierarchal abstraction. The conclusion of the experiment was that by using the hierarchical architecture would result on a better understandability at the architecture-level.

VI. Software Architecture Measurement Validation

It is commonly accepted that a software metric should be validated following two different validations: theoretical and empirical validations. This will ensure that the metric measures the attribute that it is supposed to measure and provide evidence on the usefulness of the metric. Many existing software metrics are criticized from two standpoints, theoretically and empirically. Several researchers pointed that most software metrics were developed with no or little theoretical basis [46], [47]. Furthermore, even though some metrics are theoretically valid, they lack empirical evaluation [48].

A. Theoretical Validation

Theoretical validation makes sure that a metric is measuring what is supposed to measure. The first requirement for theoretical validation is that either the analyst has an intuitive understanding of the concept that is being measured and/or that the software engineering community has a consensual intuitive understanding of the concept. There are several frameworks for the theoretical validation of metrics. Some of them are mainly subjective, while others rely on either axiomatic or measurement theory foundations. Briand et al. [49] have discussed the application of measurement theory in software engineering. The theoretical validation is generally carried out using measurement frameworks based on property-based approaches. Property-based approaches [46] allow one to prove that a measure satisfies properties characterizing a concept (e.g., size, complexity, coupling). This approach is

TABLE I. SUMMARY OF THE DISCUSSED PAPERS

Reference	Metrics	Granularity level
Abdeen et al. [27]	Index of Inter-Package Usage (IIPU), Index of Inter-Package Extending (IIPE), Package Focus (PF), Index of Package Service Cohesion (IPSC), and Index of Package Changing Impact (IPCI)	Package, architecture
Olague et al. [32]	Software Design Instability using Entropy (SDIe) metric	Architecture
Abreu and Melo. [31]	Method Hiding Factor (MHF), Attribute Hiding Factor (AHF), Method Inheritance Factor (MIF), Attribute Inheritance Factor (AIF), Polymorphism Factor (POC) and Coupling Factor (COF)	Architecture , classes
Ebad et al. [3]	Architecture Stability Metric (ASM)	Package, architecture
Ahmed et al. [22]	Shallow Semantic Similarity Metric (SSSM),Second metric is Relationship-Based Similarity Metric (RBSM)	Component, architecture
Aversano et al. [23]	Core Design Instability (CDI) and Core Call Instability (CCI)	Package, architecture
Sethi et al. [28]	Decision volatility, design volatility, impact scope, concern scope, concern overlap and independent level	Architecture
Hassaine et al. [29]	ADvISE: Architectural Decay In Software Evolution	Architecture
Li et al [25]	System Design Instability (SDI), Class Implementation Instability (CII), and System implementation Instability (SII)	Architecture , classes
Alshayeb et al. [24]	Class Stability Metric (CSM)	Class
Chidamber and Kemerer [26]	Weighted Method per Class (WMC), Depth of Inheritance of Tree (DIT), Number of Children (NOC), Coupling Between Object Class (CBO), Response for Class (RFC) and Lack of Cohesion of Metric (LOCM)	Class
Jazayeri et al. [30]	Module size, number of modules changed, number of modules added, number of modules changed in the same sequence of release, number of programs in the same version of release	Architecture
Tonu et al. [21]	Growth rate, changes rate, coupling, cohesion	Architecture, classes
Roden et al. [35]	Total Quality Index (TQI)	Package, architecture
Raemaekers et al. [37]	Weighted number of Removed Methods (WRM), the amount of Change in Existing Methods (CEM), the Ratio of Change in New to Old methods (RCNO), and the Percentage of New Methods (PNM)	Library, method
Yu and Ramawamy [36]	Distance (source code) or Distance (structure)	Component
Ratiu et al. [38]	Stability of a class (Stab) and persistence of a design flaw (Pers)	Class
Bansiya et al. [39]	The extent-of-change measure	Architecture
Alenezi and Khellah. [6]	Aggregative System Instability	Architecture

TABLE II. SUMMARY OF THE DISCUSSED PAPERS

Reference	Goal	Methodology
Gupta and Chhabra [7]	To Propose new metrics for measurement of package level coupling.	Theoretical and Empirical
Elish [42]	To explore the relationships between five package-level metric (Size, Afferent, Efferent, Instability and Distance) and the average effort required to understand a package in O.O. design.	Empirical
Stevanetic and Zdun [45]	Systematic mapping study on software metrics related to the understandability concept of such higher-level software structures with regard to their relations to the system implementation	Systematic Mapping Study
Hwa et al. [9]	To propose a hierarchical model to assess understandability of modularization in large-scale O.O. software.	Empirical
Stevanetic and Zdun [8]	To examine the relationships between the efforts required to understand a component, measured through the time that participant spent on studying a component and component level metrics that describe components size, complexity and coupling.	Experimental
Stevanetic and Zdun [43]	To examine the effect of the level of abstraction of the software architecture representation (3 levels) on the architecture-level understandability of a software system.	Experimental
Stevanetic et. al [44]	To examine the impact of hierarchies on architectural-level software understandability.	Empirical

comprehensive framework which defines the structural properties of software system mathematically which matches with the methodology of the proposed metrics in this thesis. The theory provides an empirical interpretation of the numbers (of software measures) by the hypothetical empirical relational system.

Arvanitou el al. [50] used a property-based approach to theoretically validate their new metric, which measures the coupling and class proneness to the ripple effect. Khoshkbarforoushha et al. [51] theoretically validated their new metric using a property-based framework. Tripathi and Kushwaha [52] theoretically validated their package level coupling metric using a property-based framework. Lenhard et al. [53] theoretically validated their new metric that measures installability of service orchestrations using a property-based approach. Gupta and Chhabra [7] introduced a coupling metric at the package level and theoretically validated it through property-based approach.

B. Empirical Validation

The purpose of empirical validation is to show the usefulness of the metric in real application using real data from software projects. The goal is to show that this new metric has a concrete value in a real settings. The empirical validation

of a software metric can be done using different empirical techniques. These techniques include controlled experiments, surveys, or case studies. A controlled experiment is a rigorous and controlled study. A Survey is research performed in retrospect, when the method has been in use for a certain period of time. A Case Study is an observational study, and data are collected for a specific purpose throughout the study. Experiments provide a high level of control and are useful for validating software metrics.

Arvanitou el al. [50] compared their new measure, which measures the coupling and class proneness to the ripple effect and several coupling metrics empirically to evaluate the usefulness of their metric. Khoshkbarforoushha et al. [51] empirically validated their new metric using an experiment in how the new metric can predict design-level estimation of the potential reusability of the BPEL processes. Tripathi and Kushwaha [52] empirically validated their package level coupling metric by comparing it to other package level coupling metrics. Lenhard et al. [53] empirically validated their new metric using BPEL engines to evaluate their installability and the deployability of a set of functionally different processes. Gupta and Chhabra [7] introduced a new coupling metric at the package level and empirically validated it using package understandability.

VII. Conclusion

In this work, we have laid the foundation for researchers and practitioners to come up with better ways of measuring the software architecture quality attributes. The definition and importance of software architecture were discussed. How to evaluate these measurements were also comprehensively presented in this work. Stability and Understandability were given more focus for their importance and effect on software. Future directions include devising new metrics to measure both stability and understandability of software architecture.

References

[1] L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice*, 3rd ed. Addison-Wesley Professional, 2012.

[2] ISO/IEC/IEEE, "Systems and software engineering – architecture description," *ISO/IEC/IEEE 42010:2011(E) (Revision of ISO/IEC 42010:2007 and IEEE Std 1471-2000)*, pp. 1 –46, 1 2011.

[3] S. A. Ebad and M. A. Ahmed, "Measuring stability of object-oriented software architectures," *IET Software*, vol. 9, no. 3, pp. 76–82, 2015.

[4] M. Alenezi and F. Khellah, "Architectural stability evolution in open-source systems," in *Proceedings of the The International Conference on Engineering & MIS 2015*. ACM, 2015, p. 17.

[5] S. Sehestedt, C.-H. Cheng, and E. Bouwers, "Towards quantitative metrics for architecture models," in *Proceedings of the WICSA 2014 Companion Volume*. ACM, 2014, p. 5.

[6] M. Alenezi and F. Khellah, "Evolution impact on architecture stability in open-source projects," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 5, no. 4, pp. 24–35, 2015.

[7] V. Gupta and J. K. Chhabra, "Package coupling measurement in object-oriented software," *Journal of Computer Science and Technology*, vol. 24, no. 2, pp. 273–283, 2009.

[8] S. Stevanetic and U. Zdun, "Exploring the relationships between the understandability of components in architectural component models and component level metrics," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2014, p. 32.

[9] J. Hwa, S. Lee, and Y. R. Kwon, "Hierarchical understandability assessment model for large-scale oo system," in *Asia-Pacific Software Engineering Conference, 2009. APSEC'09.* IEEE, 2009, pp. 11–18.

[10] D. Garlan, "Software architecture: a travelogue," in *Proceedings of the Future of Software Engineering*. ACM, 2014, pp. 29–39.

[11] F. Bachmann, L. Bass, D. Garlan, J. Ivers, R. Little, P. Merson, R. Nord, and J. Stafford, *Documenting Software Architectures: Views and Beyond*. Addison-Wesley Professional, 2011.

[12] D. Garlan, J. M. Barnes, B. Schmerl, and O. Celiku, "Evolution styles: Foundations and tool support for software architecture evolution," in *Joint Working IEEE/IFIP Conference on Software Architecture, 2009 & European Conference on Software Architecture.* IEEE, 2009, pp. 131–140.

[13] C. G. P. Bellini, R. D. C. D. F. Pereira, and J. L. Becker, "Measurement in software engineering: From the roadmap to the crossroads," *International Journal of Software Engineering and Knowledge Engineering*, vol. 18, no. 01, pp. 37–64, 2008.

[14] I. Sommerville, *Software Engineering*, 10th ed. Pearson, 2016.

[15] R. Malhotra, *Empirical Research in Software Engineering: Concepts, Analysis, and Applications*. CRC Press, 2015.

[16] K. Khosravi and Y.-G. Guéhéneuc, "On issues with software quality models," in *Proceedings of the 11th Working Conference on Reverse Engineering*, 2004, pp. 172–181.

[17] M. Alenezi and I. Abunadi, "Quality of open source systems from product metrics perspective," *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 5, p. 143, 2015.

[18] A. Gopal, M. S. Krishnan, T. Mukhopadhyay, and D. R. Goldenson, "Measurement programs in software development: determinants of success," *IEEE Transactions on Software Engineering*, vol. 28, no. 9, pp. 863–875, 2002.

[19] I. Abunadi and M. Alenezi, "An empirical investigation of security vulnerabilities within web applications," *Journal of Universal Computer Science*, vol. 22, no. 4, pp. 537–551, 2016.

[20] K. Mordal, N. Anquetil, J. Laval, A. Serebrenik, B. Vasilescu, and S. Ducasse, "Software quality metrics aggregation in industry," *Journal of Software: Evolution and Process*, vol. 25, no. 10, pp. 1117–1135, 2013.

[21] S. A. Tonu, A. Ashkan, and L. Tahvildari, "Evaluating architectural stability using a metric-based approach," in *Proceedings of the 10th European Conference on Software Maintenance and Reengineering*. IEEE, 2006, pp. 10–pp.

[22] M. Ahmed, R. Rufai, J. AlGhamdi, and S. Khan, "Measuring architectural stability in object oriented software," *Stable Analysis Patterns: A True Problem Understanding with UML*, p. 21, 2004.

[23] L. Aversano, M. Molfetta, and M. Tortorella, "Evaluating architecture stability of software projects," in *20th Working Conference on Reverse Engineering (WCRE)*. IEEE, 2013, pp. 417–424.

[24] M. Alshayeb, M. Naji, M. O. Elish, and J. Al-Ghamdi, "Towards measuring object-oriented class stability," *IET Software*, vol. 5, no. 4, pp. 415–424, 2011.

[25] W. Li, L. Etzkorn, C. Davis, and J. Talburt, "An empirical study of object-oriented system evolution," *Information and Software Technology*, vol. 42, no. 6, pp. 373–381, 2000.

[26] S. R. Chidamber and C. F. Kemerer, "A metrics suite for object oriented design," *IEEE Transactions on Software Engineering*, vol. 20, no. 6, pp. 476–493, 1994.

[27] H. Abdeen, S. Ducasse, and H. Sahraoui, "Modularization metrics: Assessing package organization in legacy large object-oriented software," in *18th Working Conference on Reverse Engineering (WCRE)*. IEEE, 2011, pp. 394–398.

[28] K. Sethi, Y. Cai, S. Wong, A. Garcia, and C. Sant'Anna, "From retrospect to prospect: Assessing modularity and stability from software architecture," in *Joint Working IEEE/IFIP Conference on Software Architecture & European Conference on Software Architecture. WICSA/ECSA 2009*. IEEE, 2009, pp. 269–272.

[29] S. Hassaine, Y.-G. Guéhéneuc, S. Hamel, and G. Antoniol, "Advise: Architectural decay in software evolution," in *16th European Conference on Software Maintenance and Reengineering (CSMR)*. IEEE, 2012, pp. 267–276.

[30] M. Jazayeri, "On architectural stability and evolution," in *Proceedings of the 7th Ada-Europe International Conference on Reliable Software Technologies*. Springer-Verlag, 2002, pp. 13–23.

[31] F. B. E Abreu and W. Melo, "Evaluating the impact of object-oriented design on software quality," in *Software Metrics Symposium, 1996., Proceedings of the 3rd International*. IEEE, 1996, pp. 90–99.

[32] H. M. Olague, L. H. Etzkorn, W. Li, and G. Cox, "Assessing design instability in iterative (agile) object-oriented projects," *Journal of Software Maintenance and Evolution: Research and Practice*, vol. 18, no. 4, pp. 237–266, 2006.

[33] B. Kitchenham, S. L. Pfleeger, and N. Fenton, "Towards a framework for software measurement validation," *IEEE Transactions on Software Engineering*, vol. 21, no. 12, pp. 929–944, 1995.

[34] H. Zuse, *A framework of software measurement*. Walter de Gruyter, 1998.

[35] P. L. Roden, S. Virani, L. H. Etzkorn, and S. Messimer, "An empirical study of the relationship of stability metrics and the qmood quality models over software developed using highly iterative or agile software processes," in *Seventh IEEE International Working Conference on Source Code Analysis and Manipulation SCAM 2007*. IEEE, 2007, pp. 171–179.

[36] L. Yu and S. Ramaswamy, "Measuring the evolutionary stability of software systems: case studies of linux and freebsd," *IET Software*, vol. 3, no. 1, pp. 26–36, 2009.

[37] S. Raemaekers, A. van Deursen, and J. Visser, "Measuring software library stability through historical version analysis," in *28th IEEE International Conference on Software Maintenance (ICSM), 2012*. IEEE, 2012, pp. 378–387.

[38] D. Rapu, S. Ducasse, T. Gîrba, and R. Marinescu, "Using history information to improve design flaws detection," in *Proceedings. Eighth European Conference on Software Maintenance and Reengineering, 2004.* IEEE, 2004, pp. 223–232.

[39] J. Bansiya, "Evaluating framework architecture structural stability," *ACM Computing Surveys (CSUR)*, vol. 32, no. 1es, p. 18, 2000.

[40] R. C. Martin, *Agile software development: principles, patterns, and practices*. Prentice Hall PTR, 2003.

[41] M. Alshayeb, "The impact of refactoring on class and architecture stability," *Journal of Research and Practice in Information Technology*, vol. 43, no. 4, p. 269, 2011.

[42] M. O. Elish, "Exploring the relationships between design metrics and package understandability: A case study," in *IEEE 18th International Conference on Program Comprehension (ICPC)*. IEEE, 2010, pp. 144–147.

[43] S. Stevanetic and U. Zdun, "Empirical study on the effect of a software architecture representation's abstraction level on the architecture-level software understanding," in *14th International Conference on Quality Software (QSIC), 2014*. IEEE, 2014, pp. 359–364.

[44] S. Stevanetic, M. A. Javed, and U. Zdun, "The impact of hierarchies on the architecture-level software understandability-a controlled experiment," in *24th Australasian Software Engineering Conference (ASWEC), 2015*. IEEE, 2015, pp. 98–107.

[45] S. Stevanetic and U. Zdun, "Software metrics for measuring the understandability of architectural structures: a systematic mapping study," in *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2015, p. 21.

[46] L. C. Briand, S. Morasca, and V. R. Basili, "Property-based software engineering measurement," *IEEE Transactions on Software Engineering*, vol. 22, no. 1, pp. 68–86, 1996.

[47] N. Fenton, "Software measurement: A necessary scientific basis," *IEEE Transactions on Software Engineering*, vol. 20, no. 3, pp. 199–206, 1994.

[48] M. Alshayeb and W. Li, "An empirical validation of object-oriented metrics in two different iterative software processes," *IEEE Transactions on Software Engineering*, vol. 29, no. 11, pp. 1043–1049, 2003.

[49] L. Briand, K. El Emam, and S. Morasca, "On the application of measurement theory in software engineering," *Empirical Software Engineering*, vol. 1, no. 1, pp. 61–88, 1996.

[50] E.-M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, and P. Avgeriou, "Introducing a ripple effect measure: A theoretical and empirical validation," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 2015, pp. 1–10.

[51] A. Khoshkbarforoushha, P. Jamshidi, M. F. Gholami, L. Wang, and R. Ranjan, "Metrics for bpel process reusability analysis in a workflow system," *IEEE Systems Journal*, vol. 10, no. 1, pp. 36–45, 2016.

[52] A. Tripathi and D. Kushwaha, "A metric for package level coupling," *CSI Transactions on ICT*, vol. 2, no. 4, pp. 217–233, 2015.

[53] J. Lenhard, S. Harrer, and G. Wirtz, "Measuring the installability of service orchestrations using the square method," in *IEEE 6th International Conference on Service-Oriented Computing and Applications (SOCA)*. IEEE, 2013, pp. 118–125.

WE-MQS-VoIP Priority: An enhanced LTE Downlink Scheduler for voice services with the integration of VoIP priority mode

Duy-Huy Nguyen*†, Hang Nguyen*, and Éric Renault*
*SAMOVAR, Télécom SudParis, CNRS, Université Paris-Saclay
9 rue Charles Fourier - 91011 Evry Cedex, France
†Faculty of Information Technology, Hanoi University of Mining and Geology
Duc Thang, Bac Tu Liem, Hanoi, Vietnam

***Abstract*—The Long Term Evolution (LTE) is a high data rates and fully All-IP network. It is developed to support well to multimedia services such as Video, VoIP, Gaming, etc. So that, the real-time services such as VoIP, video, etc. need to be optimized. Nevertheless, the deployment of such live stream services having many challenges. Scheduling and allocating radio resource are very important in LTE network, especially with multimedia services such as VoIP. When voice service transmitted over LTE network, it is affected by many network impairments where there are three main factors including packet loss, delay, and jitter. This study proposes a new scheduler which is based on VoIP priority mode, Wideband (WB) E-model, QoS- and Channel-Aware (called WE-MQS-VoIP Priority scheduler) for voice services in LTE downlink direction. The proposed scheduling scheme is built based on the WB E-model and Maximum Queue Size (MQS). In addition, we integrate the VoIP priority mode into our scheduling scheme. Since the proposed scheduler considers the VoIP priority mode and user perception, thus, it improves significantly the system performance. The results demonstrate that the proposed scheduler not only meets QoS demands of voice calls but also outperforms Modified Largest Weighted Delay First (M-LWDF) in terms of delay, Packet Loss Rate (PLR) for all number of user (NU) and excepting NU equals 30, respectively. For Fairness Index (FI), cell throughput, and Spectral Efficiency (SE), the difference among the packet schedulers is not significant. The performance evaluation is compared in terms of Delay, PLR, Throughput, FI, and SE.**

***Keywords*—AMR-WB; Wideband E-model; VoLTE; VoIP priority mode; User perception**

I. Introduction

In the digital device market, there are many smart mobile phones such as iPhone, iPad, Android, etc. which is enough powerful to support various types of multimedia communications such as VoIP, Video, Gaming, etc. This means there need be a high data rate network to well support these services. LTE network is proposed by the Third Generation Partnership Project (3GPP) [1]. It has high data rates, low latency, and is fully packet-switched. In downlink direction, LTE uses Orthogonal Frequency Division Multiple Access (OFDMA). This allows supporting inter-symbol interference and selecting fading. Basic components of LTE network consists of a base station (called eNodeB or eNB) and User Equipments (UEs) in addition to a gateway [2]. The eNB station combines with core network through some standard complicated protocols. A basic scheduler is carried out by mobile network operators in both eNB and UE for both downlink as well as uplink directions. However, according to the 3GPP, there are no firm specifications for scheduling technique in LTE network. One of the most important modules of packet scheduling is Radio Resource Management (RRM) which decides users will be scheduled. The scheduler should care throughput, service policies to subscribes [3].

In LTE network, a full table defined which contains thresholds of delay and packet loss rate for different service classes [1]. In this table, several or all can be performed by service providers. This table divides resource types in the LTE network into two groups, those are Guaranteed Bit Rate (GBR) and Non-GBR. Table I represents the service classes in LTE network where voice service is a Guaranteed Bit Rate (GRB) service which has the second priority just after IP Multimedia Subsystem (IMS) signaling. However, in order to guarantee voice over LTE (VoLTE) quality is an extreme challenge.

TABLE I: Different service classes with QoS demands in LTE network

Bearer Type	Priority	Packet Delay (ms)	Packet Loss Rate	Example services
Guaranteed Bit Rate (GBR)	2	100	10^{-2}	Voice call
	4	150	10^{-3}	Video call
	3	50	10^{-3}	Real-time online gaming
	5	300	10^{-6}	Video streaming
Non-GBR	1	100	10^{-3}	IMS signaling
	6	300	10^{-6}	Video, TCP-based services (e.g. www, e-mail, chat, FTP, etc.)
	7	100	10^{-6}	Voice, video, interactive gaming
	8	300	10^{-3}	Video, TCP-based services (e.g. www, e-mail, chat, FTP, etc.)
	9		10^{-6}	

In order to meet different QoS requirements for these service classes, some packet schedulers have been proposed. According to [4], the scheduling strategies for LTE downlink are divided into five groups including:

- (1) Channel-unaware strategies;
- (2) Channel-aware/QoS-unaware strategies;
- (3) Channel-aware/QoS-aware strategies;
- (4) Semi-persistent scheduling for supporting VoIP flows;

- (5) Energy-aware strategies.

For voice traffics which are very sensitive to delay and PLR, thus, the Channel-aware/QoS-aware strategies are also very essential for them. In addition, VoLTE is really a VoIP service with the QoS guaranteed, and it is transmitted over a heterogeneous LTE network, thus, it need to have a special priority. Therefore, Semi-persistent scheduling is very essential and suitable for VolTE service. Several well-known scheduling algorithms for group 3 as FLS [5], M-LWDF [6], and EXP/PF [7]. In these schedulers, there is only FLS which guarantees bounded delay for real-time flows, the remaining schedulers transmit data of user in a Transmission Time Interval (TTI) by assigning a chosen priority metric. Nevertheless, due to the lack of delay and PLR thresholds, they are not suitable for supporting simultaneously real-time and non real-time traffic [2]. For group 4, authors in [8] proposed a priority mode for VoIP traffic over 3G LTE. In their scheduler, VoIP priority mode is executed when there is VoIP packet in the queue. This means VoIP packet is scheduled before any other traffic. In [9], authors proposed a new semi-persistent scheduler. Their scheduling scheme is combination of VoIP priority mode with user coupling. This allows using efficiently system capacity. Authors in [10] proposed a new scheduling algorithm which takes into account parameters of VoIP for voice over LTE network. This algorithm is developed based on the VoIP priority mode in [8] which the metric is changed.

In this paper, we present a new downlink scheduling scheme for voice services in LTE network with the integration of VoIP priority mode. Since VoLTE is deployed in an All-IP network, thus, there need to have a special priority for it. This paper is the extension of the WE-MQS scheduler which was proposed in [11] by modifying resource allocation scheduling method. In order to do this, we propose to integrate VoIP priority mode [8] with the essential modifications. We used the WB E-model to predict MOS and use this score as a main factor in the metric. Besides, we see that, the MQS factor has significant effects on the system performance. In the LTE-Sim [12], this factor is fixed equal to 0. This means the MQS is infinite. So that, in the scheduling process, the MQS is not considered. However, in fact, the MQS should be a finite value because if the MQS value is infinite then the delay will increase and the congestion could be increased. Therefore, the MQS needs to be considered as a essential factor in the metric of the scheduling algorithms.

The proposed scheduler selects UEs that based on their priorities which are computed according to the following factors: the maximum MOS, the minimum remaining queue size, the maximum delay, the channel condition. This means for the UE which has the higher MOS, the lower remaining queue size, the higher maximum delay, and the higher channel condition will have the higher priority. We evaluate the performance of the proposed scheduler with the M-LWDF scheduler in a heterogeneous traffic including VoIP, Video, and non real-time service which is called INF-BUF user. The simulation results were implemented in the LTE system simulator (called LTE-Sim) [12] and were compared in terms of Delay, PLR, Cell throughput, and FI for the number of user from 10 to 50.

The remainder of this paper is described as follows: Overview of the system model is described in section II. In section III, we present the proposed scheduling scheme. The simulation results and performance evaluation are analysed in section IV. The conclusion and future work are shown in section V.

II. The system model

A. VoLTE traffic flow

1) Radio protocol stack: The speech frame is encapsulated by network protocols consisting of Real-time Transport Protocol (RTP), User Datagram Protocol (UDP) and Internet Protocol (IP). And then, it will be packetized by other radio protocols including Packet Data Convergence Protocol (PDCP), Radio Control Link (RLC) and Medium Access Control (MAC) at the corresponding layers. At each radio layer, the corresponding header will be added into the speech packet. This leads to data overhead. Therefore, to reduce the data overhead, Robust Header Compression (RoHC) is deployed at the PDCP layer. This allows saving bandwidth as well as enhancing voice transmission in LTE network. RoHC will presses the header size of IP packet from 40 bytes (with IPv4) or 60 bytes (with IPv6) down to 1 to 4 bytes [13]. In addition, at MAC layer, Hybrid Automatic Repeat Request (HARQ) technique is utilized to retransmit in case of FEC (Forward Error Correction) fails error correction. This allows each speech packet to be retransmitted at least from one to three times. The number of retransmissions depends on the error correction or is configured. The implemented model of VoLTE protocol stack used in LTE-Sim [12] is represented on Figure 1.

2) Source codec: VoLTE uses AMR-WB for source codec. It is a voice codec which has been developed by European Telecommunications Standards Institute (ETSI). Details of this codec is described in [14]. AMR-WB codec utilizes a sampling rate of 16 kHz, audio bandwidth is in range of 50-7000 Hz. It includes 9 different codec modes of 0-8 which correspond to 9 source bit rates from 6.6 to 23.85 Kb/s. Each mode generates a compressed speech frame evry 20 ms. The bits in this frame are ordered according to their importance. They are grouped into three classes with reduced importance called Class A, Class B and Class C. The number of bits in each class depends on codec mode. AMR-WB packet size depends on the bit rate (mode) such as described in Table II.

TABLE II: Packet sizes of AMR-WB modes

Parameter	Bit rate (kbps)								
	23.85	*23.05*	*19.85*	*18.25*	*15.85*	*14.25*	*12.65*	*8.85*	*6.6*
Payload size (bits)	477	461	397	365	317	285	253	177	132
Frame size (bits)	488	472	408	376	328	296	264	192	144
RTP header (bits)	96	96	96	96	96	96	96	96	96
Packet size (bits)	**584**	**568**	**504**	**472**	**424**	**392**	**360**	**288**	**240**

In LTE network, AMR-WB codecs are configured into 3 configurations [15] as follows:

- Configuration A: 6.6, 8.85, and 12.65 Kb/s (Mandatory multi-rate configuration);
- Configuration B: 6.6, 8.85, 12.65, and 15.85 Kb/s;
- Configuration C: 6.6, 8.85, 12.65, and 23.85 Kb/s.

These configurations are used to simplify notation of bit rate between UE and eNB, thus, will simplify implementation

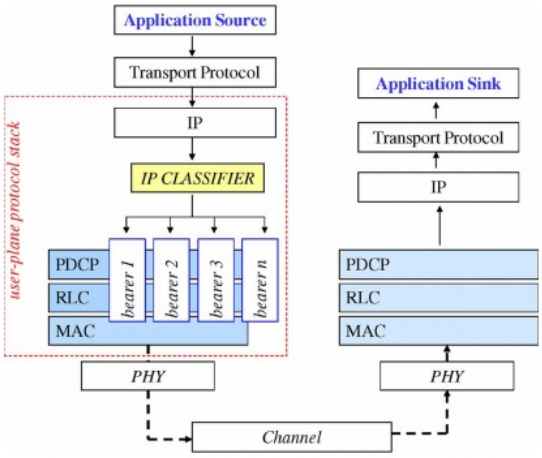


Fig. 1: Implemented model of VoLTE protocol stack [12]

and testing. The remaining bit rates are used for other purposes. In order to choose a bit rate, the receiver measures quality of radio channel. The channel quality indicator (CQI) is used for this purpose. It is defined as an equivalent carrier-to-interference (C/I) ratio. The C/I ratio then compared to a set of predefined thresholds to decide which mode to be used. Switching among modes in a configuration depends on algorithm of rate control. The criterion for mode switching is threshold value of C/I ratio. These threshold values depend on the channel condition, frequency hopping scheme, network configuration and other factors. Furthermore, network conditions change over time, so that, even well-selected adaption thresholds will not be best.

3) LTE frame structure: Figure 2 represents the structure of LTE frame for the downlink air interface [16]. In LTE downlink, a frame has length of 10 ms and is split into 10 sub-frames in time domain. This means each sub-frame has length of 1 ms and is split into 2 slots where each slot corresponds to 0.5 ms. In frequency domain, each slot consists of a number of resource blocks (RBs) (from 6 to 10 RBs). Each slot contains 6 or 7 Orthogonal Frequency Division Multiplexing (OFDM) symbols in normal cyclic prefixes and extended cyclic prefixes, respectively. Each time slot in frequency domain is split into bands of 180 kHz which consists of 12 consecutive sub-carriers. Each RB is a basic exchanging information unit in LTE downlink direction. This means RB is a radio resource which is available for user and is defined in both frequency and time domains. In a slot, the number of RBs depends on bandwidth of LTE network [17]. A sub-frame corresponds to a TTI which is minimum transmission unit. Each TTI contains at least one transport block per UE. The size of RB is as same as for all bandwidths [18].

B. Wideband E-model

WB E-model is a calculative model that is developed and is standardized by ITU-T [19]. The main purpose of this model is to predict quality of wideband audio. The main parameter of this model called R-factor. Its value is in range of 0-129. In order to perform user perception, R-factor is then translated into MOS score. WB E-model is defined such as in the following equation:

$$R_{wb} = R_{0,wb} - I_{s,wb} - I_{d,wb} - I_{e,eff,wb} + A \quad (1)$$

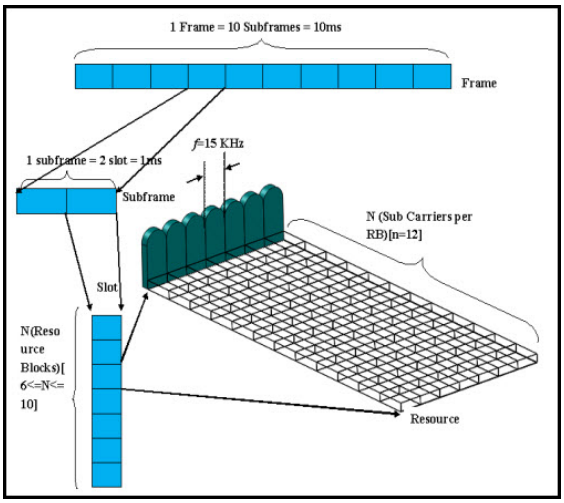


Fig. 2: Resource grid [16]

In which:

- $R_{0,wb}$: Basic ratio of signal to noise;
- $I_{s,wb}$: Simultaneous impairment factor which is sum of all impairments that may occur more or less simultaneously with voice transmission. Its default value is 0;
- $I_{d,wb}$: Factor of delay impairment factor, representing all impairments due to delay of voice signals;
- $I_{e,eff,wb}$: Factor of equipment impairment, covering the effect of signal distortion due to low bit rates of the codec and packet losses of random distribution;
- A: Advantage factor which represents the fact that some users can accept a degradation of quality due to mobility of mobile networks. Its default value is 0.

In above parameters, $I_{d,wb}$ and $I_{e,eff,wb}$ correspond to the effects of end-to-end delay and packet loss while $R_{0,wb}$ and $I_{s,wb}$ don't depend on performance of LTE network. The R_{wb} factor is mapped to MOS score as follows [19]:

For $R = R_{wb}/1.29$

$$MOS = \begin{cases} 1, \textbf{if } R < 0 \\ 1 + 0.035 \times R + 7 \times 10^{-6} \times R \times \\ (R - 60) \times (100 - R), \textbf{if } 0 \leq R \leq 100 \\ 4.5, \textbf{otherwise} \end{cases} \quad (2)$$

The relationship among R-factor, user perception, and MOS socre is shown in Table III.

TABLE III: Relationship among R-factor, user perception and MOS score

R	User perception	MOS
$90 \leq R < 100$	Very satisfied	4.3-5.0
$80 \leq R < 90$	Satisfied	4.0-4.3
$70 \leq R < 80$	Some users dissatisfied	3.6-4.0
$60 \leq R < 70$	Many users dissatisfied	3.1-3.6
$50 \leq R < 60$	Nearly all users dissatisfied	2.6-3.1
$R < 50$	Not recommended	< 2.6

R_{wb} factor is then translated into MOS score accoridng to Equation (2). Next, MOS is translated into user levels of users.

The value of $R_{0,wb}$ factor for wideband audio in equation (1) equals 129 [20], thus, equation (1) can be rewritten as follows:

$$R_{wb} = 129 - I_{d,wb} - I_{e,eff,wb} \quad (3)$$

In order to calculate the R_{wb} factor, we must to compute the values of $I_{d,wb}$ and $I_{e,eff,wb}$ factors. The $I_{d,wb}$ factor is determined as follows [21]:

$$I_{d,wb} = 0.024 \times D_{e2e} + 0.11 \times (D_{e2e} - 177.3) \times H(D_{e2e} - 177.3) \quad (4)$$

Where $H(x)$ is the Heavyside function:

$$H(x) = \begin{cases} 0, \textbf{if } x < 0 \\ 1, \textbf{otherwise} \end{cases} \quad (5)$$

In equation (4), D_{e2e} represents end-to-end delay of speech packet. It can be computed via some functions in LTE-Sim. The $I_{e,eff,wb}$ is determined according to packet loss. According to [20], $I_{e,eff,wb}$ is counted as the following formula:

$$I_{e,eff,wb} = I_{e,wb} + (129 - I_{e,wb}) \times \frac{P_{pl}}{P_{pl} + B_{pl}} \quad (6)$$

In which:

- $I_{e,wb}$: Impairment factor without any packet loss;
- P_{pl}: Packet loss rate, is also computed via some functions in LTE-Sim;
- B_{pl}: Codec specific factor that characterizes its robustness against packet loss.

The values of $I_{e,wb}$, B_{pl} for each mode of AMR-WB codec are represented in Table IV [20].

TABLE IV: Values of $I_{e,wb}$ and B_{pl} for modes of AMR-WB codec

AMR-WB mode	Bitrate (bps)	$I_{e,wb}$	B_{pl}
0	6.6	39	12.8
1	8.85	25	13.5
2	12.65	11	13
3	14.25	10	14.1
4	15.85	7	13.1
5	18.25	5	12.5
6	19.85	4	12.3
7	23.05	1	13
8	23.85	6	12.2

The R_{wb}-factor is then mapped to the MOS via equations (2) . MOS is one of the important factors for the metric in the proposed scheduler.

C. Semi-persistent scheduling

This is a hybrid method of Dynamic scheduling and Persistent scheduling [9]. VoIP packets uses a small quantity of control signaling to determine the channel quality after every fixed interval and persistently schedules the VoIP packets. This supports best to VoIP traffics due to its controlled dynamic nature and utilization of small control signaling.

D. VoIP priority mode

VoIP priority mode is proposed by Sunggu Choi and others in [8], it allocates RBs for VoIP calls before any other traffic. The limitation of this method is that when VoIP calls density is high, other traffics are not allocated. However, the authors already solved this problem by complementing a duration procedure. It is controlled dynamically to adjust consecutive TTIs according to total drop ratio of the packets measure at the eNodeB. In this mode, RBs are allocated based on Channel Adaptive Fair Queuing [22]. The metric of the scheduler in VoIP priority mode is determined based on queue length and Signal to Interference-plus-Noise Ratio (SINR) and is calculated as the following equation:

$$w_{i,j} = Q_l(i) \times \gamma(i) \quad (7)$$

Where (Q_l) is queue length and (γ) is SINR of active VoIP call i. Formula (7) indicates that UE which has the longer queue length and the better channel quality then will have the higher priority. It can be said that, the VoIP priority mode is useful when the density of VoIP calls is high. However, the downside is that other traffic to be starved. Therefore, the duration is deployed. It is controlled dynamically and depends on total of the drop ratio of the packets measured at the eNodeB. A predefined minimum and maximum drop ratio is utilized to adjust VoIP priority duration. Specifically, the maximum count of VoIP priority duration is increased when the drop ratio is less than the minimum threshold, and the maximum count of VoIP priority duration is decreased when the drop ratio exceeds the maximum threshold because in this case there is not enough resources to allocate. If the drop ratio is in range of the min to the max threshold then the duration is kept a constant. For more details of the duration, refer to [8].

III. The Proposed Packet Scheduler

In this proposal, we consider the characteristics of VoIP service. This service is sensitive to packet loss and delay, thus, scheduling process should consider various factors. Since VoLTE is transmitted over a fully packet-based network, thus, it needs to be guaranteed QoS to ensure user satisfaction. In this paper, we used the metric which was defined in [11] besides apply the VoIP priority mode for the proposed scheduling scheme.

Firstly, we define a new metric for the proposed scheduling scheme as follows: MOS is a parameter which represents user perception, thus, it should appear in the metric of scheduling algorithms. The higher MOS, the higher user satisfaction. MOS needs be automatically calculated at the receiver and is retransmitted to the eNB via feedback technique. For the MQS, according to our knowledge, there are no articles which mention about it. We think that, this factor has strong effects on the system performance. In the LTE-Sim [12], this factor is fixed equal to 0. This means the MQS is infinite. Hence the MQS is not considered in the scheduling process. However, in fact, the MQS should be finite. If the MQS value is infinite then the delay will increase and the congestion could be thus increased. Therefore, the MQS should be considered as a necessary factor in the metric of the scheduling algorithms.

The lack of LTE-Sim software is that it only supports G.729 codec while VoLTE uses AMR-WB codec. G.729 has only an unique mode which has the bitrate of 8 kbps and the packet size of 32 bytes generated in each 20 ms while AMR-WB has 9 modes. In fact, modes of AMR-WB are changed according to channel condition (i.e. C/I ratio). So, the user perception is calculated at scheduled instant. In the proposed scheduling scheme, we proposed to complement AMR-WB codec into LTE-Sim by reconfiguring some parameters and modifying essential source codes. With the presence of AMR-WB, we can simulate VoLTE traffic more easily.

The primary idea of the proposed scheduler is the consideration of user satisfaction (MOS) and the MQS factor (called also $Q_{i,max}$) included into the metric of the scheduling algorithm. This means the higher MOS and the lower $(Q_{i,max} - Q_i)$ values, the higher priority for the UE. The fixed maximum time $D_{HOL,i}$ and the maximum probability δ_i are included in the Equations (4) and (6) to calculate the factors of $I_{d,wb}, I_{e,wb}$, respectively. The metric in the proposed scheduling scheme for voice users is defined as follows:

$$w_{i,j} = \frac{MOS_i \times (Q_{i,max} - Q_i)}{\tau_i} \times \frac{r_{i,j}}{\bar{R}_i} \quad (8)$$

Where:

- $Q_i, \tau_i, r_{i,j}$ and $\bar{R}_i$ are similar to ones in the previous formulas;
- $Q_{i,max}$: The MQS of the user *i*. This value can be obtained in bytes via some functions in LTE-Sim [12].

For video and non real-time services, we propose to use the metric of the M-LWDF scheduler. The $w_{i,j}$ is a matrix that offers priority for each RB_j assigned to UE_i. It is calculated based on the MOS, the remaining queue size $(Q_{i,max} - Q_i)$, the maximum time τ_i and the channel condition. MOS is computed at the receiver and is feedbacked to the eNB for making scheduling decision. MOS included in the metric will fully exploit the user perception.

In fact, the AMR-WB mode is dynamically calculated and optimized at the AMR-WB encoder according to channel quality using rate adaptation control algorithm detailed in [15]. The limitation of LTE-Sim is that it supports only G.729 codec for VoIP. Therefore, the proposed scheduler can not get the mode chosen from AMR-WB encoder at Application layer. In order to overtake this issue, we proposed a procedure which allows to choose AMR-WB mode from C/I ration that is available in LTE-Sim. With the proposed procedure, the proposed scheduler can chooses dynamically source codec mode according to channel quality. The threshold values of C/I ratio is chosen according to [23] and [24]. The procedures for choosing AMR-WB mode and calculating the metric of the proposed scheduler are detailed in [11]. The procedure of Update AMR-WB packet size is used to update packet size according to channel condition and is used for all schedulers while the procedure of w_{ij} calculation is used only in the proposed scheduler for calculating the metric.

Secondly, we integrate the VoIP priority mode into the proposed scheduling scheme such as represented on Figure 3. The algorithm priority mode is only enabled when there is VoIP user in the queue. In order to negative effects on other traffics, the duration of VoIP priority mode is deployed. This is detailed in [8].

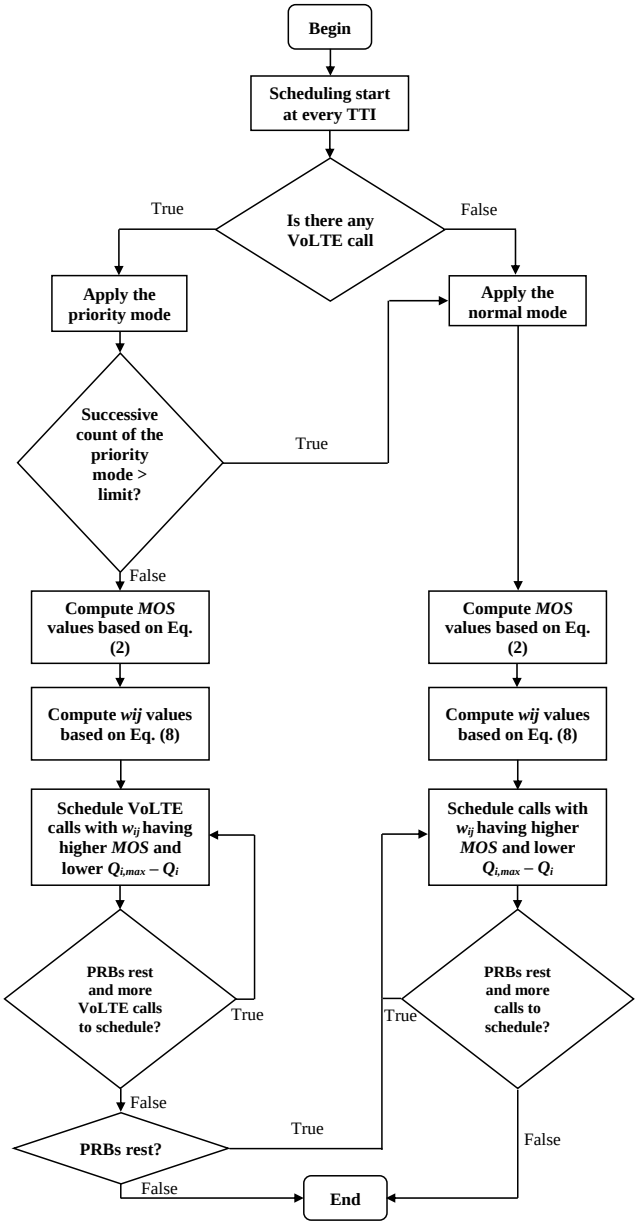


Fig. 3: The proposed scheduling algorithm

IV. Simulation environment and Performance evaluation

A. Simulation environment

1) Traffic model: In the proposed scenario, the eNodeB lies at center of a macrocell which uses an ommi-directional antenna with a bandwidth of 10 MHz. Each UE utilizes a VoIP flow, a Video flow, and a INF-BUF flow. at the same time. For the VoIP flow, a AMR-WB voice stream with bit-rates of 6.6, 8.85, and 12.65 kbps (Configuration A of AMR-WB codec which is mandatory and default for VoLTE service) were used. The voice traffic is a bursty application which is modeled with an ON/OFF Markov chain [25]. For the video traffic, a trace-based application which generates packets using real video trace files with a bit-rate of 242 kbps was used [26] and it is

also available in [12]. To attain a real simulation of a video streaming, we used an encoded video sequence "*foreman.yuv*". The encoded spatial resolution CIF 352×288 with 300 fps has been utilized for the entire simulation.

For the LTE propagation, we use a loss model which is formed by four different models as follows: Path loss, Multipath, Penetration and Shadowing [27].

- Path loss: $PL = 128.1 + 37.6 \times log(d)$, with d is the distance between the UE and the eNB in km;
- Multipath: Jakes model;
- Penetration loss: 10 dB;
- Shadowing: Log-normal distribution with mean 0 dB and standard deviation of 8 dB.

2) Simulation parameters: In this study, we investigate and evaluate the performance of M-LWDF and the proposed schedulers in LTE downlink direction. The reason of this choice is that in [11], we proposed the metric for non-VoIP users using the metric of M-LWDF scheduler. The simulation process is performed in a single cell with interference, the number of users is in range of 10..50. UEs move randomly with a speed of 30 km/h. In the proposed simulation scenario, we consider each user using a VoIP, a Video, and a INF-BUF flow. This means the proposed scheduler is evaluated in a heterogeneous LTE network with mobility. For assessing the performance of the system, we use LTE-Sim [12] software. This is a open source framework for researchers and academic community. The basic parameters used in the simulation are represented in the Table V.

TABLE V: Basic parameters for simulating

Simulation Parameters	Values
Simulation duration	100 s
Frame structure	FDD
Cell radius	1 km
Bandwidth	10 MHz
Video bit-rate	242 kbps
AMR-WB bit-rates	6.6, 8.85, 12.65 kbps
User speed	30 km/h
Number of users	10, 20, 30, 40, 50 UEs
Maximum delay	0.1 s
MQS	10^5 bytes
Maximum VoIP packet drop rate	5 %
Minimum VoIP packet drop rate	2 %
Traffic model	VoIP, Video, and INF-BUF
Packet Schedulers	M-LWDF, WE-MQS, and WE-MQS-VoIP Priority

B. Performance evaluation

For assessing the performance of the proposed scheduler, we compare it to the M-LWDF scheduler, and to the scheduler proposed in [11] (called WE-MQS scheduler or normal mode). We assess the performance in terms of delay, PLR, cell throughput, FI and SE. The analysis of the simulation results are represented in the following subsection.

1) Delay: End-to-end delay is the duration required for a packet to be transmitted from source to destination. Figure 4 illustrates the delay of VoIP flow. Such as shown on this figure, the priority mode of the proposed scheduler has the lowest delay when compared to the normal mode and the M-LWDF scheduler. Both modes of the proposed scheduler slightly increase when the NU increase while the M-LWDF heavily increases. In can be said that, when the VoIP priority mode is integrated, the delay decreases significantly.

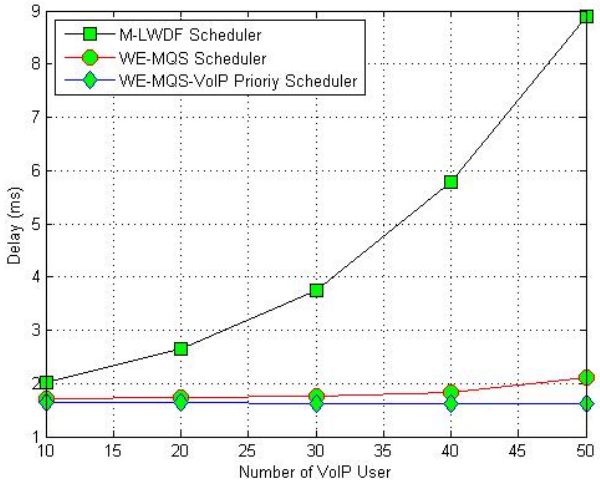


Fig. 4: Delay vs number of VoIP user

2) Packet Loss Rate: PLR represents the failure of one or more packets which are not transmitted successfully to destination. Figure 5 represents the PLR of VoLTE traffic. Normally, the PLR increase when the NU increases. In this study, we assess the system performance in a heterogeneous LTE network with mobility. The results on Figure 5 are quite special, specifically for the normal mode, the PLR decreases when the NU increases. This may be due to the not stable in a real system. In general, the priority mode has the lowest PLR in comparison with two remaining others except when the NU equals 30.

For the delay and PLR, it can be concluded that the priority mode is very suitable for VoIP service because it has the best performance.

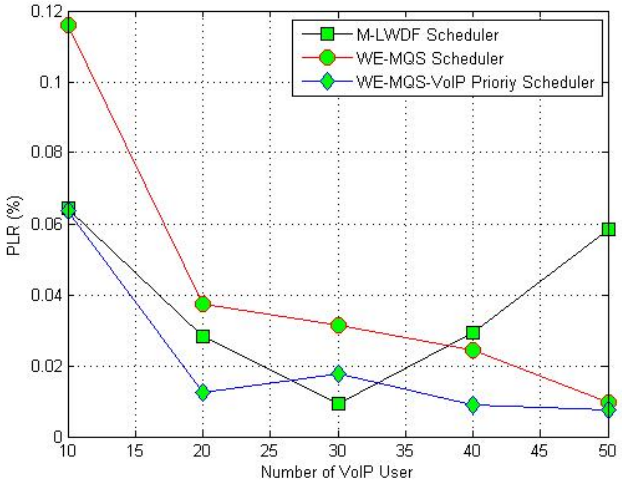


Fig. 5: Packet loss rate vs number of VoIP user

3) Cell throughput: Throughput is a measurement of how many units of information a system can process in a given amount of time. As shown on Figure 6, for the VoIP flow, the cell throughput of all the schedulers increases when the NU

increases. The priority mode has the best throughput when the NU in range of 10..30 and slightly reduces when the NU is more than 30, 40 when compared to the M-LWDF and to the normal mode, respectively.

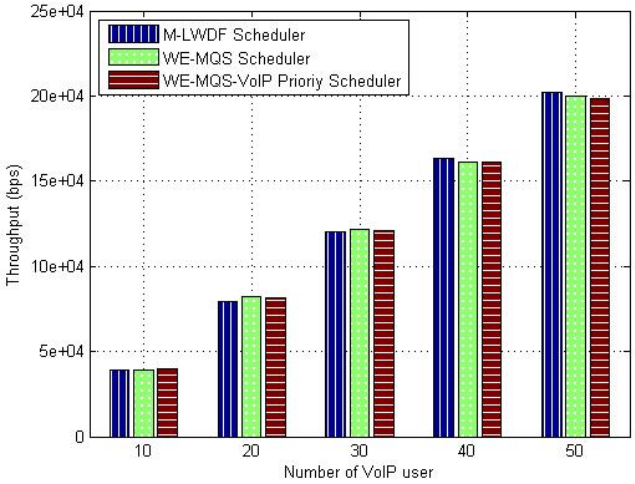


Fig. 6: Throughput vs number of VoIP user

4) Fairness index: FI is a main demand that is used to guarantee a minimum performance to the edge-cell users. For the VoIP flow as shown on Figure 7, the FIs of all schedulers are not stable when the NU increases. Normally, the FI decreases when the NU increases. The VoIP priority mode has the best FI when the NU is less than 30 and has the loewest FI when the NU is more than 30. However, the difference is not significant. The normal mode is always of the middle of two others.

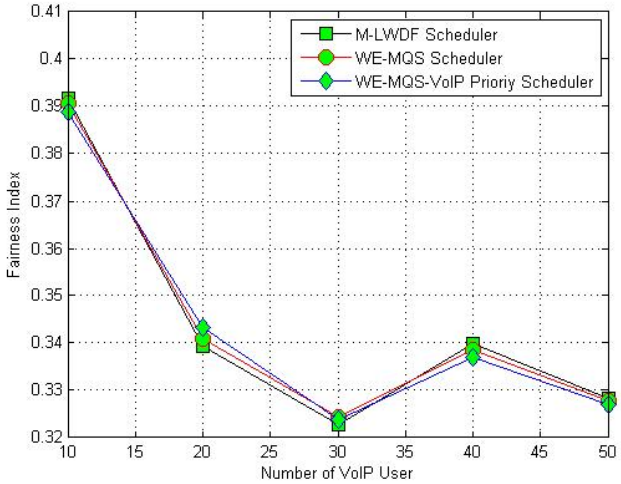


Fig. 7: Fairness Index vs number of VoIP user

5) Spectral efficiency: SE expresses successful usage of radio resources. It is a major principle of the scheduler. SE aims at performance measurements for entire cell. As shown in the Figure 8, the normal mode almost has the lowest SE for all of the NU while the VoIP priority mode has the same SE when compared to the L-MWDF scheduler when the NU is less than 40. When the NU is more than 40, the M-LWDF scheduler has the higher SE in comparison with the VoIP priority mode. However, this increase is not very significant.

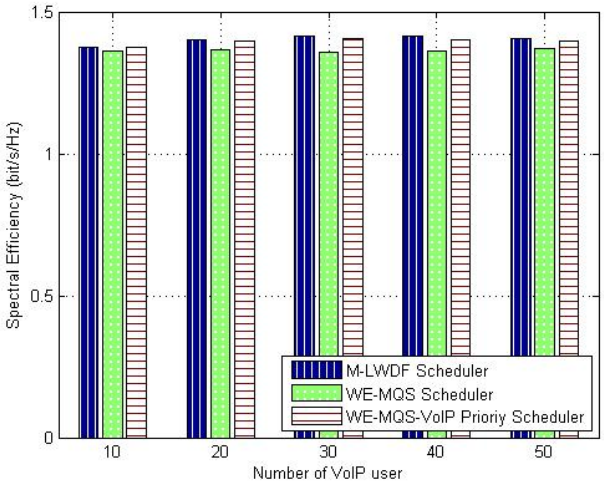


Fig. 8: Spectral efficiency vs number of VoIP user

V. Conclusions and Future Work

In this paper, we present a new Channel-, QoS- and QoE-Aware scheduling scheme with the integration of the VoIP priority mode for voice users in the LTE downlink direction. The main idea in the proposed scheduler is the consideration of user perception (via the MOS score) and the MQS factors into the metric in the proposed scheduler and the presence of VoIP priority mode for VoIP packets. The metric is based on the MOS score, the remaining queue size, the fixed maximum time, and the channel quality. The simulation results show that the proposed scheduler meets QoS requirements for voice services. In addition, it outperforms the M-LWDF scheduler and the normal mode for delay and PLR. With the integration of the VoIP priority mode, the proposed scheduler is enabled, thus, it has the lowest delay and PLR. For the throughput, FI, and SE, in general, it nearly has the same performance when compared to the M-LWDF scheduler. The advantage of the proposed scheduler is that it takes the user satisfaction and the remaining queue size into account and the presence of the VoIP priority mode, thus, it enables the priority for VoIP packets than other traffics. In addition, the proposed scheduler integrates the AMR-WB codec which is mandatory for VoLTE. This overcomes the limitation of the LTE-Sim that it supports only G.729 codec for VoIP application.

It can be said that when considering the MOS and the MQS as factors for the metric in the proposed scheduler and the integration of the VoIP priority mode, the system performance has been improved significantly. Through all simulation results, it can be said that the proposed scheduler has the best performance for VoIP users. Therefore, it is can be said that the proposed scheduler is very suitable and efficient for voice services in the LTE downlink direction. For the future work, we will build a framework to measure voice quality for the proposed scheduler.

References

[1] 3GPP, http://www.3gpp.org.

[2] S. Ali and M. Zeeshan, "A utility based resource allocation scheme with delay scheduler for lte service-class support," in *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*. IEEE, 2012, pp. 1450–1455.

[3] M. Alasti, B. Neekzad, J. Hui, and R. Vannithamby, "Quality of service in wimax and lte networks [topics in wireless communications]," *Communications Magazine, IEEE*, vol. 48, no. 5, pp. 104–111, 2010.

[4] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 2, pp. 678–700, 2013.

[5] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-level downlink scheduling for real-time multimedia services in lte networks," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 1052–1065, 2011.

[6] P. Ameigeiras, J. Wigard, and P. Mogensen, "Performance of the m-lwdf scheduling algorithm for streaming services in hsdpa," in *Vehicular technology conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, vol. 2. IEEE, 2004, pp. 999–1003.

[7] J.-H. Rhee, J. M. Holtzman, and D.-K. Kim, "Scheduling of real/non-real time services: adaptive exp/pf algorithm," in *Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual*, vol. 1. IEEE, 2003, pp. 462–466.

[8] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "Mac scheduling scheme for voip traffic service in 3g lte," in *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th*. IEEE, 2007, pp. 1441–1445.

[9] S. Saha and R. Quazi, "Priority-coupling-a semi-persistent mac scheduling scheme for voip traffic on 3g lte," in *Telecommunications, 2009. ConTEL 2009. 10th International Conference on*. IEEE, 2009, pp. 325–329.

[10] R. Musabe, H. Larijani, B. Stewart, and T. Boutaleb, "A new scheduling scheme for voice awareness in 3g lte," in *Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011 International Conference on*. IEEE, 2011, pp. 300–307.

[11] H. N. Duy-Huy Nguyen and É. Renault, "We-mqs: A new lte downlink scheduling scheme for voice services based on user perception," *International Journal of Computer Applications*, vol. 142, no. 10, pp. 28–36, 2016.

[12] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating lte cellular systems: an open-source framework," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 2, pp. 498–513, 2011.

[13] C. Bormann, C. Burmeister, M. Degermark *et al.*, "Robust header compression (rohc)," RFC 3095, June, Tech. Rep., 2001.

[14] I. Rec, "G. 722.2 (2003) wideband coding of speech at around 16kbit/s using adaptive multi-rate wideband (amr-wb)," *International telecommunication union, Geneve, Switzerland*, 2003.

[15] 3GPP, "Inband Tandem Free Operation (TFO) of speech codecs; Service description; Stage 3," 3rd Generation Partnership Project (3GPP), TS 28.062, 12 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/28062.htm

[16] R. Musabe and H. Larijani, "Cross-layer scheduling and resource allocation for heterogeneous traffic in 3g lte," *Journal of Computer Networks and Communications*, vol. 2014, 2014.

[17] F. Semiconductor, "Long term evolution protocol overview," *White Paper, Document No. LTEPTCLOVWWP, Rev 0 Oct*, 2008.

[18] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," 3rd Generation Partnership Project (3GPP), TS 36.211, 03 2010. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36211.htm

[19] I. Rec, "G. 107.1, wideband e-model," *Int. Telecomm. Union, Geneva*, 2011.

[20] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment factor framework for wide-band speech codecs," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1969–1976, 2006.

[21] C. Olariu, M. O. Foghlu, P. Perry, and L. Murphy, "Voip quality monitoring in lte femtocells," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*. IEEE, 2011, pp. 501–508.

[22] L. Wang, Y.-K. Kwok, W.-C. Lau, and V. K. Lau, "Channel adaptive fair queueing for scheduling integrated voice and data services in multicode cdma systems," *Computer Communications*, vol. 27, no. 9, pp. 809–820, 2004.

[23] A. Technologies, Available:http://rfmw.em.keysight.com/rfcomms/refdocs/gsm/default.htm#gprsla_amr_bse_config.html#CBDDEBCD, May 2010.

[24] ——, Available:http://rfmw.em.keysight.com/rfcomms/refdocs/gsm/gprsla_wb_amr_bse_config.html, May 2010.

[25] C.-N. Chuah and R. H. Katz, "Characterizing packet audio streams from internet multimedia applications," in *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 2. IEEE, 2002, pp. 1199–1203.

[26] M. Reisslein, L. Karam, and P. Seeling, "H. 264/AVC and SVC Video Trace Library: A Quick Reference Guide http://trace. eas. asu. edu," 2009.

[27] 3GPP, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," 3rd Generation Partnership Project (3GPP), TR 25.913, 12 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25913.htm

Sentiment Based Twitter Spam Detection

Nasira Perveen
Dept. of Comp. Science
Bahauddin Zakaria Univ. Multan
Pakistan

Malik M. Saad Missen
Dept. of Comp. Science and IT
The Islamia University of Bahawalpur
Pakistan

Qaisar Rasool
Dept. of Comp. Science
Bahauddin Zakaria Univ. Multan
Pakistan

Nadeem Akhtar
Dept. of Comp. Science and IT
The Islamia University of Bahawalpur
Pakistan

***Abstract*—Spams are becoming a serious threat for the users of online social networks especially for the ones like of twitter. twitter's structural features make it more volatile to spam attacks. In this paper, we propose a spam detection approach for twitter based on sentimental features. We perform our experiments on a data collection of 29K tweets with 1K tweets for 29 trending topics of 2012 on twitter. We evaluate the usefulness of our approach by using five classifiers i.e. BayesNet, Naive Bayes, Random Forest, Support Vector Machine (SVM) and J48. Naive Bayes, Random Forest, J48 and SVM spam detections performance improved with our all proposed features combination. The results demonstrate that proposed features provide better classification accuracy when combined with content and user-oriented features.**

***Keywords*—sentiment analysis; spam detection; twitter**

I. Introduction

Spam is a real threat to usefulness of the web. Spammers mask their content as useful or relevant content and hence is delivered to the user. The legitimate users consume this spam data considering it relevant to their information needs. Clay Shirky [2] remarked that a communication channel isnt worth its salt until the spammers descend.

Spams are not easy to stop. For several years, email services like Gmail, Microsoft and others have been successfully detecting spam emails but still spam emails are in circle on the web. These services have been reporting that email spamming has been up to 90 to 95 percent of the total email exchanges [3], [4], [5]. Even after successful detection of spams, companies are unable to stop spammers which ensures about the economical benefits spammers get when they trap a user clicking on a spam link. The severity of the threat posed by spamming has increased with the emergence of online social networks and twitter is one of the most popular online social network which has been highly affected by spam. twitter spamming is more threatening because its more targeted towards the trending topics of the twitter and hence bit easier to get penetrated especially because of hash-tag operator. Another fact that makes twitter a rather easier and fruitful target for spammers is its variety of audience. twitter users span across all sectors of life i.e. it can be the teachers or students, celebrities or politicians, marketers or customers or even general public. They belong to all age groups but most widely age group that uses twitter is between 55 to 64 years. There are about 60% users that access twitter from their cell phones [1]. twitter has 288 million monthly active members that make it widely growing social networking site. There are around 400 million tweets posted on daily bases, the average posts on twitter is 208 tweets per users account.

Due to this continuous distribution of information, a user faces many problems with search results that shares recurring and irrelevant information. This also can be very worrying at the times since a user has to scroll through the all information in direction to get an overall view of topic. Spam detection on the twitter network is difficult due to the noticeable usage of URLs, abbreviations, informal language and modern language concepts [6]. Old-style methods of detecting spam information fall short here. To date, study has been available on many techniques for detecting spams on twitter and blogs by using different features. After knowing the existing importance of spams on twitter, we take inspiration or motivation from this user need and decided to design and develop improved techniques to detect spams on twitter.

In this paper, we propose a spam detection approach for detecting spam tweets. This approach is based on sentimental features of a tweet. The idea is to exploit the philosophy that spammer use to force a user to click on a particular link. They definitly seek help of some motivational words (like 'the best web site', 'excellent service', etc) to make people believe in a certain tweet (examples of some spam tweets given in the table I. Results show that this exploitation of sentimental features proves fruitful.

TABLE I: Some Spam Tweet Examples

you'll laugh when you see this pic of you... tinyurl.coX/blah
you look like you lost weight in this video.. t.cX/blah
Was this blog you posted really necessary? tniX.biz/ad08 some kind of joke?
viagra,cialis,soma,tramadol and more. no prescription. ti.co/blah
Gain over 1,000 followers a week by using: ti.co/blah
wow this really works! i found out who stalks me :P go to 0rX.com/blah

The rest part of the paper is organized as the following

[1] http://blog.digitalinsights.in/social-media-facts-and-statistics-2013/0560387.html twitters Facts 2013

sections. In section II, we highlight some of the previous works done while in section III, we discuss proposed features. In Section IV-A, we describe the data collection used for experimentation. In Section IV, we describe our experimental results and comparisons of different features combinations and the conclusion is described in Section V.

II. Related Work

In this section, we describe several work related to spam detection on twitter. As discussed above that spamming on twitter is different in technique and in nature as compared to other web spams like email spam. Sarita Yardi et al discussed this in a very detailed way in their work [8]. They describe that motivating question for spammers while spamming twitter is that in which way to target and when to target the user. And also what trending topics the spammers should to target and how long they can continue their activities with spamming techniques. Being more practical, Gianluca Stringhini et al [6] explore how the spam has entered in social network sites. They use Random Forest algorithm as a classifier with Weka framework by using features like FF ratio (first feature that compares friend requests that a user sent to the number of friends she has), URL ratio, Message Similarity, Friend Choice and Friend Number. They study how spammers operate to target the social network sites. M. Chuah and M. McCord in [9] discuss some content and user based features as these features are not similar among legal users and spammers.

Zi Chu et al in [10] described that previously all spam detection methods check only individual messages or account for the existence of spam. They focused on the detection of spam campaigns that supervise multiple accounts to spread spam on the twitter network. Alex Hai Wang in [11] proposed a graph model called directed graph model to discover the friend and follower relationship on twitter network. By using Nave Bayesian classifier graph based and content based features are suggested for the detection of spam tweets. In graph based features three features are used namely friends, followers and the reputation of a user is calculated for discovering spam. In content based features duplicate tweets, HTTP links, replies and mentions and trending topics computed for spam detection. In [13], Nikita Spirin studies URLs shared by users on twitter and the estimation of spam for those users who share these links in the network and utilize the information to web spam detection algorithms by proposing a new set of URL derived features for a twitter user representation. Also propose a solution for construction of automatic dataset by analyzing URLs shared by non-spam users in social media for the problem of web spam detection.

In [14] another approach is discussed for spam detection in twitter network. They study the propagation of spam in the network. And they want to find out whether there is a pattern that spammers used for spam proliferation through the network and to determine whether the accounts are either been compromised or overtaken by spammers or certain accounts are purely created for spam activities in the network. They examine the characteristics of the graph of spam tweets and run Trust Rank technique on the collected data. In [15] introduced features for spam tweets detection without earlier statistics of the user and use statistical presentation for the analysis purpose of language to identify spam in twitter topics.

Jonghyuk.S et al in [16] discussed that previously spam detection schemes were based on the features of account information like age of the account, ratio of URLs in tweet and the content similarity of tweet. These features can easily be used by the spammers for spam proliferation activities. They introduced connectivity and distance features (of relation features) for spam detection in twitter which detects spam messages by using connectivity and distance features (of relation features) among the sender of the message and the receiver of the message for checking the spam in the message which is being in progress. Their proposed distance and connectivity features are problematic to operate upon by the spammers and these (relation) features can easily be composed rapidly. Fabricio Benevenuto et al in [12] discussed the problem of detection of spammers in the twitter network as a replacement for spam tweets. The author use social behavior and content based characteristics for the detection of spammers in the twitter network. In [17] spam identification approach is proposed and evaluated for twitter trending topics. Two components of this methodology are detection of timestamp gap among the two consecutive tweets of a user and recognizing the tweet content resemblance amongst the tweets posted by the user.

III. Sentimental and Content-Based Features

We propose sentimental features (combined with content and user based features) as part of our spam detection approach for twitter. All proposed features are described in table **??** in detail.

TABLE II: Features and their Descriptions

Feature	Description
Negative Words Count	Total negative words in a tweet. Negative score computed through SentiWordNet3.0 [18]. It is the sentimental feature.
Negative Words Ratio	It is calculated on the bases of all negative words in a tweet converted in to ratio using equation Negative words Ratio= $\frac{TotalNegativeWords}{TweetLength} \times 100$
Negative Score	Negative Score values are calculated on the bases of sum of all negative words scores of a tweet.
Positive Words Count	Positive Words Count values are calculated on the bases of all positive words in the tweet computed through SentiWordNet3.0 [18]. it is also the sentimental feature.
Positive Words Ratio	Positive Words Ratio value is calculated on the bases of all positive words in a tweet converted in to ratio. Values are calculated on the bases of following formula: Positive words Ratio= $\frac{TotalPositiveWords}{TweetLength} \times 100$

Continued on next column

Continued from previous column

Feature	Description
Positive Score	Positive Score values are calculated on the bases of sum of all positive words scores of a tweet.
Subjectivity Score	Subjectivity Score values are calculated from the tweet on the bases of following formula: Subjectivity Score = Positive Score - Negative Score
Adjectives	This value is calculated on the bases of all adjectives in a tweet with a sentimental value greater than a fixed threshhold. Adjectives are extracted from a Tweet using Part-of-Speech Tagging. It is also a sentimental feature.
Verbs	This value is calculated on the bases of all verbs used in a tweet with a sentimental value greater than a fixed threshhold. Verbs are extracted from Tweet on bases of Part-of-Speech Tagging. It is also a sentimental feature.
Adverbs	This value is calculated on the bases of all adverbs used in a tweet with a sentimental value greater than a fixed threshhold. Adverbs are extracted from Tweet on bases of Part-of-Speech Tagging. It is also a sentimental feature.
Smiles ☺	Smiles values are calculated on the base of all smiles ☺used in a tweet. It is the emotional sentimental feature.
High Smiles (;))	High Smiles values are calculated on the base of all smiles (;)) used in a tweet; High Smiles are extracted from tweet on bases of emotional sentiments (;)); it is also the emotional sentimental feature.
Sad Faces ☹	This is calculated on the base of all sad faces ☹used in a tweet. It is emotional sentimental feature.
Deep Sad Faces	This value is calculated on the base of all ;(used in the tweet text. It is also emotional sentimental feature.
Hashtags Percent	Hashtags percent values are calculated on base list of all Hashtags included in a tweet converted in to percentage. Values are calculated on the bases of following formula: Hashtags Percent= (Total Hashtags)/(Tweet Length) X 100

Continued on next column

Continued from previous column

Feature	Description
URLs Percent	URLs percent values are calculated on the bases of all URLs included in a tweet converted in to percentage. Values are calculated on the bases of following formula: URLs Percent= $\frac{TotalURLs}{TweetLength} \times 100$
Users Mention	It is calculated on the bases of all usernames (@username) mentioned in the tweet text; and converted in percentage. Values are calculated on the bases of following formula: Users Mention Percent= $\frac{TotalUsersMention}{TweetLength} \times 100$

Concluded

IV. Experiments and Evaluations

A. Data Collection

We downloaded tweets for 29 the most trending topics of twitter for year 2012 using APIs provided by twitter. After basic pre-processing, we are left with 29K (1K for each topic) tweets. Manual annotation of these tweets was done with spam or not-spam labels using two annotators A and B. Kappa score [7] for this annotation was found satisfactory (0.82) to proceed with the experiments. We decide to use standard metrics for measuring the usefulness of our approach and hence precision, recall, and F-measure are used.

B. Features Performance Comparison

Here we will discuss our proposed features spam detections performance by using five selected classifiers (SVM, Random Forest, Naive Bayes, Bays Network and J48). We have compared the performance of different features by making different combinations, We have discussing just one combination "all proposed features with baseline features combination" , its performance are given in Table III.

C. All Features and Baseline Features Comparisons

TABLE III: All Features and Baseline Features Accuracy

Classifiers	Baseline Accuracy (%age)	All Features (%age)	Improvement (%age)
BayesNet	90.60	89.76	-0.84
NaiveBayes	14.13	25.30	11.17
Random Forest	91.81	92.29	0.48
J48	91.87	92.34	0.47
LibSVM	91.27	91.41	0.13

Table III shows the accuracy of all features with baseline features by using 10 folds cross validation while figure 1 shows the graphical representation of the information represented in table III. As we have seen in table III result and 1, Naive Bayes spam detections performance improved with our proposed features. Naive Bayes accuracy with baseline features is 14.13%, result improved a lot with our proposed features combination with baseline features to 25.30% (i.e. 11.18% improvement). We have also got good improvement in Random Forest and J48 classifiers. Random Forest with baseline accuracy is 91.81% is improved with all proposed features to 92.29% with gives 0.48% improvements in accuracy while J48 has given 0.47% improvement. SVM has also shown some improvements in spam detection performance (0.14%).

We repeated the experiments using 70% training dataset fetched by using "Remove Percentage Weka"[2] unsupervised filter by setting percentage property to 70% (contain 20141 spam and non-spam tweets) and testing datasets (contain 6042 spam and non-spam tweets) is fetched by setting the "invert selection" properties to false.

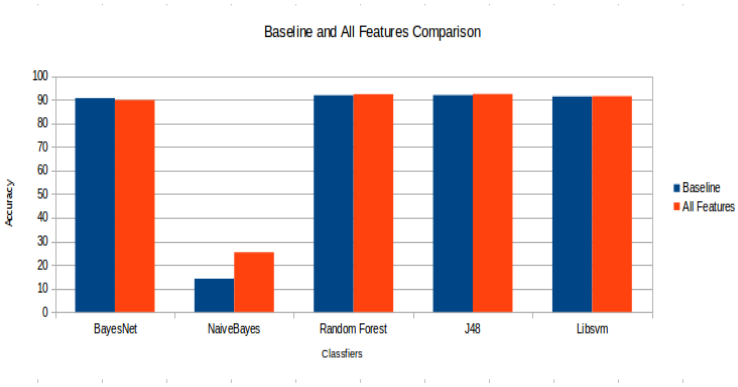


Fig. 1: All Features and Baseline Features Accuracy

Table IV displays results of these experiments while figure 2 shows the graphical representation of all and Baseline Features Accuracy with 70% Training and Testing Datasets (IV).

TABLE IV: All and Baseline Features Accuracy with 70% Training and Testing Datasets

Classifiers	Baseline Accuracy (%age)	All Features (%age)	Improvement (%age)
BayesNet	92.15	91.65	-0.50
NaiveBayes	16.56	26.68	10.12
Random Forest	91.61	92.41	0.80
J48	92.3540	92.20	-0.15
LibSVM	93.37	93.35	-0.02

As we have seen in Figure 2, Naive Bayes and Random Forest spam detections performance improved with our proposed features with 70% training and testing datasets. Naive Bayes accuracy improve further as compare to the previous experiments of 10 fold cross validation (i.e. 25.30% vs 26.68%). Random Forest has also shown some improvements in spam detection performance (0.80%).

[2]http://www.cs.waikato.ac.nz/ml/weka/Weka3

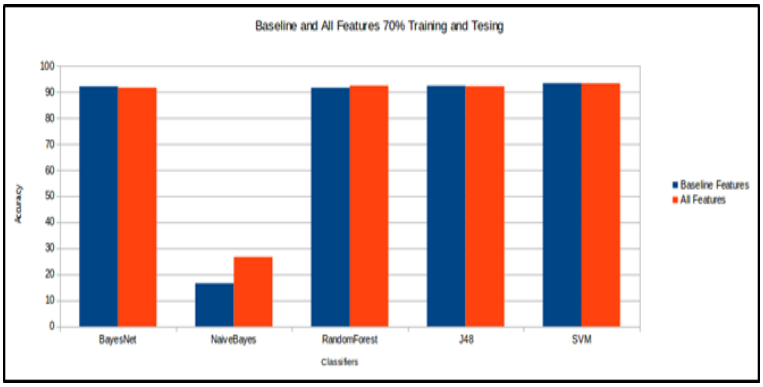


Fig. 2: All Proposed and Baseline Features Accuracy with 70% Training and Testing Datasets

Features Combination with Baseline Features Comparisons Table V shows the accuracy of all combination of features with baseline features by 10 folds using cross validation in percentage values while figure 3 shows its graphical representation.

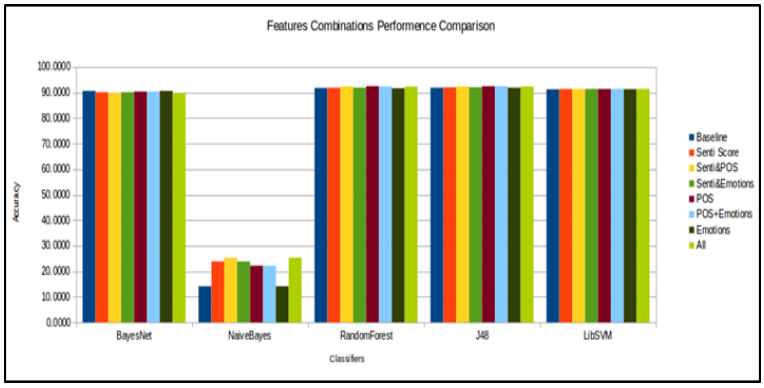


Fig. 3: All Combination with Baseline Features Accuracy

TABLE V: All Combination with Baseline Features Accuracy

Classifiers	Baseline Accuracy	Senti	Senti+POS	Senti+Emotions	POS	POS+Emotions	Emotions	All Combined
BayesNet	90.60	90.15	89.76	90.15	90.37	90.37	90.60	89.76
NaiveBayes	14.13	23.84	**25.30**	23.84	22.16	22.17	14.13	**25.30**
Random Forest	91.81	91.87	92.39	91.90	**92.48**	92.35	91.71	92.29
J48	91.87	92.05	92.36	92.05	**92.46**	**92.46**	91.87	92.34
LibSVM	91.27	91.38	91.40	**91.41**	91.37	91.40	91.32	91.41

As described in the table and figure, for Naive Bayes classifier we have got good improvement in all combinations but the best combination stands "All Combined" while Random Forest gets improvement in "POS sentimental features" combination. With J48 and SVM as we seen we are getting good performance in all features combinations.

At the end, figure 4 shows the percentage accuracy improvement of all combination of features as compared with baseline features by using 10 folds cross validation. The values are calculated by using following formula: Value = Features Combination Accuracy Baseline Features Accuracy

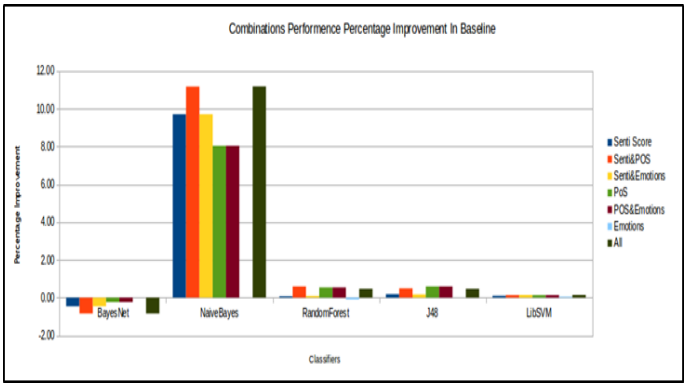


Fig. 4: Combinations Percentage Improvement Compared with Baseline Accuracy

As we have seen in figure 4, Naive Bayes shows good performance as compared with baseline features accuracy. It gained 11.18% improvement as compared with baseline features in all combinations; and with proposed (Sentimental), content and users based features. In Random Forest we have got good percentage performance in Sentimental score and POS features combination with baseline features, its improving 0.59% in spam detection performance with all features combination its just 0.47% performance improvement. In J48 as we have seen its performance improves in POS, POS and emotions combinations with baseline features both have 0.59% improvement with all features combination its just 0.47% performance improvement. SVM also have showing little bit improvement in spam detection accuracy performance its best improvement coming in combination of all proposed features with baseline features gaining 0.14% performance better then as compared with baseline features accuracy. Sentimental score and emotions features combination also have same performance output 0.14%. BayesNet have lost spam detection performance in almost all combinations

V. Conclusion

In this paper, we have suggested some sentimental and POS based features that are combined with content/user based features which can be used to differentiate between spam tweets and legitimate tweets on the twitter a popular online social networking site. Our suggested features are influenced by twitter spam detection policies and our observations of spam behaviors. By using twitter API we collected our dataset of 29 most trending topic in 2012. We proposed sentimental and some content based features which will help in identifying spam tweets and return spam filtered result set when user visit twitter with good accuracy rate. We evaluate the usefulness of our suggested features in spam detection by using five traditional classifiers like BayesNet, Naive Bayes, Random Forest, Support Vector Machine (SVM) and J48 schemes. Our experiments results shows that Naive Bayes, J48 and Random Forest classifier gives over all best performance than the other classifiers like SVM (it shows some improvements in spam detections as compared with content and user based baseline features) and BayesNet. Naive Bayes, Random Forest, J48 and SVM spam detections performance improved with our all proposed features combination. Naive Bayes accuracy with baseline features is 14.1313%, results improved a lot with our proposed features combination with baseline features to 25.3084% and it gives 11.18% performance improvement in spams detections. Random Forest baseline accuracy is 91.8118 % is also improved to 92.2914% which given 0.48% improvement. J48 baseline features accuracy is 91.8778% is improved to 92.3435% which gives 0.47% improvement. SVM baseline features accuracy is 91.2765% with combination to our all proposed features improved to 91.4156% which gives 0.14% performance improvement. By using Naive Bayes, J48 and Random Forest classifier, our suggested features can achieve 93% precision and 95% F-measure. We are leaving future work for now to evaluate our spam detection scheme using larger twitter dataset as well as other online social networking sites like Facebook.

References

[1] A. Einstein, On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat, Annalen der Physik 17, pp. 549-560, 1905.

[2] C. Shirky, Blog explosion and insiders club: Brothers in cluelessness (6 October), at http://many.corante.com/archives/2004/10/06/blog_explosion_and_insiders_club_brothers_in_cluelessness.php, accessed 14 August 2009.

[3] A. Swidler, 2009. Q309 Spam and Virus Trends from Postini (1 October), at http://googleenterprise.blogspot.com/2009/10/q309-spam-virus-trends-from-postini.html,accessed14November2009

[4] Symantec, 2009. State of spam, at http://www.symantec.com/business/theme.jsp?themeid=state_of_spam,accessed14November2009

[5] D. Waters, 2009. Spam overwhelms email messages, BBC News (8 April), at http://news.bbc.co.uk/2/hi/technology/7988579.stm, accessed 4 December 2009.

[6] Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna. Detecting Spammers on Social Networks. In Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC '10). ACM, New York, NY, USA, 1-9

[7] Viera AJ, Garrett JM. Understanding interobserver agreement: the statistic. Fam Med 2005; 37:3603.

[8] Sarita Yardi ,Danial Romero, Grant. S and danah. B Detecting Spam in a twitter Network, First monday, Volume15, Number1-4 January 2010

[9] M. McCord, M. Chuah. Spam Detection on twitter Using Traditional Classifiers 8th International Conference, ATC 2011, Banff, Canada, September 2-4, 2011

[10] Zi Chu, Indra Widjaja, and Haining Wang. Detecting Social Spam Campaigns on twitter. In Proceedings of 10th International Conference, ACNS 2012, Singapore, June 26-29, 2012.

[11] Alex Hai Wang. DONT FOLLOWME: SPAM DETECTION IN TWITTER, Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), p 1-10

[12] Fabrcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virglio Almeida Detecting Spammers on twitter, Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam ConferenceJuly 13-14, 2010, Redmond, Washington, US

[13] Nikita Spirin, Mutually Reinforcing Spam Detection on twitter and Web, Technical Report, NorthEastern University

[14] Kanak Biscuitwala, Vidya Ramesh, Kevin Tezlaf Analyzing twitter Spam, Project Report Autumn 2011,Stanford University

[15] Juan Martinez-Romo , Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis, Expert Systems with Applications, Volume 40, Issue 8, 15 June 2013, Pages 2992-3000

[16] Jonghyuk Song, Sangho Lee, and Jong Kim. 2011. Spam filtering in twitter using sender-receiver relationship. In Proceedings of the 14th international conference on Recent Advances in Intrusion Detection (RAID'11), Robin Sommer, Davide Balzarotti, and Gregor Maier (Eds.). Springer-Verlag, Berlin, Heidelberg, 301-317

[17] Puneeta Sharma and SampatBiswas. Identifying Spam in twitter Trending Topics http://www-scf.usc.edu/~sampatbi/pubs/Identifying_Spam_in_twitter_Trending_Topic.pdf

[18] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of LREC 2010

WHITE – DONKEY: Unmanned Aerial Vehicle for searching missing people.

Jaime Moreno*, Jesús Cruz, and Edgar Domínguez
Academia de Computación
Departamento Académico de Ingeniería en Comunicaciones y Electrónica
Escuela Superior de Ingeniería Mecánica y Eléctrica, Campus Zacatenco
Instituto Politécnico Nacional
07738, Mexico.

***Abstract*—Searching for a missing person is not an easy task to accomplish,so over the years search methods have been developed, the problem is that the methods currently available have certain limitations and these limitations are reflected in time location. Time location in a person search is a very important factor that rescuers cannot afford to waste because the missing person is exposed to great dangers. In people search the vision system of the human being plays a very important role. The human visual system has the ability to detect and identify objects such as trees, walls, people among others besides to estimate the distance to them, this gives the human being the possibility of moving in their environment. With the development of artificial intelligence primarily to computer vision it is possible to model the human visual perception and generate computer software needed to simulate these capabilities. Using computer vision is expected to search for any missing person designing and implementing algorithms in order to an Unmanned Aerial Vehicle perform this task, also thanks to the speed of this is expected to reduce the time location. By using of a Unmanned Aerial Vehicle is not intended to replace the human being in the difficult task of searching and rescuing people but rather is intended to serve as a support tool in performing this difficult task.**

***Keywords*—Computer Vision, Unmanned Aerial Vehicle, Search And Rescue System, Human Visual System, and Quadricopter**

I. Introduction

The search and rescue brigades aim to make relief actions immediately and adequate personnel who need the form, in the presence of an emergency. Brigade search and rescue identify and analyze the risk in their workplace also set functions for each of the members of the brigade in an emergency, search operations are always made by hunt groups of two or more brigadiers with the right equipment.

This Search And Rescue (SAR) system consists of individual elements that must work together to provide a global service[1]. The primary components are:

- A rescue coordination center to organize SAR services
- Communication within Regions Search and Rescue (RSR / SRR) and the outer SAR service.
- One or more sub-centers coordinators.
- SAR media including search and rescue units (USR / RSU) staffed with qualified personnel and specialized equipment.
- Designate a coordinator at the crash site (CLS / SMC).
- The staff mentioned above must be bilingual to better develop their activities.

Another common way to seek a missing people is employment of Search dogs, Figure 1. These kind of dogs have been the best search tools for SAR teams (search and rescue) because all dogs have highly developed senses of smell and hearing but not just any dog is useful for this activity because the dog has to be agile, fast and resist difficulties of the work done. Bloodhound breed is preferred for this activity. The search and rescue dogs are classified into two main groups, tracking dogs and air scent dogs, according to the task they was assigned.

Figure 1: Search dogs.

In addition to the search dogs today has made use of UAVs for performing search and rescue. These vehicles have the quality to cover large areas in a short time and vertical take off and landing , so they can be kept at a fixed point in the air.

II. Vertical Takeoff and Landing Vehicles

A quadricopter is an air vehicle propelled by four rotors, capable of taking off and landing vertically[2]. The quadricopters significantly reduced their size and weight, since the French aviation pioneer Etienne Oehmichen proved that it was possible the construction of theoretical helicopter with making Quadricopter Oehmichen No.2 in 1922, Figure 2. In the same year the American George Bothezat built a cuadrirrotor device but without lifting more than 5 meters above the ground, the interest in these systems was suspended for a while.

It was not until mid-twentieth century when interest in the quadricopters reemerged with projects financed by the United

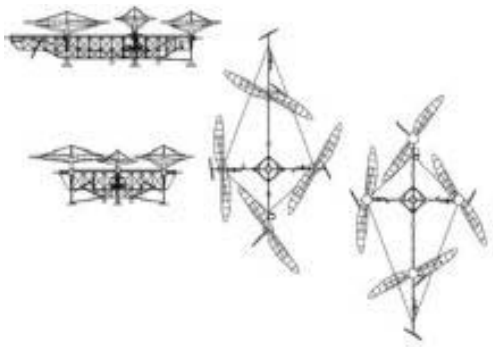


Figure 2: Quadricopter Oehmichen No.2 (1922).

States Navy, which sought the *Flying Jeep* as a means of air transport for troops in war zones, Figure3.

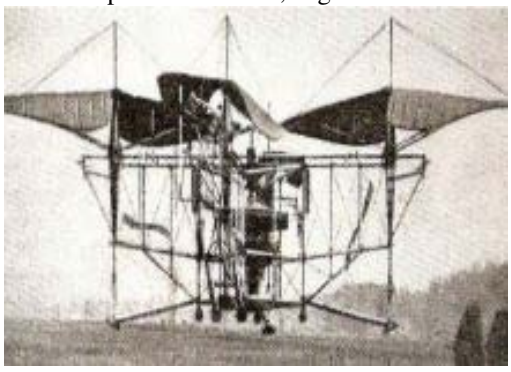


Figure 3: Flying Jeep, United States Navy.

Curtiss-Wright VZ-7 showed good maneuverability and ease of handling. The company Curtiss Wright delivered two of them in mid-1958 for testing. The aircraft performance was satisfactory, rose to 60 meters above ground level and moved to 51 km/h, but did not meet the standards of the navy and was returned to the manufacturer in 1960. That same year, Curtiss-Wright delivered another prototype X-19 aircraft. It flew for the first time in November 1963 and in August 1965 it collapsed, prompting the cancellation of the project. In 1956, another of these developments is the flight of the Convertawings quadricopter Model A, the control mechanism is much simpler compared to its predecessors, based on a differential device that balances the changes in the driving force of each rotor.

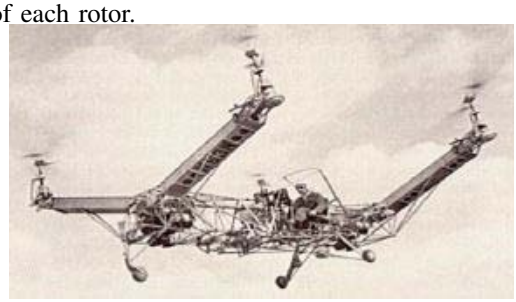


Figure 4: Convertawings quadricopter Model A.

The term UAV (Unmanned Aerial Vehicle) is common in the 90s to describe robotic aircraft and replaced the term Remotely Piloted aerial Vehicle (RPV). The interest in UAVs systems woke because a pilot aboard the aircraft is not required, allowing the use of smaller aircraft, with lower power requirements. These systems can perform dangerous missions in hostile environments or under adverse weather conditions without compromising the physical integrity of the pilot[3]. Nowadays, Searching and Rescue CENTUM company dedicated to developing engineering projects and specialty technology products, developed an unmanned vehicle with a search system in emergency situations. On the other hand, ACRE Surveying Solutions Company which is one of the main Spanish companies of topography services, leader in rental and sale of measuring instruments offers a service called thermography with UAVs for people search and rescue. Companies North Guardian UAV Services Canada and North Chile Guardian UAV Services offer real-time video and high-resolution aerial images. The UAVs are a great alternative to expensive flights with airplanes or helicopters for surveillance and rescue, they provide images for aerial photography, exploration, disaster and many other uses. Meanwhile in Mexico SkyBotica use a group of UAVs to serve in the searching work and air rescue of missing people, as well as damage assessment and mapping of natural disasters, which makes Morelos in the first bank to use the technology for this purpose. Equipping a UAV with a vision system facilitates the search and rescue of missing people and thanks to computer vision is possible to create computer algorithms that automatically carry out this task with great precision.

III. Theoretical Framework

A. Human Visual System

The eye receives light stimuli from the environment. Light passes through the cornea and reaches the pupil, the pupil contracts or expands depending on the intensity of light, if this is intense pupil contracts (miosis), if the light is dim, the pupil dilates (mydriasis), Figure 5. The Iris constriction is involuntary and is controlled automatically by the parasympathetic nervous system, the expansion is also involuntary, but depends on the sympathetic nervous system[4]. The retina contains cells called rods and cones, these sensory cells react differently to light colors and shape. The cones are concentrated in the center of the retina, while the rods are more abundant in the periphery thereof. Each cone is individually connected to the visual center of the brain, which in practice allow to distinguish two points of light separated by just a millimeter at a distance of 10 meters. Each human eye has 7 million cones and 125 million rods.

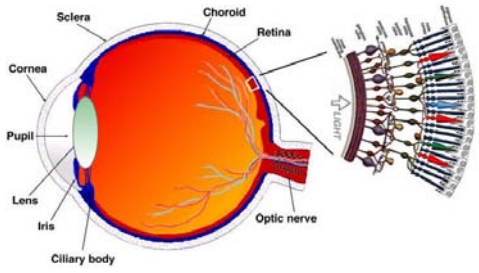


Figure 5: Physiology of Human Visual System.

Also, we want to define as Stereoscopic vision the physical and psychological faculty that human beings possess, which allows to view three-dimensional objects that are contemplated by binocular vision. On each of the retinas of both eyes, an image perspective of the same object is formed, which differ due to the different position of the views, producing the relief

effect. The distance between these two points of view, that is, the separation between the receptor organs of the human being, has an average value of 65 mm and is called interpupillary distance.

B. Digital cameras

Currently digital cameras are used to develop computer vision, because they have the properties of a digital information source: as signal to noise ratio constant and known, as well as an advanced technological infrastructure[5]. The cameras capture videos that translate into frames or pictures. A photograph serves to observe the environment and analyze it through a computer program. In a way that it can decide to perform movements to avoid obstacles, move objects, build maps, recognize objects, and others.

Figure 6: Digital Camera.

To learn how to work with images from a digital camera is necessary to know the technique used in digital photographs. This process is performed by a device called CCD (Charge-Coupled Device, device interconnected electric charges). The CCD is an integrated circuit that contains a number of capacitors coupled in a matrix form. These capacitors are light sensors of a certain frequency: the frequencies of red, green and blue which are the scheme named RGB. The CCD is composed of a matrix of a number of NxM light sensors that translate light energy into movement of electrons to cause an electric current energy.

The array of sensors is connected to a signal conditioner circuit and then to a integrated circuit that sample the signals sent by each light sensor of CCD from time to time T and then store the information in a temporary memory. Finally the memory information is sent to the software that will handle save, view or edit the captured image.

C. Computer vision

Computer vision is a branch of artificial intelligence that aims to mathematically model the processes of visual perception in living and generate computer programs that allow to simulate these visual capabilities, Figure 7.

The first attempt to solve the problems of computer vision was made by Seymour Paper in 1966. The computer vision is more complicated than most people might think. This is not the translation of lights, colors and nuances in pixels, is the translation of the pixels in abstract mathematical concepts.

Frank Rosenblatt in 1958 introduced their new algorithm, the perceptron, which is a form of a neural network[6]. Rosenblatt proved his Perceptron in the automatic classification of images. Although in research experiments, the algorithm seemed to be successful enough, it failed in field tests. Soon,

Figure 7: Visual Illusion.

and despite its natural beauty, the neural network were in disuse, although later research showed that this was largely unjustified.

After this early catastrophe, which stagnated scientific progress in artificial intelligence and computer vision, researchers focused mainly on solving image processing problem .The Image processing consist in pixel-level operations. Later in the 90s neural networks appeared again under the name of convolutional neural network. Despite its success, neural networks were still unable to perform the toughest tasks with three-dimensional objects in images without restrictions. Again, they were not favored by the community of computer vision. In the decade of the 90 and 00 was the true birth of modern computer vision. A sudden plethora of methods were proposed for dealing problems of computer vision generic stalwarts, such as object classification, object detection and segmentation, face recognition, etc. After the golden age, computer vision reached today, where it can finally begin to fulfill its prehistoric promises. Today, the modern version based on neural networks, deep learning, have the capability to classify the content of an image in a very precise way.

One of the modern computer vision approaches is the Viola-Jones algorithm [7]. This algorithm has a low cost in hardware, and consists of two main parts: cascade classifier, which ensures rapid discrimination and a trainer based on Adaboost classifiers. Viola Jones has a probability of 99.9% true positive and false positive probability of 3.33%, and in contrast to other algorithms used in methods of invariant characters,it process only the information present in a grayscale image. It does not use directly the image but instead uses a representation of the image called integral imaging. To determine whether a face is found in a image, the algorithm divides the integral image into subregions of different sizes and uses a series of classifiers (classifiers cascade), each with a set of visual features. In each classifier it determines whether the sub-region is a face or not. Using this algorithm saves considerable time because will not be processed subregions of the image that is not known with certainty that contain a face and only invest time in those subregions which may contain a face[8].

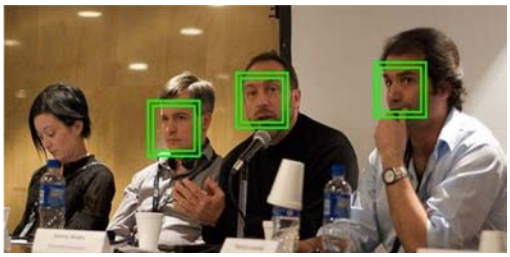


Figure 8: Algorithm Viola and Jones.

D. Principle of operation of Tetramotor

A four-engined is a six degrees of freedom system (x, y, z, pitch, roll and yaw), multivariable and tightly coupled. The main forces and moments acting on a four-engined are produced by their rotors, Figure 9. Two pairs of motors rotate in opposite directions to balance the total torque of the system[9].

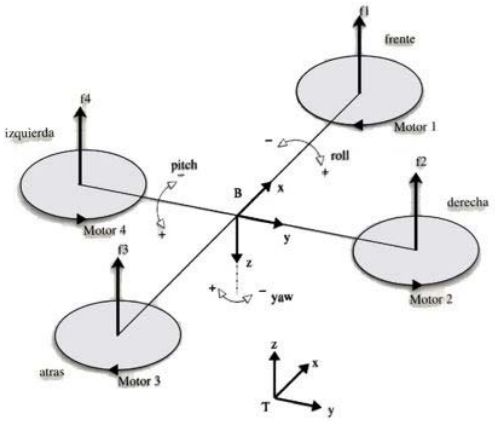


Figure 9: Principle of operation of Tetramotor.

Brushless motor:

A brushless electric motor (brushless) is an electric motor which does not use brushes to make the change of polarity in the rotor[10]. Brushless motors consist of a moving part which is the rotor, where the permanent magnets and a fixed part called the stator or casing. To make work a brushless motor is necessary to use an electronic speed control (ESC). An electronic speed control (electronic speed control or ESC) is an electronic circuit whose function is to vary the speed of an electric motor, direction and possibly to act as a dynamic brake. The ESC's are commonly used in motors electrically operated by radio control, with the variety most commonly used for brushless motors, providing a three-phase low voltage source electronically generated. An ESC may be a separate unit that plugs into the acceleration control channel in the receiver or may already be incorporated in this.

Transceiver:

A transceiver is a device that has a transmitter and a receiver that share parts of its own circuit. When the transmitter and receiver do not have common parts of the electronic circuit, it is known as transmisor-receiver. Since certain circuit elements are used for both transmission and reception, the comunicaton that provides the transceiver can only be half-duplex, which means that signals can be sent in both directions, but not simultaneously.

Microcontroller:

A microcontroller is a digital integrated circuit that can be used for very different purposes due to is programmable. It consists of a central processing unit (CPU), memories (ROM and RAM), input lines and output (peripherals), a microcontroller has the same basic function blocks of a computer. A microcontroller can be used for many applications, some of them are: management of sensors, controllers, games, calculators, sequencer lights, electronic locks, motors control. To use a microcontroller, its functions must be specified by software with programs which indicate the actions that the microcontroller must perform.

IV. WHITE – DONKEY: Unmanned Aerial Vehicle

A. General Algorithm

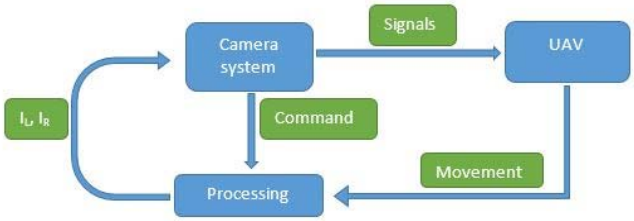


Figure 10: General Algorithm.

The proposal of this work comprehend 3 main stages, Figure 10:

1) Camera system, and Algorithm 1.
2) Processing, Algorithm 2.
3) Unmanned Aerial Vehicle, Algorithm 3.

Algorithm 1: Camera system

Input: Command, Movement
Output: I_L, I_R
1 Receive a command to take the picture.
2 Take capture of the environment that depends on the UAVs movement.
3 Delivery a left image and right image of the environment.

B. Description

After studding the different methods for object detection using techniques of computer vision and the principle of operation of the main elements that make up the UAV, is made the design and assembly of the vision system , the construction and UAV's configuration, the design of algorithms to perform image processing and design the algorithms for controlling the UAV.

Algorithm 2: Processing

Input: I_L, I_R
Output: Signals
1 Viola-Jons Algorithm is applied in I_L in order to detect people.
2 With the stereo pair images (I_L, I_R) the disparity map is calculated and stereoscopic vision is generated.
3 Telemetry is made Using stereoscopic vision in order to calculate the distance to the person from the cameras.
4 Four signals are generated from the telemetry.

Algorithm 3: Unmanned Aerial Vehicle

Input: 4 Signals
Output: Movement
1 Each received signal controls a different UAV‘s movement:
2 Pitch,
3 Roll,
4 Yaw, and
5 Rudder.

1) Capture stage: After checking the correct operation of the transmission between the camera and the receiver using Honestech VHS to DVD 3.0 SE software within MatLab software environment, certain properties of the video object are set according to our needs. Properties to set are as follows:

1) FramesPerTrigger.
2) ReturnedColorspace.
3) FrameGrabInterval.

Due to the video object always uses the same properties and for not performing the same procedure each time that cameras are used these settings are entered into a function called *configCam1* for the camera 1 and *configCam2* for camera 2 . Stereoscopic vision system is made using the new modeling and printing 3D technology in order to construct a base where the two cameras will be set, with a distance of 10cm between them, Figure 11.

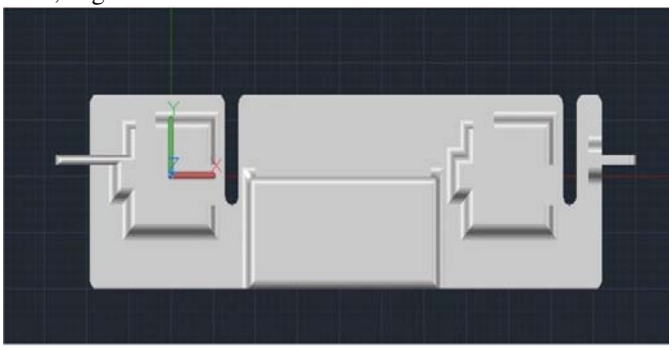


Figure 11: Stereoscopic vision system.

Now it is necesary to calibrate the cameras that make up the vision system, ie get the intrinsic and extrinsic parameters of both cameras and stereo system parameters, Figure 12. To calibrate the vision system the Zhang method is used[11], this method is already implemented in the MatLab software, so to use it simply is typed the *cameraCalibrator* command in the MatLab command window,once the calibration procedure is completed this tool returns an object with the parameters of the stereoscopic system.

Figure 12: Calibration the vision system by means of the Zhang method[11].

2) Stage processing: Once the vision system is ready, to detect a person in an image the Viola-Jones algorithm is used which is already implemented in the MatLab R2015.

3) Person and face Detection algorithm:

1) Start.
2) Set the camera 1
3) Open and configure the computers serial port
4) Take a picture of the environment and storage it in the variable I
5) Define and configure the object detector using the *bodyDetector* constructor as input the variable I and name it *bboxBody*.
6) Define and configure the object detector using the *faceDetector* constructor as input image the variable I and name it *bboxes*.
7) If the content of *bboxes* or *bboxBody* is greater than zero draw a rectangle on the image using the size and coordinates of *bboxes* also draw a rectangle using the size and coordinates of *bboxBody*.
8) If there is no content in *bboxes* or *bboxBody* send a specific character from the computers serial port
9) End.

4) Algorithm to light a LED if a person is detected, it is performed in the MSP430G2553 microcontroller:

1) Start.
2) Set inputs ports.
3) Set outputs ports.
4) Configure the serial communication.
5) If the UCA0RXBUF register of the MSP430G2553 microcontroller receives a specific character turn a LED on.
6) If the UCA0RXBUF register of the MSP430G2553 microcontroller receives a specific character turn the led off.
7) End.

What follows now is to measure the distance to the detected person, is proposed the following algorithm .

1) Start
2) Load the cameras parameters, *camStereoParams*
3) Take a picture with the right camera , storage it in *frameRigth*
4) Take a picture with the left camera, storage it in *frameLeft*

5) Rectify the stereo images *frameLeft* and *frameRigth* using the object *camStereoParams* and storage the new images in *frameLeftRect* and *frameRightRect*
6) Convert *frameLeftRect* and *frameRightRect* to grayscale and storage them in *frameLeftGray* and *frameRigthGray*
7) Calculate the disparity map between *frameLeftGray* and *frameRigthGray*
8) Reconstruct the 3D scene
9) Detect person and face in the *frameRightGray* image .
10) If a face or a person is detected, send a specific character from the computer‘s serial port.
11) Find the centroid of the detected person.
12) Find the coordinates of the *centroids* in the 3D world.
13) Find the distance to the camera in meters.
14) If the distance is less than two meters send a specific character from the computer‘s serial port.
15) If the distance is greater than two meters send a specific character from the computer‘s serial port.
16) If a person is no detected send a specific character from the computers serial port.
17) End.

5) Control Interface: Due to the complexity of design, the time it takes to create an UAV from scratch and be able to control it from a computer, it was decided to use the *kk2.0* control board, Figure 13. This card is used in various aerial vehicles, such as:

1) Tricopter
2) Quadcopter +
3) Quadcopter X
4) Hexcopter +
5) Hexcopter X
6) Octocopter +
7) Octocopter X
8) Aero 1S Aileron
9) Aero 2S Aileron
10) Flying Wing
11) Singlecopter 2M 2S
12) Singlecopter 1M 4S

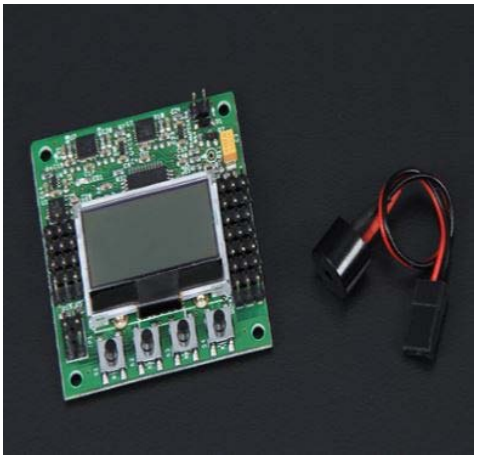


Figure 13: Control Interface.

The *kk2.0* control board has an IMU of 6 degrees of freedom that is quite sensitive, it is the MPU6050 which consists of a gyroscope and an accelerometer, this card is controlled with the HobbyKing *hk-T6A* receiver, the problem of this receptor lies in frequency because it is the same in which the cameras transmit, causing too much interference, making impossible to detect persons in a frame. The solution to this huge problem is to replace the receiver and the control circuit, so a new control circuit is used instead in order to deliver the same signals as the receiver delivers to the control board.

Chanel	Function	TH	TL
1	Aileron	1.54 mS	16.79 mS
2	Elevator	1.54 mS	16.79 mS
3	Throttle	1.02 mS	17.33 mS
4	Ruder	1.54 mS	16.79 mS
5	Aux	2 mS	16.32 mS

Table I: Measured values of each channel that the receiver delivers

From Figure 14, in Blue is the signal generated by the receiver and in yellow is the signal generated by the Microcontroller.

Figure 14: Signal generated by the receiver and the Microcontroller, blue and yellow, respectively.

Printed Circuit Board (PCB) design to generate the signals that control the UAV‘s movements, Figure 15.

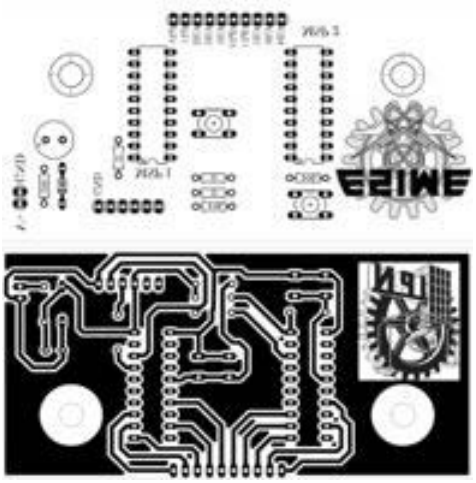


Figure 15: Printed Circuit Board of the UAV.

6) User interface: For the use of people search software be more user friendly, a Graphical User Interface (GUI) is designed with the *GUIDE* tool that the MatLab software provides. This *GUIDE* is composed of three main windows , the window 1 shows what records the right camera, the window 2 shows what records the left camera and in the window 3 the person detection process is shown, in this window the face of the detected person is shown, Figure 16.

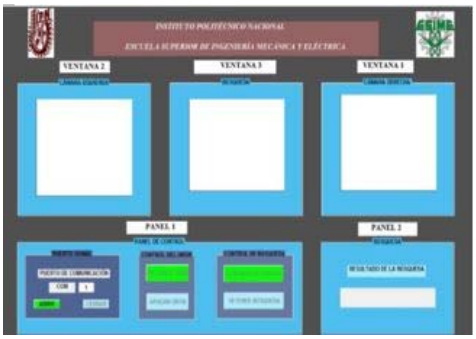
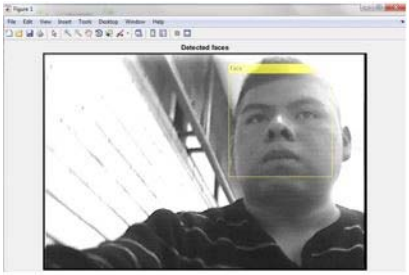
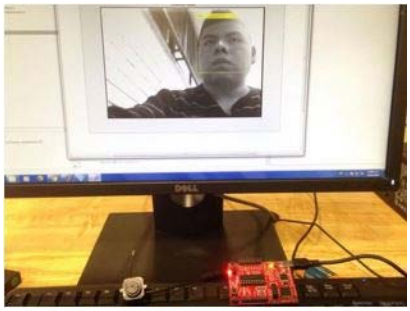


Figure 16: WHITE – DONKEY: User interface.



(a)



(b)

Figure 17: Face Detection, (a) Preview and (b) Algorithm integrated into the *LaunchPad*.

V. Results

A. Turn a LED on if a face is detected in a Image

Figure 17 shows when a face is detected in a image, a LED turns on, which is already integrated into the *LaunchPad* development tool.

B. Measure the distance to the person from the cameras.

Being ready the stereoscopic vision system this was placed focusing to a person(Figure 18), after that, the program implemented in the MatLab software is run and as a person was found in the left frame the distance to this was measured.

C. Connection to the Unmanned Aerial Vehicle

Connection and Disconnection the transmitter with the receiver on the UAV for controlling it from a computer. Figure 19 show (a) Armed or Connection Mode of the Unmanned Aerial Vehicle, (b) Safe or Disconnection Mode of the Unmanned Aerial Vehicle, and (c) Unmanned Aerial Vehicle connected and in operation.

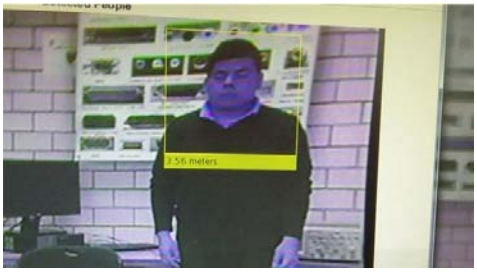


Figure 18: Measure the distance to the person from the cameras.

D. Matlab GUIDE

From Figure 20, the GUIDE in initial conditions, shows only in active state the serial port control and gives the option to introduce an ID corresponding to the port to be used. When the serial port is opened successfully, the button to power the UAV is enabled, which enables communication with this, besides the search start button is enabled too.

From Figure , in order to begin the searching the UAV begins to take off vertically and subsequently to turn on its own axis. In the central window of the GUIDE is shown graphically whether or not there is a person, besides the search panel indicates it verbatim.

Once the UAV found a person in the surroundings, it moves toward it, if the distance to the person is less than 2m the UAV stops for the security of the person found, otherwise the UAV is in progress, Figure 22. Showing the face of the detected person in the central window is made with the aim of the software operator decides whether the face of the person found matches any person reported as lost.

Acknowledgment

This work is supported by National Polytechnic Institute of Mexico (Instituto Politécnico Nacional, México) by means of Project No. 20160786, the Academic Secretary and the Committee of Operation and Promotion of Academic Activities (COFAA) and National Council of Science and Technology of Mexico (CONACyT) by means of grant No. 204151/2013.

References

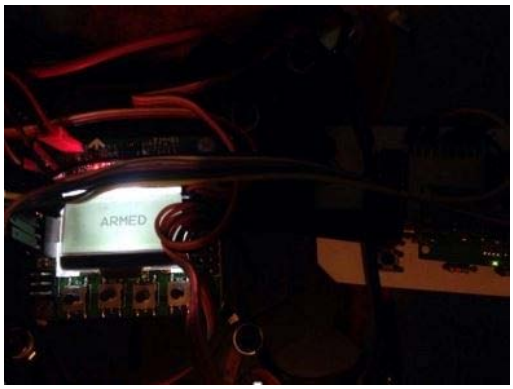
[1] F. Xiao, Y. Jin, Y. Yin, and Y. Li, "Design and research of marine search and rescue simulation system," in *Information Technology and Computer Science (ITCS), 2010 Second International Conference on*, July 2010, pp. 372–376.

[2] O. Lawlor, M. Moss, S. Kibler, C. Carson, S. Bond, and S. Bogosyan, "Search-and rescue robots for integrated research and education in cyber-physical systems," in *e-Learning in Industrial Electronics (ICELIE), 2013 7th IEEE International Conference on*, Nov 2013, pp. 92–97.

[3] S. Bertrand, T. Hamel, and H. Piet-Lahanier, "Stabilization of a small unmanned aerial vehicle model without velocity measurement," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, April 2007, pp. 724–729.

[4] Gomai, A. El-Zaart, and H. Mathkour, "A new approach for pupil detection in iris recognition system," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 4, April 2010, pp. V4–415–V4–419.

[5] Q. Weigang, "On-line yarn evenness detection using ccd image sensor," in *2011 Chinese Control and Decision Conference (CCDC)*, May 2011, pp. 1787–1790.



(a) Armed Mode

(b) Safe Mode

(c) UAV flying

Figure 19: Connection to the Unmanned Aerial Vehicle, (a) Armed or Connection Mode of the Unmanned Aerial Vehicle, (b) Safe or Disconnection Mode of the Unmanned Aerial Vehicle, and (c) Unmanned Aerial Vehicle connected and in operation.

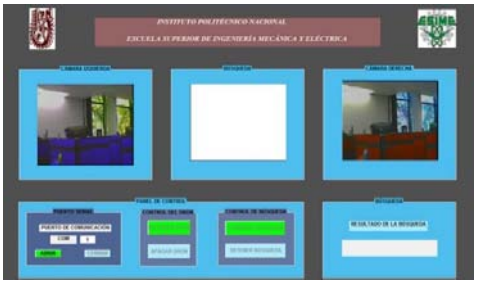


Figure 20: GUIDE in initial conditions.

Figure 21: GUIDE in initial conditions.

Figure 22: When the UVA found a person in the surroundings.

[6] G. Nagy, "Neural networks-then and now," *IEEE Transactions on Neural Networks*, vol. 2, no. 2, pp. 316–318, Mar 1991.

[7] Q. Li, U. Niaz, and B. Merialdo, "An improved algorithm on viola-jones object detector," in *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, June 2012, pp. 1–6.

[8] S. R. Lang, M. H. Luerssen, and D. M. W. Powers, "Evolutionary feature preselection for viola-jones classifier training," in *Engineering and Technology (S-CET), 2012 Spring Congress on*, May 2012, pp. 1–4.

[9] E. A. Paiva, J. C. Soto, J. A. Salinas, and W. Ipanaqu, "Modeling and pid cascade control of a quadcopter for trajectory tracking," in *2015 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, Oct 2015, pp. 809–815.

[10] L. E. Unnewehr, P. Piatkowski, and G. Giardini, "A brushless dc motor for vehicular ac/heater applications," in *Vehicular Technology Conference, 1976. 26th IEEE*, vol. 26, March 1976, pp. 8–15.

[11] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.

Wyner-Ziv Video Coding using Hadamard Transform and Deep Learning

Jean-Paul Kouma

Ulrik Söderström
Department of Applied Physics and Electronics
Umea˚ University, Sweden

***Abstract*—Predictive schemes are current standards of video coding. Unfortunately they do not apply well for lightweight devices such as mobile phones. The high encoding complexity is the bottleneck of the Quality of Experience (QoE) of a video conversation between mobile phones. A considerable amount of research has been conducted towards tackling that bottleneck. Most of the schemes use the so-called Wyner-Ziv Video Coding Paradigm, with results still not comparable to those of predictive coding. This paper shows a novel approach for Wyner-Ziv video compression. It is based on the Reinforcement Learning and Hadamard Transform. Our Scheme shows very promising results.**

***Keywords*—Wyner-Ziv; video coding; rate distortion; Hadamard transform; Deep learning; Expectation Maximization**

I. Introduction

Video compression schemes such as MPEG4 and H.264 [1] are the current state-of-art, where correlation between or among frames are exploited at the encoder side. Such schemes usually achieve high compression with a fairly low complexity at the very expense of a high complexity encoder. Compression schemes like MPEG4 or H.264 are suitable for scenarios where the encoder has enough power computation, like video-on-demand servers.

Mobile phones are today the de facto device for communication. People want to do more and more with their mobile phone. They want to be able to have a real-time video communication experience comparable to that of computers. Unfortunately, current video compression technologies [1] barely permit it: **encoder complexity** is the bottleneck. Either comparable frame rate can not be achieved or conversation cannot last long because of battery is scarce. In either case quality of experience will be dropped.

The video community is aware of that issue as a great deal of research has been conducted since the emergence of camera-based mobile devices. The common insight toward tackle the issue is the so-called *Wyner-Ziv Video Coding (WZVC)* or *Distributed video coding*.

WZVC is the consequence of information-theoretic bounds established in the 1970s. First by Slepian and Wolf for distributed lossless coding [2], and then Wyner and Ziv for lossy coding with decoder side information [3].

Let $\{(X_k, Y_k)\}_{k=1}^{\infty}$ be a sequence of independent drawing of a pair of dependent variables (X, Y) taking values in the finite sets $\mathbb{X}$ and $\mathbb{Y}$, respectively. The decoder has access to the side information Y. Illustration is shown in figure 1. Wyner and Ziv suggest that whether or not the side information Y is available at the encoder, X can be compressed - to Z and decoded to $\hat{X}$ - at a rate $R_{X|Y}(D)$ where $D = \mathbb{E}[d(X, \hat{X})]$ is an acceptable distortion.

In WZCV, unlike that of predictive coding paradigms (i.e. H.264), individual frames are encoded separately but decoded conditionally. According to [3], the compression effectiveness of WZVC schemes should be comparable to that of predictive coding. A typical WZVC setup is shown in figure 2 where both terminals are lightweight modern mobile phone capable of decoding MPEG4 frames for example. The corresponding Wyner-Ziv decoder is thought to be powerful computer capable of exploiting statistics between frames and output MPEG4 streams in real-time, using much more complex algorithms.

Most of the conducted studies in the area of WZCV have been using binary codes. Major contributions come from Standford University [4] and UC Berkeley [5]. Both methods followed a common pattern; those methods were first developed to perform in pixel domain and later in transform domain (namely Discrete Cosine Transform). Those methods suffered from three major drawbacks:

- The overhead of working in binary domain - since DCT pixels or alternatively transform coefficients have to be converted back and forth from and to bit planes during decoding process.
- Rate control - All the pixels values, alternatively transform coefficients need to be converted to binary with the same amount of bits, making the rate control difficult.
- The decoding algorithms used - either generative or discriminative - were somewhat too simplistic and did not work well in practice.

We propose a new practical compression scheme, based on Hadamard transform and reinforcement learning. In contrast to previous works, our method deals with non-binary codes. The encoding is relatively of low complexity with an inherent rate control. We also show that our algorithm outperforms that of the state of art [4] and [5] and really is comparable to predictive coding schemes.

II. Low complexity video encoding

The encoding challenge is to implement an encoder with lower encoding complexity than that of predictive video coding methods [1] and still achieve comparable codec **effectiveness**.

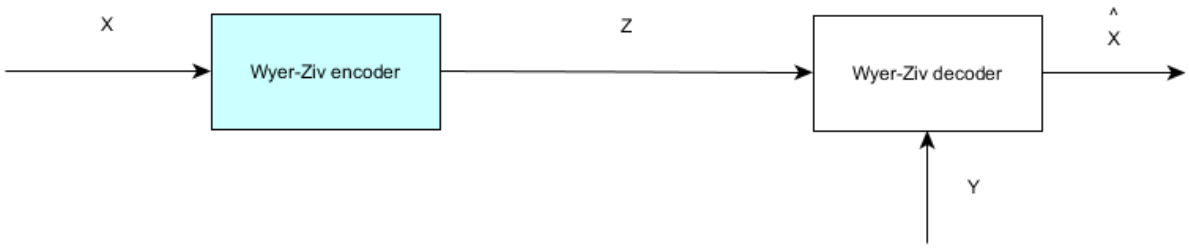


Fig. 1. Source coding with side information available at the decoder

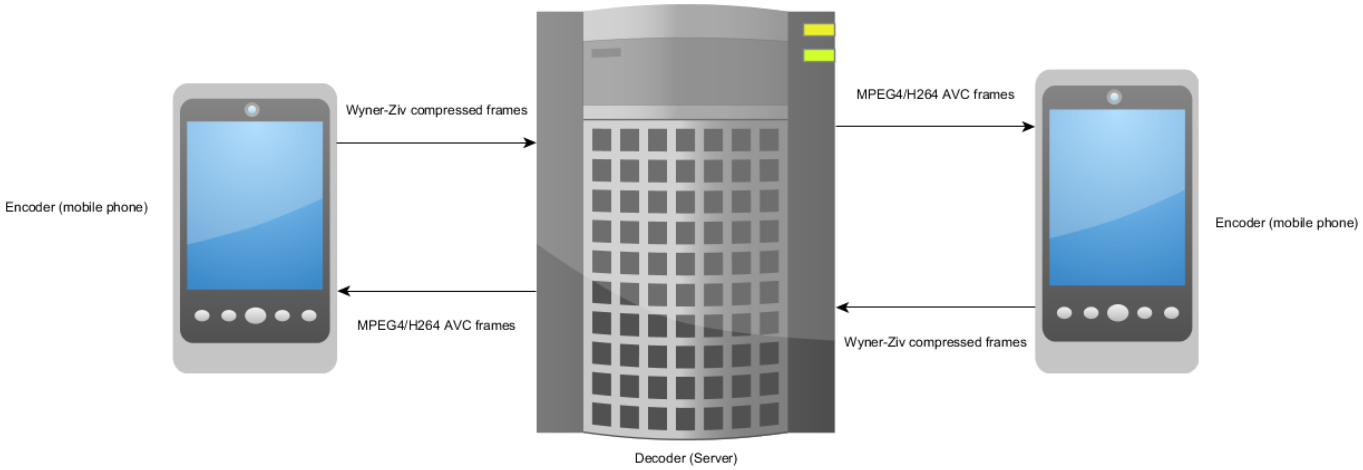


Fig. 2. Wyner-Ziv setup of video compression

A. Problem setup

Let Y and X be two consecutive video frames. X is to be encoded without the knowledge of Y. The challenge is to compress X at the bit rate R **bit/pixel**, where $R = H(Y|X) =$. $H(Y|X)$ is the entropy or the amount of information (in bits) needed to represent the frame X conditioned on that of Y. Usually, Y and X are highly correlated after motion compensation [6], $H(X|Y) < H(X)$. In practice the challenge is to encode X at a rate even lower than $R_\epsilon < H(Y|X)$ - leading to **lossy compression** - and still achieve reconstruction with satisfying fidelity, as suggested by Ziv et al [3].

B. The encoder

Let Z^* be the compressed frame or Wyner-Ziv frame from X. In our case, compression with compression ratio $n : m$ is achieved simply projecting the row version of frame $X \in \mathbb{R}^{1\times n}$ onto the $n : th$ first dimensions of the **orthogonal Hadamard vector basis** $G \in \mathbb{R}^{n\times m}$, where $m < n$. $Z^* = X \times G$. The final stream is $Z = [Z^* \quad \sigma^2]$ where σ^2 is the variance between the current frame X and the previous frame X_{-1}. Its given by

$$\sigma^2 = \sum_{n=0}^{N} X[n] - X_{-1}[n-1] \tag{1}$$

C. The Hadamard Transform

The Hadamard transform is an orthogonal transform that has been used in numerous image coding applications [7], [8]. The transform matrix of dimension 2^k for $k \in \mathbb{N}$ is given by the following recursive formula

$$H(2^k) = \begin{bmatrix} H(2^k-1) & H(2^k-1) \\ H(2^k-1) & -H(2^k-1) \end{bmatrix}$$

and

$$H(2) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

III. Decoding

A. Problem setup

The aim here is to reconstruct the encoded frame Z to X. At time t, the decoder has knowledge of the incoming Wyner-Ziv frame Z and previously decoded frame Y.

To be able to reconstruct X, Z has somehow to contain enough information about Z, possibly together with Y; that is the case in predictive coding schemes, where Z carries information about Y, usually motion vectors (pixel configuration for that matter) and Huffman or Arithmetic coded residuals of the motion-compensated frames [1]. Since we do not really

know the real pixel setup X we will be left to "guessing" it out of the pixel space $\mathbf{\Lambda}$. Mathematically, the problem could be formulated as follow: at a time t, the decoder **observing** Z and Y, aims to find the **"best"** pixel configuration $\hat{X}$. . The term **"best"** used here actually points out that X is **unobserved**. This formation leaves us to a **Maximum Likelihood** problem

$$\hat{X} \leftarrow \arg\max_{\Lambda} P(\hat{X}, Z, Y) \tag{2}$$

B. Maximum Likelihood

Decoding X by simply estimating doing $X = Z \times G^{-1}$ is obviously not optimal, but it may be worth mentioning why at this point of the study. As X grows in size, pixels in it decrease in term of correlation. Consequently, the coefficients in Z won't explain X well. Estimating X become thus equivalent to solving an under-determined system of equations - fewer equations than unknowns.

To find the "best" estimate of X we have to model an optimal Maximum Likelihood Estimator (MLE). That is, designing the Maximum Likelihood estimator to "capture" as much decoding information in Y and Z as possible, that is capable of estimating the following joint quantity

$$MLE(\hat{X}) = P(\hat{X}, Z, Y; \Theta) \tag{3}$$

where Θ is the generative model. Hidden Markov Mode (HMM)l [9] and Reinforcement Learning models (such as Q-Learnining) [10] are two good candidates for such problem. But modelling such MLE problem could be rather complex if applying either HMM or Q-learning due to the dimensionality of the tuples. Fortunately Q-learning has a variant, using function approximators and experience replay [11], that has shown to deal well with high dimensions.

C. Q-learning

Q-learning is a deep learning technique. Generally spoken, the learning model tries to learn the optimal so-called action-selection policy. We are given an agent, states **S** and a set of actions per state **A**. At a time t, the agent receives a reward r_t by executing an action a_t being in state s_t. The goal of the agent is to maximize its total reward by learning optimal action for each state; that is the cumulative discounted long-term reward $Q(a, s)$, starting from the current state. During learning process, the $Q(a, s)$ value is updated as follow

$$Q_{t+1}(a_t, s_t) \leftarrow Q_t(a_t, s_t) + \tag{4}$$

$$\alpha \left[r_{t+1} + \lambda \arg\max_{a} Q_t(s_{t+1}, a) - Q(s_t, a_t) \right] \tag{5}$$

Due to high dimensionality equation (4) cannot be applied directly to our ML problem. Instead we use a variant called *Q-learning with experience replay* [11].

Q-learning with experience replay, the agent's state-action pair is stored in a data set and re-sampled, with respect to some significance criteria, in some later episodes in conjunction with other selected, usually randomly, data set of state-action pairs.

Our Q-learning scheme is endorsing the same intuition and motivation albeit somewhat different in design; The following setup is adopted

- the output of our Q-learning scheme - Q-values - is two dimensional, as opposed to other schemes [11] outputting one-dimensional.
- An episode ends at either n iterations or p dB of PSNR. Whichever comes first. For example $n = 20$ or $p = 45$.
- The reward is delayed, i.e. until episode ends

Each column of the Q-values corresponds to the probability distributions - that are assumed to be Gaussian - over the co-located pixels candidates of X; at every position $X[i]$ in X, we consider $\lfloor \mathbf{2}\sigma \rfloor$ pixel candidates

$$X_j[i], \quad -\sigma \leq j \leq \sigma, \quad j \in \mathbb{Z}$$

Each pixel candidate $X_j[i]$ has an initial probability

$$P_{ij}^0 = \frac{1}{\sigma\sqrt{2\pi}} e^{-(j)^2/2\sigma^2}$$

Our Q-learning design is illustrated in figure 3.

D. Maximum Likelihood through Experience Replay

Recall that the idea behind this whole Q-learning business is to fit a likelihood function. We aim to find out how to capture information out of the previously reconstructed frame, an encoded stream (that is syndrome and mean-squared error) so as decoding is as effective as possible. A Maximum Likelihood function does permit us to estimate the (degree of) truthfulness of a pixel combination - possibly with some measure of confidence of interval.

Even though likelihood could be measured for every pixel configuration of X in Λ estimating the best configuration still remains intractable as X is of high dimension. We use ***Expectation-Maximization (EM)*** [9] to estimate the Maximum Likelihood. In our setup, the **E-step** correspond to the estimation of the Q-values, while the **M-step** choose the pixel combination that maximizes their probabilities:

1) E-step: In the E-step the derived Q-values at each iteration are used to perform a probability update of the side information as follow

$$P_{ij}^{t+1}(x_i j) = \frac{q_{ij} P_{ij}^t}{\sum_{i=0}^{I} q_{ij}}$$

where i is the ith side information and j is the jth distribution from the ith side information

2) M-step: The best pixel combination with respect to recent probability distribution update is selected according to

$$\hat{x_i} \leftarrow \arg\max P_{ij}(x_{ij})$$

The Learning process is depicted in figure 4

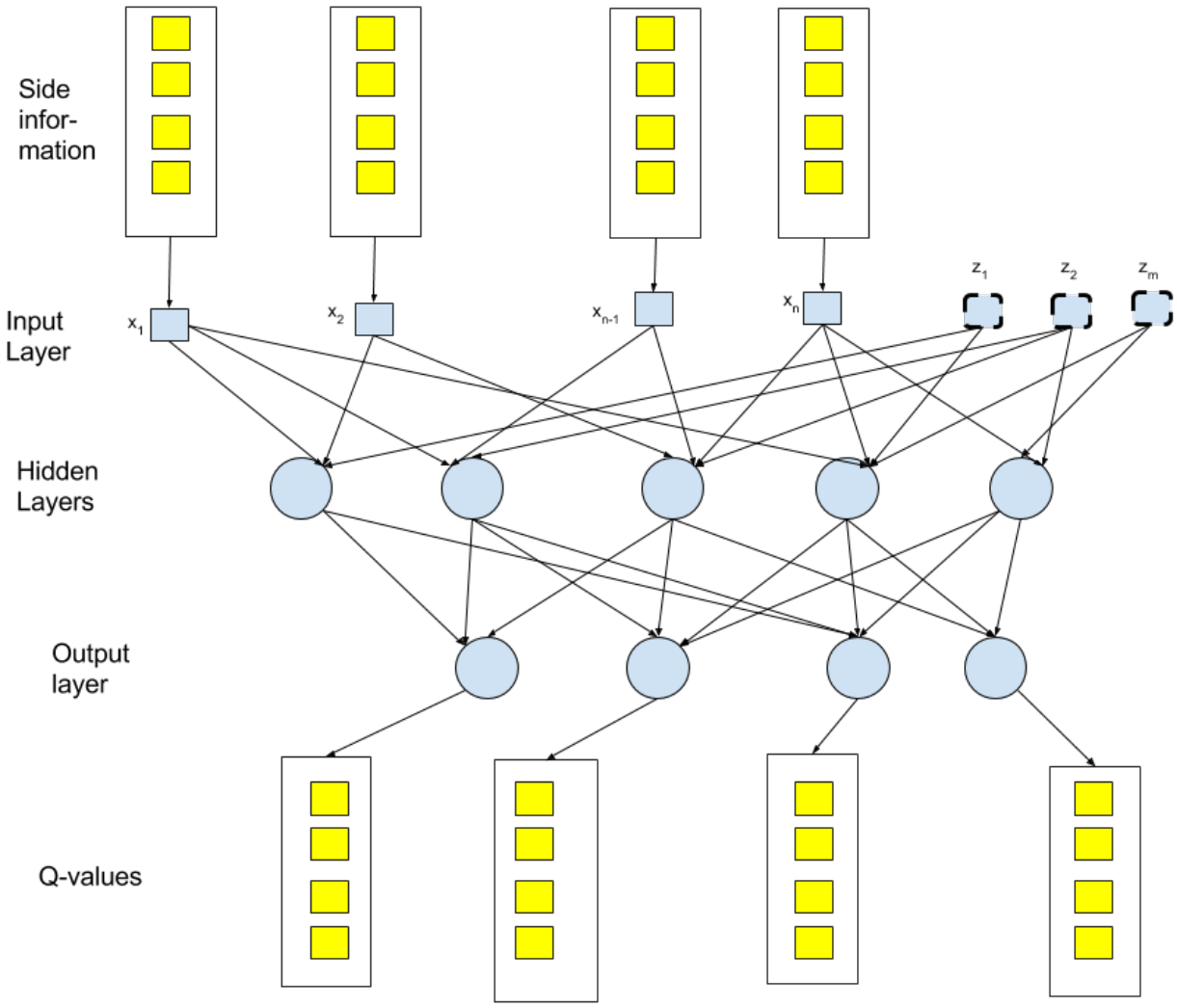


Fig. 3. Setup for Neural Network based Q-learning

IV. Experiment Results and Discussion

As mentioned in section III-C, the size of the Function Approximator's output is related the variance information. Thus a side information can in theory be up to $n = 256$ pixel away from its counterpart. This means that the size of our Q-values is $M \times N \times 256$, where M is the frame height and N is the width. That is, for grayscale QCIF video sequences with $M = 144$ and $N = 176$, as used in our experiments, the size of our Q-values will be $144 \times 176 \times 256 =$ **6 488 064**. That means the Q-values alone require a more than **50 Gigabytes** of RAM memory! As were running the experiment on a personal computer with 8 GB $(4 \times 2GB)$ of RAM, we figured that only around 1024×20 Q-values could fit at a time. We performed therefore a "cherry-picking" procedure for the sake of assessing the effectiveness of our novel algorithm.

Recall that we aim to decode a frame X given its side information Y. To fit the likelihood function, a set of QCIF video sequences were used as training samples. X and Y were divided in blocks and paired up $x_1, x_2, x_3, ..., x_K$ and $y_1, y_2, y_3, ..., y_K$ respectively. The pair (x_i,y_i) was selected as a training sample if their variance was less than 100. That means pixels in x_i is are at most 10 intensities away from pixels in y_i. The length of the block were chosen to 1024 as the width of the Hadamard matrix has to be a power of 2. Recall that the encoder sends z_i and the decoder only has access to y_i and z_i and tries to estimate x_i. Thus $[x_i z_i]$ will be the input to our Function Approximator.

During training phase, we used minibatches of size 1000, while adopting a constant $\epsilon - greedy$ algorithm of 0.1. The input was scaled between -1 and 1 prior entering the Function Approximator. We used 5 hidden layers - with $tanh$ activation function - with 200 nodes per layer. Figures 5 and 6 show the learning ability in terms of rate distortion of the Function Approximator though iterations/episodes. We notice the increase of the PSNR at each episode.

The same "cherry-picking" procedure was used for testing purpose, since we were computationally limited. We tested the algorithm on QCIF video frames for the sequences *Salesman*

		Variance information				
		$\sigma^2 = 20$	$\sigma^2 = 50$	$\sigma^2 = 100$	$\sigma^2 = 150$	$\sigma^2 = 200$
Compression ratio	**1024:3**	38,60	32,60	31,88	29,82	24,73
	1024:10	43,85	37,47	36,50	33,81	29,26
	1024:20	62,66	54,92	45,13	40,12	34,64

TABLE I. TABLE SHOWING THE RATE DISTORTION PERFORMANCE FOR SALESMAN VIDEO SEQUENCE

		Variance information				
		$\sigma^2 = 20$	$\sigma^2 = 50$	$\sigma^2 = 100$	$\sigma^2 = 150$	$\sigma^2 = 200$
Compression ratio	**1024:3**	38,60	32,60	31,88	29,82	24,73
	1024:10	43,85	37,47	36,50	33,81	29,26
	1024:20	62,66	54,92	45,13	40,12	34,64

TABLE II. TABLE SHOWING THE RATE DISTORTION PERFORMANCE FOR HALL VIDEO SEQUENCE

and *Hall Monitor*. Frames blocks with variance information ranging from *around* 10 to *around* 200 were selected. The rate distortion performances are given in tables I and II for the Salesman and Hall video sequences. It is important to notice that even though the Function Approximator is trained on variance information less than 100, we tested our algorithm on variance greater than 100. The reconstruction quality is still good to very good for compression ration 1024:10 and 1024:20, respectively. Compression ratio 1024:3 was also to assess the compression limit. The idea was to check if we could still achieve reasonable distortion by minimizing the number of bits to send.

For each frame block, the encoder generates 3 syndrome coefficients (3 integers = 3 bits) and 1 variance information (1 double =1 bits), 25 blocks and 12 fps. A full frame has 25 blocks. Arguably, compression ration 1024:3, 1024:10 and 1024:20 could thus be comparable to 33, 100 and 200 kbps respectively. This insight shows the high potential of our scheme for **low complexity, low bitrate and low distortion** video coding.

V. CONCLUSION

We have presented a new and practical Video compression scheme based on the Wyner-Ziv framework [3]. The novelty in our scheme lies mostly in the integration of Q-learning in the decoding process. The Wyner-Ziv coding problem has been subject to a great deal of research for at least the past 15 years. The mainstream of in Wyner-Ziv Video Coding has been based on binary codes [4], [12], [13]. Our algorithm is the first really dealing with non-binary codes. A second advantage is its inherent scalability. Previous schemes, such as punctured codes [12] have used different methods for rate control. Non-binary codes have he advantage of reducing the computation complexity at the both at the encoded and decoder, since calculations do not have to be performed at a bit level as in [14] for example.

We also showed that the Wyner-Ziv problem - at least in our case - can be solved using Q-learning algorithm as Likelihood Estimator with a inherent embedding of the EM framework.

However, due our computational limitation, we assessed the algorithm in a ”cherry-picking” manner. The results shown are very good. We arguably showed that our Video Coding scheme was that of **low complexity, low bitrate and low distortion** is specially meaningful for lightweight devices such as surveillance cameras, mobile phones or probably Google^{TM} Watches or Google^{TM} Glasses in the near future when provided with cameras.

Our encoding scheme is of a very low complexity compared to that of motion estimation based video encoders. The bulk of computation is shifted from the encoder to decoder. The decoder is thought to be powerful server station. This is of a great advantage, especially on lightweight devices, such as mobile phones.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the h. 264/avc video coding standard,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, 2003.

[2] D. Slepian and J. K. Wolf, “Noiseless coding of correlated information sources,” *Information Theory, IEEE Transactions on*, vol. 19, no. 4, pp. 471–480, 1973. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1055037

[3] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1055508

[4] A. Aaron, R. Zhang, and B. Girod, “Wyner-ziv coding of motion video,” pp. 240–244, 2002.

[5] R. Puri, A. Majumdar, and K. Ramchandran, “Prism: a video coding paradigm with motion estimation at the decoder,” *Image Processing, IEEE Transactions on*, vol. 16, no. 10, pp. 2436–2448, 2007.

[6] J.-P. Kouma and H. Li, *Large-Scale Face Image Retrieval: A Wyner-Ziv Coding Approach*. InTech, 2011, ch. 2, p. 29. [Online]. Available: http://www.intechopen.com/books/new-approaches-to-characterization-and-recognition-of-faces/large-scale-face-image-retrieval-a-wyner-ziv-coding-approach

[7] C. Kim, H.-H. Shih, and C.-C. J. Kuo, “Fast h. 264 intra-prediction mode selection using joint spatial and transform domain features,” *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 291–310, 2006.

[8] Z. Liu, L. Li, Y. Song, S. Li, S. Goto, and T. Ikenaga, “Motion feature and hadamard coefficient-based fast multiple reference frame motion estimation for h. 264,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 620–632, 2008.

[9] A. P. Dempster, N. M. Laird, and D. B. Rdin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[10] C. J. C. H. Watkins, “Learning from delayed rewards,” Ph.D. dissertation, University of Cambridge England, 1989.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[12] J. Sun and H. Li, “Trellis-based reconstruction in wyner-ziv codec for video,” in *IASTED International Conference on Internet and Multimedia Systems and Applications*, 2005.

[13] A. Aaron and B. Girod, “Wyner-ziv video coding with low-encoder complexity,” in *Proc. Picture Coding Symposium*, 2004.

[14] A. Aaron, S. Rane, E. Setton, and B. Girod, “Transform-domain wynerziv codec for video,” in *in Proc. SPIE Visual Communications and Image Processing*, 2004, pp. 520–528.

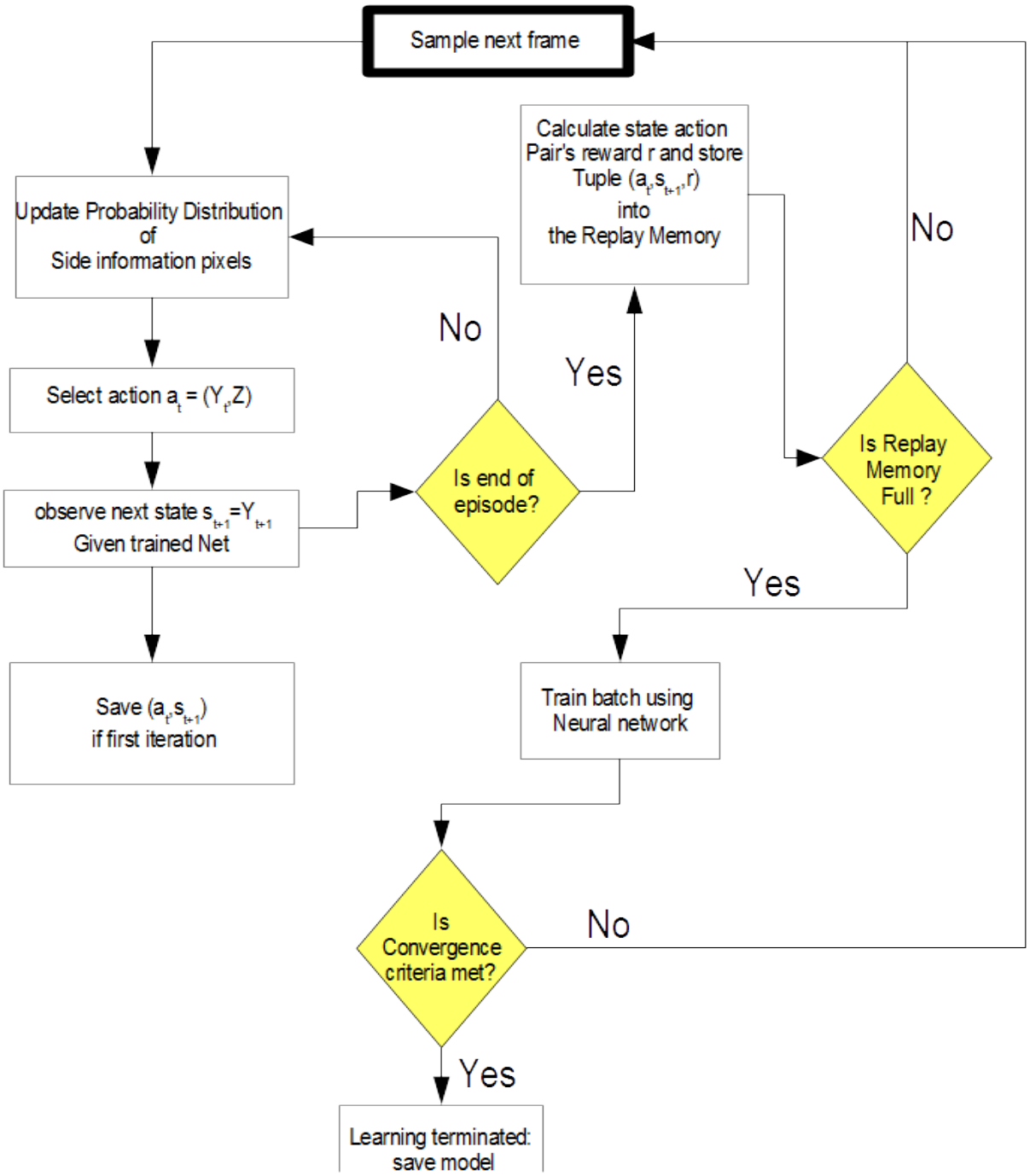


Fig. 4. Learning process of our video codec

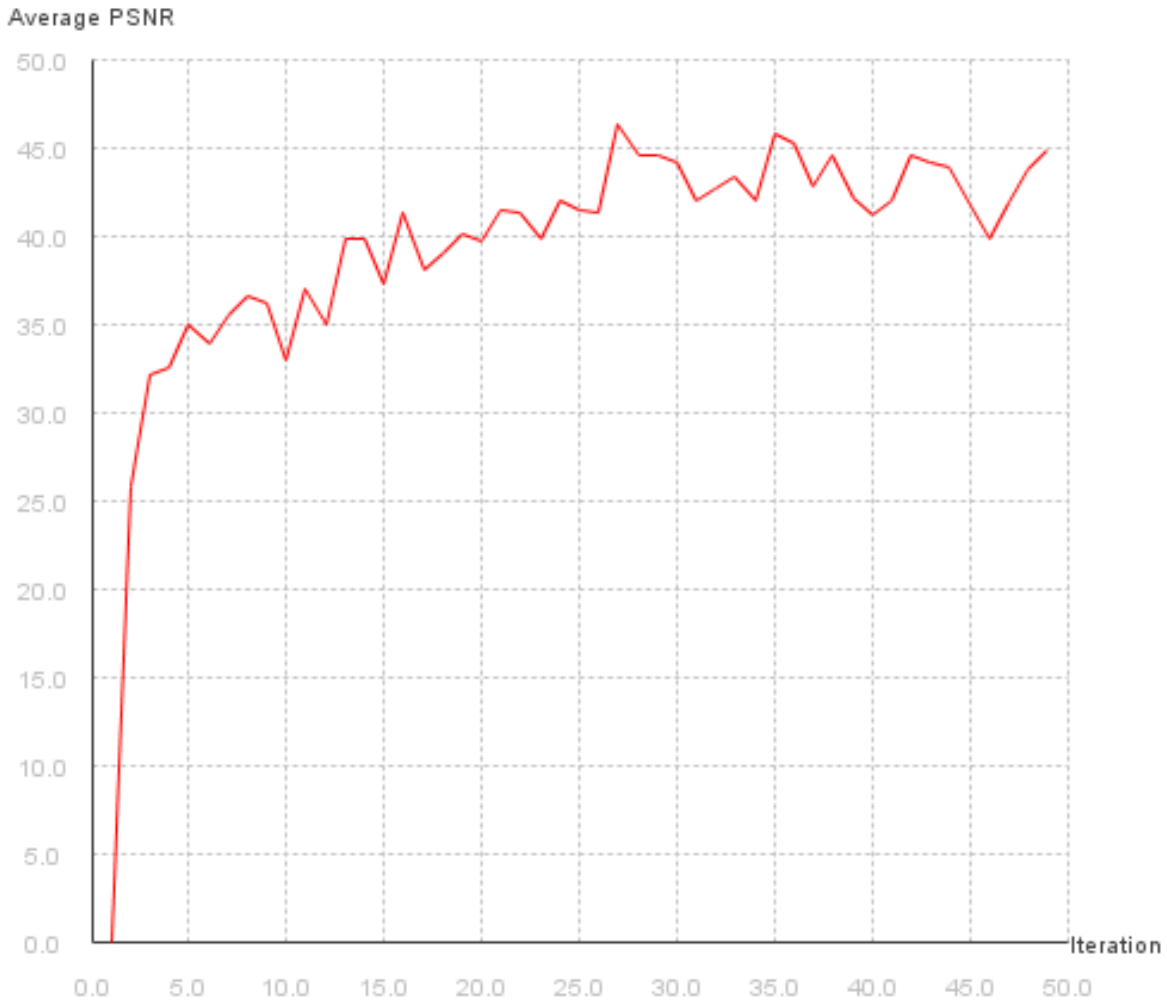


Fig. 5. Average Learning ability of our Function Approximator through episodes: *Compression ratio 1024:3*

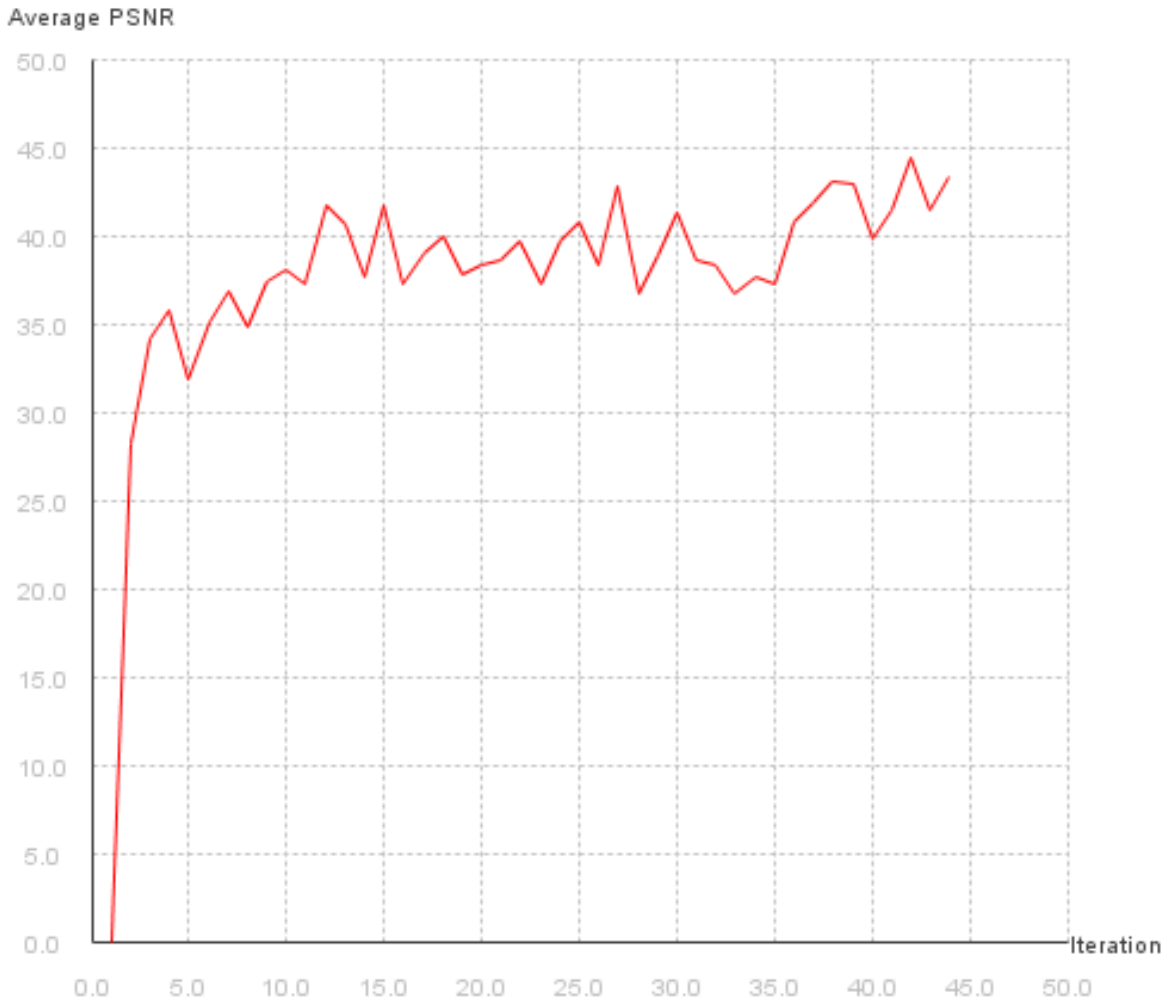


Fig. 6. Average Learning ability of our Function Approximator through episodes: *Compression ratio 1024:5*

Quartic approximation of circular arcs using equioscillating error function

Abedallah Rababah
Department of Mathematics,
Jordan University of Science and Technology
Irbid 22110 Jordan
Email: rababah@just.edu.jo

***Abstract*—A high accuracy quartic approximation for circular arc is given in this article. The approximation is constructed so that the error function is of degree 8 with the least deviation from the x-axis; the error function equioscillates 9 times; the approximation order is 8. The numerical examples demonstrate the efficiency and simplicity of the approximation method as well as satisfying the properties of the approximation method and yielding the highest possible accuracy.**

***Keywords*—Bézier curves; quartic approximation; circular arc; high accuracy; approximation order; equioscillation; CAD**

I. Introduction

The use of parametric representation of curves is convenient in the field of CAD. Especially, the parametric methods approach allows us to make use of some of the properties that are not available in the approximation of functions. For example, in [16], the idea that a parametric representation of a curve is not unique has been used to improve the order of approximation by polynomial curves of degree n from $n+1$ to $2n$. Also, the parametric form makes use of the geometric properties of the curve in design and modelling. In this paper, given the circular arc $c : t \mapsto (\cos(t), \sin(t)),\ -\theta \leq t \leq \theta$, where $\theta \in [-\pi, \pi]$, see Fig. 1, the geometric symmetries of the circle will be utilized to properly select the Bézier points in order to represent the quartic Bézier curve that has high order of approximation of 8 and possesses "the best" features.

A circle can be represented using rational Bézier curves and can be approximated by polynomial curves. Therefore, approximating a circular arc by polynomial curves with highest possible accuracy is a very important issue. It is needed for the construction of any CAD system. To approximate the circle c, there is a need to find a parametrically defined polynomial curve $p : t \mapsto (x(t), y(t)),\ 0 \leq t \leq 1$, where $x(t), y(t)$ are polynomials of degree 4, that approximates c with "minimum" error. Many researchers have tackled this issue using different norms and methods, see [2], [3], [4], [5], [6], [9], [10], [14], [16], [18]. For details and numerical comparisons with these works, see section 6. The proper function to measure the error between p and c is the Euclidean error function:

$$E(t) := \sqrt{x^2(t) + y^2(t)} - 1. \tag{1}$$

The square root limits the possibility of further progress. Thus, to avoid radicals, the squares of the components of the parametrization to the circle are used. So, the Euclidean error function $E(t)$ is replaced by the following error function

$$e(t) := x^2(t) + y^2(t) - 1. \tag{2}$$

This replacement makes sense because both $E(t)$ and $e(t)$ attain their roots and reach their extrema at the same parameters.

More precisely, **the approximation problem** in this paper is to find $p : t \mapsto (x(t), y(t)),\ 0 \leq t \leq 1$, where $x(t), y(t)$ are of degree 4, that approximates c and satisfies the following conditions:

1) p minimizes $\max_{t\in[0,1]} |e(t)|$,
2) $e(t)$ equioscillates 9 times over $[0, 1]$,

Note that condition (2) implies that p approximates c with order 8. We impose a priori these conditions, because they will be used to determine the values of the parameters that are used for geometric design of the circular arc.

The term approximation order is used in the context of Lagrange interpolation: p approximates c with order m if there exists parameters $t_1, t_2, \ldots, t_m$ in $[0, 1]$ satisfying $p(t_i) - c(t_i) = 0$, for all $i = 1, 2, \ldots, m$. This is a special case of the more general definition of order of approximation for the Hermite type including derivatives at the interpolated points.

We let the angle θ be as large as possible in order to approximate the largest circular arc with this specified error. Thereafter, the angle θ has to be scaled by a factor that also combined with a reduction in the uniform error, see the last conclusions and open problems' section.

This paper is organized as follows. Some preliminaries are given in section 2. The quartic Bézier curve of least deviation is presented and proved in section 3, and the properties are presented in section 4. All possible quartic Bézier curves of least deviation are presented in section 5. Conclusions are given in section 6.

II. Preliminaries

The notations $(x(t), y(t))$ and $\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ are used to represent parametric equations, and similarly points will be used in this article.

In this paper, the curve $p(t)$ is given in Bézier form, see Fig. 2 for possible Bézier points of quartic Bézier curve. The Bézier curve $p(t)$ of degree 4 is given by, see [11]

$$p(t) = \sum_{i=0}^{4} p_i B_i^4(t) =: \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}, \quad 0 \leq t \leq 1, \tag{3}$$

where p_0, p_1, p_2, p_3 and p_4 are the Bézier points, and $B_0^4(t) = (1-t)^4,\ B_1^4(t) = 4t(1-t)^3,\ B_2^4(t) = 6t^2(1-t)^2,\ B_3^4(t) = 4t^3(1-t)$ and $B_4^4(t) = t^4$ are the Bernstein polynomial basis of degree 4.

Since it is intended to represent the arc with a polynomial curve with minimum error, it is not important if the errors occur at the endpoints or anywhere else; it is important to maintain this disruption as low as possible there where the error occurs. In some other schemes, it is necessary that the approximating Bézier curve is G^k-continuous at the end points, see [20]. To represent a circular arc, the Bézier points are selected to take advantage of the symmetry properties of the circle. As the scheme in this paper is built on the idea of minimizing the error over all of the segment $[0, 1]$, therefore, the right choice for the beginning control point p_0 is as follows $p_0 = (-\alpha_0 \cos(\theta),\ -\beta_0 \sin(\theta))$, where values of α_0 and β_0 could but should not be the same. For symmetry reasons, the right choice for the end control point p_4 is as follows $p_4 = (-\alpha_0 \cos(\theta),\ \beta_0 \sin(\theta))$. Let $p_1 = (\gamma,\ -\zeta)$ then by symmetry $p_3 = (\gamma,\ \zeta)$. For symmetry reasons, the point p_2 must be on the x-axis, and thus it has the form $p_2 = (\xi,\ 0)$. Using the substitution $\alpha = \alpha_0 \cos(\theta),\ \beta = \beta_0 \sin(\theta)$, then the proper options for the Bézier points should be, see Fig. 2,

$$p_0 = \begin{pmatrix} -\alpha \\ -\beta \end{pmatrix}, \quad p_1 = \begin{pmatrix} \gamma \\ -\zeta \end{pmatrix}, \quad p_2 = \begin{pmatrix} \xi \\ 0 \end{pmatrix},$$

$$p_3 = \begin{pmatrix} \gamma \\ \zeta \end{pmatrix}, \quad p_4 = \begin{pmatrix} -\alpha \\ \beta \end{pmatrix}. \tag{4}$$

In order to have the Bézier curve p begin in the third quadrant, go counter clockwise through fourth and first quadrants and end in the second quadrant as the circular arc c, the following conditions should be fulfilled

$$\alpha, \beta, \gamma, \zeta > 0,\ \xi > 1. \tag{5}$$

The Bézier curve $p(t)$ in (3) with the Bézier points in (4) is arranged as follows

$$p(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}, \quad 0 \le t \le 1. \tag{6}$$

$$= \begin{pmatrix} -\alpha \left(B_0^4(t) + B_4^4(t)\right) + \gamma(B_1^4(t) + B_3^4(t)) + \xi B_2^4(t) \\ \beta \left(B_4^4(t) - B_0^4(t)\right) + \zeta \left(B_3^4(t) - B_1^4(t)\right) \end{pmatrix}.$$

There are 5 parameters $\alpha, \beta, \gamma, \zeta, \xi$ that will be used to have the polynomial curve p comply with the conditions of the approximation problem by substituting $x(t)$ and $y(t)$ into $e(t)$ and solving the resulting equation using a computer algebra system. Thereafter, it is shown that these values satisfy the approximation conditions; this is carried out in the following section.

III. The Quartic Bézier curve of least deviation

In the following theorem, the values of $\alpha,\ \beta,\ \gamma,\ \zeta,\ \xi$ that meet the terms of the approximation problem are given.

Theorem 1: The Bézier curve (6) with the Bézier points in (4), wherein

$$\begin{aligned} \alpha = \alpha^* &:= 0.91658426813952, \\ \beta = \beta^* &:= 0.40949454135449, \\ \gamma = \gamma^* &:= 0.00389865026306327, \\ \zeta = \zeta^* &:= 2.164585487675, \\ \xi = \xi^* &:= 2.9773929563972596 \end{aligned} \tag{7}$$

fulfils the following three conditions: p minimizes the infinity norm of th error function $\max_{t\in[0,1]} |e(t)|$ and approximates c with order 8, and the error function $e(t)$ equioscillates 9 times in $[0, 1]$. The error functions satisfy:

$$-\frac{1}{2^7} \le e(t) \le \frac{1}{2^7}, \quad -\frac{1}{2^7(2-\epsilon)} \le E(t) \le \frac{1}{2^7(2+\epsilon)}, \tag{8}$$

where $\epsilon = \max_{0\le t\le 1} |E(t)| \approx 2^{-8}, \forall t \in [0, 1]$.

Proof: Substituting $x(t)$ and $y(t)$ from equation (6) into the error function $e(t)$ in (2) and doing thereby some simplifications yields to the following formulation

$$\begin{aligned} e(t) &= t^8 \left(4\alpha^2 + 32\alpha\gamma + 64\gamma^2 - 24\alpha\xi - 96\gamma\xi + 36\xi^2\right) \\ &+ t^7 \left(-16\alpha^2 - 128\alpha\gamma - 256\gamma^2 + 96\alpha\xi + 384\gamma\xi - 144\xi^2\right) \\ &+ t^6 \left(40\alpha^2 + 16\beta^2 + 272\alpha\gamma + 448\gamma^2 - 192\alpha\xi - 624\gamma\xi + 216\xi^2 - 64\beta\zeta + 64\zeta^2\right) \\ &+ t^5 \left(-64\alpha^2 - 48\beta^2 - 368\alpha\gamma - 448\gamma^2 + 240\alpha\xi + 528\gamma\xi - 144\xi^2 + 192\beta\zeta - 192\zeta^2\right) \\ &+ t^4 \left(72\alpha^2 + 68\beta^2 + 320\alpha\gamma + 272\gamma^2 - 180\alpha\xi - 240\gamma\xi + 36\xi^2 - 240\beta\zeta + 208\zeta^2\right) \\ &+ t^3 \left(-56\alpha^2 - 56\beta^2 - 176\alpha\gamma - 96\gamma^2 + 72\alpha\xi + 48\gamma\xi + 160\beta\zeta - 96\zeta^2\right) \\ &+ t^2 \left(28\alpha^2 + 28\beta^2 + 56\alpha\gamma + 16\gamma^2 - 12\alpha\xi - 56\beta\zeta + 16\zeta^2\right) \\ &+ t \left(-8\alpha^2 - 8\beta^2 - 8\alpha\gamma + 8\beta\zeta\right) + (\alpha^2 + \beta^2 - 1). \end{aligned}$$

The last one is a polynomial of degree 8. The substitution of the values of $\alpha = \alpha^*,\ \beta = \beta^*,\ \gamma = \gamma^*,\ \zeta = \zeta^*$ and $\xi = \xi^*$ from (7)-(9) and doing some simplifications leads to

$$e(t) = 256\ t^8 - 1024\ t^7 + 1664\ t^6 - 1408\ t^5 + 660\ t^4 - 168\ t^3 + 21\ t^2 - t + \frac{1}{128}, \quad t \in [0, 1].$$

Making the substitution $t = \frac{u+1}{2}$ reduces the error function to the following form

$$e(u) = \frac{1}{128} - \frac{1}{4}\ u^2 + \frac{5}{4}\ u^4 - 2\ u^6 + u^8, \quad u \in [-1, 1].$$

We know that the last polynomial is the monic Chebyshev polynomial $\tilde{T}_8(u),\ u \in [-1, 1]$, which is the unique polynomial of degree 8 that equioscillates 9 times between $\pm\frac{1}{2^7}$ for all $u \in [-1, 1]$ and has the least deviation from the x-axis, see [23]. This shows that p satisfies the conditions of the approximation problem. Now it is time to show the error formula for $E(t)$. The error function $e(t)$ minimized is related to the Euclidean error $E(t)$ by the following formula

$$\begin{aligned} e(t) &= x^2(t)+y^2(t)-1 \\ &= (\sqrt{x^2(t)+y^2(t)}+1)\ (\sqrt{x^2(t)+y^2(t)}-1) \\ &= (2+E(t))\ E(t). \end{aligned}$$

Thus

$$E(t)=\frac{e(t)}{2+E(t)}.$$

Substituting the bounds of $e(t)$ gives

$$-\frac{1}{2^7(2-\epsilon)} \le E(t) \le \frac{1}{2^7(2+\epsilon)},$$
$$\text{where } \epsilon = \max_{0\le t\le 1}|E(t)| \approx 2^{-8}, \quad t\in[0,1].$$

This proves Theorem 1.

The circular arc and the approximating Bézier curve are graphed in Fig. 3. The difference between the curve and the approximation is not recognizable by the human eyes; Fig. 4 shows the corresponding error.

One would not, in general, expect a quartic polynomial to approximate almost $8/9$th the circle more accurately than this approximation. In the following section, the properties of the approximating Bézier curve are given.

IV. Properties of the quartic Bézier curve

In this section, some of the properties of the roots and the extrema of the error functions are verified. These properties characterise the approximating quartic Bézier curve. The first one is about the roots of the error functions $e(t)$ and $E(t)$ that are given in the following proposition.

Proposition I: The zeros of the error functions $e(t)$ and $E(t)$ are:

$$t_1=\frac{1}{2}(1+\cos(\frac{\pi}{16}))=0.9904, t_2=\frac{1}{2}(1+\cos(\frac{3\pi}{16}))=0.9157$$
$$t_3=\frac{1}{2}(1+\sin(\frac{3\pi}{16}))=0.7778, t_4=\frac{1}{2}(1+\sin(\frac{\pi}{16}))=0.5976,$$
$$t_5=\frac{1}{2}(1-\sin(\frac{\pi}{16}))=0.4025,$$
$$t_6=\frac{1}{2}(1-\sin(\frac{3\pi}{16}))=0.222215,$$
$$t_7=\frac{1}{2}(1-\cos(\frac{3\pi}{16}))=0.08427,$$
$$t_8=\frac{1}{2}(1-\cos(\frac{\pi}{16}))=0.009607.$$

These roots also satisfy

$$t_i+t_j = 1, \quad \text{for} \quad i+j=9.$$

Proof: The substitution of t_i in $e(t)$ gives $e(t_i)=0,\ i=1,\dots,8$. These are all zeros, since $e(t)$ is a polynomial of degree 8. The error function $E(t)$ has the same roots as $e(t)$ because $E(t)=0$ iff $\sqrt{x^2(t)+y^2(t)}=1$ iff $x^2(t)+y^2(t)=1$ iff $e(t)=0$.

The approximating quartic Bézier curve p in Theorem 1 and the circular arc c intersect at the points $p(t_i)=c(t_i),\ i=1,\dots,8$.

In the following proposition, the extreme values are given.
Proposition II: The extreme values of e(t) and $E(t)$ occur at

$$\tilde{t}_0=1,\ \tilde{t}_1=\frac{1}{2}(1+\cos(\frac{\pi}{8}))=0.9619,$$
$$\tilde{t}_2=\frac{1}{2}(1+\frac{1}{\sqrt{2}})=0.8536,\ \tilde{t}_3=\frac{1}{2}(1+\sin(\frac{\pi}{8}))=0.6913,$$
$$\tilde{t}_4=\frac{1}{2},\ \tilde{t}_5=\frac{1}{2}(1-\sin(\frac{\pi}{8}))=0.3087,$$
$$\tilde{t}_6=\frac{1}{2}(1-\frac{1}{\sqrt{2}})=0.1465,\ \tilde{t}_7=\frac{1}{2}(1-\cos(\frac{\pi}{8}))=0.0380602,$$
$$\tilde{t}_8=0.$$

These parameters satisfy the equality:

$$\tilde{t}_i+\tilde{t}_j=1, \quad \text{for} \quad i+j=8.$$

Proof: Differentiating $e(t)$ gives a polynomial of degree 7. Substituting $\tilde{t}_1,\dots,\tilde{t}_7$ gives $e'(\tilde{t}_i)=0,\ i=1,\dots,7$. Since $e'(t)$ is of degree 7 then these are all interior critical points. Checking at the end points adds $\tilde{t}_0=1,\ \tilde{t}_8=0$ to the critical points. Since for $t\in[0,1]$: $1-\frac{1}{128}\le x^2(t)+y^2(t)\le 1+\frac{1}{128}$, thus $\sqrt{x^2(t)+y^2(t)}\neq 0,\ \forall t\in[0,1]$. Differentiating $E(t)$ and equating to 0 gives $\frac{e'(t)}{\sqrt{x^2(t)+y^2(t)}}=0$ iff $e'(t)=0$. Thus $e(t)$ and $E(t)$ attain the extrema at the same values. This completes the proof of the proposition.

The difference in the values of $E(\tilde{t}_i)$ for odd and even i's is because $e(t)$ equioscillates between $\pm\frac{1}{128}$ and $\frac{1}{2^7(2-\epsilon)} \le E(t) \le \frac{1}{2^7(2+\epsilon)}$, where $\epsilon = max_{0\le t\le 1}|E(t)|$.
Proposition III: the values of $e(t)$ and $E(t)$ at $\tilde{t}_i$'s are given by:

$$e(\tilde{t}_{2i})=\frac{1}{128}, i=0,\dots,4, \qquad e(\tilde{t}_{2i+1})=\frac{-1}{128}, i=0,\dots,3.$$
$$E(\tilde{t}_{2i})=0.003899, i=0,\dots,4,$$
$$E(\tilde{t}_{2i+1})=-0.003914, i=0,\dots,3.$$

Therefore,

$$\frac{-1}{128}\le e(t)\le\frac{1}{128}=2(0.003906),\ t\in[0,1],$$
$$-0.00391391\le E(t)\le 0.003899,\ t\in[0,1].$$

Proof: Direct substitution in the error functions leads to the equalities. The details of the proof of the proposition are left to the reader.
As a consequence of Theorem 1, we have the following proposition regarding the error at any $t\in[0,1]$.
Proposition IV: For every $t\in[0,1]$, the errors of approximating the circular arc using the quartic Bézier curves in Theorem 1 are given by:

$$e(t)=256t^8-1024t^7+1664t^6-1408t^5+660t^4-168t^3+21t^2-$$
$$t+\frac{1}{128},\ \forall t\in[0,1].$$

Proof: Direct consequence of Theorem 1. The details of the proof of the proposition are left to the reader.

Using the relation between $E(t)$ and $e(t)$, we get:

$$E(t) \tilde{=} 128t^8 - 512t^7 + 832t^6 - 704t^5 + 330t^4 - 84t^3 + \frac{21}{2}t^2 - \frac{1}{2}t + \frac{1}{256}, \quad \forall t \in [0,1].$$

To get the solution in Theorem 1, some conditions were imposed on $\alpha,\ \beta,\ \gamma,\ \zeta$ and ξ in (5). These conditions give the Bézier curve with $\alpha = \alpha^*, \quad \beta = \beta^*, \quad \gamma = \gamma^*, \quad \zeta = \zeta^*$ and $\xi = \xi^*$ that represents the circular arc from third to second quadrants passing through the fourth and first quadrants generated counter clockwise, see Fig. 3. However, if these conditions on $\alpha,\ \beta,\ \gamma,\ \zeta$ and ξ are removed, there will be other possible solutions. These are given in the following section.

V. All possible quartic Bézier curves

The following theorem lists all the possible Bézier curves.

Theorem 2: By removing the conditions on $\alpha, \beta, \gamma, \zeta$ and ξ in (5) and reinvestigating the approximation problem, then the problem has 32 solutions; 24 of these solutions are complex and the other 8 solutions are real; 4 real solutions are not admissible because they have the opposite direction for the tangent; the other 4 real solutions are geometrically feasible and satisfy the conditions of the approximation problem. These solutions are sign multiple of the solution in Theorem 1 and are summarized in table 1.

Proof: The first solution has been confirmed in Theorem 1. To confirm the other 3 cases, we consider the error function $e(t)$ and do some simplifications and substitutions as in Theorem 1 to get the monic Chebyshev polynomial of degree 8. This polynomial possesses the properties of the approximation problem. The details of the proof are left to the reader.

Remarks:

1) All of the solutions in Table 1 are related to each other. The second solution coincides with the first solution, but generated clockwise. The third and fourth solutions are reflections of the first solution around the y-axis, generated counter clockwise and clockwise, respectively.
2) The sign of α is the same as the signs of γ and ξ. If the sign of α is positive then the curve begins (ends) in the second quadrant through the first and fourth quadrants and ends (begins) in the third quadrant, and if it is negative then the curve begins (ends) in the first quadrant through the second and third quadrants and ends (begins) in the fourth quadrant.
3) The sign of β is the same as the sign of ζ.

The roots and extreme values of $e(t)$ and $E(t)$ for all the solutions in Table 1 are given in the following proposition.

Proposition V: The solutions in Table 1 have the following properties:

1) The roots of the error functions $e(t)$ and $E(t)$ for all of the solutions in Table 1 are the same as in Proposition I.
2) The extreme values of $e(t)$ and $E(t)$ for all of the solutions in Table 1 occur at the same parameters that are given in Proposition II.
3) The extreme values of $e(t)$ and $E(t)$ for all of the solutions in Table 1 have the same values that are given in Proposition III.
4) The error functions $e(t)$ and $E(t),\ t \in [0,1]$ for all of the solutions in Table 1 are given by the formulas in Proposition IV.

Proof: The proofs are similar to the proofs of the similar previous cases and are left to the reader.

VI. Conclusions and Open Problems

In this article, the best uniform approximation of circular arcs with parametrically defined polynomial curves of degree 4 are explicitly given. The error function equioscillates 9 times; the approximation order is 8. Numerical examples are given to demonstrate the efficiency and simplicity of the approximation method.

The method in this paper is C^0-continuous by construction. There are methods in the literature that are G^1- and G^2-continuous, see for example [6], [9], [10], [13], [14], [16], [17], [18], [19], [22].

As future works, it is interesting to:

1) study quartic approximation with G^k-continuity, $k = 1, 2$, using equioscillating error functions and constrained Chebyshev polynomials.
2) find a way to write the Bézier points in terms of the angle θ. It would be very important to have the best approximation available for all θ perhaps by employing a semi-numerical method.
3) Apply these results in this paper to perform degree reduction of Bézier curves to get the best approximation with the minimum uniform error.
4) It would be interesting to compare our curve with the quartic exponential Euler spline defined by Schoenberg and studied by de Boor, see [7], [8], [24], [25].

Acknowledgement: The author would like to thank the referee for invaluable comments that lead to improve the paper.

References

[1] Y. J. Ahn and C. Hoffmann, Circle approximation using LN Bézier curves of even degree and its application, J. Math. Anal. Appl. (2013).

[2] Y. J. Ahn and H. O. Kim, Approximation of circular arcs by Bézier curves, Journal of Computational and Applied Mathematics, V. 81(1) (1997), 145-163.

[3] Y. J. Ahn, Y. S. Kim, and Y. Shin, Approximation of circular arcs and offset curves by Bézier curves of high degree, Journal of Computational and Applied Mathematics, V 167(2) (2004), 405-416.

[4] P. Bézier, The mathematical basis of the UNISURF CAD system, Butterworth-Heinemann Newton, MA, USA, ISBN 0-408-22175-5, (1986).

[5] J. Blinn, How many ways can you draw a circle?, Computer Graphics and Applications, IEEE 7(8) (1987), 39-44.

[6] C. de Boor, K. Höllig and M. Sabin, High accuracy geometric Hermite interpolation, Comput. Aided Geom. Design 4 (1988), 269-278.

[7] C. de Boor, On the cardinal spline interpolant to exp(iut). SIAM J. Math. Anal. 7, No 6 (1976), 930-941.

[8] C. de Boor (ed), Selected works of I.J. Schoenberg, Birkhäuser-Verlag, Basel, 1988.

[9] T. Dokken, M. Dæhlen, T. Lyche, and K. Mørken, Good approximation of circles by curvature-continuous Bézier curves, Comput. Aided Geom. Design 7 (1990), 33-41.

[10] M. Goldapp, Approximation of circular arcs by cubic polynomials, Comput. Aided Geom. Design 8 (1991), 227-238.

[11] K. Höllig and J. Hörner (2013). Approximation and Modeling with B-Splines. SIAM. Titles in Applied Mathematics 132.

[12] Z. Habib and M. Sakai, Fairing an arc spline and designing with G^2 PH quintic spiral transitions, International Journal of Computer Mathematics 90(5) (2013), 1023-1039.

[13] S. H. Kim and Y. J. Ahn, An approximation of circular arcs by quartic Bézier curves, Computer-Aided Design, V 39(6) (2007), 490-493, .

[14] S. W. Kim and Y. J. Ahn, Circle approximation by quartic G^2 spline using alternation of error function, J. KSIAM , V 17(5) (2013), 171-179.

[15] J. McCoy, Helices for mathematical modelling of proteins, nucleic acids and polymers, J. Math. Anal. Appl. 347, (2008) 255-265.

[16] A. Rababah, Approximation von Kurven mit Polynomen und Splines, Ph. D Thesis, Stuttgart Universität, Germany, 1992.

[17] A. Rababah and Y. Hamza, Multi-degree reduction of disk Bézier curves with G^0-and G^1-continuity, Journal of Inequalities and Applications 2015(1) (2015), 1-12.

[18] A. Rababah, The best uniform quadratic approximation of circular arcs with high accuracy, Open Mathematics 14 Iss. 1 (2016), 118-127.

[19] A. Rababah, Distances with rational triangular Bézier surfaces, Applied mathematics and computation 160 Iss. 2, (2005), 379-386.

[20] A. Rababah and S. Ibrahim, Weighted G^1-Multi-Degree Reduction of Bézier Curves, International Journal of Advanced Computer Science and Applications 7(2), (2016), 540-545. http://dx.doi.org/10.14569/IJACSA.2016.070270

[21] A. Rababah, L-2 degree reduction of triangular Bźier surfaces with common tangent planes at vertices, International Journal of Computational Geometry & Applications 15 (05), (2005), 477-490.

[22] A. Rababah, B.G. Lee, and J. Yoo, Multiple degree reduction and elevation of Bézier curves using Jacobi–Bernstein basis transformations, Numerical Functional Analysis and Optimization 28 (2007), 1179-1196.

[23] J. Rice, The approximation of functions, Vol. 1: linear theory. Addison-Wesley, (1964).

[24] I. J. Schoenberg, Cardinal spline interpolation and spline functions IV. The exponential Euler splines. In Linear operators and approximation. ISNM Vol. 20, Birkhäuser-Verlag, Basel (1972), 382-404.

[25] I.J. Schoenberg, Cardinal spline interpolation, SIAM, Philadelphia, 1973.

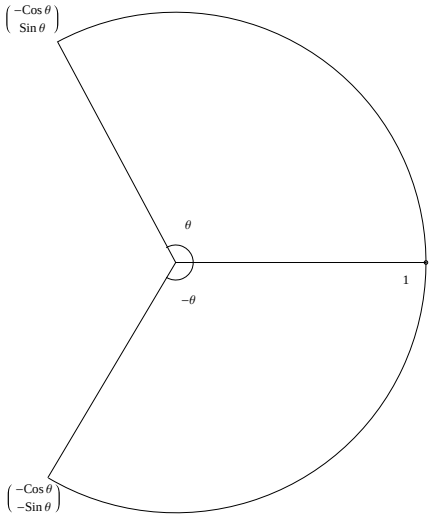


Fig. 1: A circular arc

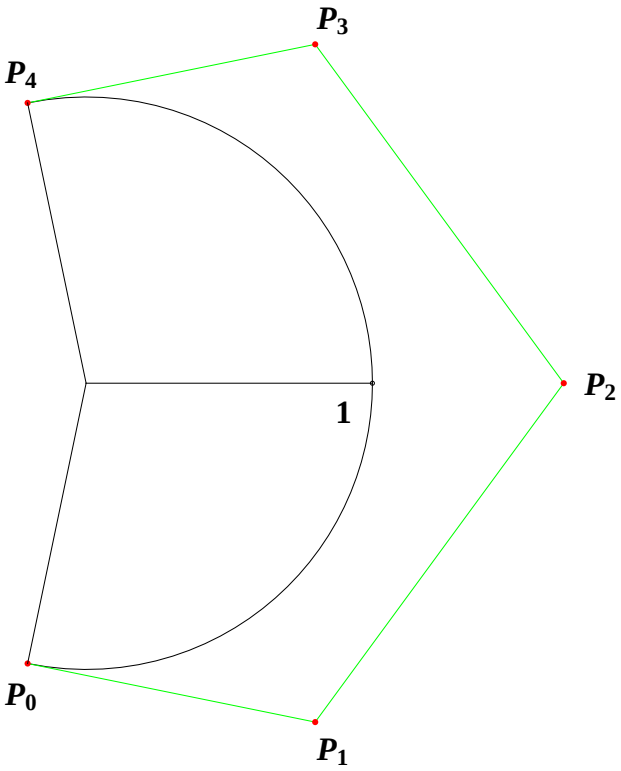


Fig. 2: Possible Bézier points of circular arc.

Solution	Sign α	Sign β	Sign γ	Sign ζ	Sign ξ	from to quadrants	generated
1st	+	+	+	+	+	3rd to 2nd	counter clockwise
2nd	+	−	+	−	+	2nd to 3rd	clockwise
3rd	−	−	−	−	−	1st to 4th	counter clockwise
4th	−	+	−	+	−	4th to 1st	clockwise

Table 1: All geometrically feasible real solutions to the approximation problem.

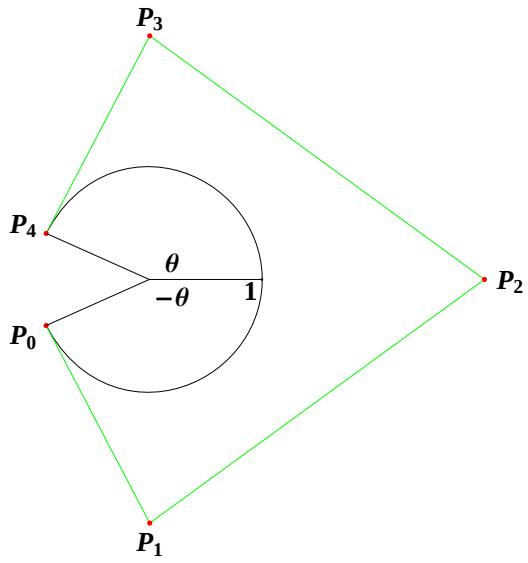


Fig. 3: Circular arc and it's quartic Bézier curve in Theorem 1.

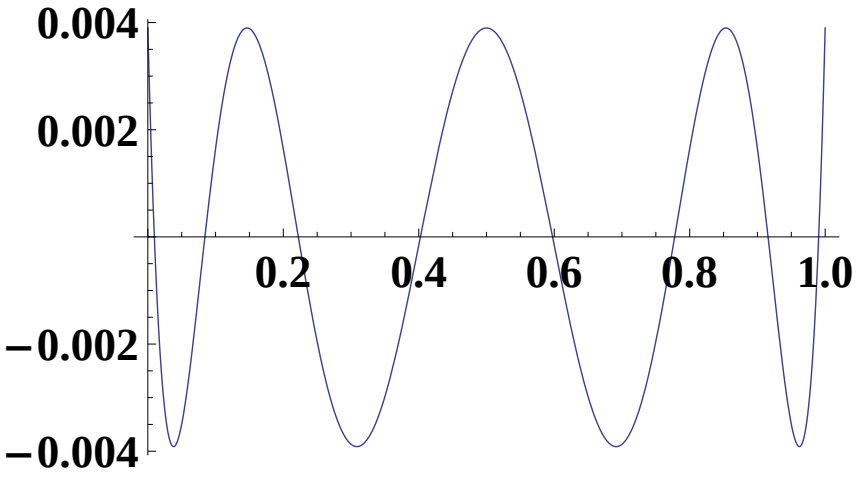


Fig. 4: Euclidean Error of the quartic Bézier curve in Theorem 1.

A Robust MAI Constrained Adaptive Algorithm for Decision Feedback Equalizer for MIMO Communication Systems

Khalid Mahmood[1], Syed Muhammad Asad[2], Muhammad Moinuddin[3], Waqas Imtiaz[4]
[1,4]Iqra National Universty Peshawar, Pakistan
[2]University of Hafr Al Batin, Kingdom of Saudi Arabia
[3]Center of Excellence in Intelligent Engineering Systems (CEIES) King Abdul Aziz University, Jeddah Kingdom of Saudi Arabia

***Abstract*—Decision feedback equalizer uses prior sensor's decisions to mitigate damaging effects of intersymbol interference on the received symbols. Due to its inherent non linear nature, decision feedback equalizer outperforms the linear equalizer in case where the intersymbol interference is sever in a communication system. Equalization of multiple input multiple output fast fading channels is a daunting job as these equalizers should not only mitigate the intersymbol interference but also interstream interference. Various equalization methods have been suggested in the adaptive filtering literature for multiple input multiple output systems. In our paper, we have developed a novel algorithm for multiple input multiple output communication systems centered around constrained optimization technique. It is attained by reducing the mean squared error criteria with respect to known variance statistics of multiple access interference and white Gaussian noise. Novelty of our paper is that such a constrained method has not been used for scheme of multiple input multiple output decision feedback equalizer resulting in a constrained algorithm. Performance of the proposed algorithm is compared to the least mean squared as well as normalized least mean squared algorithms. Simulation results demonstrate that proposed algorithm outclasses competing multiple input multiple output decision feedback equalizer algorithms.**

***Index Terms*—Decision feedback equalizer, inter symbol interference, multiple access interference, Rayleigh fading, AWGN, adaptive algorithm**

I. Introduction

Multiple input multiple output (MIMO) is a diversifying technique for mobile communication systems which uses multiple antennas at the transmitting side and receiving side for an efficient channel gain capacity. Antennas installed at both side of a communication system are utilized in such a way as to reduce the errors and boost data speed resulting in an effective communication for the users. Due to its inherent performance drawbacks, single input, single output (SISO) technology, has given way to the MIMO technology. In SISO, an antenna is utilized at transmitter as well as at receiver that results in multi path phenomina. When electromagnetic field (EM) waves are blocked by barricades, mountains, tall buildings, poles etc, these waves are dispersed, and as such use numerous ways to reach the desired destination. The late arrival of disseminated waveforms (signals) results in problems such as fading, cut outs, or sporadic reception. Another drawback of SISO is observed in mobile Internet, in which it has not only reduced the data rate as but also enhanced error propagation. Utilization of MIMO technology where multiple antennas are used at both end of the channel has mitigated the multipath wave propagation problem.

Equalization of MIMO fast fading channels is a daunting task as these equalizers should not only mitigate the intersymbol interference (ISI) but also interstream interference (ISI) and this led to the introduction of decesion feedback equlaization (DFE) [1]. A DFE uses prior sensor's decisions to mitigate damaging effects of ISI on the received symbols. Because of its nonlinear behviour, DFE outperforms linear equalizer LE) in systems where ISI is sever [2].Various equalization methods have been proposed for MIMO systems [3]–[10]. In our paper, a novel algorithm for MIMO communication systems centered on constrained optimization technique is developed. It is attained by reducing the mean squared error criteria with respect to known variance statistics of the multiple access interference (MAI) and white Gaussian noise (AWGN) [11]. It is attained by reducing the mean squared error (MSE) criteria with respect to known variance statistics of MAI and AWGN.

In multiuser environment, MAI is a restrictive aspect,so a detection receiving architecture needs to be implemented which would negate the MAI and AWGN. In adaptive filtering literature, MAI is considered as part of the interfering noise, but in reality, MAI is an amorphous AWGN. In [12] MAI is detached from AWGN, but practically, it is not a valid assumption. By using the differences in the known statistical contents of the MAI and AWGN and utilizing the combination as constraint may be used to construct an algorithm that would outperform noise only constrained algorithms.

Key contributions of our paper are:

1) An MAI and AWGN constrained MMSE criterion is utilized to develop the MIMO DFE for CDMA systems and it is obtained by reducing MMSE cost function with respect to known statistics of MAI and AWGN.
2) An adaptive DFE algorithm is established which is adaptive step size. Its updating is based on variance of MAI and AWGN.
3) Statistics (variance) of MAI and AWGN is developed in a fading channel and AWGN environment.

This paper is organized as:

After introductory section, system model is explained in section II. In section III, variance of MAI and AWGN is discussed. Proposed MIMO DFE is provided in section IV. Performance of the proposed algorithm is judged by comparing it with the other algorithms in simulation result section. Finally concluding remarks are provided in section VI.

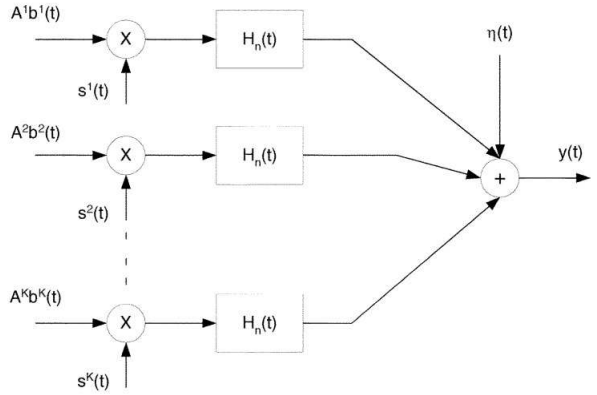


Fig. 1: Transmitter of a downlink CDMA system

II. System Model

We are using CDMA transmitter of a wireless radio setup with M transmiters and N receivers as given in Fig. 1. We are using the Rayleigh channel having impulse response between the mth transmitting antenna and nth receiving antenna of the kth signal given by

$$H_{mn}^{k}(t) = h_{mn}^{k} e^{j\theta_m} \delta(t) \quad \textbf{(1)}$$

where h_{mn}^{k} (($h_{mn}^{k} = 1$ for AWGN) channel) is the impulse response and θ_m is phase of Rayleigh channel for kth signal. The detector in mth receiver sees the signal $r_m(t)$as given below

$$\begin{aligned} \varpi_n(t) &= \sum_{n=1}^{N}\sum_{k=-\infty}^{\infty}\sum_{i=1}^{i} C^i d_m^{k,i} \xi_m^{k,i}(t) h_{mn}^{k} \\ &\quad + \varsigma_n(t), \qquad n = 1,2,3\ldots N \end{aligned} \quad (2)$$

where i is the number of subscribers, $\xi_n^{l,i}(t)$ is signature signal carrying random signature sequence of the ith subscriber given as $(k-1)T_\beta \le t \le lT_{\mathbb{C}}$. Where T_β is defined to be bit period and $T_{\mathbb{C}}$ is chip interval. Both are associated as $S_{\mathbb{C}} = T_\beta / T_{\mathbb{C}}$. $d_m^{k,i}$ is the input of the ith subscriber, C^i is the amplitude (transmitted) of the ith subscriber and $\varsigma_m(t)$ is an AWGN with zero mean and variance $\sigma_{\varsigma_m}^2$ at the nth receiving antenna. Cross correlation of signature sequences for j and l subscriber of kth signal is $\rho_k^{l,j} = \int_{(l-1)T_b}^{lT_b} \xi_m^l(t)\,\xi_m^j(t)\,dt = \sum_{i=1}^{N_c} \kappa_{k,i}^l \kappa_{l,i}^j$, where $\left\{\kappa_{k,i}^l\right\}$ is presumed to be normalized spreading sequence of subscriber l for the kth signal. A reason for assuming the spreading sequence to be normalized is to keep the auto correlation of the signature sequence unity.

The receiving side comprises of a bunch of matched filters, matched to signature signal of intended subscriber. We are using subscriber 1 is the intended subscriber. Output of the matched filter of the ith signal at nth receiver may be :

$$\begin{aligned} x_m^i &= \int_{(k-1)T_b}^{lT_b} \varpi_n(t)\,\xi_m^{k,i}(t) \\ &= \sum_{n=1}^{N} A^1 b_m^{k,1} h_{mn}^{k} + \sum_{m=1}^{M}\sum_{l=2}^{L} C^i b_m^{k,i} \rho_n^{i,1}(t) h_{mn}^{k} \\ &\quad + v_n, \qquad n = 1,2,3\ldots N \end{aligned} \quad (3)$$

In this equation, 2nd componnet is MAI and is given by

$$\begin{aligned} y_n^k &= \sum_{m=1}^{M}\sum_{i=2}^{i} C^i b_m^{k,i} \rho_m^{i,1}(t) h_{mn}^{k}, \qquad n \\ &= 1,2,3\ldots N \end{aligned} \quad (4)$$

By utilizing Cauchy-Schwartz inequality, equation (4) may be set up like

$$\begin{aligned} y_n^k &\le \sum_{m=1}^{M} C^i b_m^{k,i} \rho_m^{i,1}(t) \sum_{m=1}^{M} h_{mn}^{k} \\ &\le V_n^k \sum_{i=1}^{M} h_{mn}^{k}, \qquad n = 1,2,3\ldots N \end{aligned} \quad (5)$$

where $V_n^k = \sum_{m=1}^{M}\sum_{i=2}^{L} A^l b_m^{k,l} \rho_m^{l,k}$.

III. Variance of Multiple Access Interference (MAI) in Multiple Input Multiple Output (MIMO System)

Cross-correlation $\rho^{i,1}$ is in the range $[-1,1]$ and is shown to be [13]:

$$\begin{aligned} \rho^{i,1} &= \frac{(N_c - 2g)}{N_c}, \qquad g \\ &= 0,1,3\ldots N_c \end{aligned} \quad (6)$$

where g is a binomial random variable carrying equal probability of failure and success. It has a mean value of $E[g] = N_c/2$ and variance is $\sigma_g^2 = N_c/4$.

As channel taps are assumed to be independent of spreading sequence as well as the data sequence, interferer's components $\sum_{m=1}^{M}\sum_{i=2}^{l} A^l b_m^{k,l} \rho_m^{l,k}$, are also not dependent on each others. When all subscribers have equal received power, variance of V_n^l can be written as

$$\begin{aligned} \sigma_{V_n^k}^2 &= \sum_{m=1}^{M}\sum_{i=2}^{i} C^2 E\left[\left(d_m^{k,i}\rho_m^{i,1}\right)^2\right], \\ &= \sum_{m=1}^{M}\sum_{i=2}^{i} C^2 E\left[\left(1 - \frac{2}{N_c} g\right)^2\right], \\ &= A^2 \sum_{m=1}^{M}\sum_{i=2}^{i}\left(1 - \frac{2}{N_c}E[g] + \frac{2}{N_c^2}E\left[g^2\right]\right), \\ \sigma_{V_n^k}^2 &= \frac{C^2 N(i-1)}{N_c}, \qquad n = 1,2,3\ldots N \end{aligned} \quad (7)$$

Equation (7) is utilized in equation (5) to calculate MAI variance which may be written as

$$\begin{aligned} \sigma_{V_n^k}^2 &\le \frac{C^2 N(i-1)}{N_c} \sum_{m=1}^{M} E\left[\left(h_{mn}^{k}\right)^2\right], \qquad n \\ &= 1,2,3\ldots N \end{aligned} \quad (8)$$

From equation (8) it is evident that impairment is sever in the MIMO system compared to the SISO system and that is due to the fact that MAI variance is dependent on the MIMO system complexity.

IV. MIMO DFE Based on Constrained Optimization Technique

ISI and MAI involved in the system shown in equation (3) can be diminished by using a MIMO DFE [14]. Our proposed equalization technique comprises of M multiple input single output (MISO) DFE in parallel (Fig. 2). The mth multiple input single output (MISO) DFE which comprised of a feedforward (FFF) and a feedback filter (FBF) having L and R taps, respectively is assigned to extract the mth stream. If the sampling window is considered to be K symbols then output of N receivers is

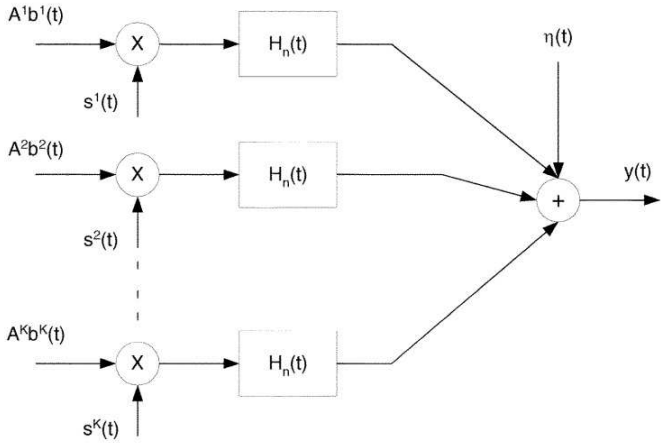


Fig. 2: MIMO DFE model

$$\mathbf{S}(k) = [s_1(k)\, s_2(k) \ldots s_N(k)]^T,$$

and

$$\mathbf{S}^k = \left[\mathbf{S}^T(k)\, \mathbf{S}^T(k-1) \ldots \mathbf{S}^T(k-K+1)\right]^T,$$

where $\mathbf{S}^k$ is an $ML \times 1$ vector. If $\mathbf{f}_n(l)$, for $m = 1,2,3 \ldots M$ represents FFF and $\mathbf{p}_n(k)$, for $n = 1,2,3 \ldots M$ represents FBF of length R then nth DFE output may be set below

$$\hat{u}_m^k = \mathbf{w}_m^H(k)\, \mathbf{o}^l, \qquad m = 1,2,3 \ldots M \tag{9}$$

where

$$\mathbf{w}_n(k) = \left[\mathbf{h}_n^T(k) \quad \mathbf{a}_n^T(k)\right]^T,$$
$$\mathbf{u}^k = \left[\mathbf{p}^{kT} \quad \check{\mathbf{b}}^{kT}\right]^T$$

and $\check{\mathbf{b}}^l$ is $MR \times 1$ vector and is a decision device's output

$$\check{\mathbf{b}}^k = \left[\check{\mathbf{b}}^{kT}(l)\, \check{\mathbf{b}}^{kT}(k-1) \ldots \check{\mathbf{b}}^{kT}(k-R+1)\right]^T,$$

and

$$\check{\mathbf{b}}^l(k) = \left[\check{b}_1(k)\, \check{b}_2(k) \ldots \check{b}_M(k)\right]^T.$$

FFF and FBF in equation (9) are updated by utilizing MAI and AWGN constrained LMS algorithm developed by [11] as

$$\mathbf{h}_m(k+1) = \mathbf{h}_m(k) + \alpha_m(k)\, e_m(k)\, \mathbf{y}^k \tag{10}$$
$$\mathbf{b}_m(k+1) = \mathbf{b}_m(k) + \alpha_m(k)\, e_m(k)\, \check{\mathbf{b}}^k \tag{11}$$

where $e_k(k) = \check{\mathbf{b}}^k - \hat{u}_m^k$ for $m = 1,2,3 \ldots M$ and $\lambda(k)$ is an adaptive learning parameter of an adaptive algorithm for MSE cost and is

$$\delta(k) = \alpha(1 + \gamma\delta(k)) \tag{12}$$
$$\lambda(k+1) = \lambda(k) + \mu\left[\frac{1}{2}\left(e^2(k) - \sigma_\alpha^2\right) - \lambda(k)\right] \tag{13}$$

where δ and μ are positive step sizes. It is evident in equation (12) that the LMS algorithm converges when $\mu = 0$. The proposed algorithm depends on the variance of MAI and AWGN shown as

$$\sigma_\alpha^2 = \sigma_{V_n^k}^2 + \sigma_{v_n^k}^2 \tag{14}$$

where $\sigma_{V_n^k}^2$is the MAI variance and$\sigma_{v_n^k}^2$ is the variance of AWGN defined earlier.

V. Discussion on Results

Simulation results are given here to judge performance of our algorithm by comparing it to LMS and normalized LMS algorithms. We have chosen a 2×2 MIMO system. Signal to noise ratio (SNR) is chosen to be twenty dB. A MIMO DFE with FFF having length five and FBF having length of two respectively are also used. Two channel settings are used while performing simulations for 4 subscribers with equal transmitted powers.

- AWGN channel
- Rayleigh channel

A. AWGN Channel Setting

In AWGN channel environment, our proposed algorithm performance is evaluated viz a viz the LMS and NCLMS algorithms using phase reversal keying and quadrature phase shift keying modulation schemes. Fig. 3 shows the performance evaluation using the phase reversal keying. As evident, our proposed MNCLMS converges faster than rest of the algorithms. It stablizes at an MSE of -18 dB in 1500 iterations. There is a bit degradation in the convergence but still has outperformed the other algorithms. A similar pattern is seen by using quadrature phase shift keying modulation scheme in Fig. 4 where there is a slight deterioration in the convergence rate but still far better then LMS and NLMS algorithms. Our proposed algorithm is converging at -18 dB in 1800 iterations

B. Rayleigh channel Setting

We are using a Rayleigh channel setting with Doppler frequency of $f_d = 250$Hz using phase reversal keying and quadrature phase shift keying modulation schemes. As seen in Fig. 5, the proposed algorithm converged faster then rest of the competing algorithms. MNCLMS algorithm achieves MSE at -19 dB in 2000 iterations. Fig. 6 depicts performance comparison of the algorithm with LMS and NCLMS algorithm for

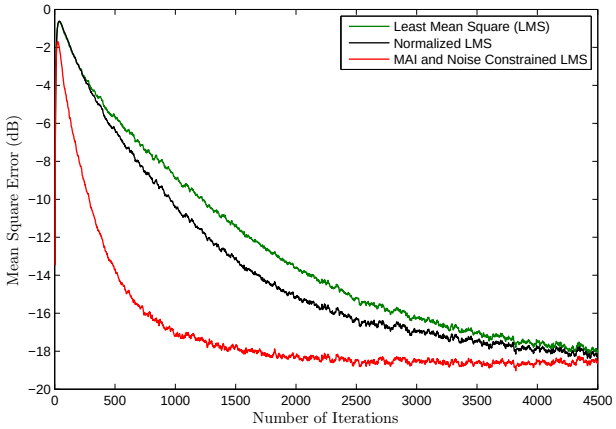


Fig. 3: Mean squared error performance comparison in an AWGN setting with 20 dB SNR and phase reversal keying modulation scheme

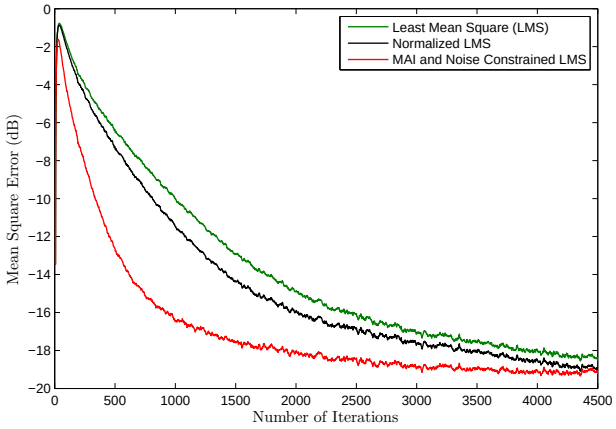


Fig. 4: Mean squared error performance comparison in AWGN setting with 20 dB SNR and quadrature phase shift keying modulation scheme

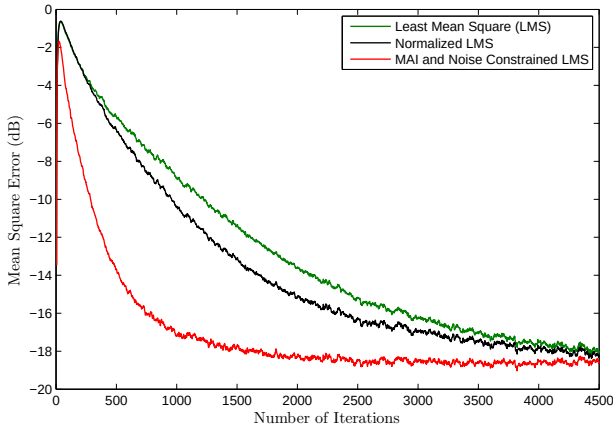


Fig. 5: Mean squared error performance comparison in Rayleigh channel environment with $f_d = 250$Hz and 20 dB SNR using phase reversal keying modulation scheme

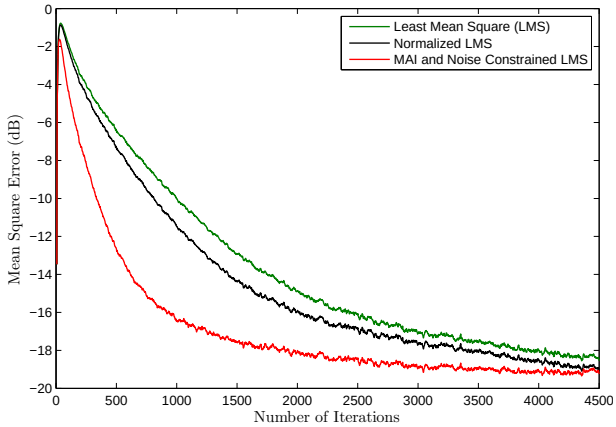


Fig. 6: Mean squared error performance comparison in Rayleigh channel setting with $f_d = 250$Hz and 20 dB SNR using quadrature phase shift keying modulation

the Rayleigh channel environment using quadrature phase shift keying modulation scheme. Our algorithm has outperformed the other competing algorithms.

VI. Conclusion

In this paper, an MAI and AWGN variance constrained algorithm for a MIMO CDMA DFE is developed. Performance of the the proposed algorithm is compared to the least mean squared as well as normalized least mean squared algorithms. Simulation results demonstrate that our algorithm has outclassed the competing algorithms.

References

[1] C. Belfiore and J. Park Jr, "Decision feedback equalization," *Proceedings of the IEEE*, vol. 67, no. 8, pp. 1143–1156, 1979.

[2] J. Proakis, "Digital communications (ise)," 2001.

[3] G. Latouche, D. Pirez, and P. Vila, "MMSE cyclic equalization," *Military Communications Conference, 1998. MILCOM 98. Proceedings., IEEE*, vol. 1, pp. 150–154, 2002.

[4] R. de Lamare and R. Sampaio-Neto, "Adaptive mber decision feedback multiuser receivers in frequency selective fading channels," *Communications Letters, IEEE*, vol. 7, no. 2, pp. 73–75, 2003.

[5] D. Palomar, J. Cioffi, and M. Lagunas, "Joint tx-rx beamforming design for multicarrier mimo channels: A unified framework for convex optimization," *Signal Processing, IEEE Transactions on*, vol. 51, no. 9, pp. 2381–2401, 2003.

[6] S. Chen, L. Hanzo, and A. Livingstone, "Mber space-time decision feedback equalization assisted multiuser detection for multiple antenna aided sdma systems," *Signal Processing, IEEE Transactions on*, vol. 54, no. 8, pp. 3090–3098, 2006.

[7] Y. Lee and W. R. Wu, "Adaptive channel aided decision feedback equalisation for SISO and MIMO systems," *Communications, IEE Proceedings-*, vol. 153, no. 5, pp. 657–663, 2006.

[8] V. Kekatos, K. Berberidis, and A. Rontogiannis, "A block adaptive frequency domain mimo dfe for wideband channels," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 3. IEEE, pp. III–197.

[9] M. Shenouda and T. Davidson, "A design framework for limited feedback MIMO systems with zero-forcing DFE," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1578–1587, Oct. 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4641967

[10] A. S. Lalos, V. Kekatos, and K. Berberidis, "Adaptive conjugate gradient DFEs for wideband MIMO systems," *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2406–2412, 2009.

[11] M. Moinuddin, A. Zerguine, and A. U. H. Sheikh, "Multiple-Access Interference Plus Noise-Constrained Least Mean

Square (MNCLMS) Algorithm for CDMA Systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 9, pp. 2870–2883, Oct. 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4490295

[12] W. Hamouda and P. Mclane, "A fast adaptive algorithm for mmse receivers in ds-cdma systems," *Signal Processing Letters, IEEE*, vol. 11, no. 2, pp. 86–89, 2004.

[13] M. Moinuddin, a. U. H. Sheikh, A. Zerguine, and M. Deriche, "A Unified Approach to BER Analysis of Synchronous Downlink CDMA Systems with Random Signature Sequences in Fading Channels with Known Channel Phase," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–13, 2008. [Online]. Available: http://www.hindawi.com/journals/asp/2008/346465.html

[14] A. Maleki-Tehrani, B. Hassibi, and J. Cioffi, "Adaptive equalization of multiple-input multiple-output (mimo) channels," in *Communications, 2000. ICC 2000. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1670–1674.

TGRP: A New Hybrid Grid-based Routing Approach for Manets

Hussein Al-Maqbali
Department of Information Technology
Ministry of Transport and Communications
Muscat, Oman

*Khaled Day
Department of Computer Science
Sultan Qaboos University
Muscat, Oman

Mohamed Ould-Khaoua
Department of Electrical and Computer Engineering
Sultan Qaboos University
Muscat, Oman

Abderezak Touzene
Department of Computer Science
Sultan Qaboos University
Muscat, Oman

Nasser Alzeidi
Department of Computer Science
Sultan Qaboos University
Muscat, Oman

***Abstract*—Most existing grid-based routing protocols use reactive mechanisms to build routing paths. In this paper, we propose a new hybrid approach for grid-based routing in MANETs which uses a combination of reactive and proactive mechanisms. The proposed routing approach uses shortest-path trees to build the routing paths between source and destination nodes. We design a new protocol based on this approach called the Tree-based Grid Routing Protocol (TGRP). The main advantage of the new approach is the high routing path stability due to availability of readily constructed alternative paths. Our simulation results show that the stability of the TGRP paths results in a substantially higher performance compared to other protocols in terms of lower end-to-end delay, higher delivery ratio and reduced control overhead.**

Keywords*—MANETs; routing protocols; NS2 simulation; performance evaluation

I. INTRODUCTION

A Mobile Ad-hoc Network (MANET) is defined as a collection of autonomous mobile nodes which communicate in the absence of access points. A node in a mobile ad-hoc network works as a host and as a router to serve multi-hop wireless communication and usually has limited power resources. Routing in a mobile ad hoc network is a challenging task since the network's topology changes frequently due to mobility. A node sends control packets to discover destinations and to establish and maintain routes. Since the channel bandwidth and power are often limited, route establishment should be done with minimum control packets and minimum usage of bandwidth and energy.

Many routing protocols have been proposed for MANETs [1] such as topology-based routing protocols (e.g. DSDV [2], AODV [3] and DSR [4] [5]), position-based routing protocols (e.g. Compass [6] and Greedy [7]) and grid-based routing protocol (e.g. GRID [8], EC-GRID [9]). The topology-based MANET routing protocols suffer from low scalability because of the high number of overhead messages and high network latency, especially with high node mobility. The availability of cheap instruments for estimating the position of nodes in a network, like Global Positioning System (GPS) receivers, has motivated many researchers to develop position-based routing protocols for MANETs [10] [11] [12]. Position-based routing protocols can eliminate the need to maintain routes. They use the knowledge of the nodes locations to route packets. Position-based protocols assume that any node is aware of its position, the position of its neighbors as well as the position of the destination. A node can discover its position using a location mechanism such as GPS [13]. It can discover its neighbors' locations by using periodic messages. The nodes use location services to discover destination nodes locations [14].

In position-based routing protocols, each node has an identifier (id) and a current geographic position. Typically in grid-based routing protocols, the physical area is divided into a logical two-dimensional (2D) grid. The logical 2D grid structure allows using cell-by-cell routes where there is a cell-head node in each cell to handle routing. Cell-based routing enhances the scalability of the routing protocol [15]. One node is elected as a cell-head in each grid cell and it has the following responsibilities: (1) forward route discovery requests to its neighbor cells; (2) transmit data packets to neighboring cells; and (3) maintain the routes that pass through its cell.

Each of the three types of existing routing protocols topology-based, position-based and grid-based has limitations. For instance, topology-based protocols generate a large amount of traffic when the network topology changes frequently due to mobility [13]. Position-based protocols suffer from a local minima problem which leads to non-guaranteed message

delivery [16]. Furthermore, position-based protocols mostly depend on location services [13] such as Home Agent [17] and Grid Location Service [14] to discover geographical locations of destinations. Another limitation of existing grid-based routing protocols is that they use an election approach for selecting cell-head (gateway) nodes which leads to high control packet overhead and high end-to-end delays. In this paper, we propose a new hybrid (proactive and reactive) routing approach which highly utilizes the logical grid environment in order to reduce the number of control packets and increase path stability. The new approach divides the routing in two layers: a proactive layer, where shortest-path trees are constructed and maintained, and a reactive layer, where destination nodes are tracked making use of the constructed shortest-path trees. In the proactive layer information about occupied grid cells is exchanged among nodes. A grid cell is referred to as "occupied" when there is at least one mobile node located in the inner margins of the cell; otherwise it is a "non-occupied" cell. Moreover, if a non-occupied cell becomes occupied, a special control packet (Empty_to_Non-Empty control packet) is flooded (using cell-based flooding) to inform all nodes in the network about this event. In cell-based flooding only one node in each cell (the cell-head) participates in broadcasting the packet to neighboring cell-heads. Similarly when an occupied cell becomes non-occupied, all nodes are informed using a cell-based flooding of a special control packet (Non-Empty_to_Empty control packet). The reactive layer is used to seek for a destination. Any node that wants to establish a connection with an unknown node (not registered in a local Node Table); it starts by sending a Route Request (RREQ) packet to seek for the destination location using cell-based flooding. The proactive layer information is saved in all nodes in the MANET environment. This information enables a cell-head to build a shortest path tree from the cell where it is located to all grid cells. If there is a change in the information about the occupied or non-occupied cells, the tree is reconstructed. The proactive mechanism runs without interrupting the data packets propagation. The reactive layer information (destination nodes location) is also saved at all nodes in a local Nodes Table.

We have conducted an intensive simulation-based performance evaluation of the proposed TGRP protocol and measured the average message delivery ratio, the normalized control overhead and the average end-to-end delay. We have extended the NS2 network simulator, which has been widely used in the literature for studying the performance of MANET routing protocols [15] [18] [19], to evaluate the performance of TGRP and compare it with the performance of GRID protocols. We have studied the performance of TGRP under a variety of network densities and cell sizes. The results show that TGRP outperforms GRID in terms of end-to-end-delays and delivery ratio. Furthermore, TGRP competes well with GRID in terms of control packet overhead. This paper is a revised and expanded version of our previous conference paper presented in [20].

The rest of this paper is organized as follows. Section 2 describes the proposed TGRP structure including its proactive layer and reactive layer mechanisms. Section 3 presents a simulation-based performance evaluation of TGRP compared to grid-based protocols. Section 4 concludes the paper.

II. The Tree-based Grid Routing Protocol (TGRP)

Like other grid-based protocols, TGRP divides the physical area into a logical two-dimensional grid of equal size cells (see Figure 1). In TGRP, a packet travels from a source node to a destination node by hopping from cell to cell making use of a previously constructed shortest path tree to decide at each hop the next cell to go towards the destination cell. A selected cell-head in each cell is responsible of forwarding the packets via that cell. The union of the cell-heads forms a backbone of the MANET.

Each node maintains four tables, namely Neighbors Table, Occupied Cells Table (OCT), Nodes Table and Tree Table (TT). The Neighbors Table in a given node contains lists of neighboring nodes. Each list represents one of the neighboring cells and lists the ids of all nodes in that cell. The Occupied Cells Table contains addresses of all occupied cells (i.e., the non-empty cells) whereas the Tree Table, which is the routing table, contains the occupied cells addresses and the next hop (next cell) info to reach them. The next hops are obtained by building a shortest path tree using an efficient algorithm. The Nodes Table contains the ids of all nodes and their locations (cell addresses). A cell address is a pair of (x, y) coordinates in the logical grid assuming the address of the bottom left corner cell is (1, 1).

Figure 1, Figure 2, Table 1 and Table 2 show respectively an example of a MANET environment, an Occupied Cells Table, a shortest path tree created at node 1 and its representation in a routing table (a Tree Table).

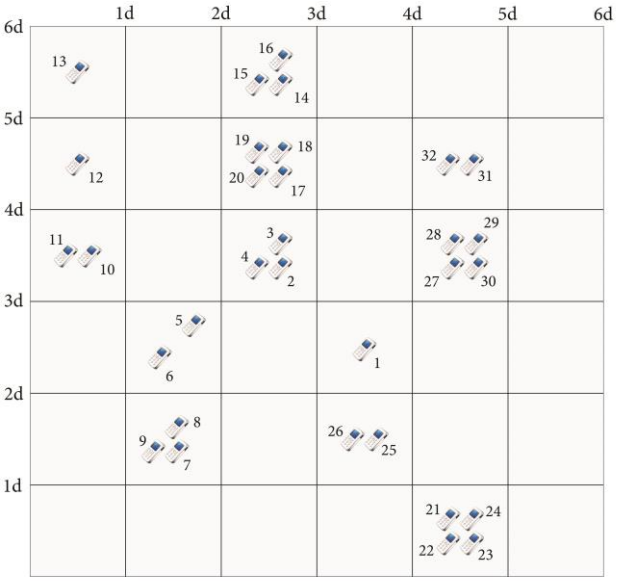


Fig. 1. An example of a grid-based MANET environment

TABLE I. Occupied Cells Table of Figure 1

(5,1)	(2,2)	(4,2)	(2,3)	(4,3)	(1,4)	(3,4)	(5,4)	(1,5)	(3,5)	(5,5)	(1,6)	(3,6)

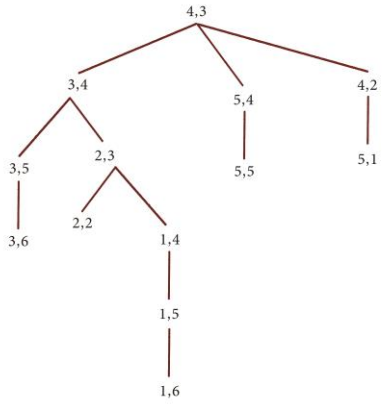


Fig. 2. A shortest path tree at Node 1 of Figure 1

TABLE II. OCCUPIED CELLS TABLE OF FIGURE 1

Destination	(3,4)	(5,4)	(4,2)	(3,5)	(2,3)	(3,6)	(2,2)	(1,4)	(1,5)	(1,6)	(5,5)	(5,1)
Next Hop	(3,4)	(5,4)	(4,2)	(3,4)	(3,4)	(3,4)	(3,4)	(3,4)	(3,4)	(3,4)	(5,4)	(4,2)

A. Control Packets

TGRP uses control packets to build and maintain its tables. The TGRP proactive layer creates a shortest path tree at each node by using information available in the Occupied Cells Table. Thus, it is important to ensure consistency of the information in the Occupied Cells Table of all nodes in the environment. We do that by using dedicated control packets. These control packets are used to register new occupied cells and to delete empty cells. The following control packets are used for this purpose:

a) Empty_to_Non-Empty (ENE) packet: when a node enters an empty cell, it floods (using cell-based flooding) an Empty_to_Non-Empty (ENE) packet. This packet contains the cell address of the entered cell. The nodes which receive this packet update their OCT tables by adding the cell address of the entered cell.

b) Non-Empty_to_Empty (NEE) packet: when a cell-head leaves a cell and there is no other node left in that cell, it floods (using cell-based flooding) a Non-Empty_to_Empty (NEE) packet. This packet contains the address of the cell that has become empty. Any node recieving this packet updates its OCT table by deleting from it the cell that has become empty.

*c) EXIT packet:*when any node moves out of its current cell to a neighboring cell, it transmits an EXIT packet to tell the neighbors about its new location.

d) INFO packet: when a cell-head node receives an EXIT packet from a node in its cell, it replies (using unicasting) by sending an INFO packet which contains the Neighbors Table. The sender of the EXIT packet replaces its Neighbors Table with the received one.

The reactive layer uses the following control packets:

e) Route Request (RREQ) packet: A Route Request packet piggy-backs the geographic location (cell address) of the source node to be recorded by all reachable nodes in the MANET. Every cell-head node will rebroadcast this packet after piggy-backing its geographic location.

f) Destinaion Location (DLOC) packet: when a destination node receives a RREQ packet, it floods (using cell-based flooding) its cell address location using a Destination Location (DLOC) packet to all nodes in the network. Any node receiving this DLOC packet updates its Nodes Table.

B. Building and Maintaining Shortest Path Trees

At each node a shortest path tree rooted at that node is constructed and used to build a Tree Table which guides the routing of data packets towards their destination cells. The shortest path tree is constructed using the information about occupied cells gathered during the proactive layer of the protocol. The tree is a breadth-first search tree of the graph of occupied cells. The vertices of this graph are the occupied cells and two occupied cells are connected by an edge if they are neighboring cells in the grid. The breadth-first search algorithm is outlined in Figure 3. It uses a FIFO queue data structure storing a list of occupied cells not yet visited in the search. The breadth-first search algorithm starts by enqueueing the root cell in the initially empty queue and then loops dequeueing at each iteration one cell from the queue and enqueueing its unvisited adjacent occupied cells until the queue becomes empty. In each iteration the links between the dequeued cell and enqueued adjacent cells are recorded in the Tree Table as (next hop) links on the shortest path tree.

Build Tree Algorithm:

Initialize Queue to empty
Initialize Tree Table to empty
Register this node's cell (root cell) in the Tree Table
Enque this node's cell (root cell)
While Queue is not empty

- Dequeue one cell *C* from Queue
- For each neighboring cell *C'* of *C* do the following:
  - If *C'* is an empty cell (not listed in the OCT table) then ignore it
  - If *C'* is already registered in the routing table then ignore it
  - If *C'* is occupied (listed in OCT) and not registered in the routing table then enqueue *C'* in Queue and register it in the Tree Table (*C* is the next hop from *C'* towards root)

Fig. 3. Construction of the shortest path tree and tree table.

The time complexity of this construction is O(*V*+*E*) where *V* is the number of occupied cells and *E* is the number of (occupied cell, adjacent occupied cell) links. In the worst case, *V* is equal to the total number of grid cells and *E* is less than 4*V* (since each cell has at most 8 adjacent non empty cells). The shortest path tree construction algorithm is therefore O(*N*) where *N* is the number of occupied grid cells.

The shortest path tree has to be rebuilt when there is a change in the occupied/empty status of the cells. To reduce the

number of times the tree is rebuilt, we use a valid Boolean flag indicating whether the current tree is valid or not depending on changes in the Occupied Cells Table. A node does not rebuild the tree until it has to route a data packet and the valid flag is false.

C. Cell-Based Flooding

Any node that decides to initiate a route discovery (if there is no information about the destination neither in the Neighbors Table nor in the Nodes Table) broadcasts a route request (RREQ) to all its adjacent cells. There are two possible cases for a node receiving the RREQ:

Case 1: Cell-Head Node: A cell-head node receiving the RREQ has to flood (cell-based flooding) a route reply (RREQ) packet. In cell-based flooding only one node in each cell, the cell-head, participates in the flooding. If any node receives a previously processed RREQ (detected by checking the node sequence number in its Nodes Table), the node discards it and does not forward it. RREQ piggybacks the location (cell address) of the previous hop node which is used to update the Neighbors Table.

Case 2: Non Cell-Head Node. It broadcasts (using cell-based flooding) a Destination Location (DLOC) packet if it is the destination; otherwise it records the cell location information of the previous forwarding node and then discards the RREQ packet.

Only one node in each cell (the cell-head) participates in rebroadcasting the RREQ. This mechanism is called cell-based flooding (as opposed to total flooding used in AODV for example). Once the source node receives the Destination Location DLOC packet, it starts forwarding data packets to the destination using the shortest path tree next hop links recorded in the Tree Table. If the destination node moves out of its cell to a new one, it should broadcast (using cell-based flooding) a Destination Location DLOC packet. This will not affect the ongoing transmission of data packets to the destination.

D. Cell-Head Selection in TGRP

After building the Neighbors Table with the most recent information, the selection of the cell-head node becomes simple and fast. The node with the highest id in a cell is implicitly chosen as the cell-head of that cell without any additional overhead.

All the nodes in the MANET have the same OCT table which leads to avoid using any special packets for maintaining cell-heads. Notice that in the GRID protocol a RETIRE packet is sent by a cell-head when it leaves its cell to another cell. Cell-heads are used in TGRP as a backbone for the cell-based flooding mechanism and there is no need for the RETIRE packet.

The cell-head could be chosen by any mechanism such as selecting the node with highest id as we do here, (or lowest id). To avoid overloading the nodes with high ids (or low ids), a node could create and use a random number as a varying id. In the unlikely case of two equal random numbers drawn at two mobile nodes located in the same cell any of them can be chosen as cell-head of the cell because both nodes have the same OCT table.

E. Operations of the Proactive and Reactive Layers in TGRP

Any node joining the MANET executes an initialization phase in which it starts by determining its geographic location. It continues monitoring its location periodically until it leaves the MANET.

TGRP is divided into two layers: a proactive layer and a reactive layer; each layer has its own mechanisms. The main function of the proactive layer is to maintain up-to-date the Occupied Cells Table and the Tree Table. All nodes in the MANET environment should have the same copy of the Occupied Cells Table (see Figure 4) which is maintained using the proactive layer control packets. Figure 4 outlines the Occupied Cells Table maintenance mechanisms.

Empty to Non-Empty

If a node enters an empty cell (Neighbors Table shows no nodes in the entered cell) then:

- Cell-based flooding of an Empty-to-Non_Empty packet
- Any node receiving the Empty-to-Non_Empty packet performs the following:
  - modify the Occupied Cells Table
  - valid = false (routing tree has to be rebuilt)

Non-Empty to Empty:

If a cell-head leaves a cell and no node is left in the previous cell (based on the Neighbors Table) then:

- Cell-based flooding of a Non-Empty-to-Empty packet
- Any node receiving this packet performs the following:
  - modify Occupied Cells Table
  - valid = false (routing tree has to be rebuilt)

Exit from Cell:

- If a node enters a new cell then it sends EXIT packet to all nodes in its transmission range
- When the cell-head of the new cell receives the EXIT packet it unicasts an INFO packet to the new node

Build Tree:

- If a node has to route a packet and the valid flag is false then run the **Build Tree** algorithm
- Valid = true

Fig. 4. The operations of the proactive layer in TGRP

The reactive layer is used to track the location of a destination and to ensure this information is known by all nodes in the MANET. Figure 5 outlines the mechanisms of this layer.

Routing of a Received Data Packet

- If no information about the destination in the Nodes Table then cell-based flooding of RREQ
- If cell-head receives RREQ request (first time) then retransmit RREQ
- If non-cell-head receives RREQ request and it is not the destination node then do nothing

- If a destination node receives RREQ then cell-based flooding of Destination Location packet

Destination Location

- If the destination moves to another cell then cell-based flooding of a Destination Location packet

Fig. 5. The operations of the reactive layer in TGRP.

III. A Simulation-Based Performance Evaluation

In order to investigate the effect of using the hybrid TGRP protocol, we have extended an original implementation of NS2 (version 3.4) with implementations of the protocols GRID and TGRP in order to evaluate and compare their performance. The performance evaluation of TGRP has been conducted using the simulation model and parameters outlined in Table 3. The evaluation study analyzes the impact of the network density and node mobility speed on the three performance metrics packet delivery ratio, normalized control overhead and average end-to-end delay for two different cell sizes.

We use the term cell size to refer to the cell side length. As shown in Table 3, two cell sizes 141 meters and 190 meters will be tested in our experiments to show the impact of cell sizes on TGRP and GRID protocols.

TABLE III. Simulation parameters

Communication type	CBR
CBR sending rate	4 packets per second
Simulation area	1000m x 1000m
Simulation protocols	TGRP, GRID
Mobility model	Steady-state random waypoint
Number of nodes	60, 100, 200, 300, 400
Nodes average speed	2, 4, 6, 8, 12, 16, 20 (meters/second)
Average pause time	2 (Delta = 1 seconds)
Number of connections	30 connections
Transmission range	300 meters
Physical link bandwidth	11 Mbps
Number of simulation trials	40 times
Simulation time	1000 seconds
Cell side length (cell size)	141, 190 meters

A. Impact of Network Density

This section presents results of studying the impact of network density on the performance of TGRP compared to GRID. The network density has been varied by deploying 60, 100, 200, 300 and 400 mobile nodes in a fixed geographic area of dimensions 1000m × 1000m. The nodes in the network move according to the steady-state random waypoint mobility model with average speed of 6 meters per second. The number of connections between randomly selected peer sources and destinations has been fixed to 30, all established during the simulation time. Each source node in a connection sends four packets per second to the corresponding destination and each packet is of size 512 bytes.

F. Delivery Ratio:

Figure 8 shows the effect of the network density on the packet delivery ratio of the TGRP and GRID protocols for different cell sizes. It reveals that TGRP with a grid cell size of 141 meters exhibits a better performance compared to GRID in terms of packet delivery. The stability of the path in TGRP allows nodes to keep pumping data packets without interruption. This stability comes from the existence of alternative paths in the OCT table. Anytime there is a change in the OCT table, a new shortest path tree is constructed proactively by the affected nodes without affecting the ongoing communications.

The mobility of a source node does not affect the path. As long as the destination node remains in the same destination cell, the packets can still be routed correctly based on the information in the Node Table. The source node can continue sending packets while moving around in the MANET environment. There is no need to inform all other nodes about its new location when it moves to a different cell. It just sends an EXIT packet to its neighbor nodes and rebuilds its shortest path tree.

The stability of paths also comes from avoiding the use of RETIRE packets and cell-head re-election when cell-heads move between cells. Cell-heads just broadcast EXIT packets when they move to different cells. When a cell-head A moves to a new cell, the cell-head B of that new cell sends a small INFO packet to A containing the list of nodes located in the neighboring cells

The increase of the delivery ratio in TGRP also comes from the fact that cell-heads in GRID protocols discard some data packets when they are not involved in routing packets for any reason such as a change in the path due to the moving of a source node to a different cell or because of received ERROR packets. This is not the case in TGRP because all nodes have the capability to forward packets to destination nodes and all nodes have consistent information about paths to destination nodes.

The simulation results show that TGRP works better with a cell size 141 meters than 190 meters. This is because of the lack of consistency between the transmission range (300 meters) and the cell size 190 meters. For example, with a cell size of 141 meters and a transmission range of 300 meters, the adjacent neighbors of a cell are approximately covered by the transmission range which will allow nodes in this cell to communicate with each other. In the example of Figure 6, a grid size of 141 meters is assumed. In this figure node 20 wants to send a data packet to node 1 through node 4. Node 20 sends its packets to node 4. Node 4 discovers from its Neighbors Table that the destination node 1 is reachable in the next hop. Node 4 delivers the packet to the destination node 1. If however a cell size of 190 meters is assumed, then when node 4 tries to forward the packet received from node 20 to node 1, the packet is dropped because the transmission range of node 4 does not cover node 1 (see Figure 7).

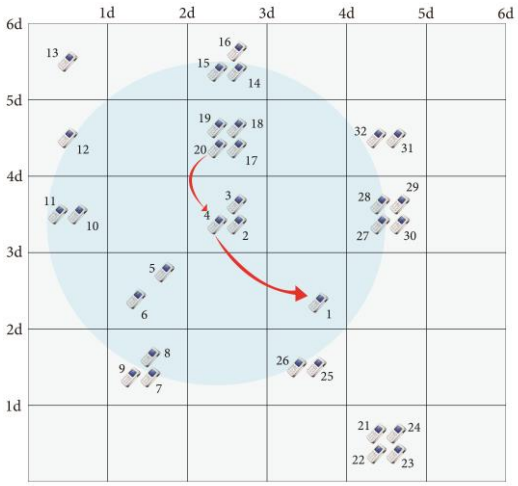


Fig. 6. Grid environment with cell size 141 meters

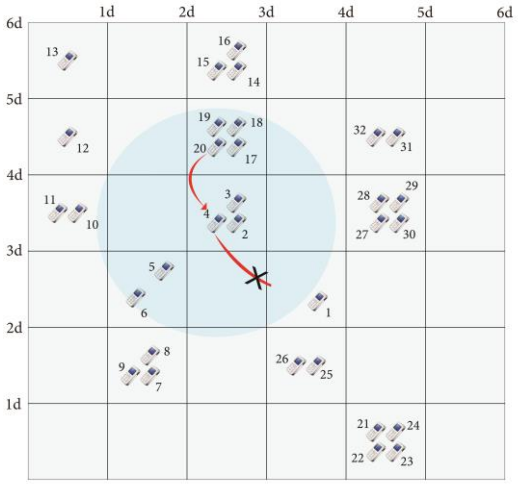


Fig. 7. Grid environment with cell size 190 meters

We observe from the results of Figure 8 that TGRP outperforms GRID protocols in terms of delivery ratio. This can be justified by the fact that the stability of the paths resulting from using the proactive layer information (availability of alternative paths) leads to an increase in the number of delivered packets to their destinations. Figure 8 also shows that TGRP incurs a higher delivery ratio with size 141 meters compared to 190 meters.

Figure 8 shows that sparse environment (density = 60, 100 nodes) has a negative effect on the delivery ratio of TGRP whereas dense environment has a positive effect due to the ability of building paths to destinations when the density is high. Moreover, the delivery ratio with 300 nodes is slightly better than the delivery ratio with 400 nodes due to the fact that 400 nodes broadcast more EXIT packets which affects the propagation of data packets.

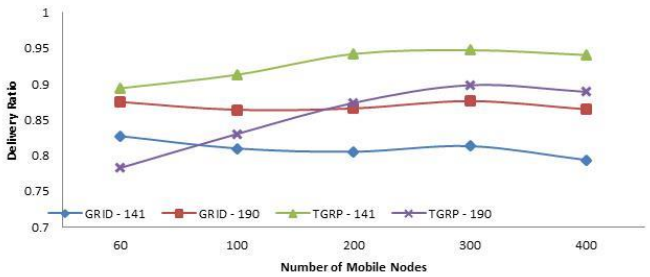


Fig. 8. Delivery ratio vs. number of mobile nodes for TGRP and GRID. Mobility speed = 6 m/number of connections = 30, CBR packet rate = 4 packets/sec.

G. Control Overhead:

In this section, we present simulation results measuring the normalized control overhead which is defined here as the number of generated control packets per delivered data packet. Figure 9 shows that TGRP is more scalable than GRID in terms of normalized control overhead when varying the number of nodes (and hence the network density). TGRP has exhibited an inverse relationship between network density and normalized control overhead whereas GRID has exhibited a positive relationship.

The proactive layer control packets (Empty-to-Non-Empty and Non-Empty-to-Empty) are affected inversely with the increase of density. Increasing the number of nodes reduces the need to send those packets because those packets are sent when the status of a cell changes from empty to non-empty or vice versa. These changes are reduced with an increased number of nodes in the cells. Thus, in TGRP, the number of control packets needed to deliver data packets is reduced with increased number of nodes.

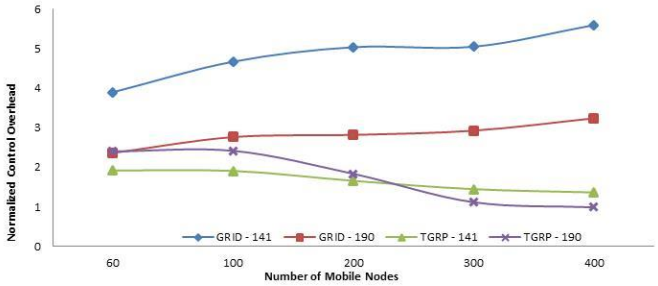


Fig. 9. Normalized control overhead vs. number of mobile nodes for TGRP and GRID. Mobility speed = 6 m/s, number of connections = 30, CBR packet rate = 4 packets/sec.

When the density is less than 200 nodes, TGRP with cell size 190 meters needs more control packets to submit data packets compared to TGRP with cell size 141 meters. This is due to the disconnection problem illustrated in Figure 7. When the density is higher than 200 nodes, the disconnecting of paths is reduced due to the existence of more nodes in cells which leads to reduce the number of control packets.

H. End-to-End Delay:

Figure 10 shows that TGRP exhibits much better performance than GRID in terms of end-to-end delay. The TGRP proactive layer provides high path stability which results in a tremendous improvement in the average end-to-end delay compared to GRID (over 80% improvement compared to GRID).

The mechanism of path maintenance in TGRP does not affect the pumping of data packets. The maintenance of the paths is done in the proactive layer (in the background). There are always chances of existence of alterative paths. All these factors lead to the superiority of TGRP. Figure 10 also reveals that the increase in the density of the network leads to improve the average end-to-end delay for all protocols. In spite of the big difference between protocols, they all exhibit the same behavior.

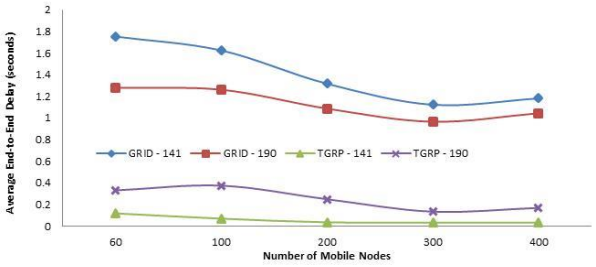


Fig. 10. Average end-to-end delay (seconds) vs. number of mobile nodes for TGRP and GRID. Mobility speed = 6 m/number of connections = 30, CBR packet rate = 4 packets/sec

I. Impact of Node Mobility

To study the effect of mobility on TGRP compared to GRID, we set a fixed number of 200 nodes and a fixed number of 30 connections and vary the node mobility speed. The nodes are placed over a network area of 1000m x 1000m using steady-state random waypoint mobility model with a variety of node mobility speeds as shown in Table 4.

TABLE IV. PARAMETERS FOR NODE SPEED

Average Node Speed (meters/second)	2	4	6	8	12	16	20
Delta of Node Speed (second)	1	2	4	6	10	12	18

J. Delivery Ratio:

Figure 11 reveals that TGRP with cell size 141 meters is more scalable and stable in terms of delivery ratio when varying the node mobility speed. The delivery ratio of TGRP is not affected by the increase of the node mobility speed. There is no election mechanism that effects the propagation of data packets. TGRP with cell size 190 meters has the same behavior as TGRP with cell size 141meters but it has lower delivery ratio because of the disconnection problem illustrated in Figure 7.

TGRP depends on its proactive layer control packets to ensure consistency of the Occupied Cells Table and hence the stability of the routing paths. With a high number of mobile nodes (high density), the number of control packets remains approximately the same when increasing the speed. A path to destination is not broken by the moving of nodes between cells. As long as the empty/non-empty status of the cells is not affected, the routing paths (cell-based paths obtained from the shortest-path tree) remain valid. There are only two control packets which are affected by increasing the mobility speed of the nodes: Destination Location packet and EXIT packet. The proactive layer increases the stability of the connected paths which leads to reduced effect of the mobility speed on the delivery ratio.

GRID with cell size 190 meters works better than with cell size 141 meters. The paths are constructed with less number of hops and less number of RETIRE packets.

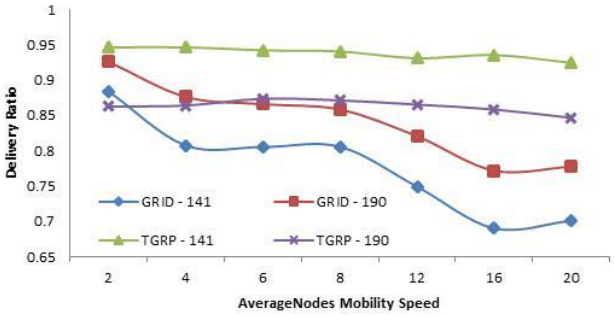


Fig. 11. Delivery ratio (%) vs. averagenode speed for TGRP and GRID. Number of CBR connections = 30 , number of nodes = 200, CBR pacekt rate = 4 packets/sec.

Overall, TGRP shows high scalability in terms of delivery ratio compared to GRID when increasing the node mobility speed. It shows improvement by at least 10% compared to GRID. It is slightly affected by the increasing of mobility speed. The scalability comes from the stability of the cell-based paths in the TGRP protocol. The paths are cell-based paths and not node-based paths.

K. Control Overhead:

Figure 12 shows that TGRP has consumed 40% less control packets than GRID to deliver data packets. There are only three types of control packets which are needed more with higher node mobility: Destination Location packets, EXIT packet and INFO packet. The number of Empty-to-Non-Empty packets and Non-Empty-to-Empty packets is affected by the number of nodes in the environment but not much by the mobility speed. If there are enough mobile nodes, there is little need for these control packets since changes in the empty/non-empty status of cells are unlikely to take place. If however the number of nodes is small then higher mobility causes more empty/non-empty status changes and hence the number of status change control packets can be large.

TGRP has used less control packets to deliver data packets compared to GRID when the average mobility speed is less than 8 meters per second. TGRP with cell size 141 meters needs about one to two control packets to deliver one data packet whereas it needs more than 2 control packets to deliver one data packet when the mobility speed exceeds 8 meters per

second. TGRP with cell size 190 meters uses approximately the same number of control packets with different mobility speeds to deliver one data packet. That is because, a grid with cell size 141 meters has more cells than a grid with cell size 190 meters which leads to have more Empty-to-Non-Eempty and Non-Empty-to-Empty packets.

Figure 12 also reveals that GRID is affected negatively by increasing the mobility speed in terms of normalized overhead. Increasing the node mobility speed leads to increasing the number of broken connections and therefore the number of control packets used to reestablish the broken connections.

Overall, TGRP is more stable than GRID in terms of control overhead when the mobility is increased.

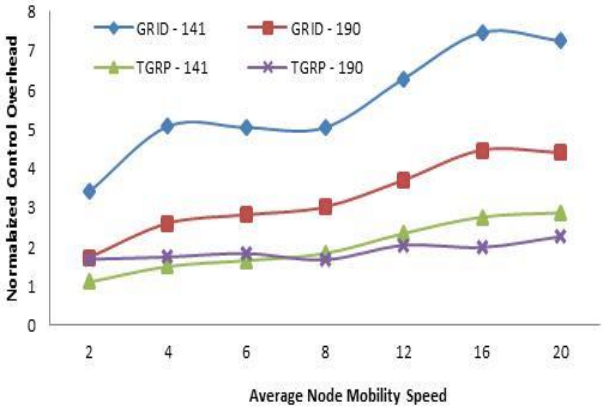


Fig. 12. Normalized control overhead vs. Average Node Mobility Speed for TGRP and GRID. Number of CBR Connections = 30, number of nodes = 200, CBR packet rate = 4 packets/sec.

L. End-to-End Delay:

Figure 13 shows that TGRP with cell size 141 meters outperforms by far GRID in terms of end-to-end delay for different node mobility speeds. It shows that it is a very stable and scalable protocol in terms of end-to-end delay. It shows an improvement of about 90% compared to GRID.

The availability of alternative paths (provided by recalculating the shortest path trees in the proactive layer) leads to have very stable connections. In addition to that, the paths (cell-based paths) are not affected by node mobility as long as the cells forming the paths remain occupied. Figure 13 also reveals that increasing the mobility speed affects negatively GRID in terms of end-to-end delay. It leads to have broken connections which requires in GRID to re-establish the connections using more control packets.

At low mobility speed, TGRP with cell size 141 meters has the lowest average delay among the evaluated protocols. Overall, Figure 13 also shows the effect of node mobility on the average end-to-end delay for the GRID protocol.

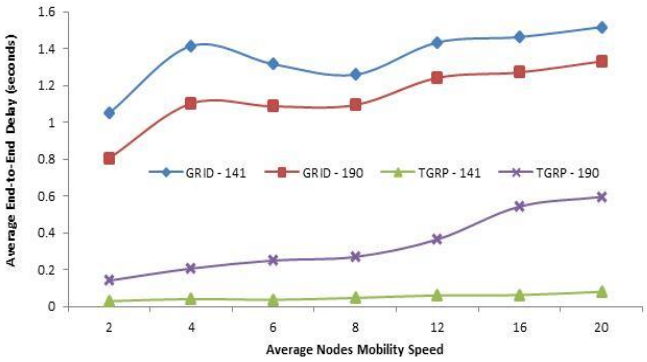


Fig. 13. Average end-to-end delay vs. average node speed for TGRP and GRID. Number of CBR connections= 30, number of nodes = 200, CBR packet rate = 4 packets/sec.

IV. CONCLUSION

We have proposed and evaluated the performance of a new routing protocol called Tree-based Grid Routing Protocol (TGRP), which uses a new hybrid proactive and reactive routing approach in grid-based MANETs. In TGRP, there is a proactive layer which builds and maintains a table called Occupied Cells Table and builds from it shortest path trees between occupied cells. There is also a reactive layer in TGRP that is responsible for discovering the location of destination nodes by exploiting the constructed shortest path trees. The performance of the proposed TGRP protocol has also been studied and compared with the performance of GRID using extensive simulation experiments. The performance has been evaluated in terms of end-to-end delay, delivery ratio and control overhead for a variety of network density and node mobility conditions with two different cell sizes. The results have shown that TGRP scales better than GRID in terms of delivery ratio and control overhead and it is by far superior to GRID in terms of end-to-end delay.

REFERENCES

[1] H. Walia, E. M. Singh and D. R. Malhotra, “A Review: Mobile Ad Hoc Routing Protocols”, International Journal of Future Generation Communication and Networking vol. 9, no. 2 (2016), pp. 193-198.

[2] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers", Symposium on Comm. Architectures and Protocols, 1994.

[3] C. E. Perkins and E. M. Roycr, "Ad hoc on demand distance vector (AODV) routing (internet-draft)", Mobile Ad-hoc Netwrok (MANET) Working Group, IETF, 1998.

[4] J. Broch, D. B. Johnson and D. A. Maltz, "The dynamic source routing protocol for mobile ad hoc networks (Internet-draft)", Mobile Ad-hoc Network (MANET) Working group, 1998.

[5] D.B. Johnson and D.A. Maltz, “Dynamic Source Routing in Ad hoc Wireless Networks”, Mobile Computing, Kluwer Academic Publishers, pp. 153-181, 1996.

[6] E. Kranakis, H. Singh and . J. Urrutia, "Compass routing on geometric networks", 11th Canadian Conference on Computational Geometry (CCCG’99), 1999.

[7] G. G. Finn, "Routing and addressing problems in large metropolitan-scale internetworks", Technical Report ISU/RR-87-180, USC ISI, Marina del Ray, CA, 1987.

[8] W.-h. Liao, J.-p. Sheu and Y.-c. Tseng, "Grid: A Fully Location-Aware Routing Protocol for Mobile Ad Hoc Networks", Telecommunication Systems, vol. 18, pp. 37-60, 2001.

[9] C.-M. Chao , J.-P. Sheu and C.-T. Hu, "Energy-Concerving Grid Routing Protocol in Mobile Ad Hoc Networks," in Proc. of IEEE 2003 Int'l Conference on Parallel Processing, 2003.

[10] Shahab Kamali, Jaroslav Opatrny, "A postition Based Ant Colony Routing Algorithm for Mobile Ad-hoc Netwroks", Academy Publisher, Journal of Networks, pp. vol. 3, no. 4, 2008.

[11] G. Finn, "Routing and addressing problems in large metropolitan-scale internetworks", Technical Report, University of Southern California, 1987.

[12] E. Kranakis, H. Singh, and J. Urrutia, "Compass routing on geometric networks", in 11th Canadian Conference on Computational Geometry (CCCG'99), 1999.

[13] A.E. Abdallaha, T. Fevensa, J. Opatrnya, and I. Stojmenovic, "Power-aware semi-beaconless 3D georouting algorithms using adjustable transmission ranges for wireless ad hoc and sensor networks", Ad-Hoc Networks, vol. 8, no. 1, pp. 15-29 , January 2010.

[14] K. Omer and D. Lobiyal, "Performance Evaluation of Location Update Schemes for MANET", The International Arab Journal of Information Technology, vol. 6, no. 3, 2009.

[15] W.-H. Liao, Y.-C. Tseng, J.-P. Sheu, "Grid: A Fully Location-Aware Routing Protocol for Mobile Ad Hoc Networks", Telecommunication Systems, vol. 18, pp. 37-60, 2001.

[16] A.E. Abdallah, T. Fevens, J. Opatrny, "High delivery rate position-based routing algorithms for 3D ad hoc networks", Computer Communications , vol. 31, no. 4, 2008.

[17] Stojmenovic I., "Home Agent Based Location Update and Destination Search Schemes in Ad Hoc Wireless Networks", Advances in Information Science and Soft Computing, 2002.

[18] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: the incredibles", ACM SIGMOBILE Mobile Computing and Communications Review, vol. 9, pp. 50-56, 2005.

[19] M. Bani-Yassein, M. Ould-Khaoua, L. M. Mackenzei, and S. Papanastasiou, "Performance Analysis of Adjusted Probabilistic Broadcasting in Mobile Ad Hoc Networks", International Journal of Wireless Information Networks, vol. 13, pp. 127-140, April 2006.

[20] H. Almaqbali, K. Day, M. Ould-Khaoua, A. Touzene, N. Alzeidi, "A New Hybrid Grid-Based Routing Approach for MANETs", Proc. of CETC2013 Conference on Electronics, Telecommunications and Computers, 5-6 December 2013, Lisbon, Portugal, pp. 81-89.

Balanced Distribution of Load on Grid Resources using Cellular Automata

Amir Akbarian Sadeghi
Department of Computer Engineering
South Tehran Branch,
Islamic Azad University
Tehran, Iran

Ahmad Khademzadeh
Research Institute for Information & Communication Technology
Tehran, Iran

Mohammad Reza Salehnamadi
Department of Computer Engineering
South Tehran Branch,
Islamic Azad University
Tehran, Iran

***Abstract*—Load balancing is a technique for equal and fair distribution of load on resources and maximizing their performance as well as reducing the overall execution time. However, meeting all of these goals in a single algorithm is not possible due to their inherent conflict, so some of the features must be given priority based on requirements and objectives of the system and the desired algorithm smust be designed with their orientation. In this article, a decentralized load balancing algorithm based on cellular automata and fuzzy logic has been presented which has capabilities needed for fair distribution of resources in Grid level.**

Each computing node in this algorithm has been modeled as a Cellular Automata's cell and has been provided with the help of fuzzy logic in which each node can be an expert system and have a decisive role which is the best choice for tasking in dynamic environment and uncertain data.

Each node is mapped to one of the VL, L, VN, and H, VH conditions based on information **exchange on certain time periods with its neighboring nodes and based on fuzzy logic and tries to estimate the status of the other nodes in subsequent periods to reduce communication overhead with the help of Fuzzy Logic and the decision making to send or receive task loads is done based on the status of each node. So an appropriate structure for the system can greatly improve the efficiency of the algorithm. Fuzzy control does not use search and optimization and makes decisions based on inputs which are effective parameters of the system and are mostly based on incomplete and nonspecific information.**

Keywords—computing Grid; load balancing; cellular automata; fuzzy logic

I. INTRODUCTION

The need for high computational power and organizational limitations have created a new type of shared computing environment, which is called computing grid. Computing Grid is a computing infrastructure that makes effective access to high performance with computing resources possible. End users and applications see this environment as a large virtual computing system. Systems that are connected by Grid may be distributed globally and be running on different hardware platforms and operating systems and belong to different organizations. In a short definition, Grid can be considered as a system for distributed resource sharing in a large scale and indeed without borders. Requests should be divided evenly among the available resources in order to globally enhance the global throughput of computing grid. Management of resources is one of the major issues in this environment. Resource management is a major and infrastructure Grid component of environment. The overall objective of resource management is effective timing to run programs that need to use resources in Grid environment. In a general definition. The purpose of load balancing algorithms is uniform distribution of the load on resources and maximizing resource efficiency as well as reducing the overall running time which means the difference between the most and least productive resources should be minimal. The load balancing problem for Grid environment in which equitable distribution of resources is one of the most important issues is also considered as a basic necessity. The desirable characteristics of a load balancing solution include: Comparability, versatility, stability, clarity of vision of program's user, capability of fault tolerance and minimal overhead costs imposed on the system. The load balancing methods are generally divided into centralized and decentralized, static and non-static, cyclic or non-cyclic, and has a threshold and no threshold. Cellular automata answers this question that How complex systems can be studied? There is the ability to predict the next state of cells in this system based on the status of each cell and its adjacent cells which can help in proper distribution of load among nodes. Given that the distribution of load needs awareness of mentioned conditions and considers the functionality of each resource in a computing grid and cellular automata has this feature which means it can predict current and future situation of each resource, the load distribution is done in a balanced way based on the needs and abilities and capabilities of each of these resources [14-19].

This article tries to execute distribution of load in Grid resources by evaluating the effectiveness of cellular automata and fuzzy logic which have capabilities required in the fair distribution of load and grade level decision-making. The load balancing algorithms has been provided in this article based on Cellular automata- and use of fuzzy rules. The remainder of the paper is organized as follows: In section 2, definition of concepts such as Grid, load balancing, cellular automata and fuzzy logic. Section 3 describes our proposed algorithm in detail. Section 4 discusses our simulation and results of evaluation. Finally, section 5 concludes this paper.

II. DEFINITION OF CONCEPTS

Advances in areas technical constantly need to have faster computing but computer hardware manufacturers have reached fundamental limitations in the physical speed [1]. Electronics and hardware advances in technology alone cannot meet the demand for increased computing speed. Parallel processing is the emerging response to this problem in which different parts of a task are simultaneously tasking on several processors [2, 3].

Although writing code that is flexible enough to be split among several processors is generally more difficult for programmers but the tendency toward parallel processing hardware and software has increased [2]. Instead of limiting the time of implementation of a program running on a processor, parallel processing task load is divided among several processors and allow this issue to be solved through team work, thus parallel processing has become a viable alternative to the circuit and faster processors which can only reduce the time of initial cycle time of the single processor [3]. Reduced costs powerful computers along with advances in computer networking technologies have increased the tendency for the use of large-scale parallel systems and distributed computing systems. In fact, recent studies in the field of computing architecture has led to emergence of a new computing paradigm which is computing Grid [4]. A computing Grid creates a hardware and software infrastructure which is: Reliable, consistent, pervasive, and has inexpensive access to high performance computing [5]. This technology is a type of distributed system that supports the sharing and coordinated use of resources, independently from physical type and their location in virtual dynamic organizations which is the same shared goal. Nowadays, a variety of Grid systems are manufactured with various definitions and facilities which have different objectives. Thus, providing a single definition that covers all aspects of grid computing technology in not easy nor true. Various experts have provided different definitions according to different pursued goals with different views towards this technology and its various applications.

Ian Foster who was the main inventor of Grid and founder of Globus defines Grid as follows [6]:

"Grid technology is seeking to create the possibility of large-scale and controlled resource sharing which is flexible and is after creating protocols, services and software packages".

Gird is defined as follows in IBM Company which is among pioneers of Grid:

"Grid is a set of distributed computing resources in a local area network or a wide area network which seems like a computer and virtual computing system for end-user or applications. Its main goal is creating dynamic virtual organizations through sharing resources using coordinated and safe methods among users, universities and organizations".

A computing grid is a grid computing infrastructure which provides access to advanced computing resources which features such as being: High-End Computational Resources, Dependable, Consistent, Pervasive and having Coordinated Resource Sharing and problem solving in dynamic virtual organizations is multi-organ.

Generally, Grid is a distributed system which contains following items [8]:

- Resources (software and hardware) are heterogeneous
- Resources are coordinated but are not under a centralized management
- The use of all-purpose standard protocols and interfaces
- Grid may have Multiple administrative domains or in other words be made of several Virtual Organizations (VO)
- Ensuring the quality of the services provided

Resource management is one of the important issues in such environment. Resource management is among major components and infrastructure of Grid environment. The overall objective of resource management is effective timing of applications which need to use available resources in Grid environment for running. In a general definition. The purpose of load balancing algorithms is equal distribution of load on resources and maximizing their performance as well as reducing the overall execution time [9]. In another definition, the load balancing algorithm is an algorithm which ultimately allows all nodes to task at once [10-11]. The issue of load balancing for Grid environment has fair distribution of load on resources as a basic necessity. The desirable characteristics of a load balancing solution include: desirable characteristics of a load balancing solution include: Scalability, Adaptability, Stability, Application Transparency, Fault Tolerant and minimum overhead imposed on the system. The mentioned specifications are greatly interdependent. For example, delays such as Computation Delay and Communication Delay have abnormal effects on the stability and thus comparability of the algorithm. Due to the many parameters involved in the problem of load balancing as well as contradictory of some mentioned features, meeting all the features in the form of a single algorithm is practically difficult or even impossible. Most of the existing methods try to satisfy one or more of the above objectives [12-14].

For better efficiency and more use of dynamic algorithms and considering that the main focus of this article is on the same set of algorithms, in general, the process of dynamic load balancing algorithms has four main routines:

A. Load Measuring routine

B. Information Exchange routine

C. Initiation routine

D. The final load balancing operation

Load measuring routine is expression of CPU load in a way that heavier load on processors will increase it and its reduction will reduce it. Since the routine is repeatedly and with great frequency in use (run) of the load balancing algorithms, the Calculation of obtaining should be as simple and as efficient as possible [17]. Information Exchange routine determine the method of collecting necessary task load for load balancing

decisions. Initiation routine decides about the time of starting load balancing. This decision-making is along with determining the ratio of efficiency to imposing overhead (which means load balancing must be effective). Load balancing methods attempt to achieve goals such as minimizing the average response time for processing or maximizing resource efficiency by running processes on distributed resources. This this goal may initially be a demand or take place after the start of its implementation. Of course, in any case, a good and efficient algorithm must consider the cost of route as well [18]. Cellular Automata (CA) is an answer to this question that how to study complex systems. Cellular automata can be a complex system in itself and yet provide appropriate methods to study complex systems like these - Complex systems – [19-20].

III. The Proposed Algorithm of Fuzzy Load Distribution using Cellular Automata (Fuzzy Load Balancing Cellular Automata)

The main idea of this project is using a cell of cellular automata to show a computational node in which the status of cell shows the status of that node. A load balancing solution can be created just using local load balancing in this method. The method of load distribution is in form of a wave motion.

All parameters that each processor considers during the proposed load balancing algorithm are described below:

M: Number of heterogeneous computing nodes in the system (P_1, P_2, ..., P_M)

x: Number of job executed in the system (j_1, j_2, ..., j_x).

T_s: Iinformation exchange time.

T_e: The estimated time period.

N_i: Buddy set of node P_i.

$S_i(T_n)$: State of node i at time T_n.

ɱ: Number of migration of a job.

$Q_i(t)$: The number of jobs waiting in the execution queue at the node Pi at time t.

w_i: Processing power at P_i.

$z(j_x)$: Size of job(x).

$TET_{i,t}$: Total waiting time for execution of waiting job at p_i queue.

$RET_{i,t}$: The remaining execution time of the job being processed at the P_i.

$LD_{i,t}$: Load of Pi at time t, comes from (1).

$$LD_{i,t} = TET_{i,t} + RET_{i,t} \quad (1)$$

$NLD_{i,t}$: Normalized average load in the buddy set of node P_i at time t.

BW_{ij}: Bandwidth communication between processors i and j

$ArrTime(j_x)$: Arrival time of job J_x.

$endTime(j_x)$: End time of job j_x.

$ETC(j_x, p_i)$: Estimated execution time of j_x at p_i, comes from (2).

$$\text{ETC}(j_x, p_i) = \frac{\text{ETC}(j_x, p_{std})}{w_i} \quad (2)$$

$T_{com}(j_x, p_i, p_j, t)$: The time required for transfer job j_x from p_i to p_j at time t.

$EFC(j_x, , p_i, p_j, t)$: Estimation of finish time of job j_x when transfer from p_i to p_j at time t.

if $\text{T}_{\text{com}}(\text{j}_\text{x}, \text{s}_\text{i}, \text{s}_\text{j}, \text{t}) \geq \text{LD}_{\text{j,t}}$

$$\text{EFC}(\text{j}_\text{x}, \text{s}_\text{i}, \text{s}_\text{j}, \text{t}) = \text{T}_{\text{com}}(\text{j}_\text{x}, \text{s}_\text{i}, \text{s}_\text{j}, \text{t}) + \text{ETC}(\text{j}_\text{x}, \text{s}_\text{j})$$

ELSE

$$\text{EFC}(\text{j}_\text{x}, \text{s}_\text{i}, \text{s}_\text{j}, \text{t}) = \text{LD}_{\text{j,t}} + \text{ETC}(\text{j}_\text{x}, \text{s}_\text{j}) \quad (3)$$

$B_x(p_i, p_j)$: Benefit of execution of the job j_x at p_j compared to execution at pi.

$$B_x = EFC(j_x, p_i, p_i, t) - EFC(j_x, p_i, p_j, t) \quad (4)$$

The general routine of the proposed Load balancing algorithm is in a way that when a new task enters the computational node, that node will decide based on cell's conditions that it should carry out this task itself or migrate it to another node.

This algorithm consists of several main routines:

- Determining the status of nodes
- Making decision to migrate the task
- Selecting the best node to carry out the task

A. Determining the status of nodes

The overall basis for all decisions is the status of that node. In fact, the essential criterion in deciding to send a task is the status of that node and the main criterion for selecting a node to perform the task is also the status of that node. Thus, determining the status of each node is crucial in load balance in the whole system.

In order to determine the status of each node and its neighbors for each running and migration, the information are needed to determine the status of nodes. There will be a large overhead in the system if a series of messages are exchanged between nodes to exchange their status. Thus, regular intervals are used to determine the status of nodes which are called information exchange periods. T_s which is greater than the period of time for running and migration of tasks is performed between nodes which estimates the status of nodes between these time periods. T_e tries to reduce communication overhead and have a more accurate decisions (Fig. 1).

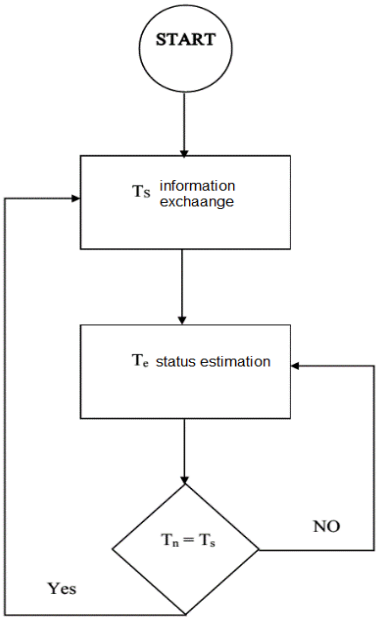


Fig. 1. Determining the status of nodes in T_n

Determining the status of the node is done in two methods:

- Information Exchange (T_s)
- State estimation (T_e)

1) Determining the status of nodes through the information exchange

Determining the status of nodes through the information exchange contains three parts of sending load information, calculating the average load and determining the status of nodes using Fuzzy Logic (Fig. 2).

a) Sending load information

All of the nodes send the information related to their load to other node in their neighboring collection in regular time periods of T_s which are called information exchange periods and receive their information. Then each node records the reviewed information from each neighboring node in neighbor table.

b) Calculate the average load

Each node calculates its average load and the load related to neighboring collection using (5) and considers the obtained index as normalized load average.

$$NLD_{i,t} = \frac{\sum_{j \in N_i} LD_{j,t}}{N \times \sum_{i \in N_i} w_i} \quad (5)$$

c) Determining the status of nodes using Fuzzy Logic

The normalized load average is considered as the point of balance and map the status of each node to one of Very light, Light, Normal, Heavy, Very heavy forms using fuzzy logic. This step is called mapping input values to fuzzy mode and the status of each node will be recorded in neighboring table.

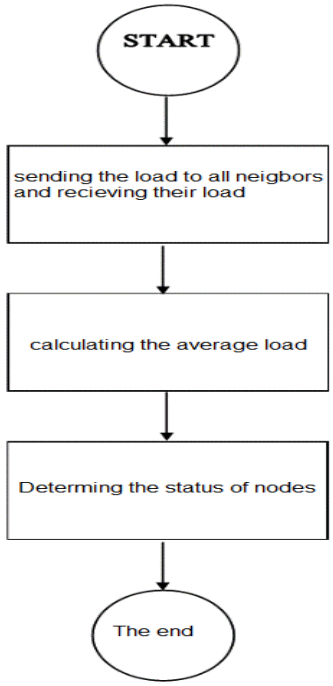


Fig. 2. Information exchange

2) Determining the status of node using estimation

If this data exchange is done in intervals with short distance, there will be a high communication overhead imposed on the system. Thus, these intervals must be increased and estimated T_e status in information exchange periods using fuzzy rules.

In order to discover these laws, the algorithm runs without estimation periods and records the data on each node. The related fuzzy rules are extracted by using Matlab software. In order to reduce communication overhead caused by data exchange, exchange periods will be increased and the status of each node using obtained rules in order to reduce errors in decision making will be estimated.

B. The decision to send task

When j_x task enters node and that node is in one of VL, L, VN states, it will be queued for processing and will be waiting to run according to respective priority and the higher rate of migration (ɱ) leads to increased running priority.

But when j_x enters node and that node is in one of H, VH states, if it is migrated from another node, it will not be accepted and rejected but if the job belongs to that node, then if the node is in VH status then it calculates the amount of its own load and the normal load and if it has a neighbor with VL status, it selects ½ of its additional load and if it has a neighbor with L status, it selects ¼ of its additional load for migration and if it has neighbors with both statuses, it sends with the same ratio to both of them. If it is in H status, it will calculate the difference between its own load and the normal load and if it has a neighbor with VL status, it will select ¼ of its additional load for migration and selects 1/8, if it has a neighbor with L status.

C. Selecting the most appropriate node to perform the task

Initially nodes with VL, L are marked in neighboring nodes with weight of 1 for L node and 2 for VL node. Then the implementation cost of performing the j_x task is estimated on specified nodes and based on $B_x(p_i, p_j)$ profit which is the difference between running cost in node p_j compared to p_i node, if this profit I positive and bigger than the threshold, a weight is given to each one of them. In this way that the higher running profit leads to higher score and the value of this profit close to the threshold will lead to having score near one. Finally, with regard to weight of node's status and the weight of running's profit in P_j node, the decision is made to send this node to P_j. the higher weight will lead to higher possibility of sending the task to that node (Fig. 3).

It should be noted that the cost of running a task may be equal in source node and the destination node which means running the task on that same node or migration it to another node may result in similar end time which has the running profit of zero and even have a little earlier end time but that task is not allowed to be migrated to that node because it imposes communication overhead on the system and the bandwidth of communication between processors is engaged even for small time.

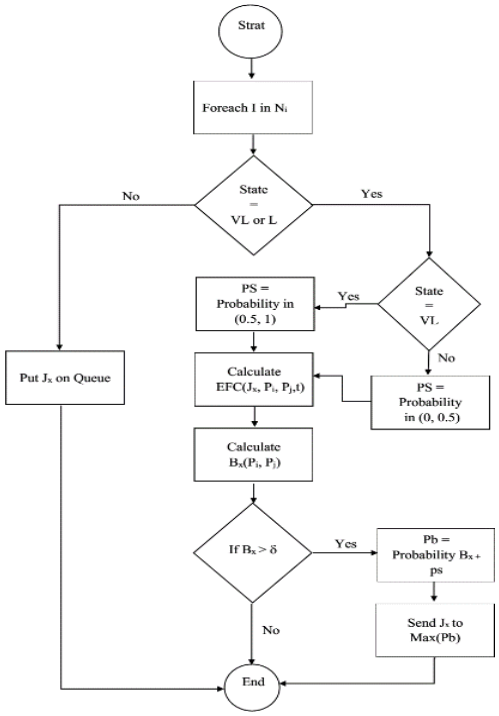


Fig. 3. Selecting the most appropriate node to perform the task

IV. Simulation and Results

Simulation is an imitation of the real world. Simulators enable program designers, instrument developers and grid developers who need to test programs, tools and services available to ensure the proper function of their program before final production and using them in real world.

The proposed FLBCA algorithm has been simulated using C# programming language in Visual Studio 2010 environment and its algorithm has been compared with MELISA algorithm.

A. Efficiency criteria

Following two criteria have been used in order to evaluate the performance of the proposed algorithm [21-23]:

1) Average response time (ART):
The average time from when the task enters the Grid until it successfully comes out of the Grid environment.

2) The average performance of computing nodes (CPU):
the ratio of working time to the total time of a system:

$$U_i = \frac{Busy_i}{Busy_i + Idle_i} \quad (6)$$

$$APU = \frac{\sum_{i=1}^{M} U_i}{M} \quad (7)$$

B. Simulation Model

This simulation is formed by 30 heterogeneous computing grids which their processing power follows random distribution in range of [1, 10] and the relation between two nodes has been formed by heterogeneous communication network in a way that their communication bandwidth is variable from 1 Mbps to 10 Mbps.

10,000 independent tasks have been used in this simulation in a way that running time of each task has been generated randomly in the range of [1, 100]. These tasks enter the system based on Poisson distribution with rate of [1, 4] and the volume of each task follows normal distribution with mean of 5 MB and standard deviation of 1 MB.

The time for information exchange (T_s) is assumed to be 20 units and the estimation time of status is considered to be 5 units

C. Simulation results

This algorithm has been evaluated in terms of performance criteria and under the effect of factors such as the number of tasks, time and period of service transition and estimation interval.

1) The effect of the works entered in the homogeneous environment

In a homogeneous environment where processing power of each CPU is 1 and the communication bandwidth between any two nodes is constant and equal to 10 Mbps, the number of tasks has been added from 0,000 to 50,000 in order to measure these factors. In these conditions where the number of tasks is 10000. The average response time is about the same among all three algorithms but with increased tasks, the efficiency of ELISA algorithm is better than MELISA algorithm and the proposed algorithm and the efficiency of FLBCA algorithm is slightly better than MELISA algorithm (Fig. 4). The total run time is about the same in all three algorithms (Fig. 5).

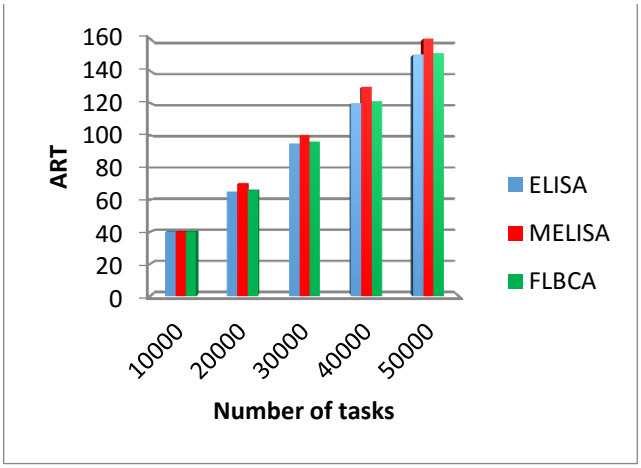


Fig. 4. The average response time in case of homogeneous

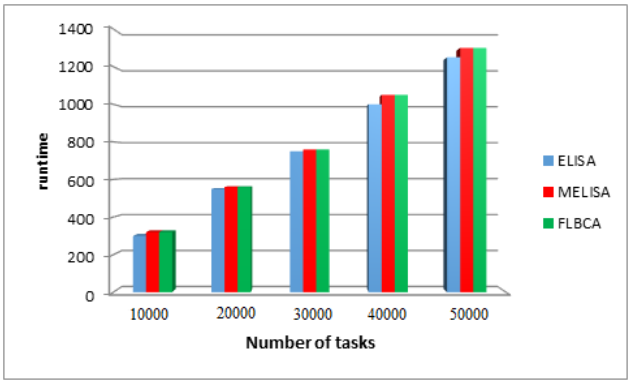


Fig. 5. The total runtime in homogeneous environment

2) The effect of tasks entered into the heterogeneous environment

Since this algorithm has been designed for heterogeneous environments, it is compared in heterogeneous environment with different processing power in the range of [1, 10] and various communication bandwidths between any two nodes in the range of 1 Mbps to 10 Mbps with ELISA and MELISA algorithms.

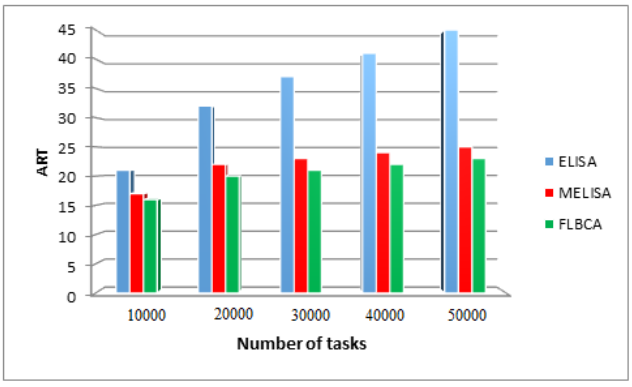


Fig. 6. Comparing the average response time in heterogeneous environment

In order to measure the effectiveness of the work, the number of tasks have been increased from 50,000 to 10,000. The average response time is much better in FLBCA and MELISA algorithms than the ELISA algorithm. The average response time is initially about the same in FLBCA and MELISA algorithms but is gets better with increasing number of tasks in FLBCA algorithm and shows better and faster decisions in fuzzy logic (Fig. 6).

The total runtime is similar among all three algorithms. The runtime is a function of the rate of entering data into the system and the runtime is about the same due to using same data (Fig. 7).

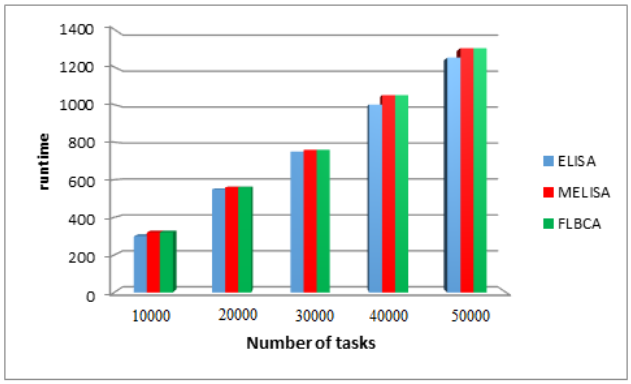


Fig. 7. Comparing the total runtime in heterogeneous environment

3) The effect of job size

This section tries to evaluate the effect of changing tasks volume from MB to 50 MB on the average response time in the proposed algorithm. The number of migration reduces and the average response time increases with increased runtime (Fig. 8).

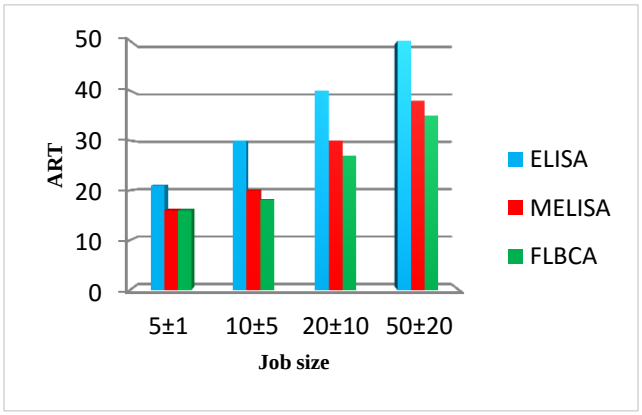


Fig. 8. Average response time for different job sizes

4) The effect of running tasks

This section tries to evaluate the effect of changing the average runtime of tasks from 10 to 150 units on the efficiency of the algorithm. The average response time increases with large rate by increasing the runtime (Fig. 9).

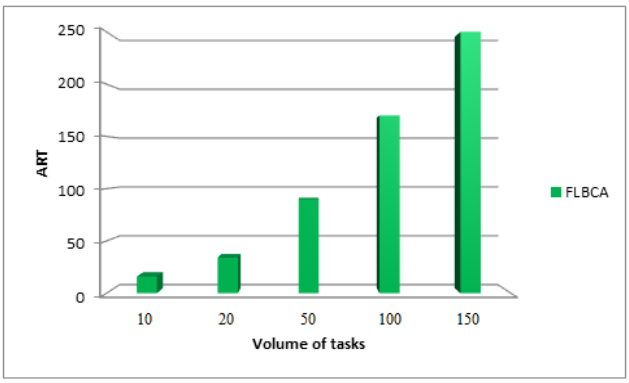


Fig. 9. The average response time for various run times

5) The effect of information exchange time

This algorithm was tested with different times in range of 2 units to 40 units in order to find a time for information exchange which is suitable in perspective of efficiency criteria (Fig. 10).

The accuracy of information reduces with increasing time of the information exchange. The distribute the load between nodes is done with less precision and average response time increases for this purpose

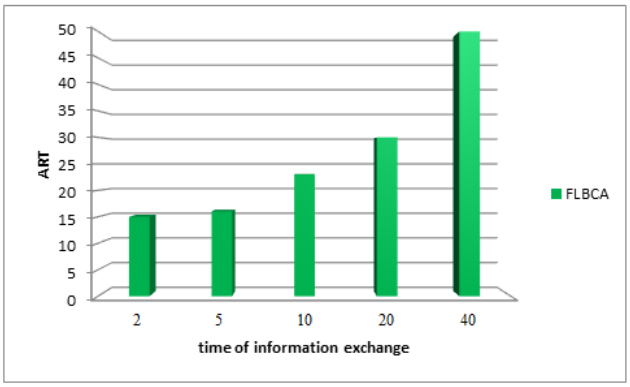


Fig. 10. The average response time for different times of exchanging information

6) The effect of state estimation time

In order to access a proper estimation time from the point of efficiency criteria, the exchange time must be considered to be 20 units by default. Then by changing the estimation time in different intervals from 2 units to 10 units, the most effective time will be found (Fig. 11).

The average response time increases by reducing estimation time due to increased computational overhead and the most optimal time is reached at times of 4 and 5 and the average response time increases again by increasing this time due to reduced accuracy of data.

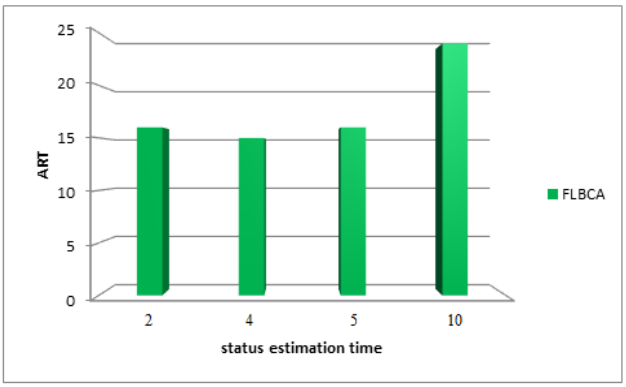


Fig. 11. Comparing the average response time for different times

V. CONCLUSION

A load balancing model has been provided in this research for grid computing environment which is called FLBCA. This algorithm tries to meet needs and characteristic required for a load balancing algorithm such as stability, versatility, transparency from the user's standpoint and minimize the communication overhead as much as possible. Since the fuzzy logic system has the ability to deal with imprecision and uncertainty. Algorithm based on fuzzy logic has been proposed for dynamic load balancing in computing grid. The main purpose of using this algorithm is increasing efficiency and productivity of Grid system which will lead to reducing organization's costs and increasing productivity and saving energy and since using energy efficiently saves it which is the top priority in our life today and helps to protect the environment and nature.

Among the tasks carried out in this paper and with respect to the fact that the issue of resource management is still discussed about in Grid as well as the importance of load balancing in it and the use of cellular automata which can be a good platform for the design of load balancing algorithms, it seems hierarchical cellular automata is an Appropriate structure for design of these types of algorithms in future which can be provided a better view of the whole condition of the Grid System. In addition to this, the use of fuzzy logic which leads to increased accuracy of decision-making in uncertain environments can be used in to improve the efficiency of parallel algorithms. The combination of fuzzy logic and cellular automata can be a good technique for a lot of parallel algorithms.

REFERENCES

[1] M. Fathi, F. Mehryari., "the principles and concepts of grid computing technology and its applications in various fields", Noorpardazan, .10, 2009 - Tehran, pp.

[2] M. Amini Salehi, H. Deldari, load balancing in the grid resources using agent-based resource management, Master's thesis, Department of Computer Engineering, Ferdowsi University of Mashhad, 2005.

[3] S. Ghanbari, balance and self-organization in the grid computing using learning automata, Master's thesis, Department of Computer Engineering, Amirkabir University of Technology, 2004.

[4] R. Tlili, Y. Slimani, A Hierarchical Dynamic Load Balancing Strategy for Distributed Data Mining, International Journal of Advanced Science and Technology, Vol. 39, February. 2012

[5] S. Adabi, reducing power consumption in wireless sensor networks based on cellular automata, Master's thesis, Department of Computer Engineering, Islamic Azad University of Science and Research Branch, 2010.

[6] L. Anand, D. Ghose, V. Mani, ELISA: An Estimated Load Information Scheduling Algorithm for Distributed Computing Systems, An International computers & mathematics with applications, 1999.

[7] L. Rostami, A. Rahmani, An adaptive Load Balancing Algorithm with use of cellular Automata for Computational Grid Systems, Euro-Par 2011 Parallel Processing, Lecture Notes in Computer Science Volume 6852, 2011, pp 419-430.

[8] A. Karimi, F. Zarafshan, A. Jantan, Anew Fuzzy Approach for Dynamic Load Balancing Algorithm, International Journal of Computer Science and Information Security, Vol. 6, No. 1, 2009.

[9] M. Marinov, Intuitinistic Fuzzy Load balancing in cloud computing, 8th Int. Workshop on IFSs, Ocy 2012.

[10] S. Mousavi Nejad, S. Mortazavi, B. Vosoughi Vahdat, Design and set optimal control and intelligent load balancing based on fuzzy logic in distributed systems, first regional conference on new approaches in computer engineering, 2011.

[11] I. Foster, and C. Kesselman, "The Grid: Blueprint for a New Computing Infrastructure" Morgan Kaufmann and Elsevier, Second Edition, USA, ISBN: 1-55860-933-4, 2004.

[12] I. FOSTER, C. KESSELMAN, M. NICK J, S. TUECKE, "Grid services for distributed system integration,« vol. 35, 6, 2002.

[13] C. J., K. E.,. L. M. Anderson D P., "SETI @ home: an experiment in public-resource computing," vol. 45 (11).

[14] S. Graupner, J. Pruyne, S. Singhal, "Making the Utility Data Center a Power Station for the Enterprise Grid," 2003.

[15] J. Liu, X. Jin, and Y. Wang, Agent-Based Load Balancing on Homogeneous Minigrids: Macroscopic Modeling and Characterization, IEEE TRANS.IONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 16 (7) 2005.

[16] L.SUN CHEUNG, A fuzzy approach to load balancing in a distributed object computing network, In Proc. Of 6th Int IEEE Conf. HPDC, 2000.

[17] T. L. Casavants and J. G. Kuhl, A taxonomy of scheduling in general-purpose distributed computing systems, IEEE Trans. Software Eng., Vol. SE-14 (2), pp. 141-154, 1988.

[18] H. Kameda, J. Li, C. Kim, and Y. Zhang, Optimal Load Balancing in Distributed Computer Systems. London, U.K .: Springer-Verlag, 1997.

[19] Z. Zeng and B. Veeravalli, Rate-Based and Queue-Based Dynamic Load Balancing Algorithms in Distributed Systems, 10th Int. Conference on Parallel and Distributed Systems, IEEE 2000.

[20] Abubakar, Haroon Rashid and Usman Aftab, Evaluation of Load Balancing Strategies, National Conference on Emerging Technologies 2004.

[21] J.Cao, Daniel P. Spooner, Agent-Based Grid Load Balancing Using Performance-Driven Task Scheduling, In Proc. of 17th IEEE Int. Parallel & Distributed Processing Symposium (IPDPS 2003), Nice, France, April 2003.

[22] A. Shaout and P. McAuliffe, Job scheduling using fuzzy load balancing in distributed system, in Proc. of 6st conf ICPAD, 1998.

[23] J. Liu ,, X. Jin, and Y. Wang, Agent-Based Load Balancing on Homogeneous Minigrids: Macroscopic Modeling and Characterization, IEEE TRANS.IONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 16 (7), 2005.

Goal Model Integration for Tailoring Product Line Development Processes

Arfan Mansoor
Software Architectures and Product Line Group
Ilmenau University of Technology
Ilmenau, 98693, Germany

Detlef Streitferdt
Software Architectures and Product Line Group
Ilmenau University of Technology
Ilmenau, 98693, Germany

Muhammad Kashif Hanif
Department of Computer Science
Government College University
Faisalabad, 38000, Pakistan

***Abstract*—Many companies rely on the promised benefits of product lines, targeting systems between fully custom made software and mass products. Such customized mass products account for a large number of applications automatically derived from a product line. This results in the special importance of product lines for companies with a large part of their product portfolio based on their product line. The success of product line development efforts is highly dependent on tailoring the development process. This paper presents an integrative model of influence factors to tailor product line development processes according to different project needs, organizational goals, individual goals of the developers or constraints of the environment. This model integrates goal models, SPEM models and requirements to tailor development processes.**

***Keywords*—Goal model; Product Line; Development Process; Process Line**

I. Introduction

Many companies rely on the promised benefits of product lines, targeting systems between fully custom made software and mass products. Such customized mass products account for a large number of applications automatically derived from a product line. This results in the special importance of product lines for companies with a large part of their product portfolio based on their product line. The success of product line development efforts is highly dependent on tailoring the development process. This paper presents an integrative model of influence factors to tailor product line development processes according to different project needs, organizational goals, individual goals of the developers or constraints of the environment. The model integrates goal models, SPEM models and requirements to tailor development processes.

Software systems developed based on the product line approach result in systems between custom made software and systems developed for a mass market. Thus, software product lines are customized mass products. The architecture of a product line consists of a core and diverse variable components. Any members of a product line are based on its core and one or more variable components. Core and variable components are pre-developed what results in the special usage of a product line. The customer simply selects and may parametrized the desired features of the future system. Based on the product line, the system (in more detail, the software application) will be automatically generated. The effort for the development of a product line core and its variable components will reach a break even point starting from four [1] up to five [2] sold applications. This is mainly due to the large development efforts for the core of the product line, the product line training needed for the developers, the migration effort for companies to go towards the product line concept and the process maturity level needed for product line development [3]. The efforts for product line specific development processes are higher than the efforts for the development of standard systems and such development processes need to be tailored towards the project environment of the development team [4], [5]. The survey of 273 software projects in[6] revealed a potential of reducing the development effort up to 21% by raising the CMM level by one. This shows the big potential of defined and tailored development processes. For the remainder of this paper the terms method and process are used according to the Software & Systems Process Engineering Metamodel (SPEM) of the Object Management Group (OMG). A method is a reusable and goal oriented procedure made of several steps, referred to as tasks. A process is a sequence of tasks together with the timing information for the sequence. Thus, a process would contain all the timed steps needed to develop a product line. As an example, a review is taken from the method library and reused at different occasions in the process to validate the documents developed along the product line development process. Ten product line case studies have been analysed in [2] out of the domains embedded, oil and gas, finances, mobile communications, telecommunications, multimedia, and the medical domain. All the case studies use a twofold development process, with a domain engineering (development of the product line itself) and an application engineering (development of applications based on the product line) phase, as shown in figure 1. Both phases are further subdivided in a requirements, a design, a realization and a testing phase. The common assets, managed in a repository, are in between both phases. They are developed in the domain engineering phase and used in the application engineering phase.

The challenges are the development methods and processes, which have been individually and manually defined by all case studies in [2] as the project proceeded. Although guidelines for the development of product lines have been developed [2], detailed recommendations for the tailoring step of a development process are still missing. It is not yet clear whether and to what degree a given development process will

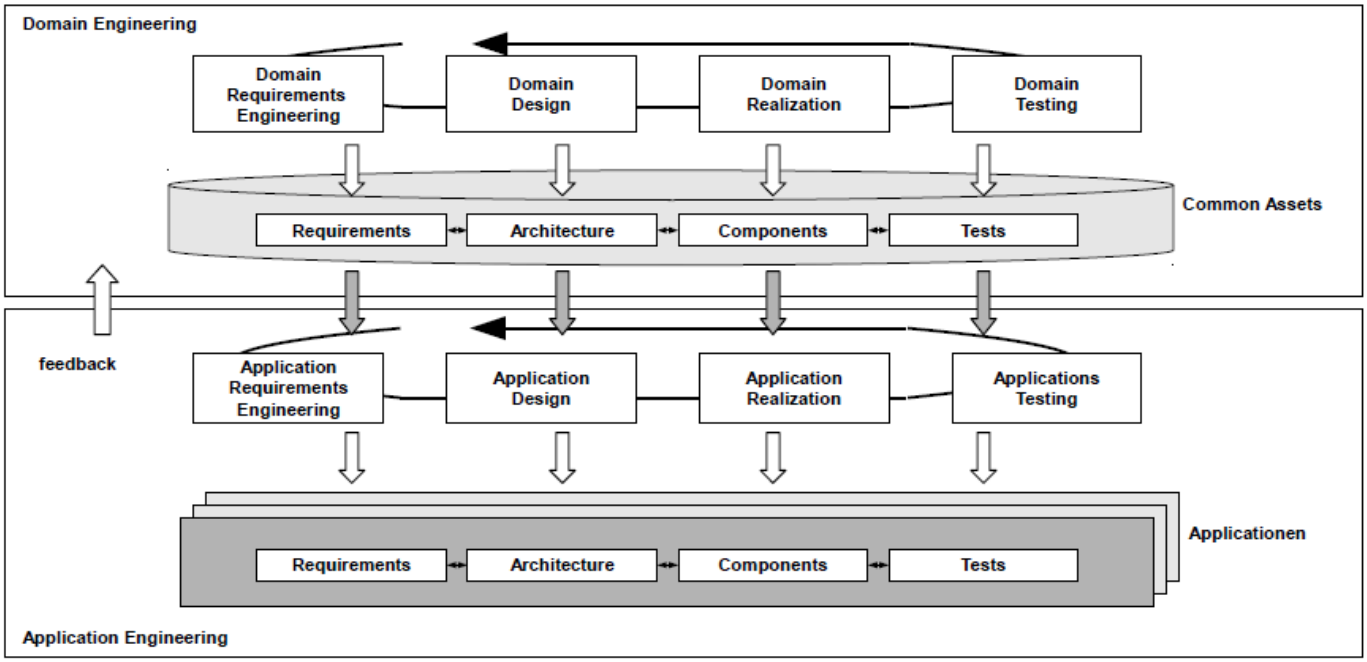


Fig. 1: Product Line Development Process

fit to its development environment. A structured approach to address this savings potential could be defined attributes together with a model to optimize the tailoring step of the development process for product lines. Therefore, a tailoring meta-model with a set of attributes to enhance the tailoring step with an optimization towards the presented attributes is presented.

The rest of the paper is organized as follow: section II discusses the need of integration model where the requirements on tailoring product line development processes are manifold. In section III requirements are divided into two main parts and the factors effecting the tailoring development processes are determined. Section IV presents a meta-model for tailoring development processes and a pseudo-code for the selection of elements. Finally, last section concludes the paper.

II. The Need of Integration Model

The product line development method PuLSE as presented in [7] is equipped with the PuLSE Baselining and Customization (PuLSE-BC) [4] procedure to tailor PuLSE towards the needs of an organization. Any tailoring decisions are bound to the variable parts of the development process. The criteria for tailoring are based on organizational and project domain issues. Such manually elicited criteria result in the variability of the development process. Pulse-BC is managing this variability in an own model. A further refinement of tailoring product line development processes is presented in [8]. Here, a product line for development processes is proposed, referred to as process line. The requirements of the development processes in this process line are based on an analysis of current and future products, projects and processes. Thus, the processes are optimized towards the products and projects, to derive a tailored development process based on the process line. Tailoring is realized with prioritized attributes, with which the resulting elements of the product, process and project analysed are ranked. An automated analysis of the underlying models is not yet realized what also hinders the efficient analysis of different scenarios in different domains. The company specific strategy and the goals of groups as well as individual developers, referred to as soft attributes are also missing. Nevertheless such attributes are important since personal factors influence the success of development process changes to a larger degree than technological challenges [9], [10],[11]. As a result, a process line model based on products, processes and project data in relation to models of the company strategy and developer goals is needed. Here, the relations of the model elements and features of the process line are highly important to be able to realize its variability [12]. In addition, there is also need of a complete model of the attributes to enable an enhanced assessment of derived development processes. Development process like the V-Model XT, SCRUM or OpenUP are targeting single system development efforts. Nonetheless parts of the methods are taken for the product line development. In [13] parts of an agile development process have been used for the product line development in a large company (SAP). Again, tailoring of development processes for product lines is an important success factor. As described in [13] but not yet accomplished, the strategic and business goals of an organization need to be part of the development process. The selection of process steps should be traceable to the business and strategic goals. Without such traces development processes cannot be fully analysed and tailored. Thus, the business goals need to be part of the above described process line. In [14] the tailorability of the V-Model XT towards product line development is analysed. Based on this work a process line was developed and a V-Model XT development processes could be derived based on the process line. Unfortunately, the selection of supporting

tools for the development process is still left to the project manager and/or developer and the selection of tools is bound to the knowledge about their advantages and drawbacks, what is currently not part of the model of process lines. The analysis of product line approaches emphasizes the relevance of tools for the success of a product line development project. All the presented approaches in this paper are based on the product line development concept shown in figure 1 and offer ideas to relate the development process to the development environment. Although, none of the approaches is able to offer a complete model of a tailorable development process together with the elements/components of the development environment. Here, the analysis and assessment of development processes need to include tools, since they strongly influence the expected effort of a product line development project. The relation of decisions to the original goals of the decisions can be realized with goal models [15]. Goal oriented business processes with variabilities are presented in [16]. Such models could be used as in [17] to analyse and assess the chances of success with the Goal-Question-Metric (GQM) method for product line development projects. For the tailoring step of a developers environment the in influential factors and attributes are still missing for process lines, but could be realized using a goal model. Thus, a comprehensive view onto product line development domain would be possible. Finally an integrative model for the description of stakeholder needs and goals in relation to the development process artifacts and the development environment specifics is needed, to be able to analyse potential influences of changing goals early in the project development.

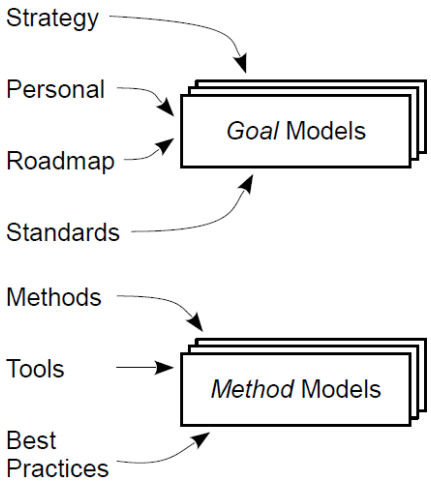


Fig. 2: Goal and Method Models

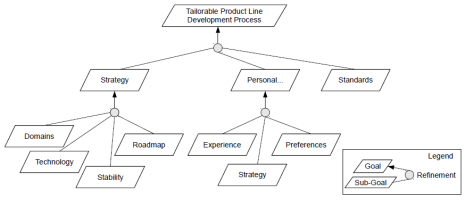


Fig. 3: Integrated Goal Model

III. Tailoring Development Processes

As stated in the previous section the requirements on tailoring product line development processes are manifold. Here, these requirements are divided in two main parts

1) The goal model based requirements
2) The method model based requirements

The following categories and parts of the two models are based on own experiences in industrial projects and lessons learned within student software development projects. First, the identification of influence factors that can be described by goal models contains soft factors, as shown in figure 2.

Based on experience, it is estimated that about 70% of the challenges throughout the software development project can be traced back to such soft factors. Thus, addressing such factors can influence the success of a project by a large degree. As shown in figure, 3, two top level factors are refined with a goal model.

The **strategy**of a company is very important when comes to the initial decision for or against a product line. Thus, the following sub-goals as refinement of the strategy are tightly connected to the product line development.

- The target **domain** or domains of the products that will be developed rule about the product line approach. New domains or domains that will be abandoned in the future need to be known and elicited in the requirements engineering phase. Of course, these requirements might have a large impact on the architecture of the product line, specifically to the core and the variabilities of the product line.
- Any strategic choice of the **technology** influences the future constraints (performance, memory, available development environment, available compilers) of the system and thus, constraints for the product line. For example, the realization of variabilities with the C language has reduced capabilities compared to C++.
- **Stability** of the strategy. For new companies this is highly relevant. The strategy is subject of a high risk for changes. Thus, this goal influences the overall feasibility of the product line development.
- The **roadmap** includes the timing for the release of product features. For each release a set of features is identified. The length (way into the future) of the roadmap influences the technological choices and the re-development of the product line. Due to technological changes, fluctuation of employees (and with them the knowledge) and unforeseen requirements the implementation of the architecture of a product line needs to be adapted to this new environment. The roadmap needs to address these large and periodic updates.

The personal factors also have a large impact onto the other elements in the goal model. The personal goals are coupled with a stakeholder model of the involved persons in a software project. Each stakeholders should have an own personal goal model reflecting his/her position towards the product line

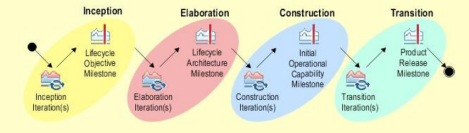


Fig. 4: OpenUP Overview

Fig. 5: Developer Role in OpenUp

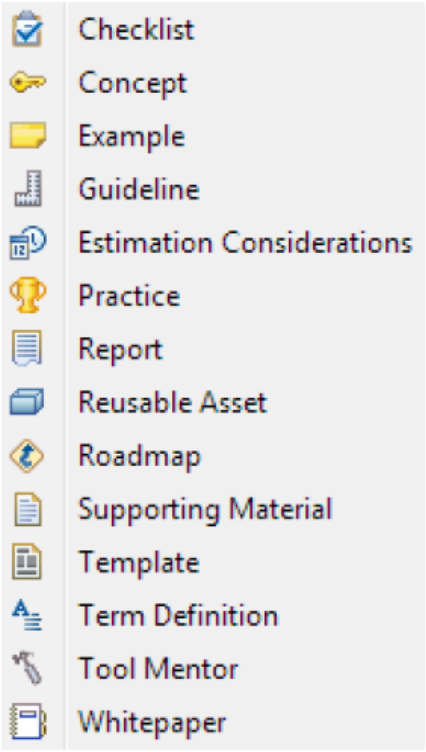


Fig. 6: OpenUP Guidance for SPEM elements

development process. Since this is a very personal information it is recommended to keep this model private but use the information in correlation with the other models (strategy and standards) as well as use the results of a model analysis as input for the periodic discussions with the management within the company and/or of the respective project.

- Each stakeholders own experience should be related to the role descriptions of the basic development processes (e.g., OpenUp, SCRUM). Besides the potentials for further personal development, such an experience level should be related to the project roles (and their skills) which are attached to each development step. For exchangeable development steps, experiences set the rules on which step to take.
- Each stakeholder has preferences for application domains or technological choices. There are also preferences for methods used along the development process or for specific templates to be used for the deliverables of the development process. These preferences will influence the choices of the method and development process parts of the product line.
- Each stakeholder might (or should) have an own strategy in contrast to the company strategy. The alignment of the strategy of all different stakeholders is impossible, due to the private nature of this information. As with the experience, the awareness of the other goals and their correlation to the own strategy is an important step towards the integration into a developer group and a good starting point to develop an own roadmap. The individual analysis of the own strategy is a good point to think about the own position in the company and/or to better understand the own position.

Standards will influence the technology goals for the strategic planning and they recommend or require technologies and/or tools. For example, the safety standard IEC61508 recommends test case generation tools. Standards could also require a specific development process structure and give recommendations or require development methods. The lower part of figure 2 shows the method models. Here, we use SPEM to describe all the needed parts of the methods, processes and best practices. As a SPEM implementation, OpenUP is shown in figure 4.

OpenUP is an open source development process for standard applications, the complete extension of OpenUP towards a product line is a future work package. Nevertheless this process is taken as tailoring example to address the above mentioned goals. The development process is split into four iterative phases. Compared to figure 1, the requirements is equivalent to the inception phase, the design is equivalent to the elaboration phase and the realization is equivalent to the construction phase. The testing steps are present in each iteration of the OpenUP process and at the first sight the testing phase in figure 1 does not match the OpenUP transition phase, but this testing phase is meant to be the final system test with an iterative testing approach as well and thus, the two models are comparable. For each of the development steps in figure 4 parts of the method steps of the OpenUP method library are taken and put together. Each task has its responsible roles attached and each role has its tasks attached. As shown in figure 5 the developer role is required to perform the five given tasks and is also responsible for the four deliverables. The last of the SPEM elements relevant for the process tailoring step are the guidances. As shown in figure 6 there are 14 guidance types which can be used to support any SPEM element, e. g., a task.

IV. Tailoring Meta-model

Based on the above mentioned relations between goal models, method/process models and requirements, the proposed meta-model as shown in figure 7. The **Element** abstracts the **Goal** model elements, the **MethodElements** of SPEM, and the **Requirement** elements found in most of the meta-models of requirements management tools like Polarion. The meta-model now allows to connect any element using links of the abstract

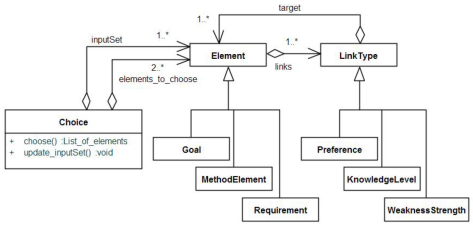


Fig. 7: Meta-model for Development Process Tailoring

1: ERMAP of elements_to_choose,rank
2: **for all** A in inputSet **do**
3: $linkSubSet =$
4: $getLinkTypesFromTo(A, elements\_to\_choose)$
5: $adjustRank(linkSubSet)$
6: **end for**
7: $select\_SPEM\_based(ermap)$

Fig. 8: choose Pseudo-code

LinkType. Currently the following link types are defined:

- Preferences - Are used to indicate a stakeholders preference for a given element (e. g., a developer might have a preference for a text editor which is part of the guidances of the process model). The preference link can have values between -100% (aversion against an element) up to +100% (this element is vitally important for a stakeholder)
- KnowledgeLevel - This link indicates the level of confidence a stakeholder might have with an element in our model. The knowledge level link is divided in two categories. The knowledge as user of an element between 0% (the stakeholder knows nothing about an element) and 50% (the stakeholder knows everything to use and work with an element). The knowledge as teacher for an element my have values between 51% (the stakeholder has taught the use of an element at least once) and 100% (the stakeholder is an experienced teacher with more than 5 years of teaching experience).
- WeaknessStrength - Any element might weaken or strengthen another element. For example, the presence of a requirement for safety in the medical domain will result in high documentation demands what in consequence will strengthen the quality of the final product and at the same time weaken a fast delivery of the product. The weakness/strength link can have values between -100% (the source element will disable/weakens the target element) up to +100% (the source element requires/strengthens the target element. Thus, the target element be comes mandatory)

To work with the product line approach, variabilities are needed, as discussed in the first sections. The variability of the process is modelled with the SPEM content variability types (contributes, extends, replaces, extends and replaces) for the elements of a SPEM model. To trigger this variability of the process model, the **Choice** is introduced in the tailoring meta-model in figure 7. This has an input set of elements influencing the choice. This input set will be updated by the ***update_inputSet()*** method whenever the choices are going to be evaluated. This method will search for elements with target links present in the ***elements_to_choose*** list and will update the ***inputSet*** list accordingly. Once the ***update_inputSet()*** method has been executed the ***choose()*** method can follow with its execution to calculate the variant based on the given input elements.

The pseudo-code in figure 8 shows how to calculate the choice of elements. First a map of elements and its ranking is created.

For all the elements in the list of input elements, the elements which have links to elements in the ***elements_to_choose*** list are filtered out . This is accomplished by the ***getLinkTypesFromTo*** method which stores its results in a list of links as subset of the original ***links*** list of the ***Element*** type. This list is then taken as input for the ***adjustRank*** method which in the current version simply adds the values for the preferences, knowledge level and weakness/strength values, to the ***ermap*** rankings discussed in the last section. Finally, a selection of choices based on the rankings and the SPEM models constraints is made. This meta-model can be extended in two ways:

1) First, any additional elements can be added to this meta-model to address future models which need to be integrated in the tailoring process.
2) Second, the link types can be extended by new links needed in the future.

V. Conclusion

In this paper, the current state of the product line development domain and the challenges are discussed when it comes to the development processes which need to be adapted to the specific needs of the development teams. Tailoring product line development processes has been identified to enable large savings for the domain engineering as well as application engineering phase of product line development projects. For an integrative approach to process line tailoring, a tailoring meta-model is proposed which includes goal models, SPEM process models as wells as requirements. With this model stakeholder specific goals can be used to support binding a variable part of the development process. This support addresses soft factors as well as concrete requirements. Future research work will be spent to further elicit attributes of different domains influencing the development process. In addition the enhancement of the few variable process steps in OpenUP towards a complete process line will also be subject of future research efforts.

VI. Acknowledgment

We acknowledge support for the Article Processing Charge by the German Research Foundation and the Open Access Publication Fund of the Technische Universit Ilmenau.

References

[1] D. M. Weiss and C. T. R. Lai, *Software Product-line Engineering: A Family-based Software Development Process*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[2] F. J. v. d. Linden, K. Schmid, and E. Rommes, *Software Product Lines in Action: The Best Industrial Practice in Product Line Engineering*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[3] L. Brownsword and P. Clements, "A case study in successful product line development," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU/SEI-96-TR-016, 1996. [Online]. Available: http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=12587

[4] K. Schmid and T. Widen, "Customizing the pulse^{tm} product line approach to the demands of an organization," in *Software Process Technology, 7th European Workshop, EWSPT 2000, Kaprun, Austria, February 21-25, 2000, Proceedings*, 2000, pp. 221–238. [Online]. Available: http://dx.doi.org/10.1007/BFb0095031

[5] V. R. Basili and H. D. Rombach, "Tailoring the software process to project goals and environments," in *Proceedings of the 9th International Conference on Software Engineering*, ser. ICSE '87. Los Alamitos, CA, USA: IEEE Computer Society Press, 1987, pp. 345–357. [Online]. Available: http://dl.acm.org/citation.cfm?id=41765.41804

[6] B. K. Clark, "The effects of software process maturity on software development effort," 1997.

[7] J. Bayer, O. Flege, P. Knauber, R. Laqua, D. Muthig, K. Schmid, T. Widen, and J.-M. DeBaud, "Pulse: A methodology to develop software product lines," in *Proceedings of the 1999 Symposium on Software Reusability*, ser. SSR '99. New York, NY, USA: ACM, 1999, pp. 122–131. [Online]. Available: http://doi.acm.org/10.1145/303008.303063

[8] O. Armbrust, M. Katahira, Y. Miyamoto, J. Münch, H. Nakao, and A. Ocampo, "Scoping software process lines," *Softw. Process*, vol. 14, no. 3, pp. 181–197, May 2009. [Online]. Available: http://dx.doi.org/10.1002/spip.v14:3

[9] M. Inoki and Y. Fukazawa, "Software product line evolution method based on kaizen approach," in *Proceedings of the 2007 ACM Symposium on Applied Computing*, ser. SAC '07. New York, NY, USA: ACM, 2007, pp. 1207–1214. [Online]. Available: http://doi.acm.org/10.1145/1244002.1244266

[10] D. Stelzer and W. Mellis, "Success factors of organizational change in software process improvement," *Software Process: Improvement and Practice*, vol. 4, no. 4, pp. 227–250, 1998. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1099-1670(199812)4:4¡227::AID-SPIP106¿3.0.CO;2-1

[11] M. Niazi, D. Wilson, and D. Zowghi, "Critical success factors for software process improvement implementation: an empirical study," *Software Process: Improvement and Practice*, vol. 11, no. 2, pp. 193–211, 2006. [Online]. Available: http://dx.doi.org/10.1002/spip.261

[12] D. Benavides, S. Segura, and A. Ruiz-Cortés, "Automated analysis of feature models 20 years later: A literature review," *Inf. Syst.*, vol. 35, no. 6, pp. 615–636, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.is.2010.01.001

[13] B. Blau and T. Hildenbrand, "Product line engineering in large-scale lean and agile software product development environments - towards a hybrid approach to decentral control and managed reuse," in *Availability, Reliability and Security (ARES), 2011 Sixth International Conference on*, Aug 2011, pp. 404–408.

[14] T. Ternite, "Process lines: A product line approach designed for process model development," in *Software Engineering and Advanced Applications, 2009. SEAA '09. 35th Euromicro Conference on*, Aug 2009, pp. 173–180.

[15] A. van Lamsweerde, *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Wiley, 2009. [Online]. Available: http://books.google.de/books?id=AYk_AQAAIAAJ

[16] E. Santos, J. Castro, J. Sánchez, and O. Pastor, "A goal-oriented approach for variability in BPMN," in *Anais do WER10 - Workshop em Engenharia de Requisitos, Cuenca, Ecuador, April 12-13, 2010*, 2010.

[17] B. Geppert and D. M. Weiss, "Goal-oriented assessment of product-line domains," in *Software Metrics Symposium, 2003. Proceedings. Ninth International*, Sept 2003, pp. 180–188.

Cuckoo Search Optimization for Reduction of a Greenhouse Climate Model

Hasni Abdelhafid
Faculty of Technology, University of Bechar
BP 417, 08000, Algeria
Email:hasni_ haf@yahoo.fr

Sehli Abdelkrim
Faculty of Technology, University of Bechar
BP 417, 08000, Algeria

Haffane Ahmed
Faculty of Technology, University of Bechar
BP 417, 08000, Algeria

Draoui Belkacem
Faculty of Technology, University of Bechar
BP 417, 08000, Algeria

***Abstract*—Greenhouse climate and crop models and specially reduced models are necessary for bettering environmental management and control ability. In this paper, we present a new metaheuristic method, called Cuckoo Search (CS) algorithm, established on the life of a bird family for selecting the parameters of a reduced model which optimizes their choice by minimizing a cost function. The reduced model was already developed for control purposes and published in the literature. The proposed models target at simulating and predicting the greenhouse environment. [?]. This study focuses on the dynamical behaviors of the inside air temperature and pressure using ventilation. Some experimental results are used for model validation, the greenhouse being automated with actuators and sensors connected to a greenhouse control system on the cuckoo search methods to determine the best set of parameters allowing for the convergence of a criteria based on the difference between calculated and observed state variables (inside air temperature and water vapour pressure content). The results shown that the tested Cuckoo Search algorithm allows for a faster convergence towards the optimal solution than classical optimization methods.**

***Keywords*—optimization; cuckoo search; greenhouses; metaheuristics; climate models.**

I. Introduction

Humanity is going through a crisis which is characterized by a number of problems, the most serious may be mentioned which is that of the highly uneven distribution of agricultural products between different nations. Indeed nearly two-thirds of humanity now living in situation of poverty and deprivation fault of the inadequate use of scientific and technological instruments. In the context of sustainable development, the fight against this serious problem of unequal distribution of agricultural products is part of a social challenge accompanied by a scientific challenge. Indeed many searches have been working for several years studying various technical and scientific means a quantitative improvement of agricultural production. Crops under shelter and specifically greenhouse crops have experienced during the last thirty years a significant expansion. The greenhouses are closed spaces, closed by translucent walls to obtain for agricultural production, better environmental conditions that natural conditions. Thus the greenhouse is a way to transform local outdoor conditions in a more favorable microclimate for plant growth. The technology (heating, cooling, control computers etc ...) allowed the improvement of greenhouses so that they become increasingly sophisticated

The objective of this work is to optimize the mathematical model[**?**] of the greenhouse to identify some physical parameters of the above model via the Cuckoo Search (CS). This new metaheuristic search algorithm, has been developed by Yang and Deb (2009).

II. Experimental data

All the experimental data used in this work have been collected between May 14 and 22 1991, in a 416m2 double roof plastic house occupied by a tomato-crop and situated near Avignon in the South of France.

III. Formulation of the problem

Optimization is a phenomenon of selecting the proper variable from given variables, defining an objective function in a given planning space, and searching for a minimum (or a maximum) value within boundary constraints. The objective of this work is to optimize a reduced greenhouse model which controlled variables are interior temperature and humidity and actuators are the vapor system, the vent opening, the soil and the air heating.

Heat and water vapour balances have been first formulated in order to get the key equations of the complete model. Then exacting equations have been added to complete the model.

TABLE I: Notations

Q_s	Air heating loads (Wm^{-2})
Q_s	Soil heating loads(Wm^{-2})
φ_l	Injected evaporative cooling by fog system (Wm^{-2})
PT_i	Water vapor saturation pressure at$T_i(hPa)$
V	Wind speed (m/s)
s	Vents opening surface (m^2)
T_i	Indoor temperature $(^\circ C)$
T_e	outdoor temperature $(^\circ C)$
P_i	Indoor pressure(hPa)
P_e	outdoor pressure(hPa)
R_g	Outside global radiation (Wm^{-2})

IV. Problem of identification

As it described with supplementary details in a precedent paper[**?**], in the case of greenhouse the identification techniques need a system approach of the mass and thermal transfers which can be qualified by four kinds of variables describing the greenhouse and its environment: the entry vector , which describes the initial conditions from which the system evolves; the output vector , or the set of state variables which can be observed and measured; the current state of the system , which includes the state variables of the system evolving as a function of time and the vectors of unknown system parameters. The dynamic behavior of the system can be described by a ensemble of tow equations:

A state equation:

$$\frac{dX(t)}{dt} = f[X(t), U(t), P] \tag{1}$$

An observation equation:

$$Y(t) = g[X(t), U(t), P] \tag{2}$$

V. Physical modeling of greenhouse climate

A. The heat and water vapour balances

In order to reduce the system order of the thermal model, an empirical approach based on considerations about the characteristic time scales of each thermal component of the system is considered, two main components are examined:

- The soil and hefty structural elements. With characteristic time scale much longer than the observation time scale. They will be jointly gathered beneath the form of a virtual thermal mass characterized by the virtual temperature T_m and thermal capacity C_m.
- The crop greenhouse superstructure and the inclosed air space, which characteristic time scale (τ_c) is feeble ($200 < \tau_c < 500s$) and enough like to the observance time ladder ($3600sor900s$),which can be typify by the temperature T_i and water vapour pressure P_i.

The equation (**??**) represente the the equation of the virtual thermal mass [**?**] :

$$C_m \frac{dT_m}{dt} = h(T_i - T_m) + Q_s + \beta R_g \tag{3}$$

where the first term on the right hand side is the heat exchanged with the greenhouse air, the second one is the soil heating flux and the last one, the solar gain directly absorbed by the thermal mass.

Ignoring air inertia in front of the hefty structure, one can perform the air thermal balance as follows [**?**]:

$$0 = \alpha R_g + Q_a + h(T_m - T_i) + K(T_e - T_i) + K_l(P_e - P_i) \tag{4}$$

where the first term on the right hand side is the solar gain, the second one the air heating, the third one the thermal exchange with the thermal mass, the fourth one is the overall heat exchange between inside and outside and the fifth and last term represents the sensible and latent heat exchanges by ventilation and leakages.

The air water vapour balance takes into consideration the crop transpiration, the water vapour added by fogging and the interchange with exterior, it can be depict by the equation(**??**) [**?**]:

$$C_l \frac{dP_i}{dt} = A\tau R_g + B(PT_i - P_i) - K_l(P_i - P_e) + \varphi_l \tag{5}$$

Whither the first term of the right hand side depicts the crop transpiration (merely described as a linear function of overall radiation and saturation deficit), the second one the interchanges by ventilation and the final one the contribution of the fog system [**?**].

B. Solving the equations

Simultaneous integration of the equations of energy ((**??**) and (**??**)) and water vapour balances (**??**) leads to a system of tree equations with tree unknowns (T_m, T_i, P_i) who may be introduc in a recursive form as a function of the past (time n), the instantaneous input vectors (R_g, T_0, V, P_0, PT_i), of the command variable (Q_s, Q_a, φ_l) and of model parameters (inclusive in the matrices line) which are partially to be identified.

The complete system can then be represented is as follows [**?**]:

$$\begin{aligned} P_{i(n+1)} &= P_{i(n)} exp(-\xi\Delta t) + (1 - exp(-\xi\Delta t)) \dots \\ &\times (\frac{rSB\gamma\tau'}{\xi}\frac{\chi}{\xi}\frac{\gamma SB}{\xi}\frac{\gamma S}{\xi}) \times (R_g, P_e, PT_i, \varphi_l)' \end{aligned} \tag{6}$$

$$\xi = \frac{(Al\sqrt{C}sV) + (Al\sqrt{C}s_0V) + d_0 + (\frac{B\gamma S}{\rho C_p})}{v} \tag{7}$$

$$\xi = (\rho C_p Al\sqrt{C}sV) + (\rho C_p Al\sqrt{C}s_0V) + \rho C_p d_0 + (B\gamma S) \tag{8}$$

$$\chi = \xi - B\gamma S \tag{9}$$

The equation (**??**) represente the Inside greenhouse temperature[**?**][**?**] :

$$\begin{aligned} T_{i(n+1)} &= \frac{h}{v}T_{m(n+1)} + (\frac{v-h}{v}\frac{\alpha}{v}\frac{1}{v}\frac{K_l}{v}\frac{-K_l}{v}) \dots \\ &\times (T_e, R_g, Q_a, P_e, P_i)' \end{aligned} \tag{10}$$

where

$$\begin{aligned} T_{m(n+1)} &= T_{m(n)} exp(-\frac{\Delta t}{\tau}) + (1 - exp(-\frac{\Delta t}{\tau}) \dots \\ &\times \delta \times (T_e, R_g, Q_s, Q_a, P_e, P_i)' \end{aligned} \tag{11}$$

and

$$\delta = (1\frac{\alpha h + \beta v}{h(k+k_s)}\frac{v}{h(k+k_s)}\frac{1}{(k+k_s)}\frac{K_l}{(k+k_s)}\frac{-K_l}{(k+k_l)}) \tag{12}$$

Fig. 1: Block Diagram of the Controlled greenhouse.

Fig. 2: Measured air temperature inside the greenhouses between May 14 and 22.

Fig. 3: Measured air water vapour pressure inside the greenhouses between May 14 and 22.

VI. PRINCIPLE OF MODEL PARAMETERS IDENTIFICATION

VII. THE RESULTS OF THE BLOCK DIAGRAM AND THE OPEN LOOP

The Fig **??** represent the block diagram of the greenhouse together with the four actuators: s , Q_a , Q_s , φ_l ; five input variables have also been considered : T_e,P_e ,R_g , V , PT_i; which are considered as disturbances in the control loop.

Some simulations have prior been carried out to study the dynamic behavior of the controlled variables (see Fig **??**). In these assay, the initial conditions for interior temperature $T_{m(n)}$ and water vapour pressure $P_{i(n)}$ are considered.

Figure **??** and Figure **??** shows the experimontal of temperature and water vapour pressure at the interior of the greenhouses during more then one week.

The elevated number of parameters to be identified leads to fix some of them, especially those which are already known with a good accuracy and particularly:

- K the overall heat loss coefficient through the greenhouse cover ($Wm^{-2}K^{-1}$) is [**?**]:

$$K = 7.6 + 0.42V \tag{13}$$

- $Al\sqrt{C}$ is a dimensionless parameter of the model of natural ventilation, it is set to: $0,2$ following previous air exchange rate studies on this particular greenhouse [**?**].
- s_0 the leakage surface (m^2) set to $s_0 = 0.7m^2$ and d_0 the leakage (m^3s^{-1}) who are separate of wind speed with $d_0 = 0.6m^3s^{-1}$ [**?**].

The six reminding parameters of the temperature and pressure balance equations to be optimized are the following:

T_{m0}: The initial thermal masse temperature $(°C)$
h : the air/ sol convective Exchange coefficient $(Wm^{-2}K^{-1})$
α: the rate absorption of the global radiation by the aerial compartment of the greenhouse.
τ : the time constant or characteristic time (s).
β: the rate of absorption of the global radiation by the thermal mass compartment of the greenhouse.
B: a parameter of the model of transpiration $(Wm^{-2}hPa^{-1})$

Table **??** gives the values of these parameters calculated by the classical algorithm [**?**]. Parameter values together with their confidence intervals were identified utilizing the classical algorithm, yet T.Boulard and B.Draoui proved that the final result was strongly dependent on the initial values of every parameter. The existence of very large interactions between variables was also shown and especially it's can be seen that one part of the variability of the results which is due to one model parameter can be attributed to another one if the two model parameters are statistically strongly correlated [**?**].

TABLE II: Identified values of the parameters T_{m0}, h , α ,τ , β , B during a one week sequence using the classical algorithm.

T_{m0}	h	α	τ	β	B
16	13.5	0.53	1058	0	3.78

TABLE III: Search space of the parameters to be identified.

T_{m0}	h	α	τ	β	B
14	0	0.2	2	0	1
28	30	0.8	1100	0.2	9.5

VIII. SEARCH SPACE OF PARAMETERS

With the classical Algorithm, the parameters to be optimized are selected, one must define also their numerical bound. For exemplar, the temperature of the thermal mass varies between $14.2°$and $29.7°$(average minimal and maximal temperatures observed in Avignon, France during May.). The search space for each parameters are given in Table **??**, Search space of the parameters T_{m0}, h , α ,τ , β , B to be identified.

IX. OBJECTIVE FUNCTIONS:

As objective function, the equations of air temperature and pressure (T_i and P_i) are considered and defined by the relations (**??**) and (**??**), the objective of this study is to minimize the difference between measured and calculated values with the selected parameters.

X. CUCKOO SEARCH

A. Cuckoo Breeding Behaviour.

CS is based on the reproduction strategy of some cuckoos species augmented by a Levy flight behaviour found in the foraging habits of other animal species.

Cuckoos are nest parasite, they lay their eggs in other birds nests and leave the host birds to incubate and rear their young.

When the Cuckoo nestling hatches, it instinctively pushes the other eggs and nestlings out of the nest. This reproductive strategy can be extremely costly for foster parents, because the reproductive success of parasitized hosts is dramatically reduced, and in most cases (depending on parasite and host species) is nil [**?**]; which, Leads to strong host adaptations to detect and reject foreign eggs or to simply abandon their nests and build new ones. As a result of these Female cuckoos developed counter-adaptations: host-egg mimesis, they occasionally specializes in using a private host species, and lay eggs that closely resemble the host eggs. This, Coevolutionary dynamics donate rise to an evolutionary antagonistic arms race between the two species With growing fitness costs of parasitism, choice for host defences increases, which in turn may force parasites to specialize and evolve fine-tuned adaptations that overcome a particular host's defences. [**?**].

B. Lvy Flights

The issue is how animals get nourishment in dynamic natural milieu where they have sparsely or no acquaintance of where resources are situated. Foraging theory foretell that Lvy steal move optimize the hit of resources random searches. [**?**] In recent years, biologists have uncovered that Lvy flights describe foraging patterns in a number of kind of animals and insects: ants, bee.... [**?**] , in the foraging of marine predators [**?**] end even in foraging movement patterns of human huntergatherers [**?**] Lvy flights, appointed by the French mathematician Paul Lvy, are a peculiar class of random walks. A random walk is a formalization of the intuitive idea of taking successive steps, each in a random direction.Thus, they are easy stochastic processes consisting of a discreet sequence of shifting events (i.e. move lengths) separated by successive reorientation events (i.e. turning angles). [**?**] The statistical dispensation of shifting lengths and changes of direction, depict the stochastic process. In particular, Lvy random walk models imply a uniform dispensation for the turning angles and a Lvy-stable dispensation for the move or flight step length. [**?**]

In Lvy flights, the lengths, l, of the steps or jumps of the walks are distributed as a power law [**?**],

$$P(l) = l^{-\lambda}, (1 < \lambda \leq 3) \tag{14}$$

Lvy flights, are typify by the being of scarce but exceedingly large steps, alternating between sequences of many short-length leaps and the same sites are revisited much fewer frequently than in patterns described by other process.

C. Cuckoo Search algorithm

CS is a population-based algorithm, in a way like to genetic algorithm. Where the solution are represented by eggs in hosts nests and the cuckoo eggs represent the further solutions, the goal is to use the new and potentially better solutions (cuckoos) to substitute the bad solutions in the nests.[**?**]

The CS can be described using following three idealized rules:

1) Each cuckoo lays one egg at a time, and dumps it in a randomly chosen nest;
2) The best nests with high quality of eggs (solutions) will carry over to the next generations.
3) The number of available host nests is fixed, and a host can discover an alien egg with probability $P_a \in [0, 1]$. In this case, the host bird can either throw the egg away or abandon the nest to build a completely new nest in a new location[**?**].

The third assumptions can be approximated as the fraction P_a of the n nests is replaced by new nests (new random solutions).

The quality or fitness of a solution can be defined in a similar way to the fitness function in genetic algorithms.

The basic steps of C.S are described in the following pseudo code [**?**]:

Begin
Objective function $f(X)$, $X = (x_1, x_2, ..., x_d)$;
Generate initial population of n host nests X_i, $(i = 1, 2, , n)$
while (t < MaxGeneration) or (stop criterion) **do**
Get a cuckoo randomly by Levy flights evaluate its quality/fitness F_i
Choose a nest among n (say, j) randomly
if $(F_i > F_j)$ **then**
replace j by the new solution;
end if
A fraction (pa) of worse nests
are abandoned and new ones are built;
Keep the best solutions
(or nests with quality solutions);
Rank the solutions and find the current best
end while
Postprocess results and visualization
End

When generating new solutions $x^{(t+1)}$ for a cuckoo i, a Lèvy flight is performed using the following equation:

$$x_i^{t+1} = x_i^t + \alpha \bigotimes Levy(\lambda) \tag{15}$$

where $\alpha > 0$ represents a step size. This step size should be related to the scales of problem the algorithm is trying to solve.

The product $\bigotimes$ means entry-wise multiplications. Lèvy flights essentially provide a random walk while their random steps are drawn from a Levy distribution for large steps.

$$Levy \sim u = t^{-\lambda}, (1 < \lambda \leq 3) \tag{16}$$

It is worth pointing out that, in the real world, if a cuckoos egg is very similar to a hosts eggs, then this cuckoos egg is less likely to be discovered, thus the fitness should be related to the difference in solutions. Therefore, it is a good idea to do a random walk in a biased way with some random step sizes.

From the implementation point of view, the generation of random numbers with Lèvy flights consists of two steps: the choice of a random direction and the generation of steps which obey the chosen Lèvy distribution. The generation of a direction should be drawn from a uniform distribution, while for the generation of steps there are a few ways, but one of the most efficient and yet straightforward ways is to use the so-called Mantegna algorithm for a symmetric Lèvy stable distribution. Here symmetric means that the steps can be positive and negative.[**?**]

In Mantegnas algorithm, the step length s can be calculated by

$$S = \frac{u}{|y|^{1/\beta}} \tag{17}$$

Where $0 < \beta \leq 2$ is an index. u and v are stochastic variables drawn from normal distributions. That is[**?**]:

$$u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2 \tag{18}$$

$$\sigma_u = \{\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma[(1+\beta)/2]\beta \cdot 2^{(\beta-1)/2}}\}^{1/\beta} \tag{19}$$

Here $\Gamma(z)$ is the Gamma function,

$$\Gamma(x) = \int_0^\infty t^{z-1}e^{-t}dt \tag{20}$$

TABLE IV: Best parameters result identified by the Cuckoo Search Algorithm.

T_{m0}	h	α	τ	β	B
23.506	3.8093	0.65352	21.098	0.001482	8.109

TABLE V: Quadratic errors of air temperature and pressure between model and experimental results according to the identification process .

quadratic error	Classical Model	CS Model
Air temperature	0.1376	0.0852
Air pressure	0.0683	0.0415

Fig. 4: Comparaison of the temperature results

Fig. 5: Comparaison of the air water vapour pressure results

XI. RESULTS

An experimental study of greenhouse was made in a period between 14 and 22 May 1991 situate nearby Avignon in South-East of France, this study allowed us to make an identification of a temperature and humidity simulation models. The greenhouse had a tomato-crop area of 416 m^2, in a double roof plastic house. Various sensors and actuators correspondant to those which are offer in the Block Diagram of **??** were established and attached to a data logger and control system based on a personal computer and a control card using a sampling interval of 1 hour. The ensemble of parameters given by the Cuckoo Search (CS) algorithm which minimizes the gap between the calculated and experimental output values of the model for a 9 days interlude on an hourly basis is presented in Table **??**. Just little seconds are requisite to identify the parameters of the reduced model. In the optimization process, the fraction probability, Pa (discovery rate) is 0.25 and the Maximum iterations is 100 and the number of nest (size of population) is 50.

The comparison of the results given by the models optimized with the Cuckoo Search Algorithm or the classical Algorithm with respect to the experimental measurements is given by Figure **??** and figure **??** gives.

The figures exemplify the nice accordance which was spotted between the observed results and the simulation found from the Cuckoo Search Algorithm, both in terms of dynamics and intensity of the signal, especially for the air water vapour pressure calculation.

In order to quantize more exactly the ratification of CS algorithm, the quadratic error are calculated between the observed and simulated results for respectively the data issued from the identified model using the Cuckoo Search Algorithm or the classical Algorithm (Table **??**), it's can be seen that, the Cuckoo Search Algorithm enhance very significantly the accuracy of the simplified greenhouse model.

XII. CONCLUSION

In this paper, the Cuckoo Search (CS) method to solve the greenhouse model parameters identification problem is investigated, the determination of the physical parameters relating the interactions between crop and climate in a horticultural greenhouse can be considerably enhanced both in terms of calculation time and accuracy of the results, via a Cuckoo Search algorithm.

This result is mainly motivating for the reason that this type of biophysical model involves the majority of the time a big number of mechanisms which are not forever exactly modeled or which modeling depends on the particular perspective of the system, such as mainly the biological one. In this case the stage of identification of the model parameters is mainly essential for selecting parameters which are both accurate and robust.

The results showed that the Cuckoo Search algorithm solution quality is better than that of classical Algorithm in most of the test cases. Furthermore, the Cuckoo Search algorithm runs quicker as compared with classical Algorithm.

The advantages of the cuckoo search comprise a uncomplicated organization, instantly accessible for useful applications, easy of realization, speed to obtain solutions and robustness.

For the future research work focuses on exploring another of bettering rendering, while trying to minimize costs, for the reason that this is a means reason for the development productivity.

REFERENCES

[1] T. Boulard and B. Draoui. *In-situ Calibration of a greenhouse climate control model including sensible heat, water vapour and CO2 balances.* IMACS/IFAC bruxelles, BELGIUM; 1995/05/09-12.pVI.A.1-1 ; VI.A.1-6

[2] T. Boulard and B. Draoui. *Natural ventilation of greenhouse with continuous roof vents: Measurements and data analysis*. Journal of Agricultural Engineering Research, (61):2736, (1995).

[3] T. Boulard, B. Draoui and F. Neirac, *Calibration and validation of a greenhouse climate control model*. Workshop: Mathematical and Control Application in Agriculture and Horticulture. Silsoe Grande Bretagne. Acta Horticulturae. 1994.

[4] T. Boulard and R. Jemaa. *Greenhouse tomato crop transpiration model application to irrigation control*. Acta Horticulturae 335. 1993. p 381-387.

[5] A. Hasni, and al, *Evolutionary Algorithms In The Optimization Of Greenhouse*,2008, Climate Model Parameters. International Review on Computers and Software, (I.RE.C0.S.)

[6] J. J Soler and al, *Change in host rejection behavior mediated by the preclatory behavior of its brood parasite*, Behavioral Ecology, Vol. 10 No. 3: 275-280, (1999).

[7] O Krüger, *Brood parasitism selects for no defence in a cuckoo host*, Proc. R. Soc. B published online 2 February 2011.

[8] Humphries and al, *Foraging success of biological Lvy flights recorded in situ*, PNAS, Vol. 109, No. 19: 71697174, (2012).

[9] C. T Brown , *Lèvy Flights in Dobe Ju/hoansi Foraging Patterns*, Hum Ecol No. 35:129138, (2007).

[10] G. M Viswanathan, *Fish in Levy-flight foraging*, NATURE, Vol. 465: 1018-1019, (2010).

[11] F. Bartumeus, *Lèvy processes in animal movement: An evolutionary hypothesis*, Fractals, Vol. 15, No. 2: 112, (2007).

[12] Yang XS, Deb S, *Cuckoo search via Lèvy flights*, Proceeings of World Congress on Nature and Biologically Inspired Computing, India, 210-214, (2009).

[13] Yang XS, *Nature-Inspired Metaheuristic Algorithms*, Second Edition, Luniver Press, (2010).

[14] B Draoui, *Caractrisation et analyse du comportement thermo-hydrique d'une serre horticole.* Thse de Doctorat de l'universit de Nice-Sophia Antipolis,(1994).

[15] Arindam Majumder and Dipak Laha, *A New Cuckoo Search Algorithm for 2-Machine Robotic Cell Scheduling Problem with Sequence-Dependent Setup Times*, Swarm and Evolutionary Computation http://dx.doi.org/10.1016/j.swevo.2016.02.001.

[16] E I Mbuyamba, *Active contours driven by Cuckoo Search strategy for brain tumour images segmentation*, Expert Systems With Applications 2016 Published by Elsevier Ltd.

[17] Li Huang, Shuai Ding, Shouhao Yu, Juan Wang, Ke Lu, *Chaos-enhanced Cuckoo search optimization algorithms for global optimization*, Applied Mathematical Modelling (2015),