SCIENTIFIC COOPERATIONS

WORKSHOPS ON ELECTRICAL AND COMPUTER ENGINEERING SUBFIELDS

22-23 AUGUST, 2014

KOC UNIVERSITY

ISTANBUL-TURKEY

PROCEEDINGS BOOKLET

ORGANIZED BY SCIENTIFIC COOPERATIONS

SCIENTIFIC COOPERATIONS PUBLICATIONS

ISBN: 978-605-86637-4-9

SCIENTIFIC COOPERATIONS
Gersan Sanayi Sitesi
2306 Sokak, No: 61, Batikent-Ankara
Tel.: +90 312 223 55 70
Fax: +90 312 223 55 71

Printing Office:

Turuncu Creative Reklamcilik Matbaacilik Tic.Ltd. Sti.

Tel:+90 312 444 11 84
Fax:+90 312 285 90 92
Web: www.turuncucreative.com
E-mail:info@turuncucreative.com

ADVISORY and REVIEWER BOARD

CONTROL ENGINEERING

Rezoug Amar, Center For Development Of Advanced Technologies, Algeria
Ankit R Patel, Gujarat Technological University
Kemal Ucak, Istanbul Technical University
Nishchal Singh Kush, Queensland University Of Technology
Mohsen Maboodi, K.n.toosi University Of Technology
José Victor Vasconcelos Sobral, Federal University Of Piaui
Mohd Ashraf Ahmad, Universiti Malaysia Pahang
Yoo, Namhyun, Kyungnam University
Fengyou Sun, Shanghai Normal University
Ljiljana Šeriæ, University Of Split
Davood Mohammadi Souran, Ministry Of Energy Of Iran
Bhupesh Jingar, National Institute Of Technology, India
Yusak Tanoto, Petra Christian University
Spyridon G. Mouroutsos, Democritus University Of Thrace
Rezoug Amar, Center For Development Of Advanced Technologies, Algeria
Yeom, Seokwon, Daegu University
Syamsul Rizal Abd. Shukor, Universiti Sains Malaysia
Prof. A Chaudhary, Mum Usa
Dinesh Kumar Sharma, Sir Padampat Singhania University, Udaipur (inda)
Nazia Badar, Rutgers University
Amin Zehtabian, Tarbiat Modares University
Touati, Computer Sce Lab. University Of Paris 8
Karla MarÃ-a Ronquillo Gonzalez, Ieee Member 90350905
Mostafa Anwar Taie, Valeo
Mohamed Saad, Automotive Sw
Ridza Azri Ramlee, Universiti Teknikal Malaysia Malacca
Prof. Ousmane Thiare, University Gaston Berger Of Saint-louis Senegal
Daniel Ramos Jaime, Comimsa
Dr. Miao Wang, Freie Universität Berlin
Helena Catarina Pereira, Phd Student
Ata Jahangir Moshayedi, Ph.d Student
Irfan Ullah, Department Of Information And Communication Engineering, Myongji University, Yongin, South Korea
Jitendra Pal, Iiita
Manoj Sharma, Bvc
Xi Yin, Institute Of Computing Technology (ict), Chinese Academy Of Sciences (cas)

COMPUTER NETWORKS AND NETWORK SECURITY

Hatem M. Bahig, Faculty Of Science, Ain Shams University
Anazida Zainal, Universiti Teknologi Malaysia
Waqar Ali Khan, Comsats Iit
Turki A. Alghamdi, Dr. Turki Alghamdi, Assistant Professor, Department Of Computer Sciences, Faculty Of Computer And Information Systems, Umm Alqura University, Makkah, Saudi Arabia
Arslan, Forman Christian College
Dr.tariq Umer, Comsats Institute Of Information Technology
Mohammed Al-shargabi, Assistant Professor, Coordinator Of Information System Department, College Of Computer Science And Information Systems Najran University.

Mehdi Bateni, Assistant Professor At Shbu
Syed Ali Zahir Bukhari, Comsats Institute Of Information Technology
Dr. Ibrahim El Rube', Associate Professor, Taif University
Prof. Kadry Ibrahim Montasser, Professor, Umm Al Qurah University
Abdul Razak, Associate Professor Of Computer Science, Jamal Mohamed College (bharathidasan University)

COMPUTER AND SOFTWARE ENGINEERING
Dr. Nedhal Al-saiyd, Associate Professor
Saira Anwar, Forman Christian College (a Chartered University)
Imen Boudali, Assistant Professor In Computer Science
Dr. Ahmad Al-hajji, Qassim University
Saif Ur Rehman Khan, Faculty Of Computer Science And Information Technology, University Of Malaya
Dr Sasikumaran Sreedharan, King Khalid University
Dr. Chiam Yin Kia, University Of Malaya, Faculty Of Computer Science And Information Technology
Laxmi Joshi, Majmaah University
Najet Zoubeir, Institut Supérieur D'informatique
Atef Ibrahim, Assistant Professor At Electronics Research Institute, Cairo, Egypt

COMPUTER SCIENCE AND INFORMATICS
Ahmad M. Hasasneh, Hebron University
Maytham Safar, Kuwait Unviersity
Ahmar Rashid, Gik Institute Of Engineering Sciences And Technology
Amelia Ritahani Ismail, Asst Prof, Dr
Barlian Henryranu Prasetio, University Of Brawijaya - Indonesia
Assoc. Prof. Dr. Mohd Nordin Abdul Rahman, Faculty Of Informatics And Computing, University Of Sultan Zainal Abidin, Campus Tembila, 22200 Besut, Terengganu, Malaysia
Saif Ur Rehman Khan, Faculty Of Computer Science And Information Technology, University Of Malaya
Muhammad Farhan, Comsats Institute Of Information Technology
Dr. Umair Ali Khan, Assistant Professor, Computer Systems Engineering Department, Quaid-e-awam University Of Engineering, Nawabshah, Pakistan.
Dr. Belal Amro, Hebron University
Syedah Zahra Atiq, Assistant Professor, Department Of Computer Science, Forman Christian College (a Chartered University)
Salah Omer Hagahmoodi, University Of Hail, Saudi Arabia
Amal Elgammal, Research Fellow At Trinity College Dublin
Nurul Azma Zakaria, Faculty Of Information And Communication Technology, Universiti Teknikal Malaysia Melaka
Khaled Ahmed Nagaty, British University In Egypt
Fathi Amer, Prof.
Hammad Majeed, Nuces

COMPUTER VISION
Mohammad Reza Daliri, Iran University Of Science And Technology (iust), Tehran, Iran
Karla María Ronquillo Gonzalez, Ieee Member 9035090
Daniel Elliott, Colorado State University
Ã ngel IvÃ¡n GarcÃ-a Moreno, Instituto PolitÃ©cnico Nacional
Muhammad Naufal Bin Mansor, University Malaysia Perlis,
Marco Block-berlitz, Htw Dresden / Germany
Fatma Susilawati, Universiti Teknologi Malaysia
George A. Papakostas, Tei Of East Macedonia And Thrace
Hamed Rezazadegan Tavakoli, University Of Oulu , Oulu

Dr Ian Anthony Williams, Digital Media Technology (dmt)
Mohammad Farshchi, Dept Of Advanced Computer Vision, Advanced Research University
S. M. R. Frshchi, Department Of Artificial Intelligence, Image Processing Labratoary
Prashanth Viswanath, Texas Instruments, India
Yogendra Narain Singh, Gautam Buddh Technical University
Vinayak L. Patil, Trinity College Of Engineering Research
Fatma Susilawati, Universiti Teknologi Malaysia
Hasan Farshchi, Dept Of Advanced Computer Vision, Advanced Research University
Muhammad Naufal Bin Mansor, University Malaysia Perlis,
John Jairo Sanabria, Universidad Industrial De Santander
Hima Patel, Ge Global Research
Dominik M. Aufderheide, South Westphalia University Of Applied Sciences
Kun Duan, Indiana University
Dr. Muhammad Usman Akram, National University Of Sciences And Technology

ENERGY, POWER AND ELECTRICAL MACHINES
Dr.chitti Babu B, Wroclaw University Of Technology Poland
Kazem Zare, University Of Tabriz
L. Suresh, Vignan's Lara Institute Of Technology & Science
Li Du, The University Of Akron
Kazem Varesi, University Of Tabriz
Mahdi Pourakbari Kasmaei, Universidade Estadual Paulista
Majid Aryanezhad, Shahid Chamran University Of Ahvaz
Mohammad Yazdani-asrami, Babol "noshirvani" University Of Technology
Yusak Tanoto, Petra Christian University
Jawad Faiz, University Of Tehran
Mohammad Mahdi Mahmoodi, Shahid Beheshti University
Behnam Mohammadi-ivatloo, University Of Tabriz
Umer Fiaz Abbasi, Universiti Teknologi Petronas
Amir Heydari, Sepid Gatch Saveh
Zaipatimah Ali, Universiti Tenaga Nasional, Malaysia

COMMUNICATION THEORY
Amin Zribi, High Institute Of Communications Technologies
Fraser Cadger, University Of Ulster
Ghanshyam Singh Thakur, Maulana Azad National Institute Of Technology
Lusheng Wang, Hefei University Of Technology
Pouria Sayyad Khodashenas, Universitat Politècnica De Catalunya
Rahul Sinha, Tcs Innovation Labs
Ying-ren Chien, National Ilan University
Dr.wafa' Slaibi Alsharafat, Al Al-bayt University
Serguei Primak, Western University
Ning Wang, Beijing University Of Posts And Telecommunications
Georges Rodriguez-guisantes, Telecom Paristech
Mikołaj Leszczuk, Agh University Of Science And Technology
Asma Mejri, Telecom-paristech, France
Arsalan Ahmad, Det, Politecnico Di Torino, Italy
Waqar Shahid Qureshi, Asian Institute Of Technology
Khurram Saleem Alimgeer, Comsats Iit
Prof. Dr. Adnan Ashraf, Unifiedcrest, Cse, Mehran University Of Engineering And Technology
Ali Hazmi, Tampere University Of Technology
Mohammed Elmusrati, Professor At University Of Vaasa - Finland

Akram Rashid, Assistant Professor Department Of Electrical Engineering Air University, Islamabad Pakistan

ARTIFICIAL INTELLIGENCE

Mu-song Chen, Department Of Electrical Engineering Of Da-yeh University, Changhua, Taiwan
Paulo Salgado, Universidade De Trás-os-montes E Alto Douro
Hassan Mustafa, Al-baha University
Yara Khaluf, Heinz Nixdorf Institut University Of Paderborn
Dustin Smith, Mit Media Lab
Abdelkader Adla, University Of Oran
Paulo J.s. Gonçalves, Polytechnic Institue Of Castelo Branco
Lizhen Dai, East China Jiaotong University
Gang Yang, Beijing University Of Technology
Mohd Herwan Sulaiman, Universiti Malaysia Pahang
Yogita Thakran, Indian Institute Of Technology
Waseem Ahmad, Aut University, New Zealand
Mohd Ashraf Ahmad, Universiti Malaysia Pahang

BIOMEDICAL ENGINEERING AND BIOINFORMATICS

Egoitz Arruti, Mondragon University
Arman Ahmadian, Sharif University Of Technology
Luís Barreto, Escola Superior De Ciências Empresariais- Polytechnic Institute Of Viana Do Castelo
Cong Bai, Zhejiang University Of Technology
Barbaresco Frederic, Thales Air Systems
Quazi Md. Alfred, Prof
Fardin Afdideh, Biomedical Signal And Image Processing Laboratory (bisipl), School Of Electrical Engineering, College Of Engineering, Sharif University Of Technology
Xiao Han, New York University
Imran Shafique Ansari, King Abdullah University Of Science And Technology (kaust)
Ali Afana, Concordia University, Montreal, Canada.
Felix Albu, Valahia University Of Targoviste
Dr. Muhammad Ishtiaq Ahmad, Beijing Institute Of Technology, Beijing, China
Marius Branzila, Technical University Of Iasi
Muhammet Fatih Bayramoglu, University Of Oulu
Muhammad Tahir Akhtar, The University Of Electro-communications, Tokyo, Japan
Harold Chamorro, Kth Royal Institute Of Technology
Alper Basturk, Erciyes University
Dr. C Bhattacharya, Defence Institute Of Advanced Technology
Chitti Babu, National Institute Of Technology
Mohammad Aazam, Kyung Hee University, South Korea
Mohammad Nasiruddin, Laboratory Of Informatics Of Grenoble
Hamzah Alzu'bi, University Of Liverpool - Phd Student
Salim Kahveci, KtÜ, Electrical&electronics Engineering, 61080, Trabzon
Iwan Adhicandra, Bakrie University, Indonesia
Sergej Andruschenko, Technical University Of Munich (tu Munich)
Giordano Cabral, Ufrpe
Mohammad Haghighat, University Of Miami
Jesus B. Alonso, University Of Las Palmas De Gran Canaria
Prof. Andrei Campeanu, Politehnica University Timisoara
Dr Costas Chaikalis, Tei Of Thessaly, Department Of Informatics Engineering
Dr. Dinesh Bhatia, North Eastern Hill University (nehu), Shillong
Ying-ren Chien, National Ilan University

Muhammad Naufal Bin Mansor, University Malaysia Perlis
Danilo Zanatta, Schlieren-zürich
Anton Popov, National Technical University Of Ukraine
Seyed Ali Amirshahi, University Of Jena
Kamil Dimililer, Near East University
Hamed Rezazadegan Tavakoli, University Of Oulu
Yalcin Isler, Izmir Katip Celebi University
Stefan Kerber, Technical University Munich,
Baher Mawlawi, Cea-leti (nuclear Center Of France)
Lahcène Mitiche, University Of Djelfa
Mircea Giurgiu, Technical University Of Cluj-napoca
Musa Peker, Karabuk University
Pravin Kumar Rana, Kth Royal Institute Of Technology
Salim Kahveci, Karadeniz Technical University
Zemouri Et-tahir, University Of Science And Technology Houari Boumediene
Uygar Tuna, Tampere University Of Technology
Ahmad Poursaberi, University Of Calgary
Kazım Yalçın ArĞa, Assistant Professor
Rengarajan Pelapur, University Of Missouri
Bouwmans Thierry, Univ. La Rochelle
David, Arroyo
Hussnain Ali, The University Of Texas At Dallas
Karim Badawi, Eth Zurich
Ali Reza Khanteymoori, Dr. Assistant Professor
Nima Jamshidi, Assistant Professor At University Of Isfahan

IMAGE PROCESSING
Egoitz Arruti, Mondragon University
Arman Ahmadian, Sharif University Of Technology
Luís Barreto, Escola Superior De Ciências Empresariais- Polytechnic Institute Of Viana Do Castelo
Cong Bai, Zhejiang University Of Technology
Barbaresco Frederic, Thales Air Systems
Quazi Md. Alfred, Prof
Fardin Afdideh, Biomedical Signal And Image Processing Laboratory (bisipl), School Of Electrical Engineering, College Of Engineering, Sharif University Of Technology
Xiao Han, New York University
Imran Shafique Ansari, King Abdullah University Of Science And Technology (kaust)
Ali Afana, Concordia University, Montreal, Canada.
Felix Albu, Valahia University Of Targoviste
Dr. Muhammad Ishtiaq Ahmad, Beijing Institute Of Technology, Beijing, China
Vania V. Estrela, Universidade Federal Fluminense (uff)
Salim Azak, Selcuk University
Marius Branzila, Technical University Of Iasi
Muhammet Fatih Bayramoglu, University Of Oulu
Muhammad Tahir Akhtar, The University Of Electro-communications, Tokyo, Japan
Harold Chamorro, Kth Royal Institute Of Technology
Alper Basturk, Erciyes University
Dr. C Bhattacharya, Defence Institute Of Advanced Technology
Chitti Babu, National Institute Of Technology
Mohammad Aazam, Kyung Hee University, South Korea
Mohammad Nasiruddin, Laboratory Of Informatics Of Grenoble
Hamzah Alzu'bi, University Of Liverpool - Phd Student

Salim Kahveci, KtÜ, Electrical&electronics Engineering, 61080, Trabzon
Iwan Adhicandra, Bakrie University, Indonesia
Sergej Andruschenko, Technical University Of Munich (tu Munich)
Giordano Cabral, Ufrpe
Mohammad Haghighat, University Of Miami
Jesus B. Alonso, University Of Las Palmas De Gran Canaria
Prof. Andrei Campeanu, Politehnica University Timisoara
Dr Costas Chaikalis, Tei Of Thessaly, Department Of Informatics Engineering
Dr. Dinesh Bhatia, North Eastern Hill University (nehu), Shillong
Ying-ren Chien, National Ilan University
Muhammad Naufal Bin Mansor, University Malaysia Perlis
Danilo Zanatta, Schlieren-zürich
Anton Popov, National Technical University Of Ukraine
Seyed Ali Amirshahi, University Of Jena
Kamil Dimililer, Near East University
Hamed Rezazadegan Tavakoli, University Of Oulu
Yalcin Isler, Izmir Katip Celebi University
Stefan Kerber, Technical University Munich,
Baher Mawlawi, Cea-leti (nuclear Center Of France)
Lahcène Mitiche, University Of Djelfa
Mircea Giurgiu, Technical University Of Cluj-napoca
Musa Peker, Karabuk University
Pravin Kumar Rana, Kth Royal Institute Of Technology
Salim Kahveci, Karadeniz Technical University
Zemouri Et-tahir, University Of Science And Technology Houari Boumediene
Uygar Tuna, Tampere University Of Technology
Ahmad Poursaberi, University Of Calgary
Ayşe Elif ÖztÜrk, Selçuk Üniversitesi
Bouwmans Thierry, Univ. La Rochelle
David, Arroyo
Hussnain Ali, The University Of Texas At Dallas
Karim Badawi, Eth Zurich

WIRELESS, MOBILE AND SENSOR NETWORKS
Dr. Arif Sari, European University Of Lefke
Adib Rastegarnia, University Of Tehran

ROBOTICS AND MECHATRONICS
Melchior, Ims-umr 5218 Cnrs, Bordeaux Inp
Mohd Ashraf Ahmad, Universiti Malaysia Pahang
Noha Sadek, The American University In Cairo & Ibm World Trade Corporation
Brian Lynch, Queen's University
Karla MarÃ-a Ronquillo Gonzalez, Ieee Member 90350905
Dilip R, Professor
S.ehsan Mirsadeghi, M.sc In Digital Electronic @ Aut(amikarbir University Of Technology)
Arif Sari, European University Of Lefke
Ashesh Vasalya, Vellore Institute Of Technology-university Vellore
Levent Bayindir, Ataturk University
Lilia Zouari, Institute Of Information Theory And Automation Of The Ascr
Najib Metni, Notre Dame University
Dr. T.c.manjunath, Visvesvaraya Technological University, Belgaum, India
Prof. Dr. Szabolcsi, Róbert, Óbuda University

Rungun Nathan, Pennsylvania State University
Hammad Khan, Mir Labs
Maysam Zamani Pedram, Sharif University Of Technology
Nalika Ulapane, University Of Technology Sydney
Spyridon G. Mouroutsos, Prof
Valerio Scordamaglia, Diies , University Of Reggio Calabria
Joao Lopes, Isw Universität Stuttgart
Abdul Waheed, University Of Sargodha, Pakistan
Jay Merja, Gujarat Technology University
Weerachai Yaemvachi, Rmutt
Bilal A. Mubdir, Sulaimani Polytechnic University
Chithrangi Kaushalya Kumarasinghe, Lecturer, Department Of Information Technology, Faculty Of Information Technology, University Of Moratuwa
Anjan Kumar Ray, University Of Ulster, Uk
Mohamed Shakir, Qatar University
Rahul Gupta, Manipal Institute Of Technology
Vicente F. Lucena Jr, Federal University Of Amazonas - Brazil
Ankit Patel, Gujarat Technological University, India
Michel Owayjan, American University Of Science & Technology (aust)
Lianbo Ma, Shenyang Institute Of Automation, Chinese Academy Of Sciences
Tarek Mohammad, University Of British Columbia
Mahdin Mahboob, Lecturer, Electronics And Telecom Engg, Ulab
Yamile Sandoval-castro, National Polytechnical Institute
Fang-ming Yu, St. John's University

SIGNAL PROCESSING
Egoitz Arruti, Mondragon University
Arman Ahmadian, Sharif University Of Technology
Luís Barreto, Escola Superior De Ciências Empresariais- Polytechnic Institute Of Viana Do Castelo
Cong Bai, Zhejiang University Of Technology
Barbaresco Frederic, Thales Air Systems
Quazi Md. Alfred, Prof
Fardin Afdideh, Biomedical Signal And Image Processing Laboratory (bisipl), School Of Electrical Engineering, College Of Engineering, Sharif University Of Technology
Xiao Han, New York University
Imran Shafique Ansari, King Abdullah University Of Science And Technology (kaust)
Ali Afana, Concordia University, Montreal, Canada.
Felix Albu, Valahia University Of Targoviste
Dr. Muhammad Ishtiaq Ahmad, Beijing Institute Of Technology, Beijing, China
Marius Branzila, Technical University Of Iasi
Muhammet Fatih Bayramoglu, University Of Oulu
Muhammad Tahir Akhtar, The University Of Electro-communications, Tokyo, Japan
Harold Chamorro, Kth Royal Institute Of Technology
Alper Basturk, Erciyes University
Dr. C Bhattacharya, Defence Institute Of Advanced Technology
Chitti Babu, National Institute Of Technology
Mohammad Aazam, Kyung Hee University, South Korea
Mohammad Nasiruddin, Laboratory Of Informatics Of Grenoble
Hamzah Alzu'bi, University Of Liverpool - Phd Student
Salim Kahveci, KtÜ, Electrical&electronics Engineering, 61080, Trabzon
Iwan Adhicandra, Bakrie University, Indonesia
Sergej Andruschenko, Technical University Of Munich (tu Munich)

Giordano Cabral, Ufrpe
Mohammad Haghighat, University Of Miami
Jesus B. Alonso, University Of Las Palmas De Gran Canaria
Prof. Andrei Campeanu, Politehnica University Timisoara
Dr Costas Chaikalis, Tei Of Thessaly, Department Of Informatics Engineering
Dr. Dinesh Bhatia, North Eastern Hill University (nehu), Shillong
Ying-ren Chien, National Ilan University
Muhammad Naufal Bin Mansor, University Malaysia Perlis
Danilo Zanatta, Schlieren-zürich
Anton Popov, National Technical University Of Ukraine
Seyed Ali Amirshahi, University Of Jena
Kamil Dimililer, Near East University
Hamed Rezazadegan Tavakoli, University Of Oulu
Yalcin Isler, Izmir Katip Celebi University
Stefan Kerber, Technical University Munich,
Baher Mawlawi, Cea-leti (nuclear Center Of France)
Lahcène Mitiche, University Of Djelfa
Mircea Giurgiu, Technical University Of Cluj-napoca
Musa Peker, Karabuk University
Pravin Kumar Rana, Kth Royal Institute Of Technology
Salim Kahveci, Karadeniz Technical University
Zemouri Et-tahir, University Of Science And Technology Houari Boumediene
Uygar Tuna, Tampere University Of Technology
Ahmad Poursaberi, University Of Calgary
Bouwmans Thierry, Univ. La Rochelle
David, Arroyo
Hussnain Ali, The University Of Texas At Dallas
Karim Badawi, Eth Zurich

CONTENTS

ACE-2014, Advances in Control Engineering

ACNNS-2014, Advances in Computer Networks and Network Security

ACSE-2014, Advances in Computer and Software Engineering

ACSI-2014, Advances in Computer Science and Informatics

ACV-2014, Advances in Computer Vision

IEPEM-2014, Innovations in Energy, Power and Electrical Machines

INCT-2014, Innovations on Communication Theory

IWAI-2014, International Workshop on Artificial Intelligence

IWBE-2014, International Workshop on Biomedical Engineering and Bioinformatics

IWIP-2014, International Workshop on Image Processing

IWMSN-2014, International Workshop on Wireless, Mobile and Sensor Networks

RARM-2014, Recent Advances in Robotics and Mechatronics

RDSP-2014, Recent Developments on Signal Processing

Analysis to find the most effective features to predict breast cancer; A data mining aproach

Fariba Tat[1], Sahar Azadi Ghale Taki[2], Sadjaad Ozgoli[3], Mahdi Sojoodi[5]
Electrical and Computer Engineering
Tarbiat Modares University
Tehran, Iran
fariba.tat92@gmail.com[1],
sahar.azadighaleh@modares.ac.ir[2], ozgoli@modares.ac.ir[3],
sojoodi@modares.ac.ir[5]

Mohammad Esmaeil Akbari[4]
Chairman of Cancer Research Center
Prof. of Department of Surgery, Shahid Beheshti University of Medical Sciences
Tehran, Iran
me-akbari@sbmu.ac.ir[4]

***Abstract*— This paper aims to present a hybrid intelligence model that uses the cluster analysis techniques with feature selection for analyzing clinical cancer diagnoses. Our model provides an option of selecting a subset of salient features for performing clustering and comprehensively considers the use of most existing models that use all the features to perform clustering. In particular, we study the methods by selecting salient features to identify clusters using a comparison of coincident quantitative measurements. When applied to benchmark breast cancer datasets, experimental results indicate that our method outperforms several benchmark filter- and wrapper-based methods in selecting features used to discover natural clusters, maximizing the between-cluster scatter and minimizing the within-cluster scatter toward a satisfactory clustering quality. The experimental dataset is based on the data gathered in a hospital in Tehran.**

Keywords—cluster analysis; cancer modeling; feature selection; cancer diagnoses.

I. INTRODUCTION

Breast cancer is the second leading cause of death after cardiovascular diseases in the world. Health professionals are seeking ways for suitable treatment and quality of care in these groups of patients. Survival prediction is important for both physicians and patients in order to choose the best way of management. Today diagnosis of a disease is a vital job in medicine. It is an essential to interpret the correct diagnosis of patient with the help clinical examination and investigations. Computer information based decision support system can play an important role in accurate diagnosis and cost effective treatment. Most hospitals have a huge amount of patient data, which is rarely used to support clinical diagnosis. It is question why we cannot use this data in clinical diagnosis and patient management? Is it possible to formulate own area based prediction system concerned with specific disease by using data mining techniques [1]? Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [2]. Basically data mining technique is concerned with data processing, identifying patterns and trends in information. In other words, data mining simply means collection and processing data in systemic manner by using computer based programs and subsequent formation of disease prediction or patient management system aid. Data mining principles have been known around for many years, but, with the advent of information technology, nowadays it is even more prevalent. Data mining is not all about the database software that you are using. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques [3]. This technique is routinely use in large number of industries like engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining [4]. Machine learning [5] [6], is concerned with the design and development of algorithms that allow computers to evolve behaviors learned from databases and automatically learn to recognize complex patterns and make intelligent decisions based on data. However the massive toll of available data poses a major obstruction in discovering patterns. Feature Selection attempts to select a subset of attributes based on the information gain [7]. Classification is performed to assign the given set of input data to one of many categories [8]. Data analysis procedures can be dichotomized as either exploratory or confirmatory, based on the availability of appropriate models for the data source, but a key element in both types of procedures (whether for hypothesis formation or decision-making) is the grouping, or classification of measurements based on either (i) goodness-of-fit to a postulated model, or (ii) natural groupings (clustering) revealed through analysis. Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification). In supervised classification, we are provided with a collection of

labeled (pre-classified) patterns; the problem is to label a newly encountered, yet unlabeled, pattern. Typically, the given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label a new pattern. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification [9].

In [10], the performance criterion of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART are compared, to find the best classifier in breast cancer datasets (WBC and Breast tissue). The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in WBC and 99.00% in Breast tissue. In [11] , the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) are compared to find the best classifier in WBC. SVM proves to be the most accurate classifier with accuracy of 96.99%. In [12], the performance of decision tree classifier (CART) with or without feature selection in breast cancer datasets Breast Cancer, WBC and WDBC. CART achieves accuracy of 69.23% in Breast Cancer dataset without using feature selection, 94.84% in WBC dataset and 92.97% in WDBC dataset. When using CART with feature selection (Principal Components Attribute Eval), it scores accuracy of 70.63% in Breast Cancer dataset, 96.99 in WBC dataset and 92.09 in WDBC dataset. When CART is used with feature selection (ChiSquared Atrribute Eval), it scores accuracy of 69.23% in Breast Cancer dataset, 94.56 in WBC dataset and 92.61 in WDBC dataset. In [13], the performance of C4.5 decision tree method obtained 94.74% accuracy by using 10-fold cross validation with WDBC dataset. In [14], the neural network classifier is used on WPBC dataset. It achieves accuracy of 70.725%. In [15], a hybrid method is proposed to enhance the classification accuracy of WDBC dataset (95.96) with 10 fold cross validation. In [16], the performance of linear discreet analysis method obtained 96.8% accuracy with WDBC dataset. In [17], the accuracy obtained 95.06% with neuron- fuzzy techniques when using WDBC dataset. In [18], an accuracy of 95.57% was obtained with the application of supervised fuzzy clustering technique with WDBC dataset.

The major contributions of the current work are twofold. First, we have developed a K-means variant that can incorporate background knowledge in the form of instance-level constraints, thus demonstrating that this approach is not limited to a single clustering algorithm. In particular, we have presented our modifications to the K-means algorithm and have demonstrated its performance on six data sets.

Second, we have used the best feature for classification so we can predict the time of cancer recrudescence which is very important for prevention of death due to cancer.

In the next section, we provide some backgrounds on the Clustering algorithm such as K-means algorithm and SVM. Next, we describe our methods and result in some tables in section 3 and 4. Finally, Section 5 summarizes our contributions [19].

II. CLUSTERING ALGORITHMS

To evaluate the effectiveness of NDR for capturing cluster structures, two classical clustering algorithms, HC and K-means, are applied to the reduced feature spaces. These two algorithms are representatives for two kinds of widely used clustering approaches. HC uses agglomerative and divisive strategies and divides data into a sequence of nested partitions, where partitions at one level are joined as clusters at the next level. The number of clusters can be determined immediately at special level upon requirements. While, K-means is one of the most widely used center-based clustering algorithms which uses a partitioning strategy to assign objects into fixed clusters. The algorithm regards data vectors as a point set in a high-dimensional space. According to the input clustering number, K-means randomly selects centroid points for each cluster and allocates each of data point into one of these clusters based on its minimum distance to these centroid points. After necessary optimizing steps, K-means can generate a good clustering solution [11],[12].

A. K-means

Clustering is an important and popular technique in data mining. It partitions a set of objects in such a manner that objects in the same clusters are more similar to each another than objects in the different cluster according to certain predefined criteria. K-means is simple yet an efficient method used in data clustering. However, K-means has a tendency to converge to local optima and depends on initial value of cluster centers. In the past, many heuristic algorithms have been introduced to overcome this local optima problem. Nevertheless, these algorithms too suffer several short-comings [22]. In this paper, we present an efficient hybrid evolutionary data clustering algorithm referred to as K-MCI, whereby, we combine K-means with modified cohort intelligence. Our proposed algorithm is tested on several standard data sets from UCI Machine Learning Repository and its performance is compared with other well-known algorithms such as K-means, K-means++, cohort intelligence (CI), modified cohort intelligence (MCI), genetic algorithm (GA), simulated annealing (SA), tabu search (TS), ant colony optimization (ACO), honeybee mating optimization (HBMO) and particle swarm optimization (PSO). The simulation results are very promising in the terms of quality of solution and convergence speed of algorithm.

B. SVM

With the development of clinical technologies, different tumor features have been collected for breast cancer diagnosis. Filtering all the pertinent feature information to support the clinical disease diagnosis is a challenging and time consuming task. The objective of this research is to diagnose breast cancer

based on the extracted tumor features. Feature extraction and selection are critical to the quality of classifiers founded through data mining methods. To extract useful information and diagnose the tumor, a hybrid of K-means and support vector machine (K-SVM) algorithms is developed. The K-means algorithm is utilized to recognize the hidden patterns of the benign and malignant tumors separately. The membership of each tumor to these patterns is calculated and treated as a new feature in the training model. Then, a support vector machine (SVM) is used to obtain the new classifier to differentiate the incoming tumors. Based on 10-fold cross validation, the proposed methodology improves the accuracy to 97.38%, when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California – Irvine machine learning repository. Six abstract tumor features are extracted from the 32 original features for the training phase. The results not only illustrate the capability of the proposed approach on breast cancer diagnosis, but also shows time savings during the training phase. Physicians can also benefit from the mined abstract tumor features by better understanding the properties of different types of tumors [23].

III. METHODS

In this paper, we have investigated two data mining techniques: Clustering and Classification. In this paper, we used these algorithms to predict the survivability rate of breast cancer data set. We have selected these two clustering and classification techniques to find the most suitable one for predicting cancer survivability rate [24]. The 24 of 76 features used in our study are listed in Table I.

TABLE I. FEATURE

parameter	*Explain*	*Parameter*	*Explain*
F	Family history	T	Size of the original tumor
MS	Marital status	N	lymph nodes
S	Smoking	N+	Cancer has been found in the lymph nodes
C	Childbirth	STAGE	A number on a scale of 0 through IV
P	Pregnancy	PATH	The type of pathology
TA	Type of abortion	GRADE	Grading is a way of classifying cancer cells
NA	Number of abortion	LVI	Lymphovascular invasion
B	Breastfeeding	ER	Estrogen receptor
H	Hormones (estrogen and progesterone)	PR	Progesterone receptor
DH	Duration of hormone use	HER2	Human epidermal growth factor receptor 2
CT	Computed tomography	P53	Tumor protein
RT	Radio therapy	KI67	Antigen identifiedby monoclonal antibody

In order to make the gathered data being in hospital in Tehran in numerical format, a coding scheme is used. This coding is depicted in Table II for coding of variables describing the general and Table III for variables coding cancer.

TABLE II. CODING OF VARIABLES DESCRIBING THE GENERAL

Gender	*Education*	*Type AB*	*FH*
Female 0	Collegiate 1	Criminal 1	Frist degree 1
Male 1	Diploma 2	Medical 2	Second degree 2
	School 3	C/M 3	Yes/unknown 3
	Guidance 4	Unknown 6	Unknown 6
	Illiterate 5	No 9	No 9

Smoking	*Fat*	*Married*	*Menopause*
Yes 1	Yes 1	Singel 1	Histectomy 1
Yes/no 2	Yes/no 2	Married 2	Natural 2
Unknown 6	Unknown 6	Divorce 3	
No 9	No 9	Widowed 4	
		Unknown 6	

TABLE III. VARIABLES CODING CANCER

Surgery	*Armpit*	*CT*	*H.Name*	*Path*
Bcs 1	AXLND/padding 1	Yes 1	Tamoxifin 1	IDC 1
Mrm 2	AXLND/darrinage 2	Neo 2	Letrozol 2	DCIS 2
Bcs/Mrm 3	SLN/darrinage 3	Unknown 6	Aromysin 3	IDC/DCIS 3
Unknown 6	SLN 4	No 9	tamox/letrozol 4	ILC 4
	AXLND 5		tamox/aromysin 5	ILC/LCIS 5
	Unknown 6		Unknown 6	Unknown 6
	SLN/AXLND 7		tamox/decapept 7	IDC/ILC 7
	Padding 8		herceptin 8	LCIS 8
	No 9			no 9
	Darrinage 10			ILC/DCIS 10
	AXLND/SLN/padding 11			
	SLN/padding 12			

In addition, the coding depicted in Table IV is used to numerate the result of test on each feature.

TABLE IV. CODES COMMON

1	yes positive
9	no negative
6	Unknown

In the next section the results of clustering and classification will be discussed.

IV. Clustering and Classification Results

We use 3 clustering for dataset that select some feature in 3 clusters .this cluster show that some feature such as p53 are more than effective to predict breast cancer. In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, sensitivity and specificity). The results were achieved using 74 features for each model, and are based on the average results obtained from the test dataset. It has been chosen 983 patients of 1621 as train data and they are clustered into three, and the results are exist in Table V. This classification algorithm uses LVI as a table and predicts cancer more precisely. Using this algorithm, accuracy numbers are listed in Table V. simulation results have been achieved using Rapid Miner.

TABLE V. RESULTS

Clustering	***Table Recurrence***
Result of data train	91.5 %
Result of data test	89.7 %

LVI as a table that shoe is the best feature for predicting effect time.

Classification	***Label: LVI***
Train data97%	Test Data91%

V. Conclusions and Future Directions

Feature selection is one of the most effective methods to enhance data representation and improve performance in terms of specified criteria, e.g., generalization classification accuracy. In the literature, many studies select a subset of salient features using supervised learning rather than unsupervised learning. When the class labels are absent during training, feature selection in unsupervised learning is integral, but its extensible application is rarely studied in the literature. The objective of this study is to select salient features that can be used to identify interesting clusters in the analysis of cancer diagnosis. Specifically, we highlight three qualitative principles that help users to analyze clinical cancer diagnosis using clusters resulting from a subset of salient features. First, the clusters built by a subset of salient features are more practical and interpretable than those built by all of the features, which include noise. Second, the clustering results provide clinical doctors with an understanding of the context of clinical cancer diagnoses. Finally, a search for relevant records based on the clusters obtained when noisy features are ignored is more efficient. These three principles rely on the discovery of natural clusters using salient features and are applicable only to unsupervised learning. To demonstrate the usefulness of these three qualitative principles, we use coincident quantitative measurements to analyze the salient features for discovering clusters. The experiments on the cancer (Diagnostic) and cancer (Original) datasets demonstrate that the selected features are effective for selecting salient features to discover natural clusters. Based on a performance evaluation using well-known validations in statistical model and cluster analysis, our analysis provides an interesting aspect in feature selection for discovering clusters.

References

[1] M. C. Tayade and M. P. M. Karandikar, "Role of Data Mining Techniques in Healthcare sector in India," *Sch. J. Appl. Med. Sci. SJAMS*, vol. 1, no. 3, pp. 158–160, 2013.

[2] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang, "Data mining curriculum: A proposal (Version 0.91)," 2004.

[3] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[4] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[5] T. M. Mitchell, "Machine learning and data mining," *Commun. ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[6] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques.," *Inform. 03505596*, vol. 31, no. 3, 2007.

[7] S. G. Jacob and R. G. Ramani, "Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques," in *Proceedings of the World Congress on Engineering and Computer Science*, 2012, vol. 1.

[8] X. Wu and V. Kumar, *The top ten algorithms in data mining*. CRC Press, 2009.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv. CSUR*, vol. 31, no. 3, pp. 264–323, 1999.

[10] S. Aruna, D. S. Rajagopalan, and L. V. Nandakishore, "Knowledge based analysis of various statistical tools in detecting breast cancer," *Comput. Sci. Inf. Technol. CSIT*, vol. 2, pp. 37–45, 2011.

[11] A. Christobel and Y, Dr.Sivaprakasam, "An Empirical Comparison of Data Mining Classification Methods.," *Int. J. Comput. Inf. Syst.*, vol. 3, No 2011.

[12] D. Lavanya and D. K. U. Rani, "Analysis of feature selection with classification: Breast cancer datasets," *Indian J. Comput. Sci. Eng. IJCSE*, 2011.

[13] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 130–136.

[14] V. N. Chunekar and H. P. Ambulgekar, "Approach of Neural Network to Diagnose Breast Cancer on three different Data Set," in *Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on*, 2009, pp. 893–895.

[15] D. Lavanya and K. Usha Rani, "ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA.," *Int. J. Inf. Technol. Converg. Serv.*, vol. 2, no. 1, 2012.

[16] B. Šter and A. Dobnikar, *Neural network in medical diagnosis: comparison with other methods*. 1996.

[17] T. Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, 1999, vol. 99, pp. 200–209.

[18] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," *Pattern Recognit. Lett.*, vol. 24, no. 14, pp. 2195–2207, 2003.

[19] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, and others undefined, "Constrained k-means clustering with background knowledge," in *ICML*, 2001, vol. 1, pp. 577–584.

[20] J. Shi and Z. Luo, "Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples," *Comput. Biol. Med.*, vol. 40, no. 8, pp. 723–732, Aug. 2010.
[21] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[22] G. Krishnasamy, A. J. Kulkarni, and R. Paramesran, "A hybrid approach for data clustering based on modified cohort intelligence and K-means," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 6009–6016, Oct. 2014.
[23] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, Part 1, pp. 1476–1482, Mar. 2014.
[24] V. Chaurasia and S. Pal, *Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability*. IJCSMC, 2014.

Teleoperation of Mobile Robot Using Event Based Controller and Real Time Force Feedback

Aamir Shahzad
Automatic Control Engineering Department
University of Siegen
Siegen, Germany
aamir.shahzad@uni-siegen.de

Hubert Roth
Automatic Control Engineering Department
University of Siegen
Siegen, Germany
hubert.roth@uni-siegen.de

***Abstract*—Event based controller has been implemented to teleoperate the real mobile robot efficiently. The system consists of master haptic device, slave robot and a communication network. On master side with the help of visual aid and real time force feedback acting on the robot the operator control and navigate the robot and receive sensory feedback. Environmental force which is acting on slave robot is modeled as virtual force based on obstacles in front of mobile robot using proximity sensors and it has been reflected to human operator in real time using perditor block. Thus the operator can feel that he is driving the robot like a car while he is present at remote location. The designed controller shows the excellent coordination between master haptic device and slave robot. The slave robot follows the master device and communication delay has no effect on the performance and stability of teleoperated robot.**

Keywords— teleoperation; event based controller; force feedback; haptic device

I. INTRODUCTION

In fact, teleoperated robots are excellent mean to work in hazardous environments where human safety is at high risk like nuclear power plants, landmines clearance and space exploration[1-4]. Also teleoperation provides solutions in cases where human operators simply can't manipulate given objects like surgery inside human body through micro-robots which is called tele-surgery. Teleoperation is finding applications in these areas because the technology can save lives and reduce cost by removing the human operators from the operation sites. However, most of these areas still need humans in the control loop because of their very high level of skills and because machine intelligence is insufficiently advanced to operate autonomously and intelligently in such complex and unstructured environments. Teleoperation has become one of the most rapidly expanding areas in mechanical, electrical, computer and control systems engineering.

Today many industries utilize robots because they offer advantage of being able to perform set routines more quickly, cheaply, and accurately than humans. Instead of using programmed routines to maneuver the robots, tele-robotics allows to operate the robot from a distance and make decisions in real time[5]. With the development of more powerful and efficient computers, the future for teleoperation seems extremely promising. On the other hand, the active research in teleoperation is being conducted using Internet as communication medium. This has happened due to the fact that the Internet has changed from a simple data transmission medium to a virtual world application like control. The system which uses real time control over the Internet has many difficulties. One of the most important difficulty is the delay due to the data packets transmission between two points over network. This delay due to its random nature plays a significant role in the stability and efficiency of the system when the commands are sent and received in real time applications. Furthermore, when the Internet began to be used for communication, packet switched networks presented the already established time-delay analysis with difficulties due to randomly varying delays, discrete-time exchange of data and loss of information. So that earlier delay related results were adapted to the new setting as it was studied in detail in [6] as well as to discrete-time setting in [7-10] and information loss in [11]. These methods found their way to several applications in handling radioactive material [12], space robotics [13,14], telesurgery [15], and recently teleoperation of mobile robots [16,17].

Moreover, several Internet based robots have been developed and studied. Reference [18] where they considered the bilateral teleoperation of a wheeled mobile robot over communication channel with constant delay to enable the user to control the mobile robot by operating a master haptic joystick. The passivity of the closed-loop system is also enforced so that, even with communication delays, humans can stably and safely teleoperate the wheeled mobile robot with force-reflection. However, this study was based on simulation framework. Reference [19] the use of a haptic interface is proposed to increase the user's perception of the workspace of the mobile robot. The passivity of the overall system is preserved, so that the stability of the virtual interaction is guaranteed. But the system behavior was not evaluated in complex tasks and also it did not take into account the significant time delay in the data transmission. Reference [20] presented the Internet-based tele-rehabilitation sharing system, whose aim is to achieve the situation where multiple stay home patients in different places can share rehabilitation instruction of one physiotherapist at the same time. However, they have done simulation which is hard to

implement and the real experiment is under planning in that research.

In this work a bilateral control of mobile robot is presented with real time force reflection to operator which is acting on mobile robot without any assumption on time delay using event based control approach. The virtual interaction force is computed on the basis of obstacles in front of the mobile robot. Thus, the live video feed and force feedback from the robot help the operator to drive robot like a car by generating linear velocity equivalent to gas pedal adjustment and heading angle equivalent to steering wheel rotation in car, commands from haptic joystick.

II. Problem Description

In fact, the teleoperation over Internet suffers with time delay. This delay happens due to latency in communication via Internet. The main effects of this delay are instability and de-synchronization. The previous researches assumed the delay time is constant or has upper bound limit[18][21]. In order to avoid these assumptions over delay an event based controller has been implemented which has no impact of delay on it. Also, a predictor block is implemented in the feedback loop that reflect real time force acting on the robot to the operator as shown in Fig. 3.

III. Non-Time Based Control for Teleoperation with Force Reflection

Different approaches have been used to stabilize the teleoperated robots. The stability in this work is ensured by using event based controller. The Fig. 1, and Fig. 2, show the conventional control block and event based control block respectively. Thereom1 explains the stability of the event based controller. The proof of this theorem has been done in [22].

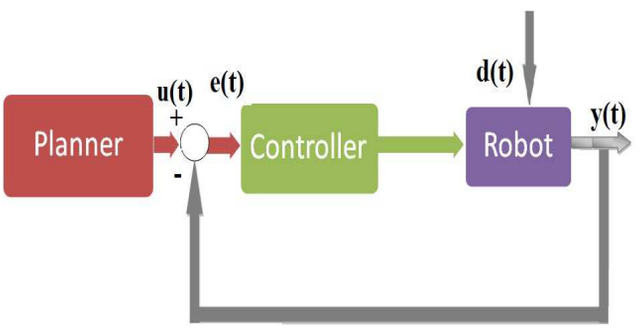


Fig. 1. Conventional Control Loop

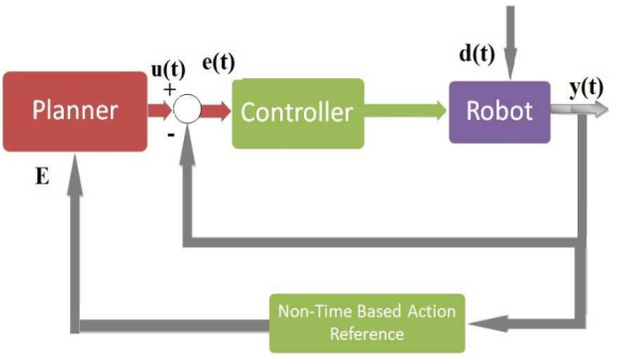


Fig. 2. Non-time based control

A. *Theorem1*

If the original robot dynamic system(without remote human/autonomous controller) is asymptotically stable with time t as its action reference and the new non-time action reference, **e=Π(y)** is a (monotone increasing) non decreasing function of time t, then the system is (asymptotically) stable with respect to the new action reference **e**. The advantage of this approach is that stability is independent of random time delay.

IV. The Control Approach

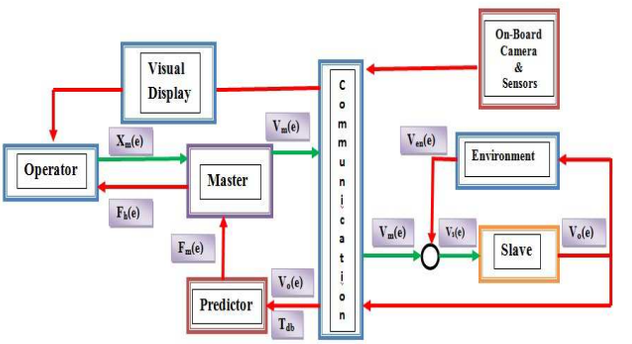


Fig. 3. Block diagram of the teleoperated system.

The telecontrol of mobile robot has been implemented as shown in Fig. 3. Haptic feedback is very crucial in telecontrol along with vision and sensory feedback to perceive the environment around the robot. The force acting on the slave robot is fed to master device so that operator can feel the real impact of force acting on slave robot. In telecontrol there is delay due to which force acting on the hapatic device is a delayed response. The force is modeled as virtual force which is acting on robot and is inversely proportional to distance to obstacle in front of robot. With the predictor block it can be made sure that the real time force is generated by using the T_{db}(delay time backward), velocity of slave and onboard proximity sensors to calculate the real time position and hence virtual force.

The operator generates joystick position $X_m(e)$, as it is given in the (1).

$$X_m(e) = \frac{F_h(e)}{C_h} \quad (1)$$

Where C_h is constant and **e** is the event.

$$X_m(e) = \begin{bmatrix} X_{mx}(e) \\ X_{my}(e) \\ X_{m\phi}(e) \end{bmatrix} \quad (2)$$

$$F_h(e) = \begin{bmatrix} F_{hx}(e) \\ F_{hy}(e) \\ F_{h\phi}(e) \end{bmatrix} \quad (3)$$

Where $F_h(e)$ given in (3) is applied force by human operator and $F_m(e)$ is reflected force. $F_{hx}(e)$, $F_{hy}(e)$ and $F_{h\varphi}(e)$ generate $X_{mx}(e)$, $X_{my}(e)$ and $X_{m\varphi}(e)$ positions of the joystick respectively.

The dynamics of joystick is given in the (4), where M_m is mass of master joystick and $V_m(e)$ is velocity of master joystick. Each event is triggered by the previous event as can be seen in the (4).

$$M_m \dot{V}_m(e+1) = F_h(e) + F_m(e) \quad (4)$$

$$V_m(e) = \begin{bmatrix} V_{mx}(e) \\ V_{my}(e) \\ V_{m\phi}(e) \end{bmatrix} \quad (5)$$

$$V_s(e) = V_m(e) - V_{en}(e) \quad (6)$$

$$V_o(e) = \begin{bmatrix} V_{ox}(e) \\ V_{oy}(e) \\ V_{o\phi}(e) \end{bmatrix} \quad (7)$$

$V_m(e)$ travels through communication channel with some delay but this delay has no effect on the performance of system since advancement in time has no effect on slave robot only advancement in event **e** stimulates the slave robot. Therefore when there is connection loss then the slave robot will stop and wait for new event. After reconnection the slave will start following master again. $V_s(e)$ mentioned in the (6) is slave input velocity and it is same as $V_m(e)$ when there is no obstacle in front of robot. $V_{en}(e)$ is reduction in Vs(e) when there is some obstacles to reduce the speed of robot. The proximity sensors mounted on the slave robot scan the environment and adjust the $V_s(e)$. $V_o(e)$ is output velocity of slave robot given in the (7).

V. Teleoperated System

A. *Hardware System*

Fig. 4, illustrates the pictorial view of the teleoperated system. It consists of force feedback joystick, client pc and AutoMerlin along with server PC. Haptic device i.e. force feedback joystick from Wingman generates the commands for robot navigation and reflects the force feedback to human operator. AutoMerlin (Auto Mobile Experimental Robot for Locomotion and Intelligent Navigation) is a four wheel car like mobile robot as shown in Fig. 5, which is equipped with dspic microcontroller and onboard sensors. The robot is driven by two dc motors and the power is transmitted equally to all four wheels. The front wheels are steered by servo motor. The controls are realized by pulse width modulation (PWM) signals from the microcontroller to the drive and steer motors, respectively. The data transfer between AutoMerlin and the server PC has been realized by UART communication using RS232 serial port.

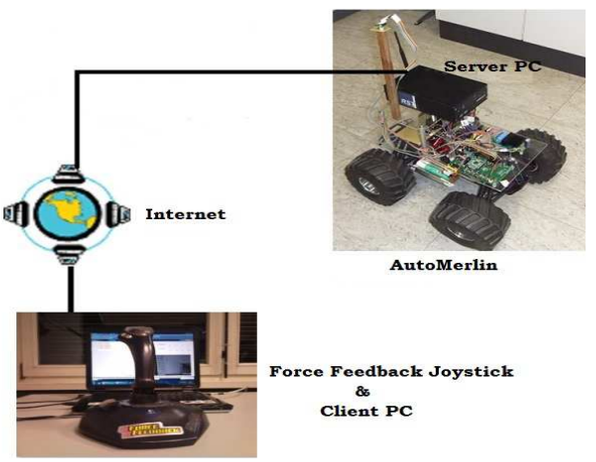


Fig. 4. Pictorial View of the system.

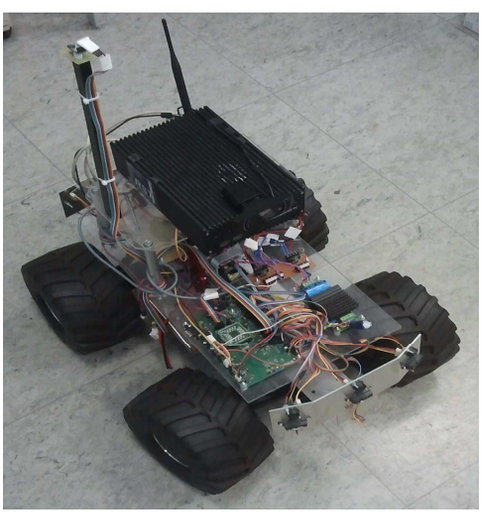


Fig. 5. AutoMerlin

B. Software Framework

Two algorithms have been developed in C# language for both client and the server. The client part is hosted in client PC and it is taking inputs (velocity and steering angle) from the joystick. Then, it has the ability to send information to the server PC via TCP/IP. It reflects force feedback to haptic device that is the force feedback joystick. Second part is the server which is hosted in the server PC. The server algorithm receives velocity commands from the client and process them according to environment e.g. if there is no obstacle then $V_m(e)$ is fed as $V_s(e)$ and if there is obstacle in the range of 0.5m then $V_m(e)$ is adjusted to reduce the $V_s(e)$, and if the object is in the range of 0.2m then the robot will stop. The server execute the velocities commands for 200ms and then wait for the next event. So each event has duration of 200ms. Client generates new commands as next event. These commands travel through network and reach server algorithm which sends them to AutoMerlin via serial port. Server algorithm receives sensor data from the robot and sends them back to the client. Also, it sends the visual feedback from the vision sensor (camera) to enhance the perception of operator about the environment around the slave robot as shown in Fig. 6.

TCP/IP is reliable and guarantees sending-receiving data, therefore it has been used for important data transfer (velocity, heading angle, and sensor data). While UDP is faster than TCP and packet loss is acceptable in video streaming, therefore it has been used for sending live video stream from robot to client.

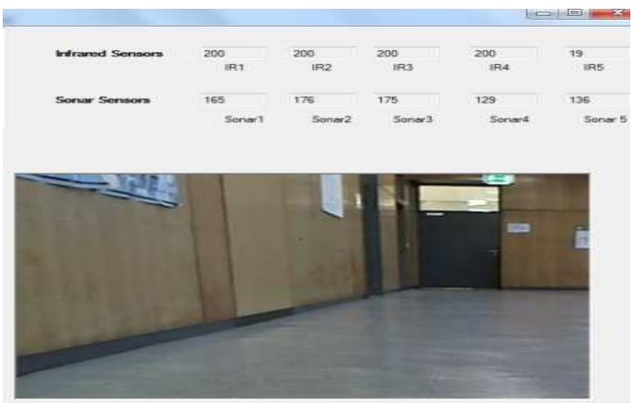


Fig. 6. Client GUI

VI. EXPERIMENTS

Some experiments were carried out over local network to check the behavior of the controller when there is perfect connection between client and server and then we made deliberate disconnection and reconnection to evaluate the performance of controller.

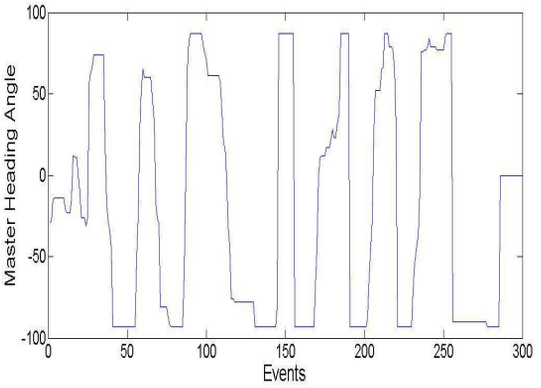


Fig. 7. The master heading angle.

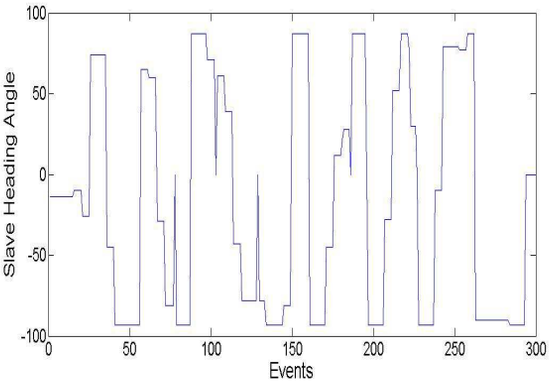


Fig. 8. The slave heading angle.

Fig. 7, and Fig. 9, show the heading angle and linear velocity of master device i.e. haptic force feedback joystick when there is perfect connection between master and slave. Fig. 8, and Fig. 10, show the slave robot heading angle and linear velocity. It is clear from these four plots that the slave robot is following the master device i.e. haptic force feedback joystick.

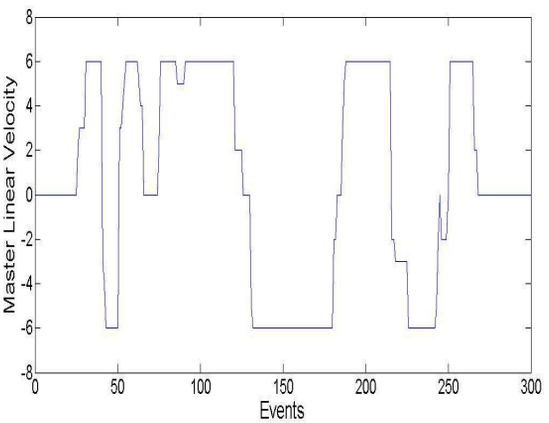


Fig. 9. The master linear velocity.

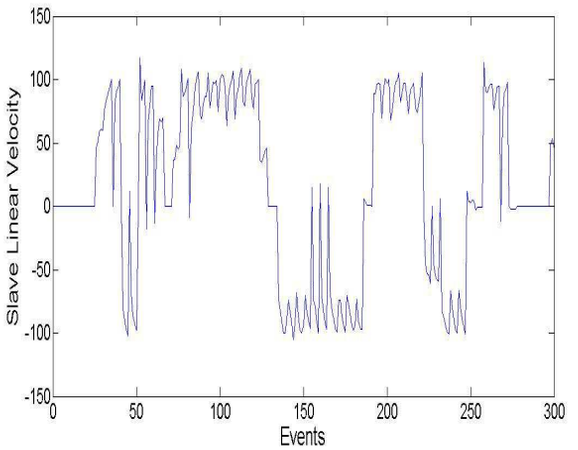


Fig. 10. The slave linear Velocity.

Fig. 11, and Fig. 12, show the master linear velocity and slave linear velocity with disconnection. In the beginning the master and slave linear velocity is zero then slave follows master until there is connection loss and both master and slave velocities become zero. When the connection is reestablished the slave starts following the master again.

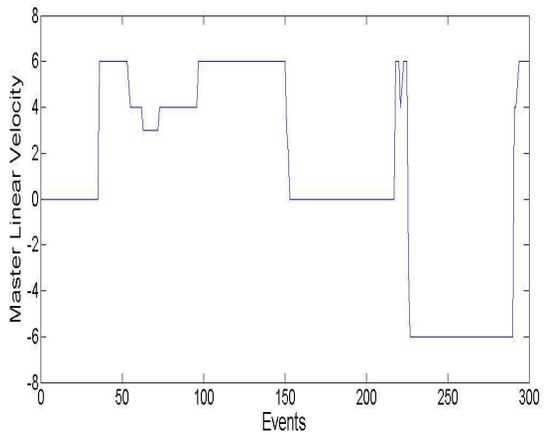


Fig. 11. The master linear velocity with disconnection.

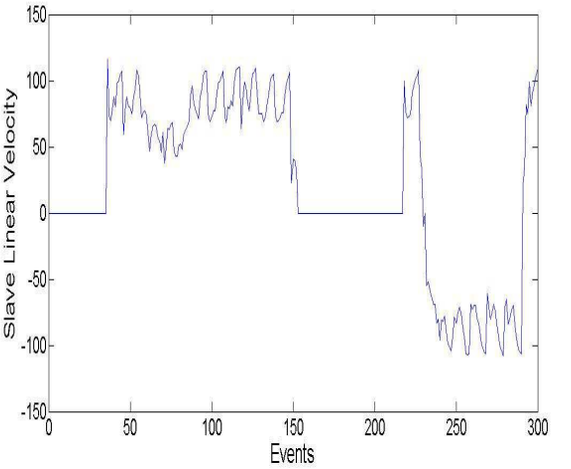


Fig. 12. The slave linear velocity with disconnection.

Fig. 13, plots the master linear velocity, Fig. 14, plots the slave linear velocity and Fig. 15, the force acting on the robot. These plots are related with each other. When there is no obstacle the slave follows the master. At event 50 there is some obstacle so velocity is reduced and force is increased as shown in Fig. 14, and Fig. 15. Slave is not following master now but instead it is following the algorithm developed in server program to avoid obstacles. So when there is maximum force the velocity is of slave is minimum regardless the velocity of master device and when the object is in the critical range of 0.2m the slave will stop and force will become maximum and it will not listen to master device.

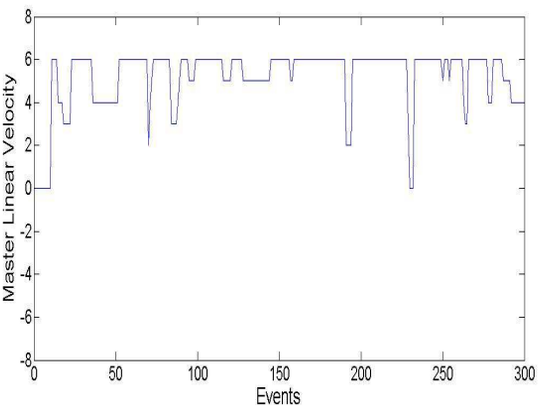


Fig. 13. The master linear velocity when there are obstacles.

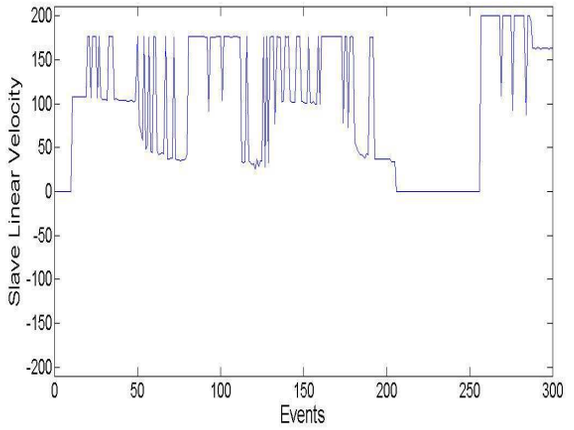


Fig. 14. The slave linear velocity when there are obstacles.

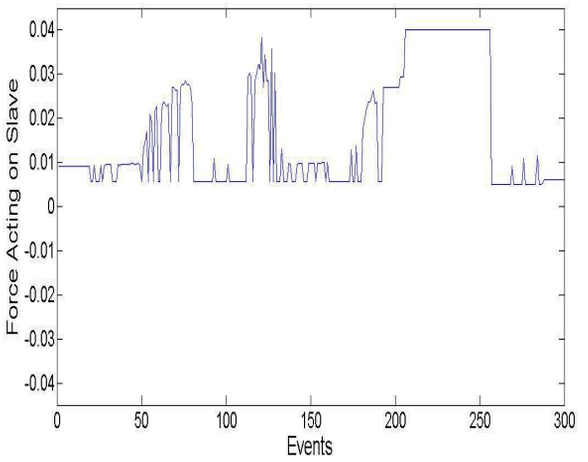


Fig. 15. The force acting on slave reflected to master device.

VII. Conclusion and Future Work

The above mentioned results have been plotted to analyze the performance of controller and the coordination between master and slave. The results are quite impressive and exhibit the excellent coordination between master and slave. In future work the map building will be added to the GUI so that the human operator can understand the environment around slave robot more precisely and the vision system will be used to detect and localize humans in the environment and then send to them the rescue robot after detection.

References

[1] O. Linda and M. Manic, "Self-organizing fuzzy haptic teleoperation of mobile robot using sparse sonar data", IEEE Transactions on Industrial Electronics, vol. 58, no. 8, august 2011.

[2] I. Farkhatdinov, J. H. Ryu, and J. An," A preliminary experimental study on haptic teleoperation of mobile robot with variable force feedback gain," IEEE Haptics Symposium Waltham, Massachusetts, USA,25 - 26 March 2010.

[3] K. M. Al-Aubidy, M. M. Ali, A. M. Derbas, and A.W. Al-Mutairi," Gprs-based remote sensing and teleoperation of a mobile robot, "10th International Multi-Conference on Systems, Signals & Devices (SSD) Hammamet, Tunisia, March 18-21, 2013.

[4] S. K. Cho, H. Z. Jin, J. M. Lee, and B. Yao," Teleoperation of a mobile robot using a force-reflection joystick with sensing mechanism of rotating magnetic field," IEEE/ASME TRANSACTIONS ON MECHATRONICS, VOL. 15, NO. 1, FEBRUARY 2010.

[5] Z. Szanto, L. Marton, P. Haller, and S. Gyorgy,"Performance analysis of WLAN based mobile robot teleoperation," IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 5-7 Sept. 2013.

[6] R. Lozano ,N. Chopra , and M. W. Spong , "Passivation of force reflecting bilateral teleoperators with time varying delay," In Mechatronics'02, Entschede, Netherlands,2002.

[7] P.Berestesky ,N. Chopra , and M. W. Spong, "Discrete time passivity in bilateral teleoperation over the Internet," In Proceedings of the IEEE international conference on robotics and automation, New Orleans, LA, USA,2004.

[8] J. H. Ryu, D. S. Kwon, and B. Hannaford , "Stable teleoperation with time domain passivity control," IEEE Transactions on Robotics and Automation, vol. 20, no. 2, April 2004.

[9] C. Secchi, S. Stramigioli, and C. Fantuzzi, "Dealing with unreliabilities in digital passive geometric telemanipulation," In Proceedings of the IEEE/RSJ international conference on intelligent robots and systems Vol.3, 2003, pp. 2823–2828.

[10] Y. Yokokohji,T. Imaida, and T. Yoshikawa, "Bilateral control with energy balance monitoring under time-varying communication delay,". In Proceedings of the IEEE international conference on robotics and automation,Vol. 3, San Francisco, CA, USA,2000, pp. 2684–2689.

[11] C. Secchi, S. Stramigioli, and C. Fantuzzi, "Digital passive geometric telemanipulation," In Proceedings of the IEEE international conference on robotics and automation Vol. 3,2003, pp. 3290–3295.

[12] W. Wang and K. Yuan, "Teleoperated manipulator for leak detection of sealed radioactive sources," In Proceedings of the IEEE international conference on robotics and automation, Vol. 2,2004, pp. 1682–1687.

[13] W. K. Yoon, T. Goshozono, H. Kawabe, M. Kinami, Y. Tsumaki, and M. Uchiyama, "Model-based space robot teleoperation of ETS-VII manipulator," IEEE Transactions on Robotics and Automation,2004.

[14] T. Imaida, Y. Yokokohji, T. Doi, M. Oda, and T. Yoshikwa, "Groundspace bilateral teleoperation of ETS-VII robot arm by direct bilateral coupling under 7-s time delay condition," IEEE Transactions on Robotics and Automation, 2004.

[15] A. J. Madhani, G. Niemeyer, and J. K. Salisbury, "The black falcon: A teleoperated surgical instrument for minimally invasive surgery," In Proceedings of the IEEE/RSJ international conference on intelligent robots and systems Vol. 2,1998, pp. 936–944.

[16] N. Diolaiti and C Melchiorri, "Teleoperation of a mobile robot through haptic feedback" In IEEE international workshop on haptic virtual environments and their applications,2002,pp. 67–72.

[17] O. J. Rösch, K. Schilling, and H. Roth, "Haptic interfaces for the remote control of mobile robots" Control Engineering Practice, 2002.

[18] L. Dongjun, M. P. Oscar, and M. W. Spong, " Bilateral teleoperation of a wheeled mobile robot over delayed communication network," Proceedings of the IEEE International Conference on Robotics and Automation, Orlando, Florida - May 2006, pp. 3298- 3303.

[19] N. Diolaiti and C. Melchiorri, "Haptic tele-operation of a mobile robot", IFAC 2003.

[20] Z. Xiu, A. Kitagawa, H. Tsukagoshi, C. Liu, and M ido," Internet-based tele-rehabilitation system bilateral tele-control with variable time delay," Proceeding of the 2006 IEEE/RSJ, International Conference on Intelligent Robots and Systems, pp. 5208-5213, Beijing, China, October 2006.

[21] M. P Oscar, L. Dongjun, M. W. Spong, I. Lopez, and C. T. Abdallah, "Bilateral teleoperation of mobile robot over delayed communication network: Implementation," Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, October 9 - 15, 2006.

[22] N Xi and T J Tarn, "Action synchronization and control of Internet based telerobotic systems," IEEE Int. Conf. on Robotics and Auto, May 1999.

Position Control of Shape Memory Alloy Actuators Using a Phenomenological Hysteresis Model

Hamid Basaeri
Graduate Student,
Center of Advanced Systems and Technologies (CAST)
School of Mechanical Engineering, College of Engineering,
University of Tehran, Tehran, Iran
h.basaeri@ut.ac.ir

Aghil Yousefi-Koma
Professor,
Center of Advanced Systems and Technologies (CAST)
School of Mechanical Engineering, College of Engineering,
University of Tehran, Tehran, Iran
aykoma@ut.ac.ir

Mohammad Reza Zakerzadeh
Assistant Professor,
Center of Advanced Systems and Technologies (CAST)
School of Mechanical Engineering, College of Engineering,
University of Tehran, Tehran, Iran
zakerzadeh@ut.ac.ir

Seyed Saeid Mohtasebi
Professor,
Center of Advanced Systems and Technologies (CAST)
School of Mechanical Engineering, College of Engineering,
University of Tehran, Tehran, Iran
mohtaseb@ut.ac.ir

***Abstract*—Shape Memory Alloy (SMA) actuators are hysteretic nonlinear materials, and thus it is difficult to effectively utilize these actuators. As a result of these effects, the position control of these type of actuators has been a great challenge in recent years. Using the phenomenological hysteresis models can compensate the hysteresis of these actuators effectively. In this paper, a feedback controller was used to control the rotation angle of a morphing wing mechanism actuated by an SMA actuator wire. Results showed that the proposed controller performed well in terms of attaining small overshoot and undershoot for square wave tracking as well as small tracking errors for sinusoidal trajectory. It has also good capability for tracking hysteresis minor loops.**

Keywords—shape memory alloy; phenomenological hysteresis model; feedback control system

I. INTRODUCTION

Since the main difficulty in controlling Shape Memory Alloy (SMA) materials is due to their nonlinear saturated hysteretic behavior during forward and reverse transformations, position and force controls of these materials have been a great challenge for practical applications during the last decade. Steady state errors and limit cycle problems are some results of this behavior when conventional controllers are used for trajectory control [1]. Also, while for slowly varying reference signals, and with properly tuned gains, feedback strategies such as Proportional–Integral (PI) control can provide adequate performance, oscillatory motions about the reference trajectory often occur for the fast reference signals [2]. Because of these reasons, recent researches on control of SMA actuators have been guided to follow the nonlinear methods.

In order to reduce the energy consumption by the SMA actuators, PWM controllers are appropriate choices as position controller. Ma and Song [3] concluded, by experimental results, that using Pulse Width Modulation (PWM) for controlling an SMA actuator efficiently saves actuation energy while keeping the same control accuracy as compared to a conventional PD controller. In ref. [4] a simple proportional controller was applied in active shape control of a flexible beam. It was shown by experimental data that in order to eliminate the steady state error of a step input signal, overshoot and actuator saturation are unavoidable. Experimental results showed that among the linear controllers, PI with anti-windup has the best results for position control of SMA actuators [5].

In another method, the hysteresis can be modeled by the use of one of phenomenological hysteresis models like Preisach model, Krasnosel'skii–Pokrovskii model and Prandtl–Ishlinskii model. As these models are defined by integral of hysteresis operators over a specified region, they have more capabilities in modeling of the systems with hysteresis behavior like SMA actuators. Due to the simplicity and availability of the standard controllers in the industrial world, using PI and PID controllers for systems with hysteretic behavior has attracted many researches. In [6], the stability properties of a plant described by a feedback interconnection of a linear system and a Bouc-Wen hysteresis model was investigated and controlled by a PID control system. In another work, the PI control system design for a plant that was purely hysteretic, i.e., without any dynamics, was discussed [7]. It offers a simple bound on the controller gains and it applies to several models. Nevertheless, ignoring the dynamic phenomena leads to satisfactory results only at low frequencies.

In this paper, the generalized Prandtl–Ishlinskii (P-I) hysteresis model is used for modeling the hysteresis of SMA actuator and the position regulation control of the actuator is

performed by using this model with a PI controller. In what follows, the generalized Prandtl-Ishlinskii model is presented and used to model a mechanism actuated by shape memory alloys. The P-I model is trained by the use of some experimental data in order to identify some unknown parameters of the model. The experimental data is obtained from the test setup consisting of a morphing wing mechanism actuated by shape memory alloy wires [11]. The identification process is implemented in order to adapt the model response to the real hysteretic nonlinearity. After that, the controller design is explained and the accuracy of this controller is studied for different inputs like square and sinusoidal waves.

II. Generalized Prandtl-Ishlinskii Model

The classical Prandtl-Ishlinskii (P-I) model uses the classical play (or stop) operator with a density function to characterize the hysteretic behavior of materials. This operator, characterized by the input u and the threshold r determining the width of the hysteresis operator, is a continuous rate-dependent operator which further details about it can be found in [8]. Assume that $C_m[0,T]$ is the space of the piecewise monotone continuous functions and the input $u(t) \in C_m[0,T]$ is monotone on each of the sub-intervals $[t_i, t_{(i+1)}]$, where $t_0 < t_1 < \cdots < t_i < t_{(i+1)} < \cdots < t_N = T$. Then the output of the generalized P-I model, $y_{generalized}$, can be obtained as follows [9]:

$$y_{generalized}(t) = \int_0^R p(\mathrm{r})\, \mathrm{S}_r[u](t)\, dr \qquad (1)$$

In this equation, $p(r)$ is an integrable positive density function, r is the positive threshold as $r_0 < r_1 < \cdots < r_i < r_{(i+1)} < \cdots < r_N = R$, and $S_r[u]$ is the generalized play hysteresis operator that is analytically expressed as:

$$\left\{ \begin{array}{c} S_r[u](0) = g_r(\mathrm{u}(0), 0) \\ \mathrm{S}_r[u](\mathrm{t}) = g_r(u(t), S_r[u](t_i)) \end{array} \right\} \qquad (2)$$

where $g_r(u,z) = max\{\gamma_l(u) - r, min(\gamma_r(u) - r, z)\}$. In the case of practical applications where a finite number of generalized hysteresis play operators is used, Eq. (1) would be expressed as:

$$y_{generalized}(\mathrm{k}) = \sum_{i=0}^{N} p(\mathrm{r})\, \mathrm{S}_{r_i}[u](\mathrm{k}) \qquad (3)$$

According to Eqs. (1) and (3), the generalized Prandtl–Ishlinskii model output depends on the shape of envelope, density and threshold functions. Generally, the shapes of these functions are defined based on the hysteresis loop of a particular material and considering that whether such material has asymmetric hysteresis loops and (or) output saturation or not. Also, the output of the mentioned functions strongly depends on the parameters of these functions. Therefore, these parameters should be obtained on the basis of some experimental data of the actuator in order to correctly predict the behavior of such materials. Due to some good properties of hyperbolic tangent functions [9], in this work, the following functions are selected for the envelope functions of the generalized play operator:

$$\gamma_r(u) = P_1 \tanh(P_2 u + P_3) + P_4 \qquad (4)$$

$$\gamma_l(u) = P_5 \tanh(P_6 u + P_7) + P_8 \qquad (5)$$

Also, the following forms are selected for the density and threshold functions [10]:

$$p_j = P_9 e^{-P_{10} r_j} \quad (j = 0,1,...,N) \qquad (6)$$

$$r_j = P_{11} j \quad (j = 0,1,...,N) \qquad (7)$$

In order to implement the generalized Prandtl–Ishlinskii hysteresis model for behavior prediction of a particular SMA actuator, first the above mentioned 11 constants, including P1, P2, . . . , P11, should be identified using the measured input–output experimental data. In the current research, this process, called training process, is performed with the MATLAB optimization Toolbox. The goal is to have minimum errors with respect to the experimental data. In this paper, the experimental data are collected from an experimental test set-up, consisting of a mechanism actuated by a shape memory alloy wire, and the details about this set-up will be explained in the following section.

III. Experimental Test Setup

The experimental setup which is shown in Fig. 1 consists of a mechanism that is appropriate for morphing wing applications. In other words, this mechanism can be used in a wing so that the wing shape can be changed at different flight conditions [11].

A PC-based experimental test setup and its associated instruments are used to investigate the capability of the generalized Prandtl–Ishlinskii model in prediction of the proposed morphing wing mechanism behavior under a SMA wire actuation. Moreover, the schematic interconnection of these components is depicted in Fig. 2. The experimental setup consists of the test-bed (the morphing wing mechanism equipped with SMA wires and two potentiometers, mounted on a test stand), a data acquisition system, a Windows-based PC, required electronic circuits (bridge circuitry with instrumentation amplifier and antialiasing filter, and voltage-controlled current amplifier circuits) and a power supply.

The main properties of the SMA wires which are used as actuators are presented in Table I. The SMA actuators are made of Nitinol (Ni-Ti) alloy which has excellent electrical and mechanical properties, long fatigue life, and high corrosion resistance and due to these properties, this material is used in many SMA actuators today [12]. Finally, the

specifications of the experimental setup which is used to verify the results of P-I model are listed in Table II.

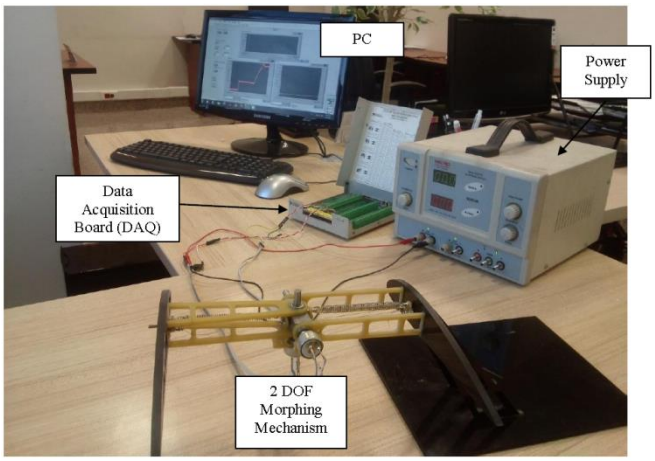


Fig. 1. A View of fabricated morphing wing test setup.

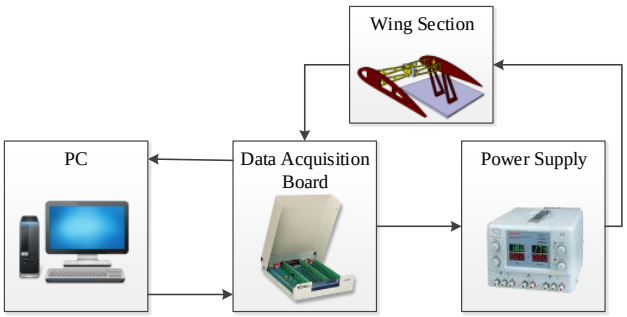


Fig. 2. Schematic diagram of the experimental setup.

TABLE I. PROPERTIES OF SHAPE MEMORY ALLOY ACTUATOR.

Coefficient	Definition	Value	Unit
d_w	Diameter	0.01	*In*
ρ	Density	6.45	g/cm^3
M_f	Martensite final temperature	43.9	^oC
M_s	Martensite start temperature	48.4	^oC
A_f	Austenite final temperature	68	^oC
A_s	Austenite start temperature	73.75	^oC

TABLE II. SOME SPECIFICATIONS OF THE EXPERIMENTAL SETUP [11].

Test-Bed		Morphing wing mechanism with 0.01" NiTi Flexinol wires & potentiometer pair, A test-stand
Data Acquisition		National Instrument, SCB-68 Noise Rejecting, Shielded I/O, Connector Block
PC	Hardware	Core2 Duo 2 GHz CPU, 2GB RAM
	Software	Windows 7, LabVIEW
Circuits		Bridge circuitry with instrumentation amplifying and anti-aliasing filter, Voltage-controlled current amplifier circuit

IV. IDENTIFICATION AND VALIDATION PROCESSES

In order to train the model, a slow decaying ramp signal which is illustrated in fig. 3 is applied to the SMA wire actuator and is defined as the input voltage. The rate of change of the input voltage is carefully chosen to be so small. To train the generalized P-I model, 642 data set which contains the major loop and 10 first order descending reversal curves attached to the major loop is used. The switching values of these descending reversal curves are selected as: [3.5, 3.1, 3, 2.9, 2.8, 2.7, 2.6, 2.5, 2.4, 2.3, and 2.2] (volt). The input voltage to the mechanism is shown in Fig. 3. For switching values less than 2.2 (volt), the change in the mechanism rotation is negligible.

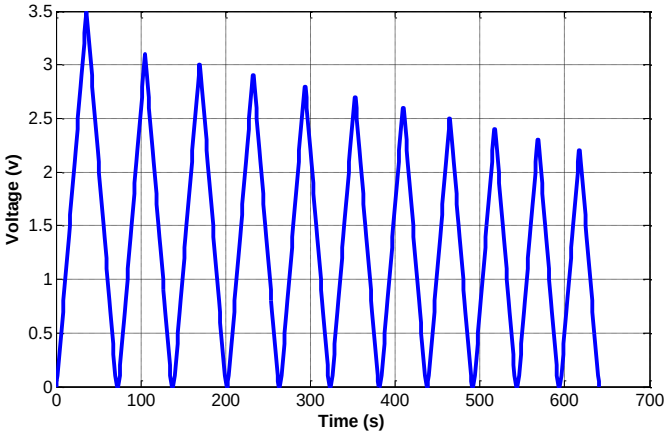


Fig. 3. The decaying ramp input voltage applied in the training process.

The experimental input-output hysteresis loops of the morphing wing mechanism with SMA wire actuator, under the above mentioned input voltage is depicted in Fig. 4. The identification process is implemented in the P-I model and, the 11 generalized P-I model parameters are identified by using MATLAB optimization Toolbox in order to minimize the error between the model output and the experimental data. These values are tabulated in Table III. Since, unlike other hysteresis models, the Prandtl-Ishlinskii model does not have exact output even for the training data, the output of the generalized P-I model in time domain under the actuation voltage profile of Fig .3, is compared with the experimental data shown in Fig. 5. This figure obviously depicts that the generalized P-I model, with selected envelope, density and threshold functions and their corresponding parameters in Table III, can effectively characterized the behavior of the morphing wing mechanism with SMA wire actuation and there are only small differences for some data.

In order to show the accuracy of modeling prediction more clearly, the maximum, mean and mean squared values of the absolute error are also presented in Table IV. Since the maximum rotation of the mechanism under the SMA wire actuation is around 16 degrees, the peak prediction error in this case is about 9.33% of the maximum output.

TABLE III. PARAMETERS OF GENERALIZED PRANDTL-ISHLINSKII MODEL IDENTIFIED BY THE MEASURED EXPERIMENTAL DATA OF SMA-ACTUATED MECHANISM.

P_1	3.3784	P_2	1.6268	P_3	-3.7776
P_4	1.5592	P_5	3.7571	P_6	1.7893
P_7	-4.0901	P_8	-1.1545	P_9	0.5657
P_{10}	0.3016	P_{11}	0.1378		

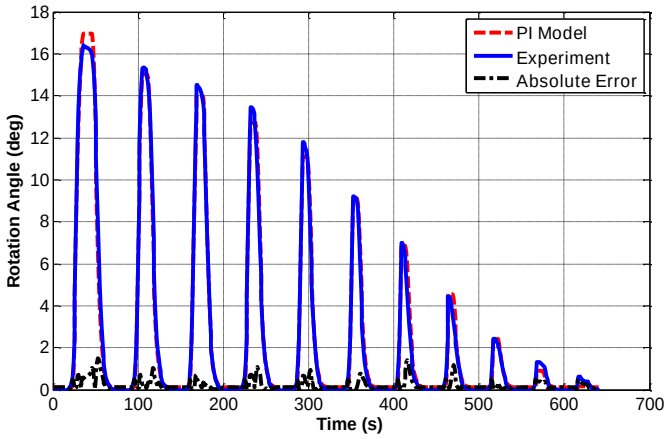


Fig. 4. Comparison between the mechanism rotation predicted by the generalized Prandtl–Ishlinskii model and the measured data.

TABLE IV. ERROR OF THE GENERALIZED PRANDTL–ISHLINSKII MODEL IN THE TRAINING PROCESS.

Mean of Absolute Error (deg)	Max of Absolute Error (deg)	Mean of Squared Error (deg)
0.21	1.52	0.11

Most of the phenomenological hysteresis models have trouble in predicting higher order hysteresis minor loops. For evaluating the ability of generalized P-I model under these circumstances, in the validation process a damped voltage profile shown in Fig. 6 is applied to the current amplifier of the SMA actuator.

The predictions of the higher order hysteresis minor loops by the generalized P-I model are compared with the experimental measured data in Fig. 7. The absolute error response with respect to measured data, in time domain, is also presented in Fig. 8. The maximum, mean and mean square values of the absolute error are also presented in Table V. The peak prediction error in this case is about 17.1% of the maximum output.

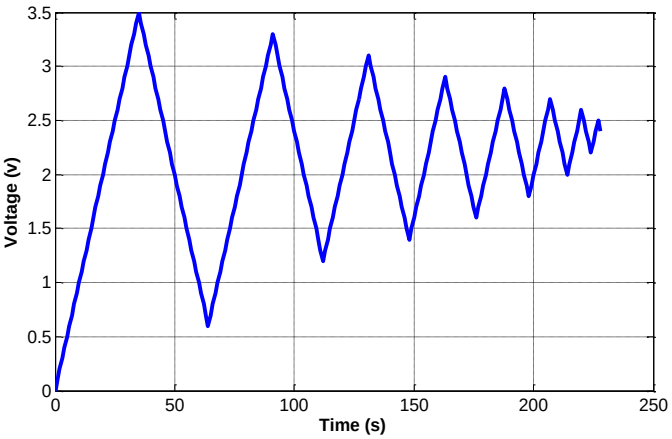


Fig. 5. The input voltage profile applied in the validation process.

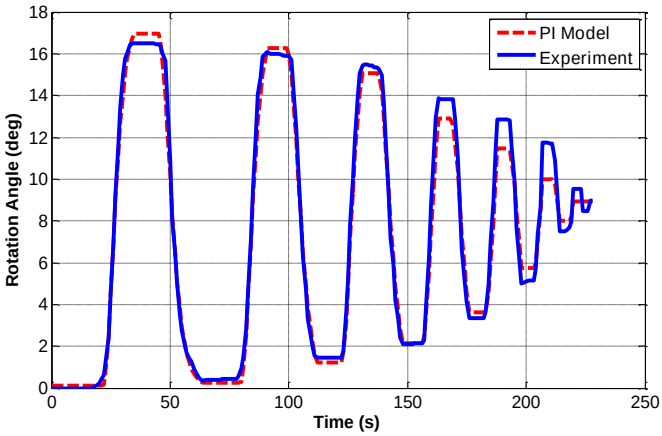


Fig. 6. Comparison between the rotation response of the generalized Prandtl-Ishlinskii model and the measured data in the validation process.

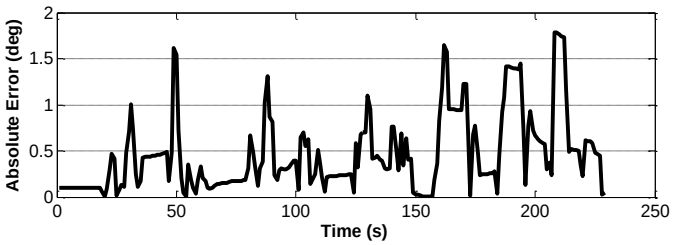


Fig. 7. Time history of absolute error between the generalized Prandtl-Ishlinskii model and experimentally measured data in the validation process.

As it was anticipated and it is observable from Fig. 7 and Table V, the generalized P-I model has adequate accuracy in predicting the higher order hysteresis minor loops particularly when, like this case, it has been only trained with some first order hysteresis reversal curves attached to the major loop.

TABLE V. ERROR OF THE GENERALIZED PRANDTL–ISHLINSKII MODEL IN THE VALIDATION PROCESS.

Mean of Absolute Error (deg)	Max of Absolute Error (deg)	Mean of Squared Error (deg)
0.46	1.79	0.38

V. CONTROL

As the generalized Prandtl–Ishlinskii model has more accuracy for hysteresis modeling of SMA actuator with respect to Preisach and Krasnosel'skii–Pokrovskii hysteresis models [13], especially for high order minor loop prediction, in this paper, the generalized Prandtl–Ishlinskii model is used for compensating the hysteresis nonlinearity of SMA-actuated morphing wing mechanism. The block diagram of the proposed controller is shown in Fig. 9. The desired rotation is used as the input. The PI controller generates the required control voltage signal for the desired trajectory tracking. The gains of the PI feedback controller, denoted by K_I and K_P respectively, are 0.1 and 0.3. The values of these gains are set in such a way that system response to step reference trajectory has the minimum overshoot as well as quick response.

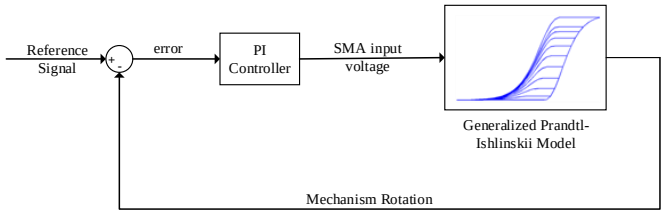


Fig. 8. Closed-loop scheme with a PI controller.

In order to show the effectiveness of the proposed control system for hysteresis compensation together with accurate position control of morphing wing mechanism, four sets of reference signals are selected for this controller verification process and the results of each test are presented later in this section.

A. Square wave tracking with fixed lower bound

As stated in Section IV, the generalized Prandtl–Ishlinskii hysteresis model is trained with the data of first order reversal descending curves attached to the ascending branch of the major loop. The purpose of the current experiment test is to verify the ability of the proposed controller in tracking a square waveform trajectory, which leads to predict some first order reversal curves behavior by the hysteresis model. The set-point values of desired rotation are chosen as: [16, 15, 13, 12, 10, 8, and 6] (volt). Fig. 10 shows the responses of the proposed control system. Fig. 11 also illustrates the SMA voltage over time in this test.

As it is obvious, the proposed control system can accurately track the reference trajectory with minimum oscillation and also eliminates the steady state errors.

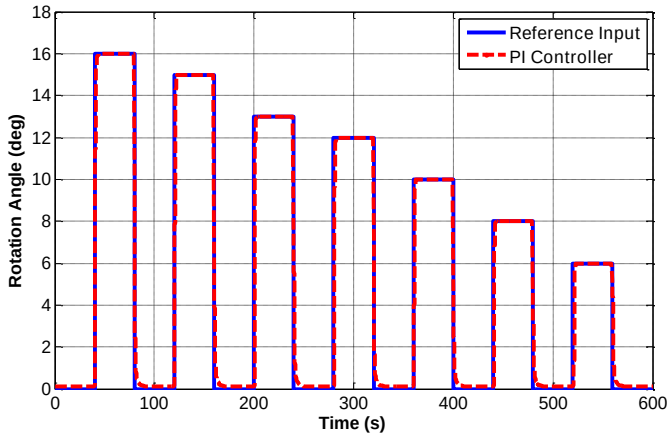


Fig. 9. Tracking control of a square wave with fixed lower bound.

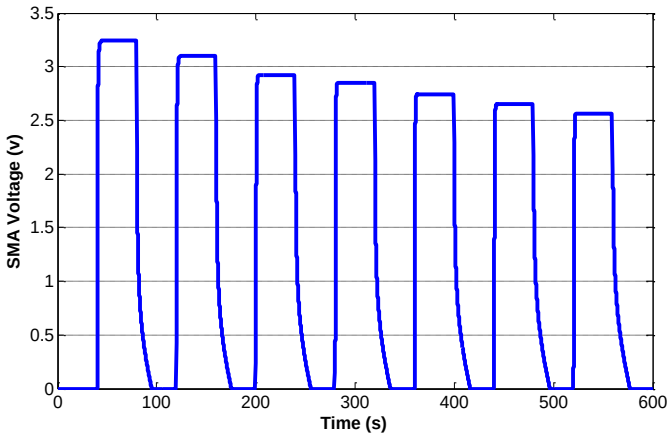


Fig. 10. SMA voltages during tracking of a square wave with fixed lower bound.

B. Square wave tracking with high order minor loops

Since many hysteresis models proposed in the literature failed to track high order minor hysteresis loops, in this experiment the response of the proposed control system is verified in tracking a square waveform trajectory with variable amplitude which leads to some minor hysteresis loop prediction. The set-point values of desired deflections are chosen as: [17, 2, 15, 4, 13, 7, 12, and 9] (volt). Fig. 12 depicts the performance of the developed control system in tracking such reference rotation signal. Furthermore, the control signal applied by the control system to the SMA actuator can be observed in Fig. 13.

While the generalized Prandtl–Ishlinskii model was only trained by data of some first order reversal curves of voltage–rotation hysteresis loops, the controller has moderate accuracy in tracking the desired deflection. Due to controller good performances, the steady state error has been easily eliminated by the closed-loop conventional PI controller and the system has minimum oscillations about the set points of desired signal. As a conclusion, against many control systems that have difficulties in minor hysteresis loops tracking, the

accuracy of this control system is excellent without using any data of minor hysteresis loops in the training process.

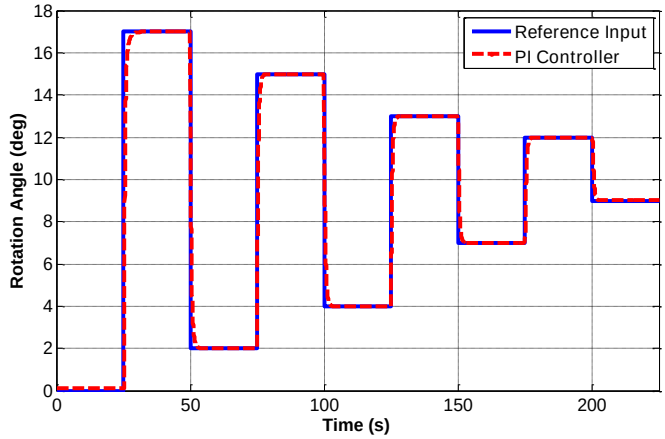


Fig. 11. Tracking control of a square wave with high order minor loops

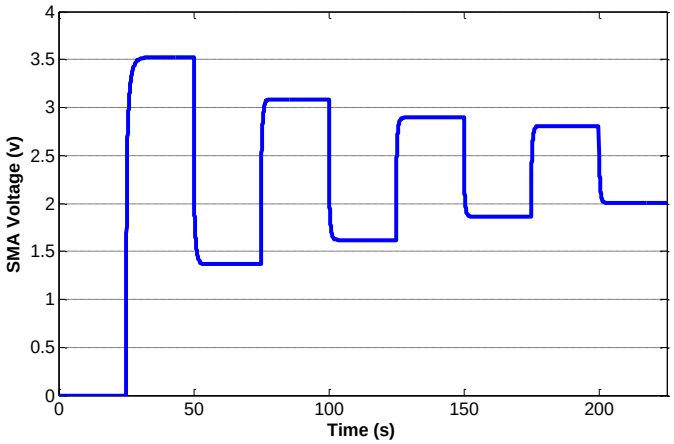


Fig. 12. SMA voltages during tracking control of a square wave with high order minor loops.

C. *Tracking of a fixed amplitude sinusoidal input*

In the current test, the reference input signal is a sinusoidal trajectory with fixed amplitude. Time functionality of this input is selected as 12sin(0.02*time - π/2)+12 which leads to a hysteresis loop prediction with no higher order minor loops in it. The result of the proposed control system is investigated and shown in Fig. 14. Also, Fig. 15 illustrates the SMA voltage over time in this test. The proposed control system can accurately track the reference trajectory with minimum oscillation. This will be clearer when the absolute value of the rotation error is depicted over time in Fig. 16. The absolute error average for the proposed control system is 0.43 degree.

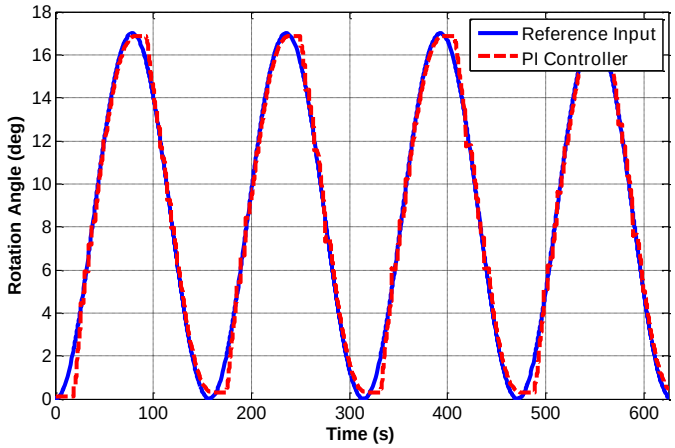


Fig. 13. Tracking control of a sinusoidal wave with a fixed amplitude.

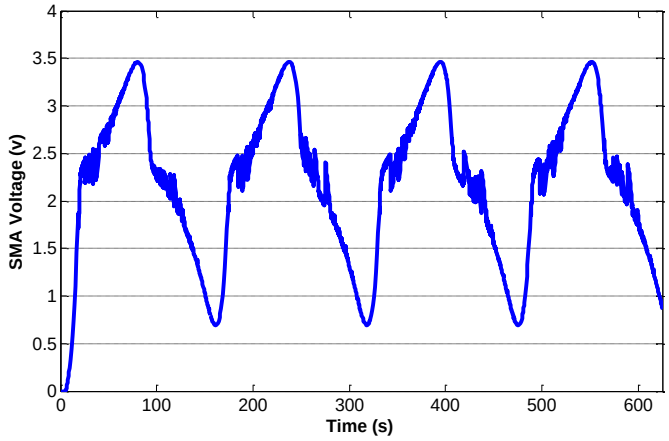


Fig. 14. SMA voltages during tracking control of a sinusoidal wave with a fixed amplitude.

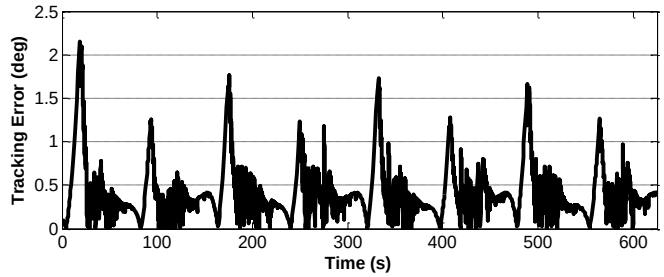


Fig. 15. Absolute of tracking error in tracking control of a sinusoidal wave with a fixed amplitude.

D. *Tracking of a decaying sinusoidal reference input*

In the last test, the reference input signal is a decaying sinusoidal trajectory which results to predicting some high order minor hysteresis loop by the hysteresis model. The time functionality of this input is selected as (12sin(0.02*time - π/2)+12)*exp(-0.005*time) which not only is a decaying input but also its variation with respect to time is almost fast. Thus, the system response to this decaying input can verify the good performance of the controller. The result of the proposed

control system is shown in Fig. 17. The system has only weak tracking in the extremum regions of the reference input in which its derivative sign changes. Moreover, the control signal applied by the control system to the SMA actuator is shown in Fig. 18, and the absolute value of the rotation error over time is displayed in Fig. 19. The absolute error average for the proposed control system is 0.28 degree. It should be noted that this worth result is obtained by only training the Prandtl–Ishlinskii hysteresis model with the data of some first order reversal curves. It means that whether the generalized Prandtl–Ishlinskii model is trained by the first order reversal curves data or is trained by higher order minor loops data, the model has good prediction of the high order minor hysteresis loops of SMA actuator.

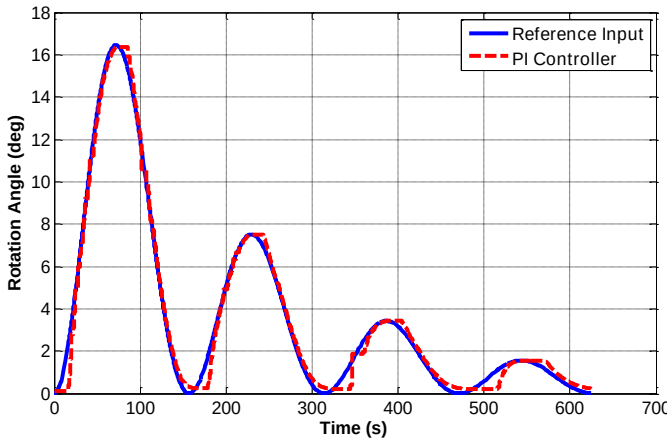


Fig. 16. Tracking control of a decaying sinusoidal wave.

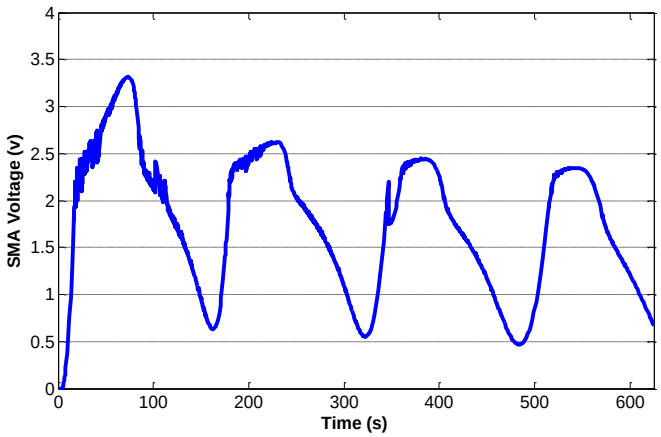


Fig. 17. SMA voltage during tracking control of a decaying sinusoidal wave.

I. Conclusion

In this paper, the generalized Prandtl–Ishlinskii model was used to model asymmetric nonlinear hysteresis behavior of Shape Memory Alloy (SMA) actuator. This model was used in a plant with a Proportional Integral (PI) controller to control a morphing wing mechanism actuated by SMA actuators. It was shown that the proposed control system has great capability in tracking square and sinusoidal trajectories and leads to low tracking error. Although the proposed control system has simple structure, it can be used for other smart structures due to the results obtained in this study. Also, it can be easily implemented for online applications and leads to good tracking error for trajectory with hysteresis loops.

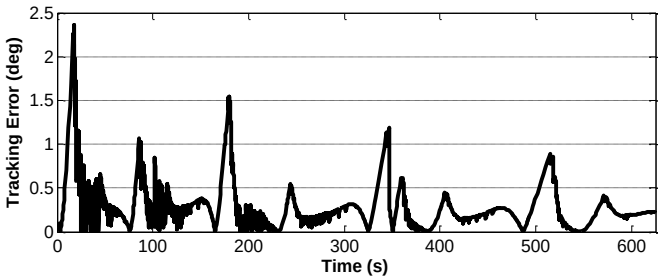


Fig. 18. Absolute of tracking error in tracking control of a decaying sinusoidal wave.

References

[1] Lee HJ, Lee JJ. “Time delay control of a shape memory alloy actuator” Smart Materials and Structures, 2004;13:227–39.

[2] Webb G, Kurdila A, Lagoudas D. “Adaptive hysteresis model for model reference control with actuator hysteresis” Journal of Guidance Control and Dynamics, 2000; 23(3).

[3] Ma N, Song G. “Control of shape memory alloy actuator using pulse width modulation” Smart Materials and Structures 2003;12:712–9.

[4] Da Silva EP. “Beam shape feedback control by means of a shape memory actuator” Materials and Design 2007;28:1592–6.

[5] Asua E, Etxebarria V, Garc´ıa-Arribas "A. Micropositioning control using shape memory alloys" In: Proceedings of IEEE conference on control applications CCA2006, Munich, Germany; 2006. p. 3229–34.

[6] F. Ikhouane and J. Rodellar, “A linear controller for hysteretic systems”, IEEE Transactions on Automatic Control, vol. 51, no. 2, pp. 340-344, 2006.

[7] S. Valadkhan, K. Morris, and A. Khajepour, “Robust PI control of hysteretic systems,” IEEE Conference on Decision and Control, Cancun, Mexico, December 9-11, 2008, pp. 3787-3792.

[8] M. Brokate and J. Sprekels, “Hysteresis and Phase Transitions”, New York, Springer, 1996.

[9] M. Al Janaideh, S. Rakheja, and C-Y Su, “A generalized Prandtl-Ishlinskii model for characterizing hysteresis nonlinearities of smart actuators,” Smart Materials and Structures, vol. 18, no. 4, pp. 1-9, 2009.

[10] H. Sayyaadi, M. R. Zakerzadeh, and M. A. V. Zanjani. "Accuracy evaluation of generalized Prandtl-Ishlinskii model in characterizing asymmetric saturated hysteresis nonlinearity behavior of shape memory alloy actuators." International Journal of Research and Reviews in Mechatronic Design and Simulation (IJRRMDS) 1, no. 3, 2011.

[11] H. Basaeri, A. Yousefi-Koma, M. R. Zakerzadeh, and S. S. Mohtasebi, “Development of a Bio Inspired 2 DOF Morphing Wing Actuated by Shape Memory Alloy”, 2nd International Conference on Manufacturing (Manufacturing 2014), Feb. 9-10, 2014. Singapour.

[12] M. Novotny, J. Kilpi, “Shape Memory Alloys (SMA)”, [online]. Available on: http://www.ac.tut.fi/aci/courses/ACI-51106/pdf/SMA/SMA-introduction.pdf.

[13] Zakerzadeh, Mohammad R., and Hassan Sayyaadi. "Experimental comparison of some phenomenological hysteresis models in characterizing hysteresis behavior of shape memory alloy actuators." Journal of intelligent material systems and structures, 2012: 23(12) 1287–1309.

An Efficient Detection Algorithm for TCP/IP DDoS Attacks

Heshem A. El Zouka
Department of Computer Engineering, College of Engineering and Technology
Arab Academy for Science & Technology and Maritime Transport,
Alexandria, Egypt
helzouka@aast.edu

Abstract — **TCP is widely known as being able to provide a reliable byte stream process for communication. It is the most widespread protocol used for exchanging data on the internet and is responsible for more than 90 percent of the world's total traffic on the internet. However, many TCP protocol were designed with little consideration given to the security implications. The TCP protocol stack, for example, could be quite vulnerable to a variety of attacks ranging from IP Spoofing to distributed denial of service (DDoS) attack. This paper presents a number of known TCP/IP attacks methods focusing in particular on password sniffing, SYN flooding, IP Spoofing, TCP Sequence number attack, TCP session hijacking, RST/FIN attacks, and the low rate TCP targeted denial of service attack. This paper also examines the drawbacks of these TCP protocols in order to provide solutions to such attacks. Finally, a real time network simulation is provided along with detailed experiments analysis to validate the efficiency of our security approaches.**

Keywords- Distributed Denial of Service; Network Security; TCP Congestion Control ; Vulnerabilities.

I. INTRODUCTION

TCP/IP is a set of protocols that is developed to support the internet & allow distant computers to share resources across network. Today, many venders are working on developing protocols that support TCP/IP. The TCP/IP protocol is reliable, robust & has become a standard for open system communications in today's networked world. However, TCP/IP protocols suffer from a number of security flaws due to the improper implementation of the applications which use these protocols. TCP, attacks have many forms including information theft, system compromises and denial of service (DOS) attacks [1], [2]. Generally, these attacks are classified into passive and active attacks depending on the motivations of the attacker. The most common form of passive attack is "*Sniffing*"; where the attacker is able to scan the network and read the data which is being transmitted. Therefore, the information which could be used to potentially gain unauthorized access to a server must be protected. On the other hand, active attacks involve writing data to network and its subsequent attempts to break the protection mechanisms in TCP suits. Indeed, any part of the data segments of TCP can be forged. For example, the source IP address can be forged, and the attack in this case is generally referred to as a "*Spoofing*" attack. Active attacks take on many forms, including DoS and distributed denial of service (DDoS) attacks [3], [4].

Therefore, The TCP/IP protocols are vulnerable to a variety of threats and attacks which may affect the throughput and delay performance in TCP networks. A program that carries out most of these attacks in is freely available on the Internet. These attacks, unless carefully utilized, can place the users of intranet at considerable risk and expose them to various types of vulnerabilities. This paper classifies a range of known attack methods focusing in particular on targeted denial of service. The paper concludes with an examination of the protocols carried by TCP/IP (including FTP for file transfer, SMTP for mail transfer, SNMP for network management and https for https for secure communication) and proposes a novel method to limit their vulnerabilities.

The rest of the paper is organized as follows: in section two, TCP/IP security protocols are surveyed including their vulnerabilities & associated attacks. In section three, the performance of a flow and congestion control mechanism is presented along with other recent DOS attacks. Section four presents a new approach for defending against DOS attacks in TCP/IP environments. Section five presents simulation results, comparing the network performance with and without the proposed defending method that protect against DOS attacks. Finally, section six summarizes the simulation results & outline recommendations for future work.

II. TCP/IP SECURITY THREATS AND ATTACKS METHODS

TCP attacks have enormous adverse impact on the internet and this explains why they present an issue of paramount importance in research communities and industrial communities. They include session hijacking, SYN flooding, IP spoofing and DDoS attacks. There will be a great need, then to handle those attacks and to provide

suitable defense methods against them. Most of the solutions that were proposed to prevent those attacks provided passive protection at or close to destinations, but on the other hand, some others provided active protection but with consequential damage. For all those reasons, it has become quite necessary to provide adequate defense mechanisms against these attacks. In this paper, various types of TCP attacks and aspects are dealt with, as well as their potential solutions to be studied and evaluated. In addition, analyzing TCP connections is one of the most important issues to address in case of Denial of Service (DoS) attacks. To establish a connection, a client sends an initial SYN segment, and upon receiving it, the server replies with a SYN/ACK segment. Then, the user returns an ACK segment to complete the three-way handshaking protocol. Finally, the communication will proceed until either the client or the server sends a FIN/RST segment. The potential for exploiting this vulnerability lies in the early stages of the connection where the attacker can generate multiple SYN segments without sending the ACK segment that completes the three-way handshaking connections. The SYN segment requests can quickly exhaust the server resources so that it cannot accept more incoming connection segments. SNY flooding is fairly similar to DoS attack where an attacker tries to identify vulnerable hosts by sending multiple SYN segments without completing the connection establishment [6]. Similarly, a port-scanning attack can happen whenever the port is open and the server responds with a SYN/ACK segment. In this attack, the port scanner completes the handshaking process by sending a RST segment and immediately closes the connection. Recently, several solutions have been provided to detect these attacks. Most of these solutions are based on a border router which connects the server to the Internet. In the border router method, each packet has to pass through three modules: flow sampling module, object module, and filter module. In sampling module, the packet is classified into a particular flow by examining its source and destination IP addresses, and other information such as port number of the source. Then, the packet passes to the object module which attempts to detect attacks by keeping necessary information updated by all active path flows. The third module is the monitoring module which conducts through the object module and periodically scans specified packets [7].

However, the obvious deficiency of these solutions is that they have to maintain connection information for all established TCP flows, which requires more memory and processing resources in each router. Secondly, considering only a few number of routers may degrade the detection capacity of all existing techniques because of the fairly low probability of scanning all essential control segments. As a consequence, the existing techniques result in poor detection rates and a high number of false positives.

While DoS attacks can be detected by careful analysis and filtered by routers, another class of DoS attacks, known as low rate TCP targeted DoS attacks, are capable of eluding traditional detection. Low rate TCP attack focuses on the bottleneck links in the network, and exploits TCP's retransmission timeout mechanism by injecting periodic bursts of packets. In this attack, the attacker tries to exploit the TCP's congestion avoidance algorithm and force TCP connections to timeout with almost zero throughout. According to Tahoe congestion avoidance algorithm, whenever a packet is lost and its acknowledgement (ACK) does not come before RTO period, it then, sets CWnD to one maximum segment size (MSS) and starts slow start phase until it reaches the threshold value. However, when CWND is above the threshold value, the state is congestion avoidance phase on which the growth of the CWND is linear. Also, when a triple ACK occurs, the threshold is set to half of value of CWND. Finally, at the occurrence of timeout connection state, the CWND is set to 1MSS and the threshold to half of this value. The whole algorithm is illustrated in Figure 1.

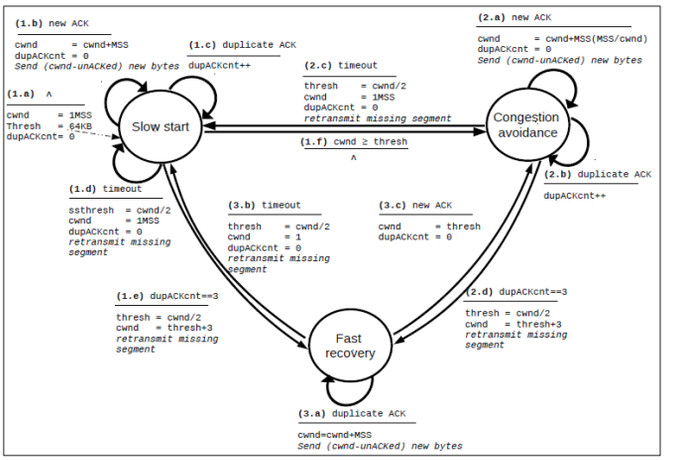


Figure 1. TCP Congestion Control

The problem with Tahoe algorithm is that it takes a complete-time period to detect a packet loss. Also, since it does not send immediate acknowledgments, but rather sends three duplicate ACKs, then, also, CWND window is halved and hence, the transmitted bit rate is reduced significantly.

So, one key question is how to calculate the length of time out value: if set too high, flows will wait long time to recover from congestion. .Similarly, if set too low, retransmissions will occur whenever packets are assumed lost while in fact the data are merely delayed. To address this problem, Paxon and Allman experimentally showed that TCP gives nearly maximal throughput if the RTO value is bounded to one second [8]. However, this fact is utilized by attacker and they set their minRTO to be one second also. Since the atteacker determines the appropriate minRTO, then, all the pakcets of one congestion window (CWND) will be retransmitted after this minRTO whenever RTO equal to minRTO and untill it experiences a dropped packet.

This causes the queues at the routers to build up producing a fload attack.

There is also another type of DoS attack, which was developed more recently and which is named Induced Low Rate TCP Targeted denial of service attack [9]. The name states that the attacker induces a server to perform low rate attack. The attacker can perform this attack by exploiting the technique of optimistic acknowledgement. This new technique involves an attacker sending a stream of acknowledgements for a data that has not yet been received.

In this scenario, a TCP receiver acknowledges the receipt of data segments sent to it by the server. A TCP server varies its transmission rate based on receiving acknowledgments of the segments it sends. An optimistic TCP acknowledgment is an ACK sent by the attacker for a segment that it has not yet received. A DoS attack exists in the potential for a recipient to send optimistic acknowledgments timed in such a way that they correspond to legitimate data segments that the sender has already injected into the network. As a result, the server believes that the transfer is progressing better than it actually is and may increase the transfer rate at which it sends data. An important impact of this condition is the induced low rate the attack introduces. An attacker exploiting this DoS attack can potentially cause victim servers to transmit much more data than the bandwidth available to the attacker causing congestion to the network.

III. SECURITY IMPLICATIONS AND CONSIDERATIONS

Several schemes have been presented recently to defend against denial of services attacks in TCP/IP environments. What we should consider here is that the attacker has already identified the drawbacks and has formed DoS attacks characterized by the use of low rate and request data to achieve the flood of server resources. This may also appear in the TCP/IP suite and its own protection mechanisms due to the security flaws as mentioned above in this paper.

In [11], [12] a survey on detection and protection schemes has been proposed. In this approach, the researchers have analyzed several DoS attack patterns of different TCP congestion controls such as TC Reno and Taho and proposed two alternative solutions. The first solution involves fair queue scheduling algorithm, where all the packets have equal opportunities to access the edge router in a manner that limits the effectiveness of DoS attack. The drawbacks of this solution are that it is not scalable and it requires potentiality in distributing routers. On the other hand, randomizing minRTO would also affect the protocol performance during non-attack periods.

At present, most traditional protection mechanisms against DoS attack are effective only when attacking signatures are detected. So, the security protocols must record more network information, so that we can enhance the performance of detection. Certainly, recording more information means that more resources will be consumed. To avoid this, Shevatker et al. [13] proposed an approach that deploys an early detection module on the edge router which connects server to the Internet. The installed detection code will examine, then ,the IP address and port number of the received packets, and check whether they are malicious or not. Therefore, this module detects attack and prevents flooding or abnormal state of network such as network congestion. However, one drawback of this approach is that it cannot maintain all information needed to detect all TCP flows in aggregate traffic due to the limited resources of the edge routers. In addition, this approach fails to detect attack when it occurs in distributed network environment where no central authority is present to verify other edge routers.

To protect the TCP implementation against TCP targeted DoS attack, a number of countermeasures have been proposed. Some schemes are designed specifically to analyze the traffic by using different analytical techniques including sampling, filtering, signature matching, and feature extraction. For example H. Sun et al. [14] proposed a coordinated router to check what sort of traffic is going out, and determine to which server it is communicating. Thus, if the targeted TCP DoS attack is discovered by the router, action can be taken to defeat the attack and the router tries to find out which port the attack is coming from and port the attack it is targeting. The effectiveness of this scheme comes from the ability of the router to detect attacks close to their sources and even go so far as to distinguish between malicious packets and legitimate ones.

Unfortunately, while achieving detection for attacks that uses a single source of attack, this scheme fails to detect attack that is done in a distributed fashion. It is also clear that the detection and prevention technique has to be applied to all routers in the network, as source can be anywhere in the network area. In addition, this detection scheme requires sampling, filtering, signature matching, and feature extraction techniques which would not be considered for large networks.

There are other schemes that have proposed in the literature to exploit the abnormal behavior of networks and their services to detect the attack, thereby mitigating the attack very effectively [15]. For example, other researchers proposed an approach to deselect these attack patterns at border router or firewall [16]. Based on the access history stored in the internal structure of the routers, each packet flow will be marked as malicious if its traffic statistics in the flow table of the router indicate that the flow takes an extended period of time. So, if the flow presents a periodic pattern with period equal to minRTO, then the burst length is greater than or equals the RTT's of other connections with the same server. However, the deficiency of this scheme is that it cannot distinguish between legitimate packets and malicious ones since the normal packets often exhibit the same behavior as the one infected [17].

Kwork et al. [18] proposed a scheme called HAWK (halting Anomaly with Weighted choking), where they record attack flows into a small table and drop packets from those flows to halt the attack. A general drawback of this

type of approaches is that they have difficulty in identifying a set of distributed Shrew attack flows where each flow occupies a small bandwidth share, while the aggregate bandwidth is large enough [19]. In particular, although HAWK can possibly tell whether a Shrew traffic exists among legitimate flows, it cannot precisely isolate malicious flows. In contrast, SAP does not attempt to identify the attack flows; it simply controls the drop rates of victim flows.

IV. Proposed Mechanism

For securing the TCP protocol against optimistic TCP acknowledgment that can cause denied of service attack, a new mechanism is presented here. It was first presented by S. Savage et al. in their paper, "TCP congestion control with a misbehaving receiver" as they briefly mentioned the possibility of using this attack to drive attack. Later, R. Sherwood et al. published a paper entitled. "Misbehaving TCP receivers can cause internet wide congestion collapse", in which they present the impact of this attack on TCP based network. In normal operation the attacker hopes to increase the date rate of packets sent by the server and, consequently, exhaust the server and prevent it from responding to legitimate users. Consequently upon receiving the acknowledgement, the attacker will intentionally increase the CWND by one MSS in attempt to increase the rate of data transferred to him in a malicious manner. Obviously, when the attacker requests large files from the target server, eventually leading it to increasing the data rate and congest the network as shown in figure 2.

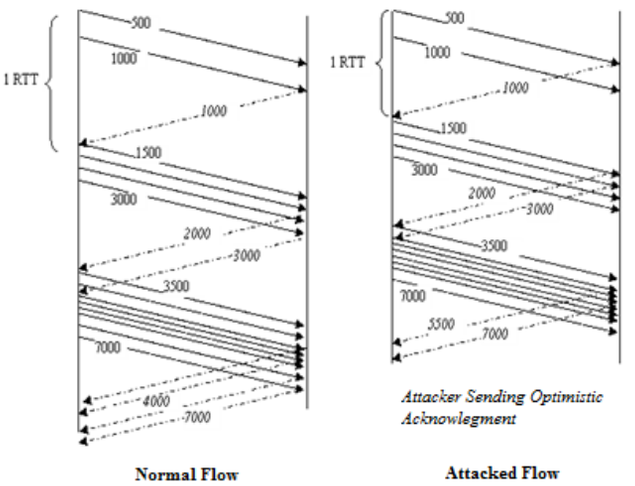


Figure 2. DoS Attack Scinario

Now once that happens for an extended period of time, the resources of the server will be consumed by sending more data segments.

In addition**,** only one solution is proposed to detect such induced DoS attack [20], and involves adding new field in the TCP header which is rather impractical solution. In this scheme, the author introduces two fields into the TCP packet format: Nonce and Nonce reply. For each data segment, the sender fills the nonce field with a unique number generated when the segment is sent. When an acknowledgment is generated by the receiver in response to a segment, it produces the nonce value by writing it into this nonce reply field by which the sender can arrange to only increase CWND in response to duplicate acknowledgments as indicated earlier in figure 1. Obviously, the implementation of TCP suite can't be modified all at once, all over the world. One possible solution was proposed by R. Sherwood [21] who testified that any server could drop packets intentionally in attempt to observe the optimistic acknowledgment attack. Therefore, if this type of DoS attack exists, the attacker will not send duplicate acknowledgment, and in this way optimistic acknowledgment is detected. The main deficiency of this solution is that it decreases the throughput efficiency of a genuine TCP flow as they have to reduce their data stream down when the victim's sever intentionally drops the packets. Other solutions involve the changes in two modules – the server which does all the detection and mitigation process, and the attacking module which has access to variety of tools to attack the victim's server. The attacker in this proposed scenario sends false acknowledgment bytes which are equivalent to the regular segment size (1MSS). Whenever a new segment with bytes less than MSS size is transmitted, a variable packet number is increased by one, and hence, the difference can be detected. Also, the solution of the low rate TCP targeted attack can be applied here, but these two approaches still have their own limitations and weaknesses as discussed in the previous paragraph in this paper.

In this paper, an algorithm which detects the induced attack in general and mitigates the optimistic acknowledgment is proposed. The potential of this proposed approach lies in the fact that it involves a little change of implementation in the victim's server, and it does not require any modification neither in the TCP packet format nor in the router. In addition, the complexity of the algorithm is in the order of packet transmission time. Therefore, the proposed approach does not require additional memory or resources. The only assumption of this scheme is that there is no segment of the previous connection which has been destructed is restored. Indeed, the previous segment can only be received after the victim's server crashes and then recovers all these TCP connections within maximum segment life time [22], which is often impractical. In this case, all other flows are considered as false positive.

In the algorithm, the victim's server generates three random numbers r1, r2, and r3 automatically to manipulate the number of packets sent and received in a way that protects the victim's server from DoS attacks in general and optimistic acknowledgment attacks in particular [23], [24], [25]. These three numbers are selected one by one, at random from n integer, and the range of each number is chosen to be in the range between 1 and 100 inclusively.

Indeed, the function rand ( ) is used to initialize the random number generator which applies a time function to generate these random numbers. The main idea behind this algorithm is randomizing the number of bytes specified in the TCP maximum segment size (MSS). On receiving an optimistic acknowledgment, the victim's server will check the acknowledgment number to see which packets are actually acknowledged. Once the MSS message is sent, the attacker won't know which packets have been sent and how much less data the server would send. The attacker will think that each segment has a fixed number of bytes, and hence creating optimistic acknowledgments which will be easily detected on the server side.

For example, if the MSS is a multiple of 512 bytes, and the default MSS advertised is 2048 bytes, the MSS is set to a different size based on the following formula : MSS (r1) = MSS – (r2+r3), where r1 represents the MSS sequence number, r2 represents the number of bytes reduced, and r3 is a counter that is increased by one each time a new segment is sent. From the receiver's perspective, the attacker replies acknowledgments prior to actually receiving the MSS segments it acknowledges. For example, if the victim's server sends data [512 : 1024], the attacker may acknowledge 1024 and immediately acknowledge 1565 without actually receiving data [1025:1565] aiming to induce DoS attack by increasing the CWND arbitrary rate. However, the detection is achieved by differentiating malicious streams of optimistic TCP acknowledgments from normal TCP acknowledgments as explained in the formula above. Therefore, and because none of the received acknowledgments have the same sequence number of the subsequent segments, the server will immediately terminate the connection and close the associated TCP socket. In this manner, the proposed algorithm could detect optimistic acknowledgment attacks efficiently and resist DoS attacks. In terms of computation time, the algorithm could also detect DDoS attacks more efficiently as the detection rate of the proposed algorithm is relatively high among distributed networks. From the receiver side, the attacker needs to guess three numbers correctly in order to escape the detection. Obviously, the probability of guessing four numbers with each number randomly chosen from 1 to 100 is "1 in a million" chance before the attacker can succeed in comprising the system.

V. Simulation Results

In this paper, the proposed algorithm for the NS2 simulator (26) is implemented and its performance against induced Dos attacks is evaluated and other various results are presented. The proposed defense mechanism is compared with the induced attack scenario in order to correlate them together and compare the performance with and without TCP DoS attacks. The performance is evaluated according to the following metrics: first, the aggregated throughput which is measured in terms of kbit/sec is delivered. Second, the loss of packets caused by the induced DOS attack is measured as a number of packets lost divided by the total packets sent during the process. In the first experiment, the bandwidth is varied from 20% to 80% and the simulation time is 5 sec in all cases,

The simulation result shows that the proposed protection mechanism achieves optimal throughput utilization when compared with unprotected TCP data transfer rate. More specifically, if the most effective bandwidth at which TCP is able to communicate over a protected scenario is 789.66 kb/sec, the analysis will show that it was 159.97 Kb/sec for non-protected scenario where the attack is conducted and still in progress. Figure 3 compares the results between a genuine TCP throughput and the throughput of the protected approach. Obviously, the proposed protection mechanism offers nearly the same TCP throughput values.

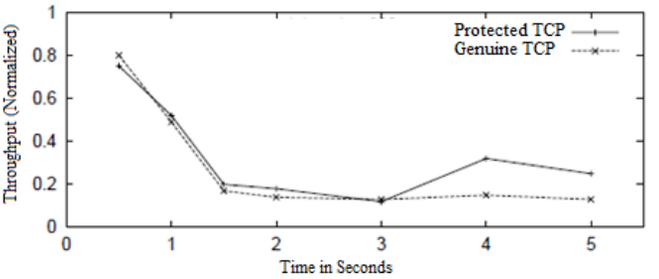


Figure 3. Comparison of results

VI. Conclusion and Future Work

In this paper, several security weaknesses in TCP/IP protocol suite were discussed such as DDoS attacks & induced TCP attacks. Other attack scenarios that exploit vulnerabilities in TCP environments were analyzed. Moreover, "long lived" flows caused by optimistic acknowledgment and induced attacks resulting in sudden increase in the TCP traffic load were estimated. In this paper, a lightweight detection scheme that can effectively identify induced TCP DDoS attack and mitigate it is proposed. This scheme does not need any new implementation of TCP suite. Also, the overall computational complexity of the scheme was affordable and bounded only by the packet transmission time. Therefore, this scheme does not need additional memory or CPU utilization and the performance analysis showed that the proposed protection scheme offers nearly the same throughput performance as genuine TCP/IP network.

The future work will study the application of the proposed scheme to a real TCP network and the evaluation of its performance against different types of DDoS attacks.

VII. References

[1] XP. Luo and RKC. Chang, "On a new class of pulsing denial-of-service attacks and the defense," Proceedings of the Network and Distributed System Security Symp. (NDSS 2005), San Diego, CA, 2005.

[2] A. Herzberg and H. Shulman, "Stealth MITM DoS attacks on secure channels," Computer Research Repository (CoRR), arXiv: 0910.3511, October 2009

[3] J. Lemon, "Resisting SYN Flood DoS Attacks with a SYN Cache," Proceeding of the 10th ACM SIGOPS European Workshop, pp. 89-97, 2002.

[4] W. Eddy, "TCP SYN Flooding Attacks and Common Mitigations," RFC 4987, August 2007.

[5] Z. Qian and Z. Mao, "Off-Path TCP Sequence Number Inference Attack," in IEEE Symposium on Security and Privacy, 2012, pp. 347-361.

[6] W. Eddy, "Defenses Against TCP SYN Flooding Attacks," Cisco Internet Protocol Juournal, vol. 9, no. 4, december 2006.

[7] V.Jacobson, "Pathchar: A tool to infer characterstics of internet paths," Network Research Group, Berkeley, CA, May 1997.

[8] V. Paxson and M. Allman, J. Chu, and M. Sargent, "Computing TCP's Retransmission Timer," RFC 6298 (Proposed Standard), June 2011.

[9] V. Anil Kumar, G. Patra, and R. Thangavelu, "On Remote Exploitation of TCP Sender for low-rate flooding denial-of-service attack," IEEE Communications Letters, vol. 13, no. 1, January 2009, pp. 46-48.

[10] N. Hubballi, S. Biswas, S. Nandi, "Towards Reducing False Alarms in Network Intrusion Detection Systeems with Data Summarization Techniques," International Journal on Security and Communicaiton Networks, vol.6, no.3, 2013, pp. 275-285.

[11] A. Kuzmanovic and E. W. Knightly, "Low-rate TCP-targeted denial of service attacks: the Shrew vs. the Mice and Elephants," ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, vol.1, 2003, pp. 75-86.

[12] A. Kuzmanovic and E. W. Knightly, "Low-rate tcp-targeted denial of service attacks and counter strategies," IEEE/ACM Trans. Netw., vol. 14, no. 4, pp. 683-696, 2006.

[13] A. Shevtekar, K. Anantharam, and N. Ansari, "Low Rate TCP Denial-of-Service Attack Detection at Edge Routers," IEEE Communications Letters, vol. 9, no. 4, pp. 363-365, april 2005.

[14] H. Sun, J. C. S. Lui, and D. K. Y. Yau, "Defending Against Low-rate TCP Attacks: Dynamic Detection and Protection," in Proceedings of IEEE Conference on Network Protocols (ICNP 2004), pp. 196-205, 2004.

[15] J. Liu and M. Crovella, "Using Loss Pairs to Discover Network Properties," in Proceedings ACM Internet Measurement Workshop, 2001, pp. 127-138.

[16] C. L. Schuba and E. H. Spaord, "A Reference Model for Firerewall Technology," in Proceedings of the 13th Annual Computer Security Applications Conferene (ACSAC), 1997, p. 133.

[17] R. Sekar, M. Bendre, D. Dhurjati, and P. Bollineni, "A Fast Automaton-Based Method for Detecting Anomalous Program Behaviors," in IEEE Symposium on Security and Privacy, Oakland, California, 2001, pp. 144-155.

[18] Y. Kwok, R. Tripathi, Y. Chen, and K. Hwang, "Hawk: Halting Anomalies with Weighted Choking to Rescue Well-Behaved TCP Sessions From Shrew DDOS Attacks," in Proceedings of the 3rd International Conference on Networking and Mobile Computing (ICCNMC 2005), Springer – Verlag, New York, 2005, pp. 423-432.

[19] C. Chang, S. Lee, B. Lin, and J.Wang, "The Taming of The Shrew: Mitigating Lowrate TCP-Targeted attack," IEEE Transactions on Network and Service Management (TNSM), vol. 7, no. 1, pp. 1-13, 2010.

[20] S. Savage, N. Cardwell, D.Wetherall, and T. Anderson, "TCP Congestion Control with a Misbehaving Receiver," ACM Computer Communicaitns Review, vol. 29, October 1999, pp. 71-78.

[21] R. Sherwood, B. Bhattacharjee, and R. Braud, "Misbehaving TCP Receivers can Cause Internet Wide Congestion Collapse," in ACM Conference on Computer and Communications Security, 2005, pp. 383-392.

[22] A. Ramaiah, R. Stewart, and M. Dalal, "Improving TCP's Robustness to Bline in-Window Attacks," RFC 5961 (Proposed Standard), August 2010.

[23] C. G. Cassandras and S. Lafortune, "Introduction to Discrete Event Systems, Springer, 2nd edition, 2008.

[24] S. Whitaker, M. Zulkernine, and K. Rudie, "Towards Incorporating Discrete-Event Systems in Secure Software Development," The 3rd International Conference on Availability, Reliability and Security. IEEE, 2008, pp. 1188-1195.

[25] K. Cheng and A. Krishnakumar, "Automatic functional test generation using the extended finite state machine model," in Proceedings of 30th Design Automation Conference, Dallas, Texas, USA, ACM Press, 1993, pp. 86-91.

[26] NS2, The Network Simulator. http://www.nsnam.isi.edu/nsnam/ [Retrieved on January, 2014].

Best Test Cases Selection Approach

Aysh Alhroob
Department of Software Engineering, Isra University
Amman- Jordan
Aysh@iu.edu.jo

Abstract— **This paper proposes an approach for selecting best testing scenarios using Genetic Algorithm. Test cases generation approach using UML sequence diagrams, class diagrams and Object Constraint Language (OCL) as software specifications sources. There are three main concepts: Edges Relation Table (ERT), test scenarios generation and test cases generation used in this work. The ERT is used to detect edges in sequence diagrams, identifies their relationships based on the information available in sequence diagrams and OCL information. ERT is also used to generate the Testing Scenarios Graph (TSG). The test scenarios generation technique concerns the generation of scenarios from the testable model of the sequence diagram. Path coverage technique is proposed to solve the problem of test scenario generation that controls explosion of paths which arise due to loops and concurrencies. Furthermore, GA used to generates test cases that covers most of message paths and most of combined fragments (loop, par, alt, opt and break), in addition to some structural specifications.**

Keywords— Test Cases; Software Specifications; UML Diagram; Genetic Algorithm.

I. INTRODUCTION

There are many forms of testing a software system. One of them is structural specifications testing [1]. On the other hand, testing of behavioural specifications of a software system is very important to evaluate the system interaction and for testing scenarios. A critical component of testing is the construction of test cases. Test cases may be used to detect any faults and problems that exist in the software design even before the program is implemented. A test case contains test input values, expected output and the constraints of preconditions and post conditions for the input values. Generation and selection of the effective test cases from UML models is one of the most challenging tasks [3].

Sequence diagram is one of the main UML design diagrams. It is used to represent the behavioural details and specifications of a software system. However, sequence diagram alone does not express liveness/progress properties or can differentiate between necessary and possible behaviour [2]. To address these limitations, UML class diagram and OCL are also used in this work to model the software and generate test cases more accurately. This paper aims to generate test cases that are able to cover the interaction faults and scenario faults. Figure 1 represents the proposed approach in this work and introduces three main components: Edges Relation Table (ERT), Test Scenarios Generation and Test Cases generation. The scenarios generation technique concerns the generation of scenarios from the testable model of the sequence diagram. Coverage of all paths is a major problem in generating test scenarios as explosion of paths which arise due to loops and concurrencies is to be properly dealt with. Each combined fragment could be one scenario or combined testing scenarios (interaction). Information Table (InfT) is proposed to enrich the testing scenarios with necessary information being extracted from class diagram [1].

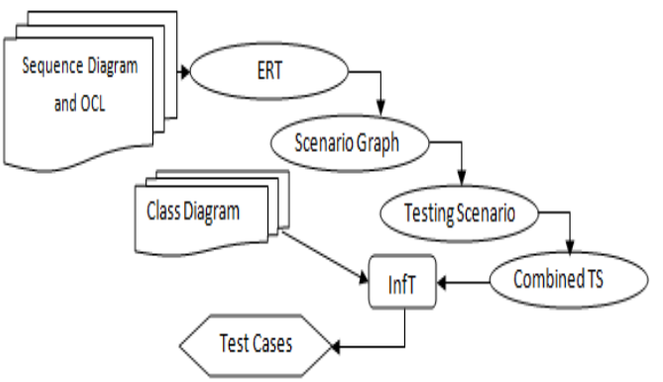


Fig. 1. Test cases generation model

This paper is organized as follows: A related work is described in Section 2. Decomposing Sequence Diagram into Edges is described in Section 3. Edge relationship table is discussed in Section 4. Section 5 describes Test cases generation approach.. Finally, Section 6 draws the conclusions and proposes future work.

II. Previous Work

The previous works provide automatic and manual methodologies to extract test cases from one primary UML diagram. These methods also use some other UML diagrams to obtain additional information. These approaches do not de ne a precise model for the generation of test paths which lead to generation of test cases. Many fault types may exist in different combinations of paths / statements. Detection of such faults in all combinations is a hard problem as the number of statements and faults increases with the size of a software system. Systematic Test cases generation technique is needed to overcome this combinatorial problem [14].

The research in [15] considered reusing of common activities in model based testing and model based development. The authors presented a case study (microwave oven) in which the executable and translatable UML system models were used for automatically generating test models in the QTronicModeling Language using horizontal model transformations.

Genetic Algorithms are used in [16] to generate and optimize test cases from UML State Chart diagram. Crossover method of GA is applied to generate the test sequence and the Mutation Analysis is used for evaluating the efficiency of the test sequence. The proposed algorithm was tested using a case study of driverless train. The authors concluded that GA can be used for test case generation; however, it can only be used effectively if there are a large number of test cases and large number of possible test sequences. Research in [17] also used GA for improved test case generation. The authors presented an approach that combine information from UML state chart diagram used as finite state machines with GA. Test cases were generated by transforming EFSMs into extended control flow graph. GA is applied to generate feasible test sequence and test data. The work is primarily motivated by three factors. First, state-based testing is widely used in protocol software and embedded system software. Second, generation of test sequences using brute force method is very expensive. Third, generation of test sequences using random test cases may not always lead to feasible test paths, these needs to satisfy the guard condition.
The authors noted that superior results were reported by GA based technique.

The authors in [18] consider the problem of automatically generating test cases for dynamic specification mining that observes program execution to infer models of normal behaviour. If a sufficient number of tests are not available, the resulting specification may be too incomplete to be useful, therefor necessitating the requirement for systematic test case generating. The novelty of their approach lies in combining systematic test case generation and typestate mining. TAUTOKO, a typestate miner is used to generate test cases that cover previously unobserved behaviour, systematically extending the execution space, and enriching the specification.

Researchers in [19] considered user-driven test case generation aimed at efficiently testing multimedia applications concerning Digital TV (DTV) receivers and Set-Top Boxes (STB). The proposed approached was experimentally tested for correlation between the detected DTV functional failures using the proposed user-driven test-case generation method and subjectively perceived failures. Improvement in testing efficiency and reduction compared to referent approaches.

The research work in [20] tackled the problem of detecting and improving the test cases after mutation a problem where artificial bugs referred to as mutants are injected into software and the resulting test cases are executed on these fault injected version. They proposed an approach, µTEST that automatically generates unit tests for object-oriented classes based on mutation analysis.

S. Ali et. Al. published a review of the application and empirical investigation of search-based test generation. Readers are referred to [21] for a detailed overview of studies concerning search-based software testing and their empirical evaluation.

III. Decomposing Sequence Diagram into Edges (Nodes)

The simplicity of sequence diagrams makes it easier for the customers to express the requirements and understand them. Unfortunately, shortage or lack of semantic content in sequence diagrams makes them ambiguous and therefore difficult to interpret [7]. Examination of the informal documentation almost resolves the ambiguity but, in some cases, these ambiguities may go undetected leading to costly software errors. To overcome this problem, two solutions could be used. Firstly, the user may provide messages with complete semantic information. Secondly, if sufficient semantic information is not available, one may interpret sequence diagrams based on some heuristics. In the proposed approach in this paper, a compromise has been done, whereby messages in a sequence diagram may be annotated with a pre/post-condition style specification expressed in OCL as shown in Figure 3. Note that this is only a small additional burden on the user since the amount of information required by our methodology in this paper is actually very small.
The specifications should include the declaration of global state variables where a state variable represents some important aspect of the system, e.g., whether or not the user has inserted valid coin into the coffee machine mentioned in Figure 2. The pre/post-conditions should then include references to these variables.

Now, there are two kinds of constraints on a sequence diagram: constraints on the state vector given by an OCL specification, and the constraints on the ordering of the messages given in the sequence diagram [4]. Edges structure generation is based on the edges vector as shown in Notation 1; edges vector is

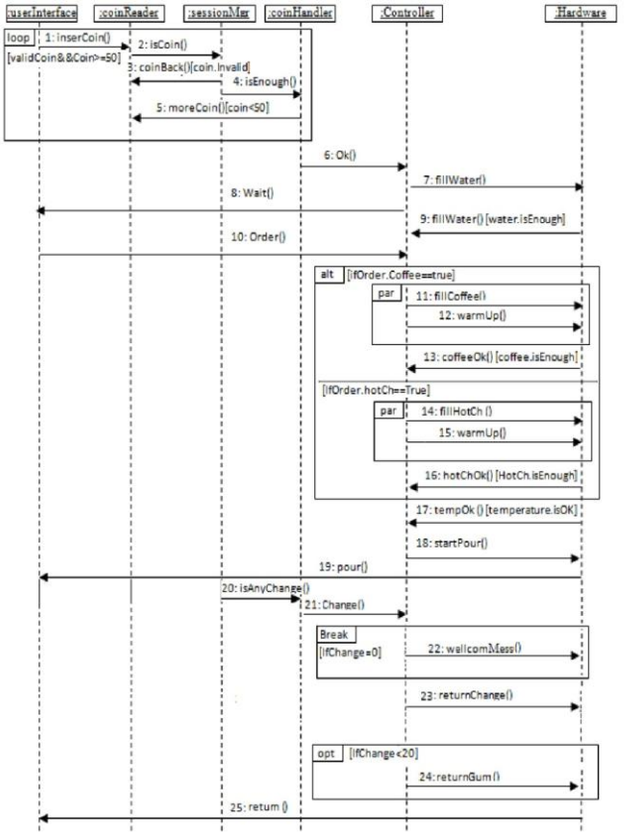


Fig. 2 Sequence Diagram for a coffee machine

responsible for merging a proper object with its message. Thus, the sequence diagram edges will be presented as follows:

$$s_0 \rightarrow m_1 \rightarrow s_0^{'}, s_1 \rightarrow m_2 \rightarrow \cdots \rightarrow m_{r-1} \rightarrow s_{r-1}^{'}, s_r \rightarrow m_r \rightarrow s_r \qquad (1)$$

Where, the m_i is a message between objects and $s_i, \acute{s}_i$, are the edges state vector immediately before and after message m_i is executed. The source and destination object of message mi are denoted by m_i^{source} and m_i^{dest}, respectively. S_i denotes either s_i or $\acute{s}_i$. $S_i^{'}$ is the *jth* element of the vector S_i. Let v_j denote the name of the variable associated with position j in the edge vector.

The initial edges vectors are generated directly from the message specifications: if m_i has precondition $v_j = y$ then let $S_i[j] := y$, and if m_i has a post condition $v_j = y$, let $S_i[j] := y$. Identifying edges is not enough to establish a coherent testing scenario. These edges are connected by many types of relationships in a sequence diagram. These relations must be identified to build an integrated testing scenario graph.

```
coinIn, coinback, coinmore, fillwater, wait: Boolean
coin: coinamonut, water: Sequence
    Insert coin(c: Coin)
      pre: coinIn = false
      post: coinIn = true and coin = c
    coinBack()
      pre: coinIn = true and coin =false
      post: coinIn = false
    moreCoin()
      pre: coinIn = true and coin = true
      post: coinIn = true
    fillWater(w:Water)
      pre: coinIn = true and coin = true coinAmount=true
      post: wait = true
```

Fig. 3 OCL style specifications.

IV. EDGES RELATIONSHIPS TABLE (ERT)

Alhroob, Dahal and Alamgir [1] proposed an automatic approach to generate hierarchy table to set the relationships among classes in a systematic way. In this paper, this hierarchy table is enhanced to the Edges Relationships Table (ERT) to set the relationships between the edges automatically.

A. Testing Scenarios Graph (TSG)

In this section, the TSG is defined and then we present the proposed methodology to generate TSG from a sequence diagram. TSG is defined as follows (Notation 2):

$$TSG = \left[N_{TSG}, \sum_{TSG} q0_{TSG}, F_{TSG}, PAR_{TSG}, SEL_{TSG}, BR_{TSG} \right] \qquad (2)$$

Where, N_{TSG} is the set of all nodes of testing scenarios; each node basically represents an event.
$\sum_{TSG}$ is the set of edges representing transitions from one state to another. $q0_{TSG}$ is the initial node representing a state from which an operation begins. F_{TSG} is the set of final nodes representing states where an operation terminates.
PAR_{TSG} is the set of parallel edges and it represents states where one operation red other operations at the same time.
SEL_{TSG} is the set of either alt or opt edges.
BR_{TSG} is the set of edges that cause a sequence operation termination.

The proposed methodology in this paper identities the set of all testing scenarios where $scn_i = scn_1, scn_2, scn_3 \ldots \ldots scn_m$ and the set of all nodes are also identified. Initially TSGcontains only the StartState; then each node of all $scn_i \in$ scn should be followed by its corresponding NextState, and the duplicates, if any, are removed. The StartState for different scenarios as illustrated in [1] is StateA and five FinalStatesare B, C, D, E and F. An operation starts with an InitialState and undergoes a number of intermediate states due to the occurrence of various edges. Initial edge(s) can be easily detected from ERT. The algorithm of the generating TSG from a sequence diagram uses the ERT as input to get the

sequence diagram testing scenario graph as the output. This process goes through the following steps:
Find out all edges.

1. Detect the initial edge after *InitialState* and check if it was visited before.
2. Detect the pre/post conditions of initial edge.
3. If it was visited before but it affects more than one edge, it is revisited again.
4. If it was visited before and does not affect more than one edge, go to next edge.
5. If the edge leads to *FinalState*, consider the path from *InitialState* to *FinalState* as one scenario.
6. If two or more edges are related by parallel relationship, connect all parallel edges by one circle notation at the beginning of process and in the end as well.

If two edges are related by option relationship, connect option edges by one decision notation in the beginning of process.

Now, the TSG represents all of the sequence diagram combined fragments and sequence messages. The initial edge is detected by identifying its preconditions and post conditions. If the precondition of an edge is not emerging from post condition of another edge, it can be said that the former is an initial edge. E1 in ERT is initial edge because it affects the other edges but it is not affected by any other edges except the loop edge. In contrast, the final edge is affected by others but it cannot affect others except the loop edges.

B. Testing Scenarios

Testing scenarios generation is the main step towards the generation of correct test cases. The proposed technique enumerates all possible paths from the start edge to the final edge of the TSG to generate test scenarios which cover all edges. Each path then is visited to determine which Combined Fragment represents it. Figure 2 represents 25 messages $m_j (j = 1,2, \dots 25)$between two objects with guard conditions from which 25 corresponding edges $E_i (i = 1,2, \dots 25)$ are obtained. The proposed methodology in this paper generates 19 testing scenarios automatically to cover all paths and operations as shown in Table 1. The variety of testing scenarios is a main challenge in this phase. Five main novel features (loop, par, alt, opt and break) are available in sequence diagram as shown in Figure 2. This causes a large number of Combined Fragments. The number of Combined Fragments directly affects the number of scenarios. Furthermore, the increase in the number of guard conditions (constraints) causes an increase in the number of testing scenarios. Each testing scenario starts with *StartState* and ends with *FinalState*. Each *FinalState* may be the end of one or more scenario. For example, *StateB* (forscn_1), StateC (forscn_2), StateD (forscn_3), StateE (for scn_4andscn_5) and StateF (for scn_6, , and$scn_7, scn_8,\ scn_9$) are FinalStates.

V. TEST CASES GENERATION APPROACH

Test cases are used to detect faults in a software system. A software system is a combination of behavioural and structural information. Sequence diagram provides behavioural information. OCL is used to determine the pre/post conditions of the edges but even then, some other additional information is missing in a sequence diagram especially in the description of method constraint and type information. This additional information may be derived from the class diagram and then can be appended to each message. Therefore, the structure of each variable *v* in the method (e.g., name, type, value, constraint) is obtained. Here, type and constraint refer to the data type and the attribute constraints associated to a variable *v*:name. The type information is used to map each variable *v*:name to a range of values *(min, max)*. This makes sure that each variable on a path from the initial edge to the final edge is mapped to a range of values. Furthermore, the constraint information is used to appropriately set the boundary values min and max. For example, the type information such as int to a variable *X* sets$(min_x, max_x) = (min_{Int}, max_{Int})$.
Where min_{Int} and max_{Int} are the boundary values. Let us consider$x \geq 5$ as the constraint associated to a variable *x*. In that case, (min_x, max_x) will be set to $5, max_{Int})$ which will be recognized as the initial domain. Figure 1 represents the class diagram determination to enrich the *TSG* by additional information.

Now, each testing scenario must be represented by one test case. For example, $scn_1$is presented by Test Case 1 in Figure 8. In this case, the variable coin has to be checked if an input is a coin or not. Actually, in this case, no variable type can be checked to ensure that. In other case studies, such as student registration system, the type of variable plays an important role to test the model. For example, for a message *enterStudentAge* (age: integer), if the user used decimal number to represent age, the test case must detect that error. If the testing scenarios in Table 1 are mapped directly to test cases, then we get 19 test cases to cover all model specifications but these many cases will not be sufficient to cover all sequence diagram features or all Combined Fragments behaviour. Recall that a test scenario is a sequence of operations starting with first edge and ending into final one, whereas a test case must examine the Combined Fragment and test whether it works well or not. The following sub-section presents the combination of testing scenarios to generate legitimate test cases.

A. Combining Testing Scenarios

The combination process is based on the similarity of functionality of scenarios. For example, the test scenarios 4, 5, 6 and 7 are similar in behavioural point of view but after E10, the edges 11, 12, 13 and 14 are fired in parallel. The work in [1] shows that each pair of edges (11, 12 AND 14, 15) are fired separately and are connected with selection node. To avoid confusion in how theses edges are related, SP node and EP

node are used in this work. Subsequently, the combination of test scenarios is specified for parallel operations. The proposed algorithm, in this paper, identifies the parallel edges in each scenario, and then combines all of the scenarios that are same but different in parallel edges. Thus, the scenarios 4

, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19 can be combined as shown in Table 2. Therefore, the numbers of test cases are reduced to 11 test cases instead of 19. This reduction is important and necessary to produce legitimate test cases because parallel operations can be covered only after combining the relevant testing scenarios. In contrast, the selection operations (alt, opt and break) and the loop testing scenarios need not be combined because these scenarios depend on the selection condition and, thus, these scenarios are independent.

B. Legitimate Test Cases

After combining testing scenarios, the focus is on generating legitimate test cases. The last step of test case generation is to convert the testing scenarios (after combinations) to actual test cases. The test scenarios which are already populated with the necessary information from two important design artefacts namely, sequence diagram and OCL. Third design artefact the class diagram has provided the information about types and constrains of parameters in a message. In sequence diagram, the method m represents the event from a sender class C1 to a receiver class C2. The method *m* in class C2 has signature which is defined in the class diagram. The method signature represents the name of the method, types of the parameter and the return type while the class attributes represent information about the instance variables such as their names and types. In addition, there may also be OCL constraints which are used to express invariants on the class attributes [8]. These invariants identify attribute constraints that are true for all instances of the class. Information Table (InfT) represents the relationships between the methods in sequence diagram and its signature in class diagram. This technique completes the full image of test cases.

For each edge *Ej* of scenario scn_i the edge method *m* is identified. For the purpose of discussion, two cases are used in this paper: a method having a guard condition (see Notation 3) and another method without guard condition. If the method *m* has no guard condition, the method is represented in test case as follows:

$$TC = [preC, I(a_1, a_2 \ldots\ldots a_n), O(d_1, d_2 \ldots\ldots d_m) posC] \quad (3)$$

where, *TC*is a Test Case.
preC = precondition of the method m.
$I(a_1, a_2 \ldots\ldots a_n)$= set of input values for the method *n*.
$O(d_1, d_2 \ldots\ldots d_m)$= set of resultant values in the object when the method is executed.
posC= the post condition of the method *m*.

In contrast, when the method has a guard condition the test case uses this condition to show the state of processing. For example, the loop Combined Fragment in Figure 2 is affected by a guard condition that forces the user to enter valid coins and the *coinAmount* is set to 50. This operation repeats until the suitable condition is reached that is acceptable according to the given guard condition and then it goes to the next step.

TABLE I. COFFEE MACHINE TESTING SCENARIOS

Scn ID	Testing Scenarios
Scn_1	$E_1 \rightarrow E_2 \rightarrow E_3$
Scn_2	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_5$
Scn_3	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8$
Scn_4	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{11} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}$
Scn_5	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{12} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}$
Scn_6	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{14} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}$
Scn_7	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{15} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}$
Scn_8	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{11} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{22}$
Scn_9	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{12} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{22}$
Scn_{10}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{14} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{22}$
Scn_{11}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{15} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{22}$
Scn_{12}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{11} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{23} \rightarrow E_{25}$
Scn_{13}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{11} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{24} \rightarrow E_{25}$
Scn_{14}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{11} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{23} \rightarrow E_{25}$
Scn_{15}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{12} \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{24} \rightarrow E_{25}$
Scn_{16}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{14} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{23} \rightarrow E_{25}$
Scn_{17}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{14} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{24} \rightarrow E_{25}$
Scn_{18}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{15} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{23} \rightarrow E_{25}$
Scn_{19}	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow E_{15} \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19} \rightarrow E_{20} \rightarrow E_{21} \rightarrow E_{24} \rightarrow E_{25}$

coinAmount<50 will be appearing in test cases until the event satisfied the guard condition. The test case with guard condition semantic in Notation 4 is similar in Notation 3, but the guard conditions *(g(v))* appeared in test case.

$$TC = [preC, I(a_1, a_2, \ldots, a_n), O(d_1, d_2, \ldots, d_m), g(v), posC]$$

Figure 4 shows the first 4 test case generated automatically to test the coffee machine model. We need to analyse whether all these test cases are legitimate.

Legitimate set of test cases refers to the ability of test case to achieve the best testing in minimum time and cost. Removing redundant test cases is one of the major challenges to obtain legitimate set of test cases. Redundant test case is one, which if removed, will not affect effectiveness of fault detection of the suite of remaining test cases [9]. In this paper, redundant test cases cannot be present because redundancy is detected automatically in the stage of TSG construction before generating testing scenarios. On other hand, if test cases are generated from only edges and nodes from sequence diagrams, the resultant suite may contain redundancy. For example, TC (4,5) and TC(6,7) contain similar edges except $[E_{11}||E_{12}]$ and $[E_{14}||E_{15}]$. So, the difference between two Combined Testing Scenarios (CTS) occurs in the branches that are affected by *alt*. Furthermore, before and after the branches, the similarity between the CTSs can be seen. Unfortunately, test cases based on the scenarios are not able to eliminate redundant edges or nodes because each scenario must be tested separately. To minimise this problem, we propose an approach to select best testing scenario in the next sub-section.

C. Best Testing Scenario Selection Approach Using Genetic Algorithm

Test cases are derived from testing scenarios. An important question is what is meant by the best testing scenario in the sequence diagram? Coverage criteria play a major in selection of the best scenario. In this section, the coverage criteria that are used in this paper are listed. These criteria are used to evaluate the test cases that have been generated previously (see Figure 4). A total of 11 test cases satisfied our criteria. We propose the following coverage criterion which is expressed by the following condition.

Loop adequacy criterion: Test cases must contain at least one scenario test case in which control reaches the loop and then check the non-executable loop (zero iteration). Another test case for testing the body of the loop that is executed at least once before control leaves the loop (more than zero iteration) [10].

Concurrent (parallel) coverage criterion: For each parallel node in TSG, test cases must include one test case corresponding to every valid interleaving of message sequences.

All messages coverage criterion: For all messages in sequence diagram, test cases must execute all message sequence paths of the sequence diagram [11].

Branch coverage: Each decision outcome within the diagram must be covered by at least one start-to-end path [26].

Selection coverage criterion: For each selection fragment (alt, opt and break), test cases must include one test case corresponding to each evaluation of the constraint.

The criteria above are covered by test cases that are generated in this paper, but what is the best test case, which covers all criteria?. When we say "best test case" we mean the one which has the highest percentage of coverage. To select the best test case, the criteria mentioned above are categorised into two evaluation steps: first one includes all of the above coverage criteria except the "All message coverage criterion" and we call it Combine Fragments Coverage Criterion (CFCC), and the second one is specified for "All messages coverage criterion". Figure 5 shows the uses of these two criteria to select the best CTS.

TABLE II. COMBINED TESTING SCENARIOS

CTSj	Testing Scenario
CTS_1	$E_1 \rightarrow E_2 \rightarrow E_3$
CTS_2	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_5$
CTS_3	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8$
CTS_4 (4,5)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{11}\|\|E_{12}] \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}$
CTS_5 (6,7)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{14}\|\|E_{15}] \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}$
CTS_6 (7,8)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{11}\|\|E_{12}] \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}\ E_{20} \rightarrow E_{21} \rightarrow E_{22}$
CTS_7(10,11)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{14}\|\|E_{15}] \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}\ E_{20} \rightarrow E_{21} \rightarrow E_{22}$
CTS_8(12,14)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{11}\|\|E_{12}] \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}\ E_{20} \rightarrow E_{21} \rightarrow E_{23} \rightarrow E_{25}$
CTS_9 (13,15)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{11}\|\|E_{12}] \rightarrow E_{13} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}\ E_{20} \rightarrow E_{21} \rightarrow E_{24} \rightarrow E_{25}$
CTS10 (16,18)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{14}\|\|E_{15}] \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}\ E_{20} \rightarrow E_{21} \rightarrow E_{23} \rightarrow E_{25}$
CTS_{11}(17,19)	$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_6 \rightarrow E_7 \rightarrow E_8 \rightarrow E_9 \rightarrow E_{10} \rightarrow [E_{14}\|\|E_{15}] \rightarrow E_{16} \rightarrow E_{17} \rightarrow E_{18} \rightarrow E_{19}\ E_{20} \rightarrow E_{21} \rightarrow E_{24} \rightarrow E_{25}$

To achieve this target, the proposed approach using GA aims to generate testing scenarios to cover maximum Edges using genetic algorithms technique. The "All message coverage criterion" depends on the weights of the edges. The weight of the edges is calculated by the following Equation:

$$E_i w = \frac{E_p}{N} \qquad (1)$$

$E_i w$ is the weight of each edge.
E_p is the edge presence probability in whole testing scenarios.
N is the total weights of all edges in all testing scenarios.

For example, the repeated use of the E10 is 8 times over all combined testing scenarios in Table 2. The weight of E10 is calculated by Equation 1 as follows,

$$E_{10} w = \frac{8}{147} = 0.54421769 .$$

GA is iterative procedures which work with chromosomes. The chromosomes are a population of candidate solutions that are maintained by the GAs throughout the solution process [9]. At first a population of chromosomes is generated

randomly. A selection operator is used to choose two solutions from the current population. The selection process used the measured goodness of the solutions (Fitness Function). The crossover operator swaps sections between these two selected solutions with a defined crossover probability. One of the chromosomes solutions is then chosen for application of the mutation process. The algorithm is ended, when a defined stopping criterion is reached. Since our approach focuses on finding a set of transitions by triggers firing, a chromosome in our approach is a sequence of triggers itself. Testing scenarios will be the first population from all possible cases. The fitness value for each chromosome is calculated from the number of Edges which is covered.

GAs operators used in proposed technique are the two point crossover and random mutation. Based on some experimentation and previous knowledge on GAs application with other problem [20], the parameters of genetic operation are set as follow:

- The crossover probability is 0.5.
- The mutation probability is 0.05.
- The size of population in each generation is 10.

Precondition	Coffee Machine displaying main screen	
Test Case 1	**Input:**	coin: inserCoin (): isCoin (coin) ="False"
	Output:	coinBack (coin)
	Postcondition:	Display main screen
Test Case 2	**Input:**	coin: coinAmount: : inserCoin (): isCoin (coin) ="True": isEnoogh (coinAmount) ="false"
	Output:	moreCoin (coin)
	Postcondition:	Display more coin message: "coinAmount<50"
Test Case 3	**Input:**	coin: coinAmount: water: : inserCoin (): isCoin (coin) ="True" isEnough (coinAmount) ="True": ok (): fillWater (water)
	Output:	Wait ()
	Postcondition:	Display wait message
Test Case 4	**Input:**	coin: coinAmount: water: coffee: HotCh: inserCoin (): isCoin (coin) ="True": isEnough (coinAmount) ="True": ok (): fillWater (water): getWater (water) ="true": Order (): Ordr.coffee () ="True" [fillCoffee (coffee) || warmUp ()]: coffeeOk (coffee) ="True": tempOk () ="True": stratPour (coffee)
	Output:	Pour (coffee)
	Postcondition:	Display wait message

Fig. 4 First 4 test cases for Coffee Machine system.

The All messages coverage criterion uses the weight of the edges to calculate the weight of a CTS. This weight is used to identify the best scenario. The intuition is that such a CTS makes best messages coverage. Equation 2 shows that the weight of a scenario is the sum of weights of all edges in that scenario. The proposed methodology shows the combined testing scenarios CTS8, CTS9, CTS10 and CTS11 have the highest weight which means that these four CTSs are the best from message coverage point of view.

$$CTS_i w = \sum_{i=0}^{n} Ew_{i+1} \qquad (2)$$

Note that we have not adopted the *All messages coverage criterion* as the only factor to select the best test scenario. Second factor (CFCC) is also used to check the Combined Fragments coverage. Based on this factor, the best CTS is that one which is able to cover the highest number of Combined Fragment's edges.

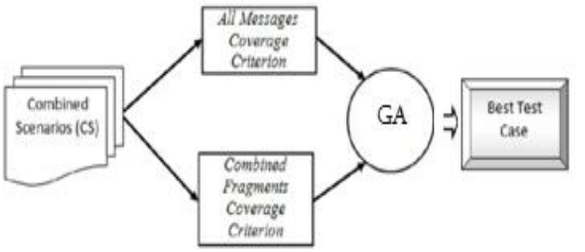


Fig. 5. Best test case selection methodology.

By referring to Figure 2, it can be noted that the *Combined Fragments* (*loop*, *alt*, *par*, *opt* and *break*) contain 13 edges (E_1, E_2, E_3, E_4, E_5, E_{11}, E_{12}, E_{13}, E_{14}, E_{15}, E_{16}, E_{22} and E_{24}). The best CTS is the one that covers highest number of these edges. Thus, the proposed approach in this paper examines the presence of Combined Fragment edges in each testing scenario. CTS_{10} covers 6 Combined Fragments edges (E_1, E_2, E_4, E_{14}, E_{15} and E_{16}) out of 13. The CTSs with the best Combined Fragments coverage are CTS_6, CTS_7, CTS_9 and CTS_{11}.

The best combined testing scenario (s) is that one which achieves highest coverage in both the criteria *CFCC* and *All messages coverage criterion*. CTS_9 or CTS_{11} can be selected as best *CTS* because these two have the highest coverage in both these criteria. Let us say CTS_{11} is the best test scenario and the corresponding test case covers 88% of *All messages coverage criterion* and 55% of *CFCC*. As noted, the *CTS*is unable to capture the coverage of enough edges in the Combined Fragment as just 55% coverage may not be satisfactory.

To improve the coverage of the Combined Fragments edges, we propose to select the second best combined testing scenario as well to support the first one. This technique is similar to that used in [1] to select second testing path to improve unit coverage.

The selection method for the second best *CFCC* scenario depends on the non-similarity of edges contained in the best *CFCC* scenarios. In [1], the non-similarity criterion is used to select the second best testing path, and the same is used in this paper as well. Based on the non-similarity degree between the best *CFCC* scenario and others, the scenarios which are eliminated have the biggest similarity degree. The CTS_6 is the highest non-similar scenario in comparison with CTS_{11} because it has 4 different Combined Fragment's edges (E_{11}, E_{12}, E_{13} and E_{22}) compared to CTS_{11}.

Consequently, both scenarios CTS_{11} and CTS_6 cover 85% of Combined Fragments edges and 97% of *All messages coverage criterion*. Together they are considered as best testing scenarios.

VI. CONCLUSION AND FUTURE WORK

An approach is proposed to generate test cases automatically from UML sequence diagram, class diagram and OCL. The approach generates efficient test cases that meet required coverage criteria. Furthermore, the relationship between edges has been recognised by the edge relationships table. This table is generated automatically to reveal the relationship of the edges as specified in the sequence diagram. The proposed methodology covers *loop*, *parallel*, *alternative*, *option*, *break* and *sequence* relationships and a number of structural specifications as well. This work provides an efficient technique to generate testing scenarios graph that is used to represent the testing scenarios of the events in a sequence diagram. The testing scenario graph represents combined fragments and shows the relationships among scenarios. The proposed methodology decomposes testing scenarios graph to combined test scenarios to achieve minimum number of testing scenarios. Information Table (**I**nf**T**) technique is proposed to provide the sequence diagram with exact method signatures that are obtained from the class diagram. Despite this, presence of the redundant edges in each testing scenario led us to develop a new technique to select best testing scenario. To achieve this, we have used *All messages coverage criterion* and *Combined Fragments coverage criterion* to evaluate the best testing scenario (and the best test case) using GA. The weakness that appeared in selecting best testing scenario is a variety of the coverage percentage that the best testing scenario can cover. This variety comes from the variety of the model designs. Generally, this technique almost covers more than 85% of Combined Fragment's edges and more than 94% of the sequence diagram messages.

The test cases that are generated automatically in this paper meet the coverage criteria. However, UML sequence diagrams do not contain all information related to verification and non-functional parameters of the software. This limitation comes from the semi-formal characteristics of UML diagrams [12], especially during the first cycle of the software production. In order to avoid its semi-formal characteristics, we propose to transform sequence diagram and class diagram to HLPN, a type of Petri Net in future. Petri Net (PN) is a graphical diagram for the formal description of the flow of activities in complex systems [13].

REFERENCES

[1] A. Alhroob, K. Dahal, and H. Alamgir, “Automatic Test Cases Generation from Software Specifications Modules," in Proceedings of the 4th IFIP TC2 Central and East European. Conference on Software Engineering Techniques CEE-SET. Springer, 2009, pp. 130-142.

[2] A. Cavarra and J. Kuster-Filipe, “Combining sequence diagrams and OCL for liveness," Electronic Notes in Theoretical Computer Science, vol. 115, pp. 19-38, 2005.

[3] A. O utt, Z. Jin, and J. Pan, “The dynamic domain reduction procedure for test data generation," Software: Practice and Experience, vol. 29, no. 2, pp. 167{193, 1999.

[4] [4]B. Li, Z. Li, L. Qing, and Y. Chen, “Test Case Automate Generation from UML Sequence Diagram and OCL Expression," in Proceedings of the 2007 International Conference on Computational Intelligence and Security. IEEE Computer Society, 2007, pp. 1048-1052.

[5] G. Booch, J. Rumbaugh, and I. Jacobson, UniedModeling Language User Guide, The (Addison-Wesley Object Technology Series), 2005.

[6] D. Aredo, “A framework for semantics of UML sequence diagrams in PVS," Journal of Universal Computer Science, vol. 8, no. 7, pp. 674{697, 2002.

[7] X. Li, Z. Liu, and H. Jifeng, “A Formal Semantics of UML Sequence Diagram," in Proceedings of the 2004 Australian Software Engineering Conference. IEEE Computer Society, 2004, p. 168.

[8] OMG, “Object constraintlanguage, version 2.2," http://www.omg.org/spec/OCL/2.2.htm [Last accessed in August 10, 2010],2010. [Online]. Available: http://www.omg.org/spec/OCL/2.2.htm

[9] N. Koochakzadeh and V. Garousi, “A Tester-Assisted Methodology for Test Re-dundancy Detection," Journal on Information and Software Technology, vol. 52, no. 5, pp. 625-640, 2010.

[10] J. Edvardsson, “A survey on automatic test data generation," in Proceedings of the 2nd Conference on Computer Science and Engineering, 1999, pp. 21{28.

[11] A. Andrews, R. France, S. Ghosh, and G. Craig, “Test adequacy criteria for UML design models," Software Testing Verication and Reliability, vol. 13, no. 2, pp. 95-127, 2003.

[12] H. Motameni, A. Movaghar, I. Daneshfar, and H. Zadeh, “Mapping to Convert Activity Diagram in Fuzzy UML to Fuzzy Petri Net," World Applied Sciences Journal, vol. 3, no. 3, pp. 514-521, 2008.

[13] Bobbio, “System modelling with Petri nets," in Systems reliability assessment: proceedings of the Ispra course held at the EscuelaTecnica Superior de IngenierosNavales, Madrid, Spain, September 19-23, 1988, in collaboration with Universidad Politecnica de Madrid. Springer, 1990, p. 103.

[14] DebasishKundu, DebasisSamanta, Rajib Mall "Automatic code generation from unified modelling language sequence diagrams", IET Software, Volume 7, Issue 1, February 2013, p. 12 – 28.

[15] Federico Ciccozzi, Antonio Cicchetti, Toni Siljamäki, "Automating Test Cases Generation: From xtUML System Models to QML Test Models", ACM proceeding in MOMPES ’10, September 20, 2010, Antwerp, Belgium

[16] PreetiGulia, R. S. Chillar, "A New Approach to Generate and Optimize Test Cases for UML State Diagram". ACM SIGSOFT Software Engineering Notes archive, Volume 37 Issue 3, May 2012, Pages 1-5

[17] Mahesh Shirole, AmitSuthar, Rajeev Kumar, Generation of Improved Test Cases from UML State Diagram Using Genetic Algorithm, Proceeding of the 4th India Software Engineering Conference, ISEC 2011, Pages 125-134

[18] ValentinDallmeier, Nikolai Knopp, Christoph Mallon, Gordon Fraser, Sebastian Hack, and Andreas Zeller, Member, "Automatically Generating Test Cases for Specification Mining", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 38, NO. 2, MARCH/APRIL 2012

[19] TarkanTekcan, Vladimir Zlokolica, VukotaPekovic, Nikola Teslic and Mustafa Gündüzalp, "User-driven Automatic Test-case Generation for DTV/STB Reliable Functional Verification", IEEE Transactions on Consumer Electronics, Vol. 58, No. 2, May 2012. Pages: 587-595

[20] Gordon Fraser and Andreas Zeller, "Mutation-Driven Generation of Unit Tests and Oracles", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 38, NO. 2, MARCH/APRIL 2012, Pages: 287-292

[21] Shaukat Ali, Lionel C. Briand, HadiHemmati, and Rajwinder K. Panesar-Walawege "A Systematic Review of the Application and Empirical Investigation of Search-Based Test Case Generation", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 36, NO. 6, NOVEMBER/DECEMBER 2010

Some Parallel aspects of the QR Decomposition Method

Halil Snopce
CST-Faculty
SEE-University
Tetovo, R. Macedonia
h.snopce@seeu.edu.mk

Azir Aliu
CST-Faculty
SEE-University
Tetovo, R. Macedonia
azir.aliu@seeu.edu.mk

***Abstract*—In this paper we discuss the parallelization of the QR decomposition of matrices. For this purpose we have analysed the method based on Given's rotation and the method based on the Householder reflection. The mathematical background is followed by the parallelization which basically uses the systolic approach.**

Keywords—QR decomposition, Given's rotation, Householder relection, Parallelization of QR decomposition, Systolic array for the QR decomposition.

I. INTRODUCTION

An important matrix problem that arises in many applications, like signal processing, image processing, solution of differential equations etc., is the problem of solving a set of simultaneous linear equations. There are two basic methods for solving such kind of systems. Direct methods and iterative methods. QR decomposition is one of the best methods for matrix triangularization which uses the so called direct method. This method uses triangularization of the coefficient matrix, followed by the use of back substitution. Most of the QR-decomposition implementations are based on three methods:

1. The Given's rotation method (also known as Jacobi rotations) used by W. Givens and originally invented by Jacobi.
2. The Gram-Schmidt method and
3. The method with the Hausholder transformations.

The Hausholder transformation is one of the most computationally efficient methods to compute the QR-decomposition of a matrix. The error analysis carried out by Wilkinson [1, 2], showed that the Hausholder transformation outperforms the Given's method under finite precision computation. Due to the vector processing nature of the Hausholder transformation, no local connections in the implementation of the array are necessary. Therefore QR decomposition by the method of Hausholder transformation is more difficult. Each rotation in the case of Given's method modifies just two rows of a matrix. This is a reason why during this method the order of rotations can be changed influencing different rows. Taking this into the consideration, the parallel processing of this method is very appropriate.

The basic idea of the QR-decomposition of a matrix is to express a given $m \times n$ matrix A in the form $A = QR$, where Q is an orthonormal $m \times n$ matrix and R is an $n \times n$ upper triangular matrix with nonzero diagonal entries.

A parallel version of Given's rotation was proposed in [3]. In [4] it is proposed an alternative way for parallelization of Given's rotation which is more efficient for larger matrices. In [5] it is given a parallel pipeline version of Given's rotation for thr QR decomposition. In [6] one can find the block version of the QR decomposition, which first transforms the matrix into the Hassenberg form and then applies Given's rotation to it. The design which is based on the method of Householder transformation is given in [9]. In [13] is presented a new algorithm for finding QR decomposition for square and full column matrix. The numerical analysis and experiment is given in [14]. In [15] is demonstrated a parallel algorithm based on the Gram-Schmidt method.

The analysis in this paper uses the givens rotation method [7] and the householder method [16]. In [7] and [8] are proposed two systolic arrays for the QR decomposition with hardware complexity $O(n^2)$ and time complexity O(n) which are based on the method of Given's rotation.

II. THE QR DECOMPOSITION BASED ON GIVEN'S ROTATION

The upper triangular matrix is obtained using sequences of Given's rotations [10] such that the subdiagonal elements of the first column are nullified first, followed by those of the second column and so forth, until an upper triangular matrix is reached. The procedure can be written in the form given below:

$$Q^T A = R$$

$$\text{where } Q^T = Q_{n-1} Q_{n-2} \ldots Q_1 \qquad (1)$$

$$\text{and } Q_p = Q^{p,p} Q^{p+1,p} \ldots Q^{n-1,p}$$

where $Q^{p,q}$ is the Given's rotation operator used to annihilate the matrix element located at row $q+1$ and column p. When we work with 2x2 matrices, an elementary Given's transformation has the form:

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \cdot \begin{bmatrix} 0 \ldots 0 & r_i & r_{i+1} \ldots r_k \\ 0 \ldots 0 & x_i & x_{i+1} \ldots x_k \end{bmatrix} = \begin{bmatrix} 0 \ldots 0 & r_i' & r_{i+1}' \ldots r_k' \\ 0 \ldots 0 & 0 & x_{i+1}' \ldots x_k' \end{bmatrix} \quad (2)$$

where c and s are the cosine and the sine of the annihilation angle, such that:

$$c = \frac{r_i}{\sqrt{r_i^2 + x_i^2}}, \qquad s = \frac{x_i}{\sqrt{r_i^2 + x_i^2}}.$$

It is not difficult to verify that the product of two rotations is also rotation. Let A be an nxn matrix. In order to transform A into an upper triangular matrix R, we can find a product of rotations $Q^T = Q_{n-1}Q_{n-2}...Q_1$ such that $Q^T A = R$. It is not difficult to show that $O(n^2)$ rotations are required. Because the number of operations in every rotation is O(n), the complexity of this algorithm will be $O(n^3)$. In general, the computational complexity of the QR decomposition is given below [15].

1. Householder: $4/3n^3 + O(n^2)$
2. Given's: $8/3n^3 + O(n^2)$
3. Fast Given's: $4/3n^3 + O(n^2)$
4. Gram-Shmidt: $2n^3 + O(n^2)$

III. Parallel Algorithm Based on Given's Rotation

In [11] it is shown that a triangular systolic array can be used to obtain the upper triangular matrix R based on sequences of Given's rotations. This systolic array is shown in Fig. 1.

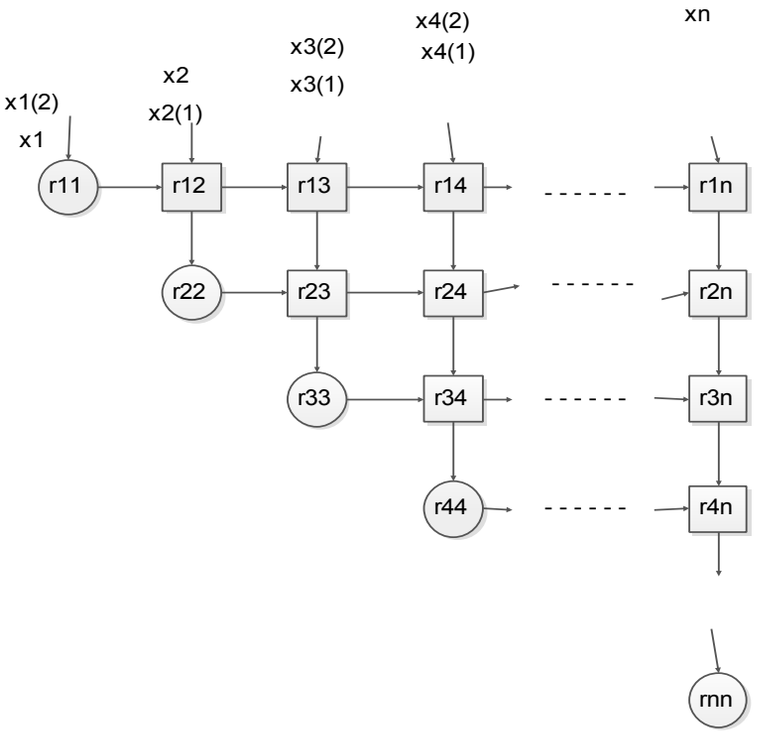


Fig. 1. Triangular systolic array for computing the upper triangular matrix R

As we can see, the array consists of two different shapes of cells. The cells in the shape of a circle (fig. 2), and the cells in quadratic shape (as in fig.3).

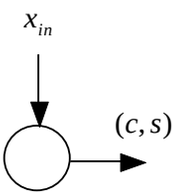


Fig.2. Input and output of the circle cell of the array in fig.1

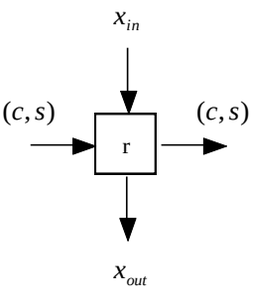


Fig. 3. Input and output of the quadratic cell of the array in fig.1

The cells of fig.2 perform according to algorithm 1:

Algorithm 1:

$$
\begin{aligned}
&If\ x_{in} = 0\ then\\
&\quad c = 1;\ s = 0\\
&otherwise\\
&\quad r' = \sqrt{r^2 + x_{in}^2};\\
&\quad c = r/r';\ s = x_{in}/r'\\
&\quad r = r';\\
&end
\end{aligned}
$$

The calculations of quadratic cells are given by the relations below:

$$
\begin{aligned}
x_{out} &= cx_{in} - sr\\
r &= sx_{in} + cr
\end{aligned}
$$

According to the relations in (1), the Q matrix cannot be obtained by multiplying cumulatively the rotation parameters propagated to the right. Accumulation of the rotation parameters is possible by using an additional rectangular systolic array.

IV. The Computation of $R^{-T}X$

We present a brief derivation of the result presented in [12], about the property that a triangular array can compute $R^{-T}x$ in one phase with the matrix R situated in that array.

Let $r_{ij} = [R]_{ij}$ and $r'_{ij} = [R^{-1}]$, where $r_{ij} = 0$ and $r'_{ij} = 0$ for $i > j$. It can be shown that:

$$
r'_{ij} = \begin{cases} \dfrac{1}{r_{ii}}; & i = j \\ -\displaystyle\sum_{k=i}^{j-1} \frac{r'_{ik} r_{kj}}{r_{jj}}; & i < j \le n \end{cases} \tag{3}
$$

Let

$$
[y_1, \ldots, y_n]^T = R^{-T} X \tag{4}
$$

Then the recursive computation of (4), where R^{-T} is a nxn matrix and X is an nxm matrix is:

$$
y_j = \sum_{i=1}^{j} x_i r'_{ij}, \qquad i = 1, \ldots, n \tag{5}
$$

In particular (because we want to use R and X to compute $R^{-T}X$), y_j can be expressed in terms of r_{ij} and x_i. By substituing the equation (4) into equation (5) we have:

$$
\begin{aligned}
y_j &= \sum_{i=1}^{j} x_i r'_{ij} = y_j = \sum_{i=1}^{j-1} x_i r'_{ij} + x_j r'_{jj}\\
&= \sum_{i=1}^{j-1} x_i r'_{ij} + \frac{x_j}{r_{jj}}
\end{aligned} \tag{6}
$$

If we continue, by transforming the relation (6), we will have:

$$
y_j = \frac{x_j}{r_{jj}} + \sum_{i=1}^{j-1} x_i r'_{ij} = \frac{x_j}{r_{jj}} - \sum_{i=1}^{j-1} x_i \sum_{k=i}^{j-1} \frac{r'_{ik} r_{kj}}{r_{jj}}
$$

And finally we get:

$$
\begin{aligned}
y_j &= \frac{1}{r_{jj}} \cdot \left(x_j - \sum_{i=1}^{j-1} x_i \sum_{k=i}^{j-1} r'_{ik} r_{kj} \right) =\\
&= \frac{1}{r_{jj}} \cdot \left(x_j - \sum_{k=1}^{j-1} \sum_{i=1}^{k} x_i r'_{ik} r_{kj} \right)
\end{aligned} \tag{7}
$$

Using the relation (5), for the final form of y_j, we get

$$
y_j = \frac{1}{r_{jj}} \left(x_j - \sum_{k=1}^{j-1} y_k r_{kj} \right) \tag{8}
$$

Finally, using the relations obtained above (where Y is the nxm matrix, R is nxn upper triangular matrix and X is an nxm matrix), the algorithm for computing $R^{-T}x$ is given:

Algorithm 2

$$
\begin{aligned}
&for\ i = 1\ to\ n\\
&y_1 = \frac{1}{r_{11}} \cdot x_1\\
&for\ j = 2\ to\ n\\
&\quad begin\\
&\quad z_j = x_j\\
&\quad for\ k = 1\ to\ j-1\\
&\qquad z_j = z_j - y_k r_{kj}\\
&\quad y_j = \frac{z_j}{r_{jj}}\\
&end
\end{aligned}
$$

The corresponding systolic array is similar as the array in fig.1. The data movement of input values x and output values y is presented in the figure 4.

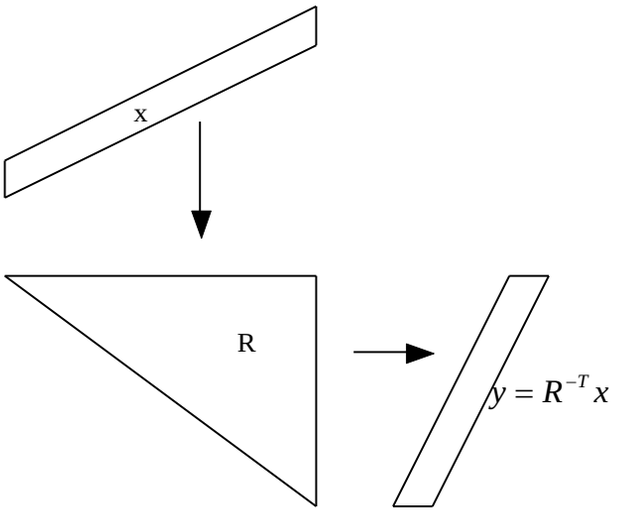


Fig. 4. Data movement of x and y in the computation of $R^{-T}x$

In the case presented above, the elements of the matrix R are stored in the triangular array. The cells of fig.2 (circle cells) perform the division part of the equation (8) (the part $1/r_{jj}$). The second part of the eq. (8) (the part $x_j - \sum_{k=1}^{j-1} y_k r_{kj}$) is performed by the quadratic cells shown in fig.3.

V. Mapping Into the Systolic Array

The number of processors in the fig. 1 can be given by the formula $\frac{p(p+1)}{2}$ for some integer number p. To give the mapping scheme of this array we assume that $m = \left\lfloor \frac{n}{p} \right\rfloor$ and that the cell on the position (i, j) is mapped on to the processor (s, k) in the corresponding network. The mapping is given by the formula:

$$s = \begin{cases} \left\lfloor \frac{i}{m+1} \right\rfloor & if \quad \frac{i}{m+1} < n \bmod p \\ \left\lfloor \frac{i - n \bmod p}{m} \right\rfloor & otherwise \end{cases}$$

And

$$k = \begin{cases} \left\lfloor \frac{j}{m+1} \right\rfloor & if \quad \frac{j}{m+1} < n \bmod p \\ \left\lfloor \frac{j - n \bmod p}{m} \right\rfloor & otherwise \end{cases}$$

The relations given above produces a uniform mapping in the case when p is divisible with n. On the other hand (when p doesn't divide n), some processors (in the first $\boldsymbol{n \bmod p}$ columns and $\boldsymbol{n \bmod p}$ rows), take a matrix which is one dimension larger.

VI. Parallel Algorthm Baed on Hausholder Reflections

Let's take the matrix $A = I - \tau u u^T$ where $u \neq 0$ and τ is a constant which is not equal to 0. The purpose is to choose τ such that A is orthogonal ($A^T A = I$). We have:

$$\begin{aligned} A^T A &= (I - \tau u u^T)^T (I - \tau u u^T) \\ &= I - 2\tau u u^T + \tau^2 u u^T u u^T \\ &= I - 2\tau u u^T + \tau^2 (u^T u) u u^T \\ &= I + (\tau^2 u^T u - 2\tau) u u^T \\ &= I + \tau(\tau u^T u - 2) u u^T \end{aligned}$$

From above, if $\tau = 2/u^T u$, then $A^T A = I$. If we take $u^T u = 1$ then $A = I - 2vv^T$, where $v^T v = 1$.

Householder reflection first implements the decomposition:

$$Q_1 A = R = \begin{bmatrix} x & x & x & \cdots & x \\ 0 & & & & \\ 0 & & A_k & & \\ \vdots & & & & \\ 0 & & & & \end{bmatrix}$$

where $Q_1 = I - 2\frac{uu^T}{u^T u}$ and A_1 is the first vector of A. The matrix Q can be obtained applying the formula $Q^T = Q_{n-1} Q_{n-2} \ldots Q_1$.

Graphical representation of the computation of A is given as below:

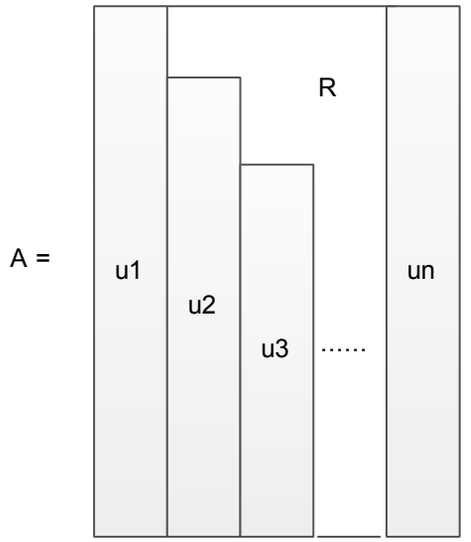


Fig. 5. Graphical representation of the computation of A

The corresponding algorithm is given in as below [16]:

Algorithm 3

for $j = 1$ *to* n *do*
  $s = 0$
  for $i = j$ *to* m *do* $s = s + a_{ij}^2$
  $s = sqrt(s); d_j = -s$ *if* $a_{jj} > 0$, *else* $d_j = s$
  $F = sqrt(s * (s + abs(a_{jj})))$;
  $a_{jj} = a_{jj} - d_j$;
  for $k = j$ *to* m *do* $a_{kj} = a_{kj} / F$;
  for $i = j + 1$ *to* n
  begin
    $s = 0$;
    for $k = j$ *to* m *do* $s = s + a_{kj} * a_{ki}$;
    for $k = j$ *to* m *do* $a_{ki} = a_{ki} - a_{kj} * s$;
  end
end

The dependence graph and the corresponding array using the projection direction [1 0 0] are given in the fig. 5 and fig. 6.

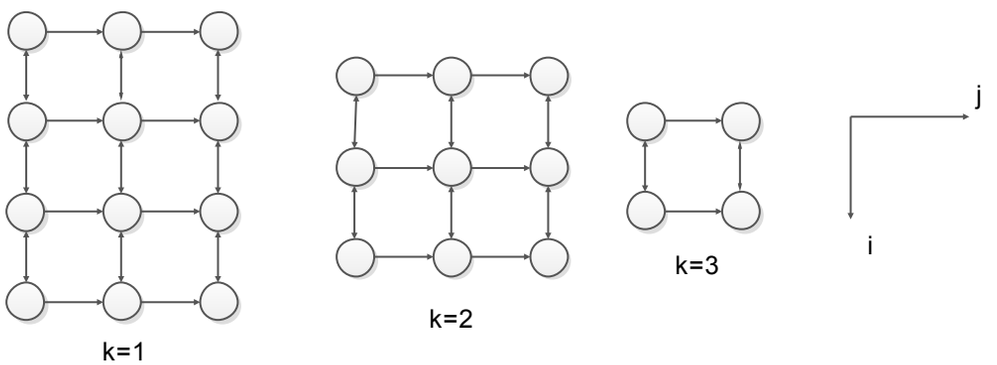


Fig. 6. Dependence graph of the systolic array for QR decomposition using hauseholder reflections

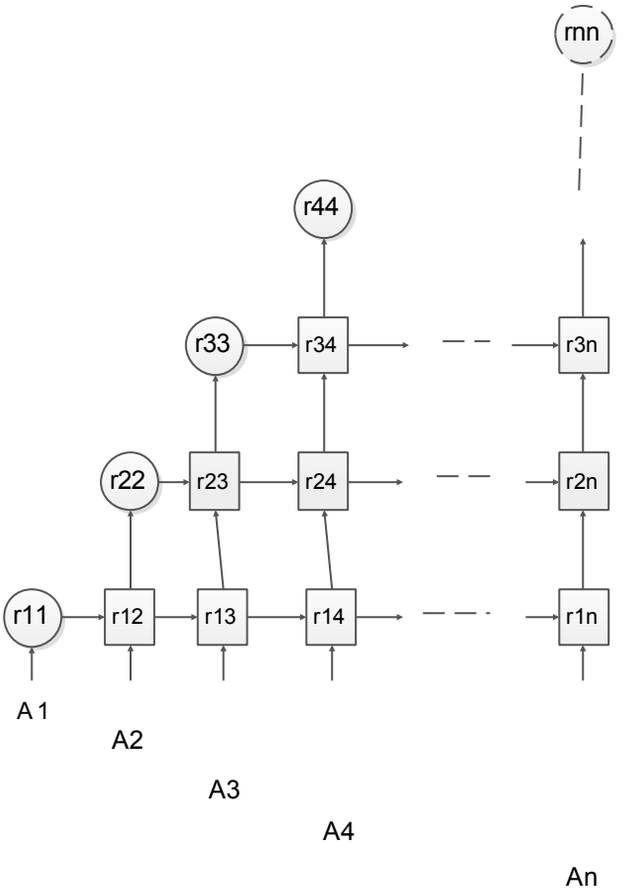


Fig. 7. Systolic array for parallel QR decomposition using hauseholder reflection

In the first step u, v(1) and Q are computed in the first column. Then there is a movement of u and Q in the direction of j axis, and then v(2), ….., v(n) are computed correspondingly in respective columns. In the case of fig. 7, A_i represents the column i of matrix A. As we can see the array is triangular array with the hardware complexity of $O(n^2)$. The array consists of two different shapes of cells. The cells in the shape of a circle (fig. 8), and the cells in quadratic shape (as in fig.9).

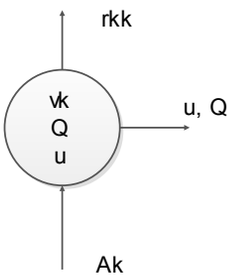


Fig.8. Input and output of the circle cell of the array in fig.7

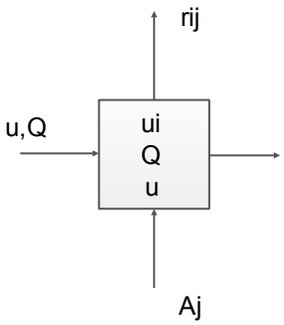


Fig. 9. Input and output of the quadratic cell of the array in fig.7

REFERENCES

[1] J.H. Wilkinson, "The Algebraic Eiginvalue Problem". Oxford, 1965.

[2] L.Johnsson, "A Computational Array for the QR Method". In Proc. 1982 Conf. Advanced Res. VLSI (M.I.T., Cambridge, MA), pp.123-129.

[3] A.H. Sameh and D.J. Kuck, " On stable parallel linear system solvers, " *J. ACM*, vol. 25, pp. 81-95, 1978.

[4] J.Modi and M. Clarke, "An alterantive givens ordering", *Numerische Matheamtik*, vol. 43, no.1, pp. 83-90, 1984.

[5] M. Hofmann and E. Kontoghiorghes, “pipeline Givens Sequences for computing the QR decomposition on a EREW PRAM,” *Parallel computing*, vol. 32, no. 3, pp. 222-230, 2006.

[6] M. Berry, J. Dongara, and Y. Kim, “ a parallel algorithm for the reduction of a nonsymmetric matrix to block upper-Hessenberg form,” *Parallel Computing*, vol. 21, no. 8, pp. 1189-1211, 1995.

[7] S. Y. Kung; VLSI array processors; Prentice Hall, New Jersey (1988).

[8] Systolic Signal Processing Systems; Edited by Earl E. Swartzlander, Marcel Decker, New York, 1987.

[9] D. Kahaner, C. Moler and S. Nash; Numerical methods and Software; Printce Hall, Englewood Cliffs, 1989.

[10] Carl G.J. Jacobi. Uber eine neue Auflosungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen. Astronomishe Nachricten, 22, 1845. English translation by G.W. Stewart, Technical Report 2877, Department of Computer Science, University of Maryland, April 1992.

[11] W.M. Gentleman and H.T. Kung, “Matrix triangularization by systolic array”, Proc. SPIE Int. Soc. Opt. Eng., vol. 298, p. 298, 1981.

[12] J.G. McWhirter and T.J. Shepherd, “An efficient systolic array for MVDR beamforming”, in Proc. Int. Conf.Systolic Array, 1988, pp.11-20.

[13] Feng Tianxiang and Liu Hongxia, “The computer realization of the QR Decomposition on Matrices with Full column rank”, *Proc. IEEE int. Conf. On computer Intellegence and Security*. NO (2009), 76-79.

[14] A. S. Nugraha and T. Basaruddin, “Analysis and Comparisons of QR Decomposition Algorithm in Some Types of Matrix”, *Proc. Of the Federated onference on Computer Science and Information Systems*, 2012, pp. 561-565.

[15] S. Roholah Ghodsi, Bahman Mehri and Mohammad Taeibi-Rahni, “ A parallel implementation of Gram-Schmidt Algorithm for Dense Linear Systems of Equations”, *International Journal of Computer Applications*, Vol. 5-No. 7, 2010, pp. 16-20.

[16] Walter Gander, “Seminar Fuer Angewandte Mathematic Eidgenoessische Technische Hochschule”, Research report No. 80-02, CH-8092 Zuerich, 2003.

Particle Swarm Optimization – Artificial Bee Colony Chain (PSOABCC): A Hybrid Meteahuristic Algorithm

Oğuz Altun
Department of Computer Engineering
Yildiz Technical University
Istanbul, Turkey
oaltun@yildiz.edu.tr

Tarık Korkmaz
Department of Computer Engineering
Epoka University
Tiran, Albania
tarikkorkmaz5@hotmail.com

***Abstract* — A metaheuristic is a search strategy for finding optimal solutions in blackbox optimization problems. Artificial Bee Colony Algorithm (ABC) is a metaheuristics optimization algorithm mimicking the mobility of bees in a bee hive. Particle Swarm Optimization (PSO) is another metaheuristics optimization algorithm inspired by the behaviour of swarms searching for food sources.**

In this work we build a hybrid algorithm that we call 'Particle Swarm Optimization – Artificial Bee Colony Chain' (PSOABCC) that repeatedly apply ideas from PSO and ABC algorithms in a loop.

To ensure that all algorithms themselves have optimal parameters in the experiments, we optimize parameters of PSO, ABC, and PSOABC algorithms with Cuckoo Search optimization algorithm.

The hybrid algorithm shows better convergence performance than the other two algorithms on the two dimensional sphere, rosenbrock, and rastrigin functions.

Keywords— Metaheuristics; optimization; artificial bee colony; particle swarm optimization; Particle Swarm Optimization – Artificial Bee Colony Chain.

I. INTRODUCTION

Metaheuristics are algorithms for finding an optimal value in black box optimization problems. Swarm based metaheuristic algorithms are class of metaheuristics inspired by intelligent behavior of swarms. Amongst swarm based metaheuristic algorithms, Particle Swarm Optimization (PSO) [4] inspires from the naturalistic behavior of birds living in flocks as they gather information on the landscape and transmit the needed information to the other spread flock of birds as in groups' for reaching the best. Artificial Bee Colony (ABC) [6] inspires from the intelligent behavior of a bee colony while searching for better food sources. Cuckoo Search (CS) [5] inspires from the behavior of cuckoos while they look for ideal host nests to lay their eggs.

In this work we build a hybrid algorithm chaining PSO and ABC algorithms. We call the hybrid algorithm we build 'Particle Swarm Optimization - Artificial Bee Colony Chain' and we abbreviate that name as PSOABCC. We compare the success of PSO, ABC, and PSOABCC on three numerical function optimization problems. As each of the PSO, ABC, and PSOABCC have their own parameters that affect their behavior, before doing the comparison, we optimize those parameters of the PSO, ABC, and PSOABC themselves. We prefer to use a metaheuristics different from the algorithms compared (PSO, ABC, and PSOABCC) for this purpose, and decide on using Cuckoo Search (CS), a new metaheuristics algorithm with promising results [5].

The rest of the paper is organized as follows: In Section II we introduce the terminology used in the paper to prevent possible confusion for the reader that is accustomed to different terms used in the literature. In Section III we discuss the Particle Swarm Optimization algorithm. In Section IV we give the pseudo code of the Artificial Bee Colony algorithm. In Section V we summarize the Cuckoo Search algorithm. In Section VI we introduce the hybrid Particle Swarm Optimization – Artificial Bee Colony Chain. In Section VII we review the test functions used. In Section VIII we discuss a method for optimizing parameters of the algorithms themselves. We conclude by the comparison and discussions in Section IX and the summary in Section X.

II. TERMINOLOGY

The algorithms we discuss in this paper find the maximum value of a given function f in a search space (e.g. food source with maximum amount of food). In accordance, we use the word 'fitness' (and also sometimes 'height' or 'f value', interchangeably) for the value of the function, and design the problems as maximization of these functions (explained in detail in section VII).

We follow [1] and use the term 'function assessment' for obtaining the value (or fitness, or height) of the function f on a given coordinate in the search space. Some other works (e.g. [2]) use the term 'function evaluation' for the same purpose. All the algorithms discussed in this work keep track of the number of function assessments done so far during its running time in a variable called $assessmentcount$, and

break out of its main loop when maximum allowed number of function assessments ($maxassessments$) is reached (e.g. when the condition $assessmentcount \leq maxassessments$ holds).

III. Particle Swarm Optimization (PSO) Algorithm

Particle Swarm Optimization [3] algorithm, whose pseudo code is given in Fig. 5, utilizes the idea of birds searching for food sources. Each bird has a velocity and a personal best position. On each step velocity of each bird is recalculated using current velocity, its personal best, and the global best as in (1). The new position of the bird on the step is calculated using this new velocity. Depending on the parameters , $psip$, and $psig$ the algorithm can behave more exploitative or explorative.

PSO(f, $maxassessments$, n):

1. // inputs are f, function to be maximized, $maxassessments$, the number of assessments algorithm allows before breaking out main loop, and n, number of particles
2. // output is g, the coordinate with the best fitness (f value, height) in the search space
3. Assign value 0 to the variable $assessmentcount$. Whenever the fitness is obtained (calling the function f), the $assessmentcount$ is incremented by 1.
4. For each particle
    a. Assign a random position x
    b. Accept that random position as the personal best p of this particle.
    c. Assign a random velocity v
5. Scan positions of all n particles and select position with the best f value as the global best g.
6. Repeat while $assessmentcount \leq maxassessments$
    a. For each particle
        i. Calculate new velocity v using (1).
        ii. Calculate new position x using $x = x + v$
        iii. If the new position x has a better fitness than personal best p, assign x to p
        iv. If the new position x has a better fitness than the global best g, assign x to g
7. Return g, the global best position

Fig. 1 the Particle Swarm Optimization (PSO) algorithm used in our comparisons

In the pseudo code in Fig. 1 new velocity of a particle is calculated by (1) where v is the velocity, w, $psip$, and $psig$ are weight scalars, and rp and rg are random values between 0 and 1.

$$v = w * v + rp * psip * (p - x) + rg * psig * (g - x) \quad (1)$$

IV. Artificial Bee Colony

Artificial Bee Colony (ABC), whose pseudo code is given in Fig. 2, is a metaheuristic that uses ideas from the behaviour of a bee hive. The algorithm starts with n food sources and corresponding n employed bees in the location of these food sources. The employed bees checkout neighbour positions and move there if those positions are better. Later, employed bees return to the hive and inform onlooker bees of the location and quality of the food sources they worked on. A food source may be chosen to be re-visited by an onlooker bee with probability proportional to its quality (fitness). When a food source is checked for better quality neighbours more than the allowed number ($limit$) and no better neighbour is found, the food source is abandoned, and a scout bee finds a random new food source instead of the abandoned one. The algorithm returns the best food source at the end.

ABC(f, $assessmentcount$, n):

1. // inputs are f, function to be maximized, $maxassessments$, the number of assessments algorithm allows before breaking out main loop, and n , number of food sources (equally, number of employed bees)
2. // output is g, the coordinate with the best fitness (f value, height) in the search space
3. // assume we have n food sources and n corresponding employed bees.
4. Assign value 0 to the variable $assessmentcount$. Whenever the fitness is obtained (calling the function f), the $assessmentcount$ is incremented by 1.
5. For each food source assign a random position x. Assume each employed bee is on one of these food sources, and measuring its quality (fitness).
6. Repeat while $assessmentcount \leq maxassessments$
    a. // Employed bee phase
    b. For each food source:
        i. Tweak the position x of the bee employed to get a new position
            1. If the new position has a better fitness accept the new position (move the bee to new position).
            2. Otherwise increase trial value of the food source

by one.

c. // Dance phase: Assume each of the employed bees are returned to the hive, and informed the quality of the corresponding food source (with a 'dance') to the onlooker bees.
d. Assign a probability r of being selected by onlooker bees to each food source. r should be such that the food source with higher fitness has more probability of being chosen.
e. // Onlooker bee phase:
f. For each employed bee (that returned to the hive and did a dance)
    i. With probability r, let its food source be visited by an onlooker bee. Onlooker bee checks fitness of a position in the neighborhood of employed bee position.
        1. If the new position is better accept the new position (move the bee to new position). Onlooker bee turns into an employed bee.
        2. Otherwise increase trial number of the food source by one.
g. // Scout bee phase.
h. For each food source position which trial number exceeds the $limit$ parameter:
    i. Let a scout bee randomly determine a new position and chose the neighborhood as the food source. Let the scout bee turn into an employed bee.

7. Scan the latest food source (employed bee) positions, and return the one with best fitness as the global best g.

Fig. 2 the Artificial Bee Colony (ABC) algorithm used in our comparisons

V. CUCKOO SEARCH (CS) ALGORITHM

Cuckoo Search (CS) [5] metaheuristic algorithm, inspires from the breeding behavior of cuckoos. Instead of building and maintaining their own nests, and raising their own babies, cuckoos lay eggs to nest of an unaware other bird (a.k.a. host nest). The host bird raises cuckoo baby as its own.

In the Cuckoo Search pseudo code given in Fig 3., each cuckoo looks for a new nest in each iteration of the main loop. If the new nest has better quality, it lays an egg there this time. The quality of host nest is assumed to be equivalent to the value of the function f to be maximized at that position in search space. Some of the host birds discover that there is an alien egg in their nest, and get rid of it. The cuckoo abandons this nest, and lays an egg to a new host it finds. The new nests are found using Levy Flights [5].

CS(f, $assessmentcount$, n):

1. // inputs are f, function to be maximized, $maxassessments$, the number of assessments algorithm allows before breaking out main loop, and n , number of nests (equally, number of cuckoos)
2. // output is g, the coordinate with the best fitness (f value, height) in the search space
3. Assign value 0 to the variable $assessmentcount$. Whenever the fitness is obtained (calling the function f), the $assessmentcount$ is incremented by 1.
4. For each cuckoo assign a random nest (position).
5. Repeat while $assessmentcount \leq maxassessments$
    a. // a new breeding season has begun
    b. For each cuckoo:
        i. Find a new nest, using Levy Flights [5].
        ii. If the new nest has a better fitness than the current nest, this year lay egg to new nest. Otherwise, lay egg to old nest.
    c. Abandon pa percent of the worst nests. The eggs on these nests are discovered by host bird.
    d. For each of the abandoned nests:
        i. Find a new nest, using Levy Flights [5].
        ii. Lay an egg to the new nest.
6. Scan the nests and return g, the nest with best fitness.

Fig. 3 Cuckoo Search (CS) algorithm used for optimizing parameters of PSO, ABC, and PSOABCC.

VI. PARTICLE SWARM OPTIMIZATION – ARTIFICIAL BEE COLONY CHAIN (PSOABCC) ALGORITHM

We build a hybrid algorithm that we call 'Particle Swarm Optimization – Artificial Bee Colony Chain' (PSOABCC) as given in Fig. 4. The algorithm starts by building random initial personal best positions for particles/bees. Then in the main

loop it improves the personal bests using PSO_PHASE (Fig. 5) and ABC_PHASE (Fig. 6) methods. When the main loop ends, the algorithm scans the latest personal best positions and returns the best of them as the global best.

The PSO_PHASE method (Fig. 5) takes a list of best positions of particles, updates them with better positions, and returns back. In contrast with the personal bests, a new random current position is assigned for each particle. The main loop of the method runs a fixed $niterpso$ times. The rest of the method behaves like the normal PSO algorithm in Fig. 1, except the output. PSO_PHASE outputs the modified set of personal bests back, whereas PSO algorithm returns a single global best position.

The ABC_PHASE method (Fig. 6) takes an initial list of employed bee (equivalently food source) positions as a parameter, improves them, and returns them back. In the PSOABCC this list of positions comes from the personal bests of the previous PSO_PHASE. This ensures that PSO_PHASE and ABC_PHASE incrementally improve the same list of personal best positions. The main loop of the method runs a fixed $niterabc$ times. The rest of the method behaves like the normal ABC algorithm in Fig. 2, except what it outputs. While the ABC_PHASE returns the list of current food source (employed bee) positions, the ABC algorithm returns the best of these food source positions as the global best.

PSOABCC(f, $assessmentcount$, n):

1. //inputs are f, function to be maximized, $maxassessments$, the number of assessments algorithm allows before breaking out main loop, and n, number of particles (equally, number of food sources, equally, number of employed bees)
2. //output is g, the coordinate with the best fitness (f value, height) in the search space
3. Assign value 0 to the variable $assessmentcount$. Whenever the fitness is obtained (calling the function f), the $assessmentcount$ is incremented by 1.
4. For each particle/employed bee assign a random initial position.
5. For each particle/employed bee assign initial position as the personal best. Let set of personal best values be P.
6. Repeat while $assessmentcount \leq maxassessments$
    a. Improve personal bests using PSO_PHASE as shown in Fig. 5: P <- PSO_PHASE(P)
    b. Improve personal bests using ABC_PHASE as shown in
    c. Fig. 6: P <- ABC_PHASE(P)
7. Scan the set of personal bests and return the one with the best fitness as the global best g.

Fig. 4 Particle Swarm Optimization – Artificial Bee Colony Chain (PSOABCC) algorithm pseudocode

PSO_PHASE(P):

8. // input: $P = \{p_1, p_2, ...\}$ (initial best positions of the particles)
9. // output: $P = \{p_1, p_2, ...\}$ (final best positions of the particles)
10. For each particle
    a. Assign a random position x
    b. Update the personal best if x has better fitness than p
    c. Assign a random velocity v
11. Select position with the best fitness as global best g.
12. Repeat while we iterated less then $niterpso$:
    a. For each particle
        i. Calculate new velocity v using (1).
        ii. Calculate new position x using $x = x + v$
        iii. If the new position x has a better fitness than personal best p, assign x to p
        iv. If the new position x has a better fitness than the global best g, assign x to g
13. Return set of particle best positions P

Fig. 5. Particle swarm optimization phase (PSO_PHASE)

ABC_PHASE(X):

1. // input: $X = \{x_1, x_2, ...\}$ (initial employed bee positions)
2. // output: $X = \{x_1, x_2, ...\}$ (final employed bees positions)
3. While we iterated less than $niterabc$:
    a. // Employed bee phase
    b. For each food source:
        i. Tweak the position x of the bee employed to get a new position
            1. If the new position has a better fitness accept the new position (move the bee to new position).
            2. Otherwise increase trial value of the food source by one.
    c. // Dance phase:
    d. Assign a probability r of being selected by onlooker bees to each employed bee. r should be such that the employed bee with position that has better fitness has more probability of being chosen.
    e. // Onlooker bee phase:
    f. For each employed bee (that returned to the hive and did a dance)
        i. With probability r, let its food source be visited by an onlooker bee. Onlooker bee checks fitness of a position in the neighborhood of employed bee position.
            1. If the new position is better accept the new position (move the bee to new position). Onlooker bee turns into an employed bee.
            2. Otherwise increase trial number of the food source by one.
    g. // Scout bee phase.
    h. For each food source position which trial number exceeds the $limit$ parameter:
        i. Let a scout bee randomly determine a new position and chose the neighborhood as the food source. Let the scout bee turn into an employed bee.
4. Return X

Fig. 6 Artificial bee colony phase (ABC_PHASE)

VII. TEST FUNCTIONS

We have evaluated algorithms on following test functions.

A. Negative Rastrigin Function

Negative Rastrigin Function is the "negative" of Rastrigin Function [1] and can be defined by (2).

$$\text{Maximize } f(< x1, \dots, xn >) = -10n - \sum_{i=1}^{n} {x_i}^2 - 10\cos 2\pi x_i \quad (2)$$

In (2) n is number of dimensions, x_i is value of the solution vector in dimension i. Global maximum value is 0 and is on (0,0,..,0) . Initialization range of the function is $[-5.12, 5.12]$.

B. Negative Rosenbrock Function (Ros)

The Negative Rosenbrock (Ros) Function is the "negative" of the Rosenbrock Function [3] and can be defined by (3).

$$\text{Maximize } f(< x1, \dots, xn >) = -\sum_{i=1}^{n-1}(1 - x_i)^2 + 100(x_{i+1}x_i^2)^2 \quad (3)$$

In (3) n is number of dimensions, and x_i is value of the solution vector in dimension i. Global maximum value is 0 and is on (1,1,...,1). The initialization range of the function is $[-2.048, 2.048]$.

C. Negative Sphere Function (Sph)

Negative Sphere Function (Sph) is the negative of Sphere Function [1] and can be defined by (4).

$$\text{Maximize } f(< x1, \dots, xn >) = -\sum_{i=1}^{n} {x_i}^2 \quad (4)$$

In (4) n is number of dimensions, x_i is value of the solution vector in dimension i. Maximum height is 0 and is on (0,0,…,0). The initialization range is $[-5.12, 5.12]$.

VIII. OPTIMIZING ALGORITHM PARAMETERS USING CUCKOO SEARCH

Metaheuristic algorithms PSO, ABC, and PSOABC themselves have parameters that affect their behavior. In order to compare each algorithm at its best success rate, we optimize the parameters of the algorithms before we compare them. For this purpose we use a fourth algorithm, Cuckoo Search (CS), given in Fig. 3, as the optimizer.

For this purpose we define the parameter optimization fitness function F in Fig. 7. In contrast to simpler fitness functions we defined in Section VII, this function has an $algorithm$ (e.g. PSO), and a simple function f (e.g. Negative Rastrigin Function discussed in Section VII.A) to be solved by $algorithm$. F takes the list of parameter values $params$ (e.g. $\{w = 0.3, psip = 0.5, psig = 0.4\}$) as coordinates, and returns a fitness value. For this, the $algorithm$ solves the simple function f m times with given parameter value set $params$, to get m different best positions. Each best position has a corresponding f value (fitness). The mean of these fitness values is returned as a means of how well the parameter value set $params$ did with $algorithm$ and f. Running m independent runs and taking the mean is necessary because all the algorithms we are interested in comparing (PSO, ABC, and PSOABCC) behave according to randomly generated values.

We give F as a parameter to CS as a function to be optimized. This way we actually optimize the parameter set of the $algorithm$. Any metaheuristic algorithm could be used for this purpose, we choose CS because a) it is a different algorithm than PSO, ABC, and PSOABCC, b) good results with CS are reported in [5].

$F(algorithm, f; params)$:

1. $ntrials$: <- 1
2. while $i \leq m$:
   a. g_i <- $algorithm(f, params)$
   b. v_i <- $f(g_i)$
3. s <- $v_1 + v_2 + \cdots + v_m$
4. $meanh$ <- $\frac{s}{m}$
5. Return $meanh$

Fig. 7 A fitness function for optimizing metaheuristics algorithms

For the PSO algorithm, we optimize the parameters w, $psip$, and $psig$ used in (5). For ABC algorithm and PSOABCC algorithm we optimize the $limit$ variable used in their scout bee phases. For PSOABCC algorithm we optimize parameters $niterpso$ and $niterabc$ that affect breaking out of main loops of PSO_PHASE and ABC_PHASE. Each algorithm has different optimal parameters for each fitness function f. Hence, we re-optimize each parameter set on each test function.

We prefer not giving the optimized parameter values here as we think the values we found may be local optima.

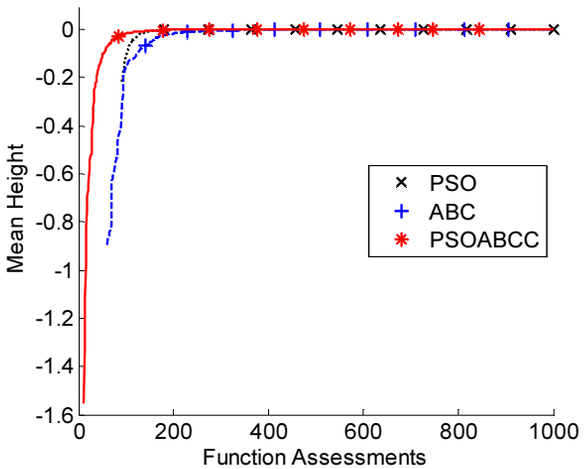
IX. OPTIMIZATION COMPARISON AND DISCUSSION

To compare optimization performances of PSO, ABC, and PSOABCC, we do 100 independent runs of each of these algorithms on each test function and plot the mean convergence graph. Each algorithm was terminated after reaching 1000 function assessments (function evaluations) to make a fair comparison. On each of the test functions we tested, the PSOABCC is clearly converges faster than the other two algorithms.

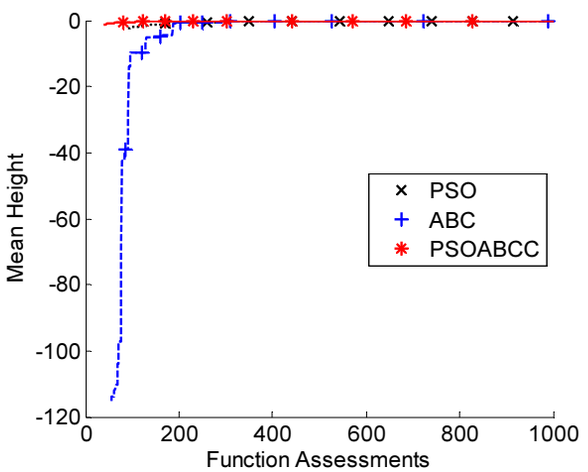
On Negative Sphere Function (Fig. 8a) which is a unimodal function, all the algorithms converge before 300 assessments, and the order of convergence is PSOABCC>PSO>ABC. Hence PSOABCC converges faster than the other two.

On Negative Rosenbrock Function (Fig. 8b) which is a unimodal function, all three functions converge around 200 function assessments. Again PSOABCC is the fastest.

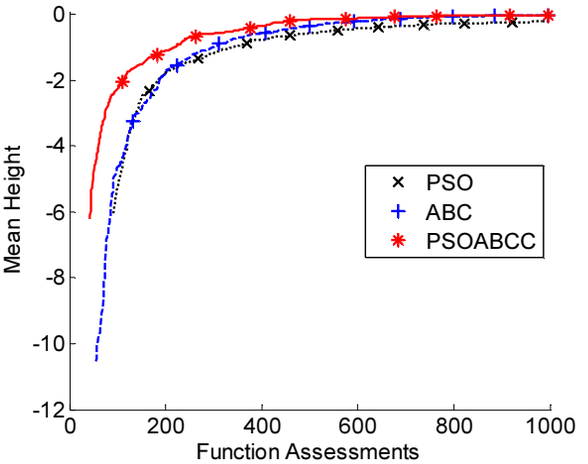
On Negative Rastrigin Function (Fig. 8c) which is a multimodal function, PSO keeps stuck in a local optima and can not reach to the global optimum even after 1000 assessments. After 800 assessments ABC catches on PSOABCC, but PSOABCC clearly converges faster.



(a) Negative Sphere Function



(b) Negative Rosenbrock Function



(c) Negative Rastrigin Function

Fig. 8 Convergence graphs of PSO, ABC and PSOABCC algorithms on (a) Negative Sphere Function, (b) Negative Rosenbrock Function, (c) Negative Rastrigin Function. Each convergence plot shows the mean height of 100 independent runs against number of function assessments (function evaluations).

X. SUMMARY

We chain the Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC) algorithms to obtain a hybrid we call Particle Swarm Optimization – Artificial Bee Colony Chain (PSOABC). We compare the three algorithms PSO, ABC, and PSOABC on three well known test functions: Negative Sphere Function, Negative Rosenbrock Function, and Negative Rastrigin Function.

We pre-optimized the parameters of the algorithms themselves on each function using another algorithm, Cuckoo Search, since each algorithm behave different according to these input parameter values.

We compare the convergence graphs of PSO, ABC, and PSOABCC on each function. The convergence graphs show that PSOABCC converges faster from the two other algorithms.

Testing the algorithm further on other test and real world problems and looking for algorithmic improvements remains as future work.

XI. REFERENCES

[1] S. Luke, Essentials of metaheuristics, 2012.

[2] X. Yan, Z. Yunlong and Z. Wenping, "A Hybrid Artificial Bee Colony Algorithm for Numerical Function Optimization," in *11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011.

[3] E. James and K. Russell, "Particle Swarm Optimization," *1995 IEEE International Conference on Neural Networks,* vol. 4, p. 1942–1948, 1995.

[4] M. Melanie, An Introduction to Genetic Algorithms, MIT Press, 1998.

[5] R. Storn and K. Price, "Differential evolution – a simple and efficient adaptive scheme for global optimization over continuous spaces," In Technical Report TR-95-012, Berkley, 1995.

[6] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm," *Journal of Global Optimization,* vol. 3, p. 459–471, 2007.

[7] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," In Technical Report TR-06, Kayseri/Türkiye, 2005.

[8] A. Eiben and J. Smith, Introduction to evolutionary computing, Springer, 2003.

[9] X.-S. Yang and S. Deb, "Engineering optimisation by Cuckoo Search," *Int. J. Mathematical Modelling and Numerical Optimisation,* vol. 1, no. 4, pp. 330-343, 2010.

[10] R. Eberhart, Y. Shi and J. Kennedy, Swarm intelligence, Morgan Kaufmann, 2001.

A Scalable Algorithm for Similar Image Detection

Andrei-Bogdan Pârvu
Computer Science and Engineering Department
Faculty of Automatic Control and Computers
POLITEHNICA University of Bucharest
Email: andrei.parvu@cti.pub.ro

Nicolae Țăpuș
Computer Science and Engineering Department
Faculty Automatic Control and Computers
POLITEHNICA University of Bucharest
Email: nicolae.tapus@cs.pub.ro

Ștefan-Teodor Crăciun
Adobe Systems Romania
Email: scraciun@adobe.com

Virgil Palanciuc
Adobe Systems Romania
Email: virgilp@adobe.com

***Abstract*—In this paper, we propose an algorithm for detecting similar images in a large scale dataset. After considering various methods as invisible watermarking and image feature detection, we developed a method based on the $Harris\ corner\ detection$ and the $Scale\ Invariant\ Feature\ Transform$ keypoints and descriptors for analyzing the properties of an image.**
Our goal is that the proposed algorithm be resilient to watermarked, scaled or cropped images, and provide a fast running time for our large dataset.
In order to be able to handle a large amount of images and associated data, we have decided to maintain the descriptors of the images in a set of KD-tree structures for fast querying. This allows an initial filtering of the data set, reducing the number of images which should be analyzed. Thus, we will use two search algorithms, one of which has a lower time complexity, but has a lower accuracy when providing the results, and a second one, which performed a more in-depth analysis of the images.

I. Introduction

Image processing has been a very important research domain the last years, the ever increasing number of images present on the Internet becoming more and more challenging to index, store and analyze.
Copyright infringement and image detection have also been big issues which have been discussed and studied for a long time, major users in the photography industry wanting to know at all times when someone is using or modifying one of their photos.
The problem discussed in this paper is considered a very difficult one, because the algorithm needs to be at the same time accurate, fast and scalable.

Our goal was to design a scalable algorithm which can index information about a large set of images, and perform queries of finding a highly similar image with a given input one.
The algorithm should be focused on copyright detection, so it should be able to determine if the input image is one of the images in the dataset, with one or several of the following transformations applied:

- watermarks
- scaling
- cropping
- various filters

As said above, the granularity of the algorithm should be set so that it should be able to distinguish between two lightly similar images (e.g. two different pictures of the Eiffel Tower, made by two different persons) and two images, one of which is obtained from the other.

The speed and scalability of the algorithm are also major factors which should be seriously taken into consideration. It should be able to handle a large number of queries and provide a small response-time per query. Of course, there is a close connection between the running time of the algorithm and its precision, connection which should be closely analyzed in order to create a balance between the two.

June 2014

II. Previous Work

There are a lot of algorithms which focus on image similarity and key feature detection, and after studying several ones, we decided on focusing on two main ones, the $Harris\ corner\ detector$ and the $Scale\ Invariant\ Feature\ Transform$. Our choice was determined on the fact we wanted to be able to control the sensitivity of the matches, but also to efficiently distribute the algorithm on several machines for a large input set.

A. Harris Corner Detection

The main idea of the $Harris\ corner\ detector$ algorithm is that, given an input image, the most predominant features that a human eye recognizes and memorizes are corners. A corner is considered to be an intersection of two edges, so, selecting a small area around the point and shifting it should result in a large variation in the intensity of the pixels in that area.
Therefore, each area in an image can be classified in three categories:

- flat, in which intensities do not vary in either direction

- edge, in which intensities don't vary in the direction of the edge
- corner, in which intensities vary in all directions

In order to determine in which category a certain area with a size of (w, h) belongs to, we will compute the variation of intensity: $E(w,h) = \sum_{x,y} w(x,y) * [I(x+w, y+h) - I(x,y)]^2$, where w is a window function, which assigns weights to pixels, and I is the intensity of a certain pixel of the grayscale image. In order to determine the corner areas, we have to maximize the function $\sum_{x,y}[I(x+w, y+v) - I(x,y)]^2$, which using $Taylor$ expansion and representing in a matrix form can be written as $E(w,h) \approx \begin{bmatrix} w & h \end{bmatrix} * \left(\sum_{x,y} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \right) * \begin{bmatrix} w \\ h \end{bmatrix}$, and, furthermore, using a substitution $E(w,h) \approx \begin{bmatrix} w & h \end{bmatrix} * M * \begin{bmatrix} w \\ h \end{bmatrix}$. Using this equation, the score of a certain area is computed as $R = det(M) - k * (trace(M))^2$. A higher score of R denotes a higher probability of the area being a corner.

B. Scale Invariant Feature Transform

1) Keypoint localization: Although the $Harris\ corner\ detection$ algorithm presented in the previous section is immune to rotation transformations of an image, it does not perform well if the image is scaled, because a high intensity change in an area of size (w, h) of an image might vary if the dimensions of the image change, but the size of the area remains the same.
Thus, D. Lowe, in 2004, presented a new algorithm for extracting keypoints and computing their descriptors, named $Scale\ Invariant\ Feature\ Transform$.
At first, a Gaussian distribution is applied on the analyzed image, which depending of the standard deviation, σ, blurs the image with a certain amount: $G(x,y) = \frac{1}{2*\pi*\sigma^2} * e^{-\frac{x^2+y^2}{2*\sigma^2}}$.
Then, the Laplacian of the image is computed, in order to highlight the regions of rapid intensity changes: $L(x,y) = \frac{\delta^2 I}{\delta x^2} + \frac{\delta^2 I}{\delta y^2}$. Combined with the previous Gaussian filter, we obtain the so called Laplacian of Gaussian: $LoG(x,y) = -\frac{1}{\pi*\sigma^4} * \left(1 - \frac{x^2+y^2}{2*\sigma^2}\right) * e^{-\frac{x^2+y^2}{2*\sigma^2}}$.
Because the LoG has a high computational cost, it is approximated with a Difference of Gaussians, which is a difference of two Gaussians with two different σ deviations, representing two different scaled images. The local extrema of the computed DoG are considered potential keypoints.

2) Computing the Descriptors: Once we have the keypoints, the corresponding descriptors are computed by taking a 16×16 neighborhood around the keypoint, and creating a 8 bin histogram for each sub-block of 4×4 size of the initial neighborhood. Thus, a keypoint descriptor will contain 128 values.

III. The Image Similarity Algorithm

The algorithm we developed is composed out of two parts:

- the analysis of a pair of images, in which we try to determine the degree of similarity between two images
- the retrieval of a small subset of likely candidates of similarity from a large image database

A. Analysis of a pair of images

In order to analyze a pair of images, we used the SIFT descriptors determined in the previous section. The two sets of descriptors are compared in order to obtain the best matches between pairs of keypoints. A distance is computed between each pair of keypoint descriptors, which is the Euclidian norm between the SIFT descriptors of the keypoints. Experimentally, we have concluded that a match between two keypoints has a high similarity if the Euclidian norm is less than 100.
For a pair of images, we first find the set of SIFT components that are part of the Harris corner-mask and then compute the best matches between these keypoints. We have decided to use the corner-mask as a filtering mechanism for the SIFT keypoints in order to be able to potentially reduce the number of descriptors of a given image. We keep only the best 10 matches, and compute the arithmetic mean between the distances of these matches; if this mean is smaller than 100, the two images are considered similar. We shall name this algorithm the $pair\ similarity\ algorithm$ and the mean between the distances $image\ pair\ score$.
Figure 1 shows the corresponding matches between two images with two different watermarks, one of which is rotated 90^o clockwise.

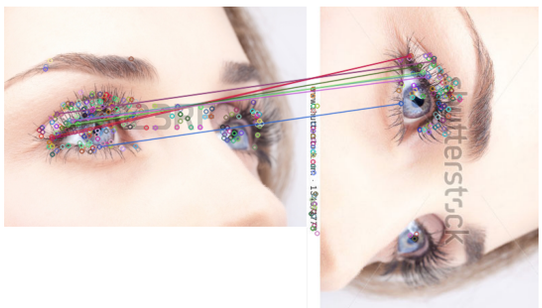


Fig. 1. Comparison between two images

B. Analysis of a set of images

Suppose that we have a set of images, and we want to compare a test image with each image in the set and detect whether there exists a similar one. Of course, the first possibility is the brute force one: we iterate through all the images and apply the $pair\ similarity\ algorithm$ described in the previous section. Although this provides a correct result, it has a complexity of $O(number\_of\_images * image\_match\_time)$. We shall name this basic algorithm as the $linear\ algorithm$. Although this algorithm is very straightforward, its complexity is undesirable if the number of images becomes large (i.e. more than 500). Moreover, most of the images in our set will likely have a big similarity distance with our searched image, so maybe we don't want to apply the full $pair\ similarity\ algorithm$.
Thus, we need to determine an efficient algorithm which can filter the initial set of images to a smaller set which contains the best possible candidates in terms of visual similarity.
The filtering algorithm is implemented as follows: we maintain a maximum number of M descriptors for each image in the initial set and create P KD-trees with the set of descriptors of the initial images. When a query for a test image arrives, we compute its descriptors and then perform a K nearest-neighbor search on each of our KD-trees. Then, we select the

top T images returned from the nearest neighbor searches and perform the $linear\ algorithm$. We shall name this algorithm the $kdtree\ algorithm$.
Of course, the important factors in the $kdtree\ algorithm$ are:

- the values of the numbers M, T and K from the above description
- the metric used in selecting the top T images from the nearest-neighbor search
- the metric used in selecting the M descriptors from the input images to form the KD-tree
- N, the number of images that form a KD-tree
- P, the number of KD-trees, which can be determined from the total number of images and N.

We propose three metrics in for selecting the filtered images from the KD-tree based on the nearest-neighbor search:

- the images with the largest number of descriptors returned from the search
- the images with the smallest average of the distance of the found descriptors
- the images with the largest distance between its descriptors and the average of all found descriptors (from all the images)

The main problem when storing the descriptors in the KD-tree is in what form to have the images during the computation. We have three possibilities:

- keep them in the original size
- resize all the images to a fixed dimension (eg. 400 $\times$ 300)
- resize all the images to a fixed width and scale the height

For the initial attempts, we have set the values of M, T and K to 40, 10 and 5 and used the largest number of found descriptors as metric for selecting the filtered images from the KD-tree.

C. Architecture of the application

The basic structure of our application is maintaining a set of Image Servers, each containing a number of R KD-trees to use for quering. Thus, when a query arrives, we shall use a map-reduce technique: the query is distributed among the Image Servers, which compute the most similar images from their associated KD-tress and send the results back to a special entity called the Map Reducer. The Map Reducer combines these results, takes the best ones and sends them back to the quering entity.

IV. Data Set and Testing

We have collected a set of 4500 images from the Internet, which we have classified in about 1600 similarity sets, each set being composed of up to five images. The images within a similarity set differ in size, having various watermarks and filters applied (these sets are known to be correct beforehand). We have inserted these images in a larger set of 100.000 images taken from the Internet and performed a queries for each of the 4500 images of the similarity sets, getting the top 10 similar images. We want to observe:

- if the algorithm finds the exact match, i.e. the image that has been queried with
- if the algorithm finds the other images which are part of the same similarity set
- the mean running time of a query
- how varying the metrics described above influences the results of the query

We evaluate the response to a query, by looking at the indices of the images from the current similarity set in the list of results returned by the query. Thus, the $index\ score$ for a certain query can be computed as the sum of these indices; the smaller the sum, the closer in the response list are the images we want to find.
For the set of 100.000 images we use $P = 20$ KD-trees, each KD-tree storing the descriptors for $N = 5000$ images.
The results of this test can be seen in the following two tables. The first one describes, for similarity sets of size 3, 4 and 5, and for the metrics described in The Image Similarity Algorithm, the average number of how many of these images are found in the list of 10 returned images.

	max nr descs	min avg	max dist to avg
3	2.47	2.37	2.52
4	3.26	3.21	3.38
5	3.98	3.63	4.14

The second table shows the $index\ score$ (described above) for the same queries and metrics.

	max nr descs	min avg	max dist to avg
3	8.32	9.29	7.84
4	12.69	13.15	11.76
5	18.35	21.36	17.10

It can be observed that the third metric, the largest distance from the descriptors of an image to the average of all found descriptors, provides the best results.

We computed the descriptors for images resized to a fixed value (height 400 and width 300), and analyzed the same metrics and scores described above. The results can be seen in the tables below, and they confirm that this image storing performs worse than the aspect-ratio one, so we have decided to continue using the first one.

	max nr descs	min avg	max dist to avg
3	2.69	2.08	2.73
4	3.55	2.60	3.62
5	4.17	2.80	4.32

	max nr descs	min avg	max dist to avg
3	6.32	11.93	5.97
4	10.37	18.46	9.84
5	16.8	28.35	15.67

As stated in The Image Similarity Algorithm, we decided to maintain multiple KD-trees in one Image Server in order to improve the quality of the heuristic search on one such KD-tree and reduce the total number of processes.
On the same set of 100.000 images, we created $P = 40$ Image Servers, with each server having 5000 images and $R = 2$ KD-trees (that means $N = 2500$ images per KD-tree), so that we could compare the *correctness scores* between this implementation and the one with $R = 1$.
These are shown in the tables below:

	max nr descs	min avg	max dist to avg
3	2.86	2.5	2.87
4	3.83	3.35	3.85
5	4.6	4.75	4.62

	max nr descs	min avg	max dist to avg
3	4.76	8.03	4.70
4	7.96	11.92	7.82
5	13.56	20.35	13.49

It can be seen that the scores are comparable with the ones obtained from running the algorithm it the two cases shown in the previous subsections, so it can be confirmed that using multiple KD-trees of a fixed (and lesser size) is in our advantage when the overall dataset becomes larger, in order to avoid a large number of Image Servers.

Also, we have constructed a single KD-tree which contains the 4500 images from the similarity set in order to test the three different metrics described in The Image Similarity Algorithm for selecting the filtered images.
We retained the *image pair score* of the returned similar images, and evaluated these three metrics by computing the following *correctness scores*:

- the mean between the *image pair scores*
- the sum of the differences between the *pair scores* of two consecutive similar images in the returned list
- the maximum *image pair score*

The goal is to minimize each of these *correctness scores*.

The results of this test are shown in the next table:

	max nr descs	min avg	max dist to avg
mean	661880	693745	647355
sum of diff	867877	1021543	856521
max score	1162618	1134470	1150899

As it can be seen, the third metric, largest distance from the descriptors of an image to the average of all found descriptors, performs the best out of the three metrics.

A. Running Time

We have tested our algorithm on a machine with 55GB of RAM, 16 quad-core processors with a frequency of 2.4GHz. There are two different running times that we have been interested in: the first is the actual time that it takes for a query result to be computed - because of the heuristic search on the KD-tree with a limited depth, this will not vary very much when the KD-tree grows in size. For a number of 5000 queries, the mean running time of a search on a KD-tree is 1.36 seconds.
The second running time is the initialization of the KD-tree, which is divided into two steps: the computation of the descriptors for the images, and the construction of the actual KD-tree.
In Figure 2 we can see the total initialization time for a KD-tree, and the time needed only for the construction of the KD-tree data structure (presuming that the descriptors are already computed).

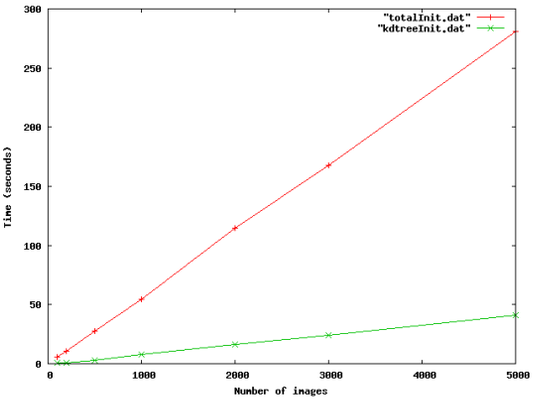


Fig. 2. Initialization runtimes

number of images	total init (seconds)	kdtree init (seconds)
100	5.34	0.48
200	10.53	0.99
500	27.71	3.17
1000	54.52	7.72
2000	114.27	16.26
3000	167.83	24.31
5000	280.67	41.20

We analyzed how a certain KD-tree Image Server with a varying number of linear processed KD-trees performs when submitting a query, in order to determine a viable value for the R parameter (the number of KD-trees that form a KD-tree Image Server).
In the table below we can observe the running time per query as we increase the number of KD-trees for with $N = 2500$ images.

R, number of KD-trees	running time (seconds)
2	2.45
3	2.75
5	2.84
10	3.01
20	3.44
40	4.41

By looking at these running times, we can determine that we can easily increase the number of KD-trees per Image Server, as long as we maintain a decent number of images as input for the *linear algorithm*, which is used after the actual KD-tree search.

Because the total number of images is divided into KD-trees of fixed dimension (in our case 5000 images), inserting a

new image into our database is constant, because it just implies creating a new KD-tree (or expanding a current one, if its dimension doesn't exceed 5000 images).

V. Conclusions and Further Work

In this paper, we have developed a scalable algorithm for finding similar images in a large database, using the Harris corner transformation, the SIFT descriptors and multiple KD-trees for storing the image data. We have explored several metrics for selecting images out of the KD-trees and have analyzed how these different metrics influence the images returned by a query.

We have analyzed the running time of the algorithm, varying the dimensions of the image database, and the associated KD-trees.

The algorithm has performed well under the current conditions, and we plan on continuing our work on it, running a larger number of queries on a larger initial image set and varying the size of the KD-trees, in order to study how this affects overall running time and accuracy of results.

We are also planning on running this algorithm with different kinds of descriptors, like SURF or GIST, and try to implement an efficient method for compressing these descriptors, which would allow a larger storage capacity in the KD-trees.

References

[1] D.G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints*, 2004

[2] M. Johnson, *Generalized Descriptor Compression for Storage and Matching*, 2005

[3] O. Chum, J. Philbin, M.Isard, A. Zisserman, *Scalable Near Identical Image and Shot Detection*, 2008

[4] A. Vedaldi, *An implementation of SIFT detector and descriptor*, 2006

[5] M Trajković, M Hedley, *Fast Corner Detection*, 1998

[6] C Harris, M Stephens, *A combined corner and edge detector*, 1988

[7] E. J. Stollnitz, T. D. DeRose, D. H. Salesin, *Wavelets for Computer Graphics*

A Location-Map Free Reversible Watermarking With Capacity Control

Wen-Shyong Hsieh
Department of Computer and Communication
Shu-Te University, Taiwan
wshsieh@stu.edu.tw

Jui-Ming Kuo
Department of Computer Science and Engineering,
National Sun Yat-sen University, Taiwan
g6326852@gmail.com

***Abstr*act—Reversible watermarking techniques enable the extraction of the embedding bits from a watermarked image in a lossless way. It exploits the high spatial correlation among neighboring pixels. Application in reversible watermarking includes military and medical images. However, images occurs overflow or underflow problems during the imbedding process. Since pixels value may be out of range between [0:255]. Most methods require a location map to solve such problems. In this study, we propose a location map free reversible watermarking algorithm. First, a prediction threshold value is computed, histogram shifting scheme based on the prediction threshold value to solve overflow and underflow problems, second, another threshold value is adopted to achieve capacity control, image quality is better in different payload length with this control. The experimental results reveal that the performance of our proposed method outperforms that proposed by FUJIYOSHI et al. For example, with the same imbedding capacity, the PSNR of our scheme is higher than FUJIYOSHI et al. by 3 dB. Furthermore, our algorithm provides higher embedding capacity compared with FUJIYOSHI et al.**

Keywords—reversible watermarking; capacity control; predicted value; histogram shifting;

I. INTRODUCTION

Because of the popularity of computers and the development of Internet, digital images and video audio become more accessible. However digital media is easy to be copied and modified, when the digital media of intellectual property owner suffer from infringement, the owner is hard to proof his Intellectual Property. In order to proof the Intellectual Property, the owner can put digital signature on the digital media. The research of verify media intellectual property rights so that rightful ownership can be declared is watermarking technique.

Depending on whether the human eye can recognize, digital watermarking technology can be divided into two categories, one is the visible watermark technology, and the other is the invisible watermark technology. The visible watermark technology embeds watermark like a translucent loge into a media, its main purpose is to declare the ownership, to prevent illegal use. However, the disadvantage is that reduce the commercial value of the media. In addition, visible watermark easily be overwritten or removed via signal processing approach.

Currently watermark technology research and development, mainly focusing on the invisible watermark technology development. Invisible watermark can be divided into two categories according to the embedding data domains, namely, the frequency domain[14] and spatial domain[1], using the frequency domain watermark is Robustness, but computational intensity requires large amount than that embedding data in spatial domain, and the capacity is lower. On the contrary, the spatial domain watermark is high capacity, but the watermarked image is fragile. In this article, we mainly introduce the spatial domain watermark scheme.

Some traditional watermarking technique does not recover original image. However, data hiding in medical and military images[5], because of their specific requirements, sensitive in image quality. Therefore, reversible watermarking has been proposed to restore the image after watermark was extracted.

There are two important objectives for reversible watermarking techniques, the embedding capacity and the watermarked image quality. It is difficult to achieve these two objectives at the same time. In general, an improved technique embeds the same capacity with lower distortion or vice versa. Reversible watermarking techniques also have to solve the location map problem caused by overflow and underflow, location map is additional burden for watermarking techniques[7], many watermarking method try to reduce the size of location map. In this paper, we proposed a novel watermarking technique that don't need location map anymore.

II. RELATED WORKS

Tian [1] proposed a reversible watermarking scheme by difference expansion (DE). He used the redundancy in the digital images to find extra storage space. His method divides image pixels into pairs, then watermark bits into each pairs through the difference expansion technique. Since Tian's method embed those pairs of pixels that will not cause overflow and underflow problem. To check the positions of watermarked pairs, need to construct a location map. The location map size is half of cover image (0.5bpp). Compression can reduce the location map size, but still a large overhead. In addition, the embedding capacity is at most 0.5bpp in single round, a higher embedding capacity is achieved by multi-layer embedding.

Alattar [3] extended Tian's method using difference expansion of vectors of adjacent pixels to watermarking bits. Location map used to identify different vectors. He simulated

results using quad-based algorithm and his method has better performance than Tian[1], because the location map need 1/N bpp without compression.

Kamstra and Heijmans [11] used the variance of neighboring pixels to sorting, due to the high correlation of image pixels, improve the performance and reduce the location map size

Ni et al. [4] developed a histogram shifting method. Firstly, scan the image to build a histogram of pixels. Next, find the pair of peak and zero points from the histogram. Embedding data into peak pixels and shifting others pixels to zero points. The advantages are low computational complexity and execution time. But the image capacity is limited by the number of the peak points.

Fallahpour et al. [5] introduce a highly efficient reversible data hiding system. Dividing the image into four or sixteen non-overlapping image tiles. Find the pair of peak and zero points of image histogram from each tile. The frequency of the peak point determined the embedding capacity. This applies in the special case like medical images. With the special properties of medical images, this method can result in 30%-200% capacity improvement.

Wang et al. [10] proposed a novel framework that design 2D reversible data hiding scheme, two prediction methods are used to compute different prediction error for one pixel then forming a planar, channel is defined to represent a slash. Peaks in each channel are selected to embed data. This method has better performance than conventional histogram shifting method and can be further extended into a multi-dimensional framework.

Thodi and Rodriguez [6] proposed a reversible watermarking method which is used prediction error expansion and histogram shifting scheme. This method more exploits the neighborhood pixels, the prediction errors are Laplacian distribution. The histogram shift technique improves the image capacity and distortion, and have the ability to embed more watermark into the zero prediction errors. Resulting in a better performance in capacity than with difference expansion (DE) [1] .

Sachnev et al. [7] proposed a reversible watermarking algorithm using sorting and prediction. The scheme sorts the prediction errors base on their local variance separately. They can embed watermark bits according to local variance in ascending order. Because the local variance is proportional to the magnitude of prediction errors. The location map size is reduced which increasing the image capacity. The performance is better in low payload. Combine with double embedding scheme and prediction using a rhombus pattern, ideal image capacity reaches 1b/pixel. Comparing with the research of Thodi and Rodriguez [6], there are Significant improvements in both capacity and image quality from investigation of Sachnev et al. [7].

Lee et al. [12] proposed a reversible watermarking scheme based on prediction and difference expansion without using a location map. To solve overflow and underflow problems, the scheme shrinkage the histogram of image pixels by narrowing down the pixels value that close to value 0 and 255 and record those pixels position as a n-bit string regarded as payload length also embedding to image. The real image capacity is affected by the size of the n-bit string. The n-bit string don't needed in extraction process, the advantage compared with location map is don't worry about where to hide a n-bit string.

In 2010, FUJIYOSHI [9] proposes a reversible data hiding method, used no image-dependent parameter or any image-dependent location map. Instead, the scheme used a threshold parameter to limit embedding range, maintain a certain degree of image capacity but lack of capacity control. The scheme sorted the watermarked positions base on the maximum absolute deviation- like parameters. The sorting technique increase the watermark embedding capacity slightly. Their performance is better than [8] , the experiment result of our proposed method will compare it in section 4.

III. PROPOSED MEHTOD

The proposed method used prediction values and rhombus pattern prediction scheme from Sachnev et al. method [7], the rhombus pattern scheme divides the original image into two planes: the half plane1 and half plane2 [10], each pixel is surrounded by the pixels of the other half plane. The two half planes are independent, in this way, the embedding capacity can reach 1 bpp(bit per pixel) in a best situation.

See Fig.1. Half plane1 is the brown grids, half plane2 is the white grids.

169	167	164	167	165	165
169	167	164	168	166	166
168	167	165	169	167	166
167	167	166	170	167	166
166	167	165	169	167	166
166	166	164	168	165	164

Fig. 1. The half plane1 and half plane2

A. *Predicted value and prediction error*

Fig. 2. The rhombus prediction pattern

This method used rhombus pattern prediction scheme with the half plane1 and half plane2. Half plane1 is the brown grids, half plane2 is the white grids. Pixel I used to embedding watermarking, for the purpose of reversibility, a predicted

value is computed use four neighboring pixels (i.e., $I_{i,j+1}$, $I_{i,j-1}$, $I_{i+1,j}$ and $I_{i-1,j}$) in white grids, see Fig. 2. Predicted values is not changed after embedding process, at extraction step, the same predicted values are computed.

With center pixel $I_{i,j}$ in half plane1 and four neighboring pixels $I_{i,j+1}$, $I_{i,j-1}$, $I_{i+1,j}$ and $I_{i-1,j}$ in half plane2, the predicted value $p_{i,j}$ is computed as follow:

$$p_{i,j} = \left\lfloor \frac{I_{i,j+1} + I_{i,j-1} + I_{i+1,j} + I_{i-1,j}}{4} \right\rfloor \quad (1)$$

For the pixels $I_{i,j}$ in half plane1, predicted value is computed by four neighboring pixels in half plane2.

According to the predicted value $p_{i,j}$,prediction error $e_{i,j}$ is computed as follow

$$e_{i,j} = I_{i,j} - p_{i,j} \quad (2)$$

In this method, prediction error $e_{i,j}$ expand to embedding message bit b

$$E_{i,j} = 2e_{i,j} + b \quad (3)$$

Where (3) combine with histogram shifting scheme shows in (10)(11).

The watermarked pixel $W_{i,j}$ is computed as follow

$$W_{i,j} = p_{i,j} + E_{i,j} \quad (4)$$

In extraction process, predicted value $p_{i,j}$ and the watermarking pixel value $W_{i,j}$ is used for extraction of embedded bit and recover original pixel value.

$$E_{i,j} = W_{i,j} - p_{i,j} \quad (5)$$

Then, the watermark bits b can be extracted as follow:

$$b = E_{i,j} \bmod 2 \quad (6)$$

The original prediction error is computed as follow:

$$e_{i,j} = \left\lfloor \frac{E_{i,j}}{2} \right\rfloor \quad (7)$$

The original pixel value is computed as follow:

$$I_{i,j} = e_{i,j} + p_{i,j} \quad (8)$$

B. Predicted threshold value

Predicted threshold value T is main idea of proposed method. According to predicted threshold value T and predicted value $p_{i,j}$, the histogram shifting scheme modifies the pixels value to avoid overflow and underflow problems caused by embedding process. When the half plane1 is used to embed data, the half plane2 is used to compute predicted threshold value, equally, when the half plane2 is used to embed data, the half plane1 is used to compute predicted threshold value.

Predicted threshold T in half palne1 (brown grids) is computed as follow:

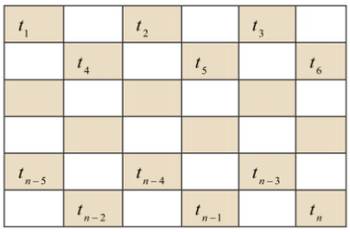


Fig. 3. Pixels of half plane1

First, the rhombus prediction pattern divide image into chessboard-like grids, when the half plane2 is used to embed date. Calculate the average of pixels value from t_1、t_2 ..... t_n where t_1、t_2 ..... t_n are image pixels value in all half plane1, see Fig. 3.

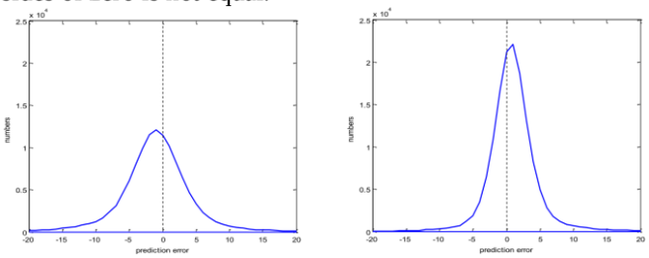
$$T = \frac{t_1 + t_2 + ... + t_n}{n} \quad (9)$$

Where T is predicted threshold value.

The predicted threshold value of half plane1 is computed using (9), where t_1、t_2 ..... t_n are replaced pixels value in the white grids.

C. Predicted threshold value and predicted value

When prediction errors are distributed based on predicted threshold value and predicted value, prediction errors on both sides of zero is not equal.



(a) $p_{i,j} < T$ (b) $p_{i,j} \geq T$

Fig. 4. The histogram of prediction errors based on predicted threshold value

The relationship between $p_{i,j}$ 、 T and are used in this method. See Fig.4, the two pictures are asymmetric. When $p_{i,j} \geq T$, the prediction errors $e_{i,j}$ have more positive values, see Fig.4 (b). When $p_{i,j} < T$, the prediction errors $e_{i,j}$ have more negative values, see Fig.4 (a).

Since T is made to calculate the average using half plane1 or half plane2 in the image. If randomly choose half pixels of the half plane to calculate the average value, this value will very close to T . If randomly choose quarter pixels of the half plane to calculate the average value, this value will be close to T . If only randomly choose four pixels of the half plane, this value may close to T , but not so obvious.

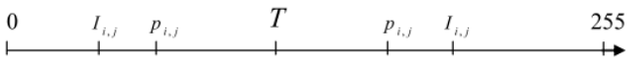


Fig. 5. predicted threshold value and predicted value

For each pixel $I_{i,j}$, the corresponding $p_{i,j}$ is high spatial correlation, because $p_{i,j}$ is the average of $I_{i,j}$ four neighboring pixels, affecting $p_{i,j}$ more closer to T than $I_{i,j}$, see Fig.5. When $p_{i,j} \geq T$, the predicted value $p_{i,j}$ have more probability to smaller than original pixel $I_{i,j}$. Similarly, when $p_{i,j} < T$, the predicted value $p_{i,j}$ have more probability to greater than $I_{i,j}$, mainly, the prediction errors $e_{i,j}$ have more negative value.

TABLE1 PREDICTION ERRORS COUNTS

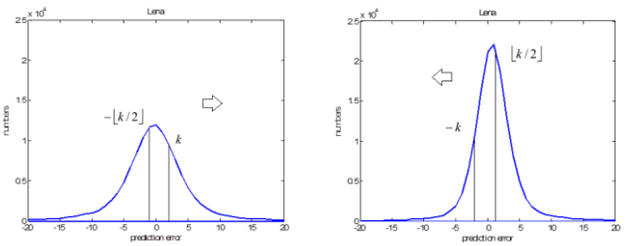
	Lena	Airplane	Cameraman	Barbara
$p_{i,j} \geq T$				
e>0	77738	104880	89302	74846
e<0	43657	37719	17908	46164
$p_{i,j} < T$				
e>0	50514	31849	23339	54917
e<0	57050	41099	37951	60829

Table1 represented four images in Fig.9, their number of prediction error. It results in line with our above described. This feature contributes less distortion when embedding combine histogram shifting scheme.

D. histogram shifting scheme

The histogram shifting scheme combined prediction error is widely used in most methods. The prediction errors close to zero that expansion to embedding watermarking, then histogram shifting scheme modify others prediction errors to avoid overlapping problems. The histogram shifting scheme feature is embedding most data in one side from the zero position of prediction error histogram according to predicted threshold value T .

As shown in Fig.6, the k is threshold value that the range of prediction errors $-\lfloor k/2 \rfloor \leq e_{i,j}$ and $e_{i,j} \leq k$ are used for embedding watermark when predicted value $p_{i,j} < T$, in Fig.6(a). The range of prediction errors $-k \leq e_{i,j}$ and $e_{i,j} \leq \lfloor k/2 \rfloor$ are used for embedding watermark when predicted value $p_{i,j} \geq T$, in Fig.6(b).



(a) $p_{i,j} < T$ (b) $p_{i,j} \geq T$

Fig. 6. The histogram of prediction errors of Lena image,

The reason to choose $\lfloor k/2 \rfloor$ and k for embedding range in histogram of prediction errors are achieve equally embedding capacity around zero both side. In this way, watermarking scheme have large embedding capacity, reduce the amount of movement of all of the pixels
Histogram shifting scheme is presented as follow:

$$E_{i,j} = \begin{cases} 2e_{i,j} + b + \lfloor k/2 \rfloor, & if\ e_{i,j} \in [-\lfloor k/2 \rfloor; k] \\ e_{i,j} + k + \lfloor k/2 \rfloor + 1, & if\ e_{i,j} > k \\ e_{i,j}, & if\ e_{i,j} < -\lfloor k/2 \rfloor \end{cases} \quad (10)$$

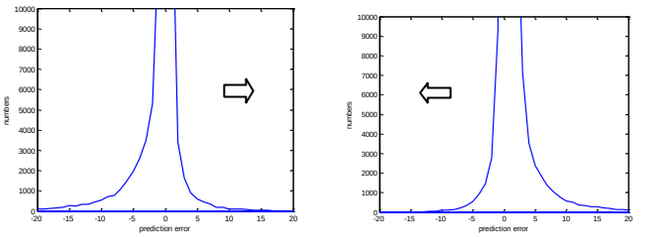
Prediction errors $e_{i,j}$ are shifted using (10) base on predicted threshold value T and prediction value $p_{i,j}$. $p_{i,j} < T$ means prediction errors $e_{i,j}$ are shifted to the right, the prediction errors $e_{i,j}$ belong to the range $[-\lfloor k/2 \rfloor; k]$ are expanded to embedding a message bit b , then add $\lfloor k/2 \rfloor$ to avoid underflow cause by expansion. For the prediction errors $e_{i,j} > k$ are shifted $k + \lfloor k/2 \rfloor + 1$ to avoid overlapping problem cause by expansion, for the prediction errors $e_{i,j} < -\lfloor k/2 \rfloor$ are stay unchanged.

$$E_{i,j} = \begin{cases} 2e_{i,j} + b - \lfloor k/2 \rfloor - 1, & if\ e_{i,j} \in [-k; \lfloor k/2 \rfloor] \\ e_{i,j} - k - \lfloor k/2 \rfloor - 1, & if\ e_{i,j} < -k \\ e_{i,j}, & if\ e_{i,j} > \lfloor k/2 \rfloor \end{cases} \quad (11)$$

Prediction error $e_{i,j}$ are shifted using (11) base on predicted threshold value T and prediction value $p_{i,j}$. $p_{i,j} \geq T$ means prediction error $e_{i,j}$ are shifted to the left.

The prediction errors $e_{i,j}$ belong to the range $[-k : \lfloor k/2 \rfloor]$ are expanded to embedding a message bit b , then add $-\lfloor k/2 \rfloor - 1$ to avoid overflow cause by expansion. For the prediction errors $e_{i,j} < -k$ are shifted $-k - \lfloor k/2 \rfloor - 1$ to avoid overlapping problem cause by expansion, for the prediction errors $e_{i,j} > \lfloor k/2 \rfloor$ are stay unchanged.
In this condition, some images like Lena or Airplane is more smooth for the prediction error $e_{i,j}$ when $p_{i,j} \geq T$, as shown in Fig. 6(b). As the result, most of the data are embedding into higher pixel value.

Because high spatial correlation among predicted value $p_{i,j}$ and original pixel value $I_{i,j}$, embedding watermarking according to $p_{i,j}$ and T can solve overflow and underflow effectively.



(a) $p_{i,j} < T$ (b) $p_{i,j} \geq T$

Fig. 7. The bottom histogram of prediction errors of Cameraman image.

Combination of the above section, see Fig.7, the amount of large value prediction errors in the embedded direction is less, and more prediction errors are on the other side of the zero point, they don't change value through embedding process. As the result, image's quality is improved when those pixels shifting to solve the overlapping problem.

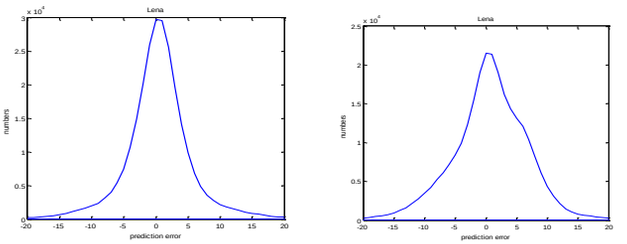
E. Modification of predicted value in half plane2

Firstly proposed method embed watermarking into half plane1, when half plane1 capacity is fully used, the half plane2 continue to embed watermarking. Because the Histogram shifting scheme (10), (11) shifts pixels value $\lfloor k/2 \rfloor$ and $k + \lfloor k/2 \rfloor$, the high spatial correlation among neighboring pixels are influenced by the histogram shift scheme in half plane1. Lead to reduced image capacity and increased image distortion in half plane2. In order weaken the influence. The predicted value $p_{i,j}$ in half plane2 is computed as follow:

1)Compute predicted value $p_{i,j}$ using (1).

$$2) \quad p_{i,j} = \begin{cases} p_{i,j} + \left\lfloor \frac{k}{2} \right\rfloor, if \ \ p_{i,j} \geq T \\ p_{i,j} - \left\lfloor \frac{k}{2} \right\rfloor, if \ \ p_{i,j} < T \end{cases} \qquad (12)$$

See Fig.8, when using equation (12) in half plane2, the prediction error distribution is better.



Compute using (12) Compute using (1)

Fig. 8. k=10, the histogram of prediction errors of Lena image

F. Maximum of threshold value k

To ensure there is no overflow and underflow during imbedding process, limit threshold value k is necessary. We can use (13) in half plane1 and half plane2 separately. The maximum threshold value k is computed as follow:

$$k + \lfloor k/2 \rfloor + 1 \leq \min \begin{cases} 255 - I_{i,j}, if \ \ p_{i,j} < T \\ I_{i,j} - 0, if \ \ p_{i,j} \geq T \end{cases} \qquad (13)$$

Table2 show the experimental results of maximum threshold value k , as long as the value k don't exceed those values for each image, the overflow and underflow problem not occur during embedding process.

TABLE2 EXPERIMENTAL RESULTS OF MAXIMUM THRESHOLD VALUE

Image	maximum threshold k
Airplane	$k < 20$
Lena	$k < 20$
Barbara	$k < 13$
Cameraman	$k < 20$

G. Embedding and Extraction

The proposed method is used prediction error and histogram shift method to embedding data. For extracting watermarking

bits and recovering original image , threshold value k and payload size P_{L1} in half plane1 or payload size P_{L2} in half plane2 should be known in extraction process. They can be transmitted through a covert channel or it may also be included in the image.

The LSB value of the first 47 image pixels value of final raw are replaced with threshold value k (7 bits), payload size P_{L1} (20bits) and payload size P_{L2} (20bits). Original 47 LSB values are collected as payload embedding into image. The final raw of image is not used for embedding data.

In the section, embedding steps and extraction steps illustrate in detailed.

I. Embedding process

In proposed method, prediction errors that are close to zero are used to embed the watermark bits, in order to achieve of no underflow and overflow problems causing by histogram shifting scheme, we adopt an threshold value k to control the embedding range of prediction errors, combined with predicted threshold value T , we use histogram shifting to avoid underflow and overflow problems. Firstly, the half plane1 is used for embed date, then the half plane2.

The embedding scheme is designed as follows:

1) *The original LSB values of the first 47 image pixels value of final raw are collected as payload embedding into image.*
2) *Compute predicted threshold value T using (9) in half plane 2;*
3) *Compute the following value:*

   predicted value $p_{i,j}$ using (1);

   prediction errors $e_{i,j}$ using (2);
4) *The embedding histogram shifting scheme is designed as follow:*

$$E_{i,j}=\begin{cases}2e_{i,j}+b+\lfloor k/2\rfloor, & if\ e_{i,j}\in[-\lfloor k/2\rfloor;k]\ and\ \ p_{i,j}<T\\ e_{i,j}+k+\lfloor k/2\rfloor+1, & if\quad e_{i,j}>k\ \ and\ \ p_{i,j}<T\\ e_{i,j}, & if\quad e_{i,j}<-\lfloor k/2\rfloor\ and\ \ p_{i,j}<T\end{cases} \tag{14}$$

$$E_{i,j}=\begin{cases}2e_{i,j}+b-\lfloor k/2\rfloor-1, & if\ e_{i,j}\in[-k;\lfloor k/2\rfloor]\ and\ \ p_{i,j}\ge T\\ e_{i,j}-k-\lfloor k/2\rfloor-1, & if\quad e_{i,j}<-k\ and\ \ p_{i,j}\ge T\\ e_{i,j}, & if\quad e_{i,j}>\lfloor k/2\rfloor\ and\ \ p_{i,j}\ge T\end{cases} \tag{15}$$

   The watermarked image pixel $W_{i,j}$ is computed using (4)
5) *Repeat step3 to step4 until the final pixel of half plane1;*
6) *For half plane2 repeat step2 to step4 again, but the predicted threshold value T is computed using (9) in half plane 1, predicted value using (12);*
7) *Threshold value k and payload size P_{L1} in half plane1 and payload size P_{L2} in half plane2 are embedded into the LSB value of the first 47 image pixels value of final raw.*

After all message bits are embedded, the watermarked image is received

II. Extracting process

In the extracting process, we need threshold value k , predicted value $p_{i,j}$ predicted threshold value T , payload size P_{L1} in half plane1 and payload size P_{L2} in half plane2. The prediction value $p_{i,j}$ and prediction threshold value T can be computed from watermarked image. Threshold value k and payload length P_{L1} and P_{L2} are extracted from the LSB value of the first 47 image pixels value of final raw

The extraction scheme is designed as follows: start with half plane2

1) *Find threshold value k , payload length P_{L1} and P_{L2};*
2) *Compute predicted threshold value T using (9) in half plane 1;*
3) *Compute the following value:*

   prediction value $p_{i,j}$ using (12) ;

   Expansion of prediction error $E_{i,j}$ using (5);
4) *The watermarked bits can be extracted using (16) and $E_{i,j}$ must move $\lfloor k/2\rfloor$ according to $p_{i,j}$ and T before extracted;*

$$b=\begin{cases}(E_{i,j}-\lfloor k/2\rfloor)\bmod 2, if\ \ E_{i,j}\in[-\lfloor k/2\rfloor;2k+\lfloor k/2\rfloor+1]and\ \ p_{i,j}<T\\ (E_{i,j}+\lfloor k/2\rfloor+1)\bmod 2, if\ \ E_{i,j}\in[-2k-\lfloor k/2\rfloor-1;\lfloor k/2\rfloor]and\ \ p_{i,j}\ge T\end{cases} \tag{16}$$

5) *The extraction histogram shifting scheme is designed as follow:*

$$e_{i,j}=\begin{cases}(E_{i,j}-b-\lfloor k/2\rfloor)/2, & if\ E_{i,j}\in[-\lfloor k/2\rfloor;2k+\lfloor k/2\rfloor+1]and\ \ p_{i,j}<T\\ E_{i,j}-k-\lfloor k/2\rfloor-1, & if\quad E_{i,j}>2k+\lfloor k/2\rfloor+1\ and\ \ p_{i,j}<T\\ E_{i,j}, & if\quad E_{i,j}<-\lfloor k/2\rfloor and\ \ p_{i,j}<T\end{cases} \tag{17}$$

$$e_{i,j}=\begin{cases}(E_{i,j}-b+\lfloor k/2\rfloor+1)/2, & if\ E_{i,j}\in[-2k-\lfloor k/2\rfloor-1;\lfloor k/2\rfloor]and\ \ p_{i,j}\ge T\\ E_{i,j}+k+\lfloor k/2\rfloor+1, & if\quad E_{i,j}<-2k-\lfloor k/2\rfloor-1\ and\ \ p_{i,j}\ge T\\ E_{i,j}, & if\quad E_{i,j}>\lfloor k/2\rfloor and\ \ p_{i,j}\ge T\end{cases} \tag{18}$$

6) *The original image pixel $I_{i,j}$ can be recovered using (8);*
7) *Repeat step3 to step6 until the final pixel of half plane2;*

8) For half plane1 repeat step2 to step6 again, but the predicted threshold value T is computed using (9) in half plane 2, predicted value $p_{i,j}$ using (1);

9) Recover the LSB value of the first 47 image pixels value of final raw;

IV. Experimental results

Four size of 512×512 grayscale images are used for simulation, as shown in Fig. 9. The test watermark is a random binary string. We simulate the algorithm proposed by FUJIYOSHI et al. [9] to compare the performance with our scheme.

Fig. 9. Original images: Airplane, Lena, Barbara, Cameraman.

The definition of the PSNR(peak signal-to-noise ratio) is shown as follow:

$$PSNR = 10 \times \log_{10}\left(\frac{255^2 \times m \times n}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left[I(i,j) - W(i,j)\right]^2}\right) \quad (19)$$

Where m , n are the size of the image.

Table3 and Table4 show the maximum capacity and corresponding PSNR(dB) in different threshold value k for four images in Fig.9. See Table5, the simulation result show that the proposed method is perform better than FUIYOSHI et al[9] in image capacity.

TABLE3 EXPERIMENTAL RESULTS OF DIFFERENT VALUE FOR IMAGES AIRPLANE AND LENA

Threshold k	Airplane		Lena	
	Maximum capacity, bits	PSNR, dB	Maximum capacity, bits	PSNR, dB
0	46377	52.63	33449	51.69
1	70453	48.27	59607	46.86
2	138364	43.34	111925	41.05
3	146688	41.50	125580	40.04
4	184610	38.41	162027	32.82

TABLE4 EXPERIMENTAL RESULTS OF DIFFERENT VALUE FOR IMAGES BARBARA AND CAMERAMAN

Threshold k	Barbara		Cameraman	
	Maximum capacity, bits	PSNR, dB	Maximum capacity, bits	PSNR, dB
0	25571	51.75	93725	52.87
1	46157	46.59	114907	50.22
2	87689	40.75	198843	44.02
3	99106	49.47	201689	43.68
4	128908	36.59	227356	39.98

TABLE5 EXPERIMENT AL RESULTS OF CAPACITY COMPARISION

	FUJIYOSHI[9] Maximum capacity, bits	Proposed threshold k =8 capacity, bits
Airplane	63932	223578
Lena	62366	212230
Barbara	32980	174432
Cameraman	10353	244733

With the same embedding capacity, the PSNR of our scheme is higher than FUJIYOSHI et al[9]. See Fig.10, the dramatic decline caused by the threshold value k , since when k =0, the maximum capacity in Airplane is 46377,see Table3, in order to get more embedding capacity, the threshold value is changed to k =1, the maximum capacity is reach to 70453, this is same as Fig.11, Fig.12, Fig.13.

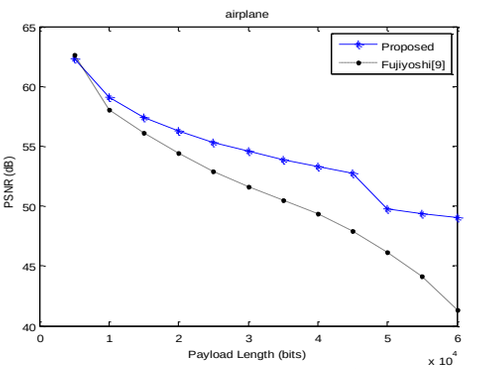


Fig.10. show the result compare to [9] for airplane image

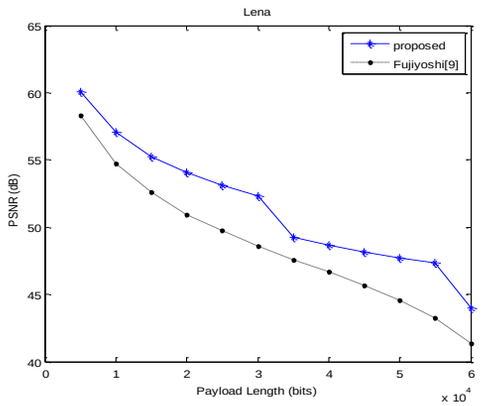


Fig.11. show the result compare to [9] for Lenae image

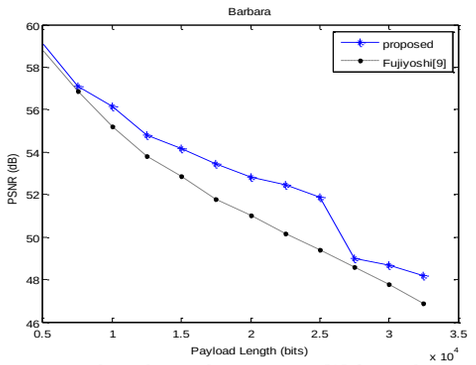


Fig.12. show the result compare to [9] for Barbara image

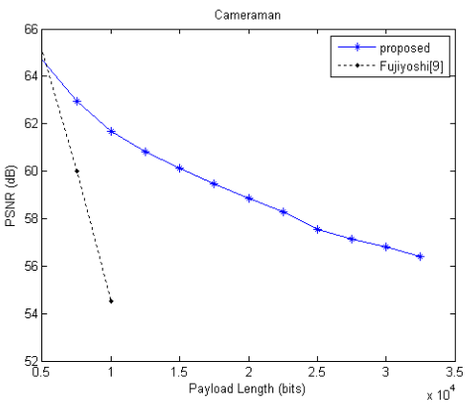


Fig.13. show the result compare to [9] for Cameraman image

V. CONCLUSIONS

In this paper, we have proposed a reversible watermarking method without using a location map to solve overflow and underflow problems during embedding process and extracting process. The concept is to find predicted threshold value and histogram shifting scheme exploit the predicted threshold to embedding data. We find a relationship between predicted threshold value and prediction error. This feature helps us achieve the paper goal of location map free watermarking method with better image quality. Capacity control achieves better image quality with different payload length that other location free method is failed to do. The image capacity and quality are great improved in the location map free scheme.

ACKNOWLEDGEMENT

This paper is the partial result of project NSC100-2632-E-366-001-MY3. We would like to thank the supporting of Ministry of Science and Technology, R.O.C., also thank the value advices of Miss Yu-Hsiu Huang and Prof. Chun-Hung Richard Lin.

References

[1] Jun Tian,"Reversible data embedding using a difference expansion,"IEEE Trans. Circuits Syst. Video Technol., vol.13, pp.890-896, Aug. 2003.

[2] A.M. Alattar, "Reversible watermark using difference expansion of triplets," in Proc. Int. Conf. Image Process., vol. 1. Barcelona, Spain, 2003, pp. 501-504

[3] A.M. Alattar, "Reversible watermark using the difference expansion of a generalized integer transform," IEEE Trans. Image Process., vol. 13, no. 8, pp. 1147-1156, Aug. 2004.

[4] Z. Ni, Y. Q. Shi, N. Ansari, and S. Wei, "Reversible data hiding," IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 3, pp. 354-362,2006.

[5] M. Fallahpour, D. Megias, M. Ghanbari, "Reversible and high-capacity data hiding in medical images" IET Image Process., 2011, Vol.5, lss. 2, pp. 190-197

[6] D. M. Thodi and J. J. Rodriguez, "Expansion embedding techniques for reversible watermarking," IEEE Trans. Image Process., vol. 16, no. 3, pp. 721-730, Mar. 2007.

[7] V. Sachnev, H. J. Kim, J. Nam, S. Suresh, and Y. Q. Shi, "Reversible watermarking algorithm using sorting and prediction," IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 7, pp. 989-999, Jul. 2009.

[8] M. Fujiyoshi, S. Sato, H.L. Jin, and H. Kiya, "A location-map free reversible data hiding method using block-based single parameter," in Proc. IEEE ICIP, 2007, pp.257-260.

[9] M. Fujiyoshi, TSUNETYO, and h. Kiya, "A reversible data hiding method free from location map and parameter memorization," in Proc. IEEE ICIP, pp. 978-1-4244-7010-5/10, 2010.

[10] Shyh-Yih Wang, Chun-Yi Li and Wen-Chung Kuo, "Reversible data hiding based on two-dimensional prediction errors" IET Image Process., 2013, Vol. 7, lss. 9, pp.805-816 doi:10.1049/iet-ipr.2012.0521

[11] R. Naskar, R.S. Chakraborty, "Reversible watermarking utilising weighted median-based predicion" IET Image Process., 2012, Vol. 6, lss. 5, pp. 507-520 doi: 10.1049/iet-ipr.2011.0244

[12] Lee, C.F., Chen, H.L., and Tso H.K., "Embedding capacity raising in reversible data hiding based on prediction of difference expansion" The Journal of Systems and Software 83, pp. 1864-1872,2010.

[13] L. H. J. Kamstra and A.M. Heijmans, "Reversible data embedding into images using wavelet techniques and sorting," IEEE Trans. Image Process., vol. 14, no. 12, pp. 2082-2090, Dec. 2005.

[14] Chang, C.C., Pai, P.Y., Yeh, C.M., Chan, Y.K. ' A high payload frequency-based reversible image hiding method ' , Inf. Sci., 2010, 180 , (11), pp. 2286 – 2298

Facial Expression Recognition Based on Facial Components Detection and HOG Features

Junkai Chen[1], Zenghai Chen[1], Zheru Chi[1], and Hong Fu[1,2]

[1]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
[2]Department of Computer Science, Chu Hai College of Higher Education, Hong Kong
Email: Junkai.Chen@connect.polyu.hk

***Abstract*—In this paper, an effective method is proposed to handle the facial expression recognition problem. The system detects the face and facial components including eyes, brows and mouths. Since facial expressions result from facial muscle movements or deformations, and Histogram of Oriented Gradients (HOG) is very sensitive to the object deformations, we apply the HOG to encode these facial components as features. A linear SVM is then trained to perform the facial expression classification. We evaluate our proposed method on the JAFFE dataset and an extended Cohn-Kanade dataset. The average classification rate on the two datasets reaches 94.3% and 88.7%, respectively. Experimental results demonstrate the competitive classification accuracy of our proposed method.**

Keywords—facial expression recognition, HOG features, facial component detection, SVM

I. INTRODUCTION

Human beings could convey intentions and emotions through some nonverbal ways, such as gestures, facial expressions and involuntary language. Facial expressions may be the most useful nonverbal ways for people to communicate with each other. Facial expressions recognition has gained a growing attention because it could be widely used in many fields such as lie detection, medical assessment, and Human Computer Interface (HCI). In fact, a widely accepted prediction is that computing will move to the background, weaving itself into the fabric of our everyday living spaces and projecting the human user into the foreground [1]. To reach this goal, computer vision and machine learning techniques have to be developed while strengthening psychological analysis of emotion.

However, facial expression recognition is an extremely challenging task. Many factors like illumination, pose, deformation and wild environment could contribute to the complexity. Moreover, facial expressions are subtle facial muscle movements, and it is challenge to detect and represent these kinds of slight changes.

Facial expressions have been studies for a long time and we have witnessed some progress in recent decades. The Facial Action Coding System (FACS), which was proposed in 1978 by Ekman et al. [2] and refined in 2002 [3], is a very popular facial expression analysis tool. FACS attempts to decompose facial expressions into different action units. Based on the combination of the action units, facial expressions could be recognized. Another approach is to recognize facial expressions directly from images.

In a direct approach, two mainstream approaches, called appearance-based and geometry-based [4], are used in facial expression recognition. Appearance-based methods apply the Gabor filters, Local Binary Pattern (LBP) texture descriptors to represent the features of facial expressions. Geometry-based methods focus on capturing the shape of faces. A shape is constituted with a group of fiducial points. These points could be regarded as the geometry features.

Many attempts have been made to recognize facial expressions. Zhang et al. [5] investigated two types of features, the geometry-based features and Gabor-wavelets based features, for facial expression recognition. They applied a two-layer perceptron as the classifier and compared the performance of the two features. Feng et al. [6] provided a coarse-to-fine classification scheme for facial expression recognition. The coarse stage included producing the basic model vectors and computing the distance from the feature vectors to the model vectors. After that, a K-nearest neighbor classifier was employed to do the final classification in the fine stage. In [7], Khandait et al. found that the width and height of the face portions were distinct features in facial expression recognition. Based on the facial elements and muscle movements, Zhang et al. adopted the salient distance features to do the facial expression recognition [8]. They extracted the 3-D Gabor features, selected the "salient" patches and matched the patches to obtain salient distance features. Shan et al. [9] considered that the Local Binary Pattern (LBP) was a good texture descriptor and could be used to represent facial expressions. They adopted a Boosted-LBP to select the most distinguished LBP features. The boosted-LBP features were employed to train the SVM and acquired a prominent recognition rate.

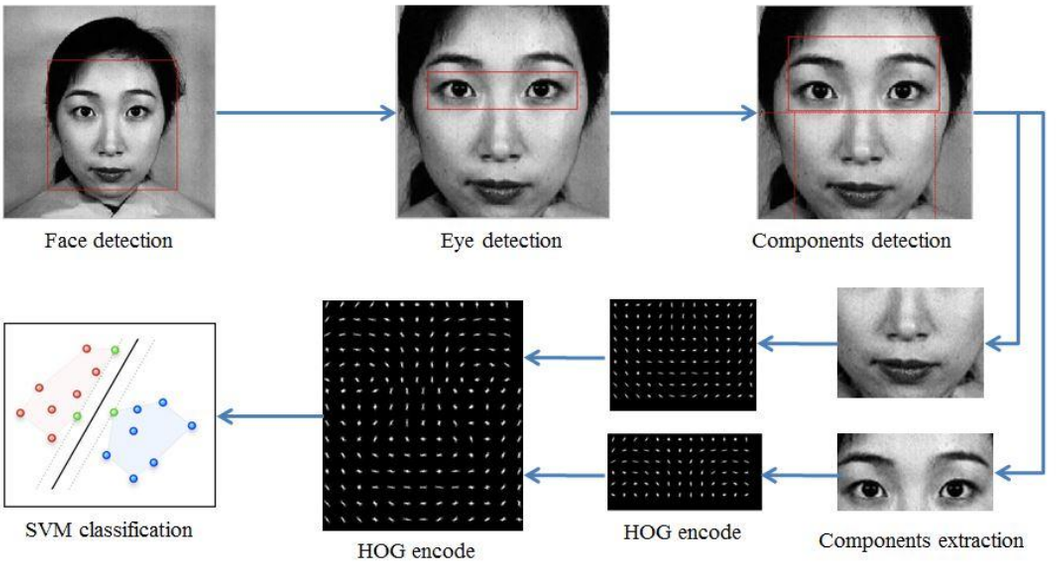


Fig. 1. The schematic overview of the propsoed system.

In this paper, we introduce an effective appearance-based method to handle the facial expression recognition problem. Given a face image, the system detects the face first and then, extracts the facial components from the face image. After that, Histogram Oriented Gradient (HOG) is extracted to encode these facial components and concatenate them into a single feature vector. These feature vectors are used to train a linear SVM.

Our work is somewhat similar with the previous work in [10]. However, there are still some differences between our work and the previous work. The previous work applied the feature descriptors on the whole face, and they explored different features including HOG, LBP and LTP. Our work considered the facial components and employed the HOG feature descriptors on the facial components. The previous work focused on the problem of facial expression recognition with registration errors. Our study paid attention to the facial components which contribute to the facial expression recognition.

The rest of this paper is organized as follows. Section Ⅱ describes our proposed facial expression resonation system, and the details of computing facial components and HOG. Experimental results and analysis are given in section Ⅲ. Concluding remarks are made in Section Ⅳ.

II. PROPOSED FACIAL RECOGNITION SYSTEM

The proposed system includes three function blocks. The first function is face detection and facial components extraction. The second function block is using HOG to encode these components. The last function block is training a SVM classifier. The schematic overview of our proposed facial recognition system is shown in Fig.1.

A. Face Detection and Facial Components Extraction

This part begins with face detection using the Viola-Jones face detector [11]. After the face region is acquired, it is necessary to extract the brows, eyes, nose and mouth from the face. We could detect the eyes first and extract the other components based on the relative positions of these components. The face images of the database we used are all of the frontal view and we know that the brows are above the eyes. We could enlarge the detected eye regions to contain the brows as well. As for the nose and mouth, we know they locate just below the eyes; it is not difficult to locate the region which contains the nose and mouth.

B. Histogram of Oriented Gradients Features

Different features including SIFT [12], Gabor filters [13], Local Binary Patterns (LBP) [14] and HOG (Histogram of Oriented Gradient) [15] have been proposed for facial expression recognition. Facial expressions result from muscle movements and these movements could be regarded as a kind of deformation. For example, the muscle movements of the mouth cause the mouth open or close, and cause brows raiser or lower. These movements are similar to deformations. Considering that HOG features are pretty sensitive to object deformations. In this paper, we propose to use the HOG features to encode facial components. HOG was first proposed by Dalal and Triggs in 2005 [15]. It is well received by computer vision community and widely used in many object detection applications, especially in pedestrian detection. HOG numerates the appearance of gradient orientation in a local patch of an image. The idea

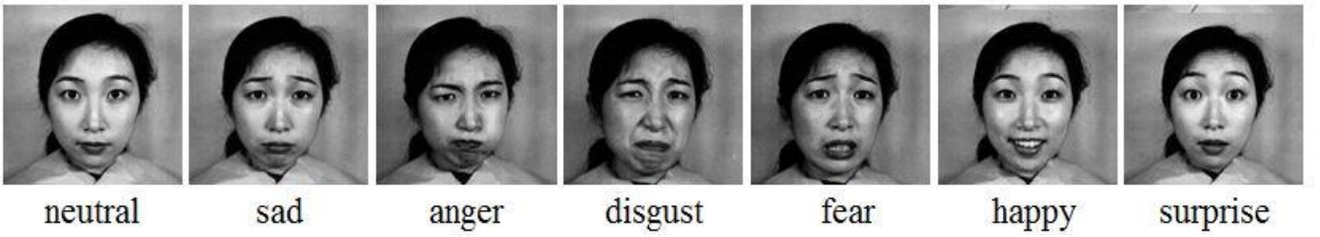


Fig. 2. The seven expressions from one subject.

is that the distribution of the local gradient intensity and orientation could describe the local object appearance and shape [15].

Compared with other features such as LBP and Gabor filters, HOG is also very useful in facial expression recognition. HOG can characterize the shapes of important components constitute facial expressions. So we apply the HOG to encode these facial components. In our experiments, we set cell size to 8×8, the number of bin size to 9, the orientation range to 0 -180.

C. *Support Vector Machine*

Support Vector Machine (SVM) has been widely used in various pattern recognition tasks. It is believed that SVM can achieve a near optimum separation among classes. In our study, we train SVMs to perform facial expression classification using the features we proposed. In general, SVM builds a hyperplane to separate the high-dimensional space. An ideal separation is achieved when the distance between the hyper plane and the training data of any class is the largest. Given a training set of labeled samples:

$$D = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in R^n, y_i \in \{-1,1\}\}_{i=1}^{p} \qquad (1)$$

A SVM tries to find a hyperplane to distinguish the samples with the smallest errors.

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \qquad (2)$$

For a input vector $\mathbf{x_i}$, the classification is achieved by computing the distance from the input vector to the hyperplane. The original SVM is a binary classifier. However, we can take the one-against-rest strategy to perform the multi-class classification. We use the LIBSVM in our experiments [16].

III. Experimental Results and Discussion

In order to evaluate the performance of our proposed approach, we utilize two commonly adopted datasets: The Japanese Female Facial Expression (JAFFE) Database [17] and the Extended Cohn-Kanade Dataset [18].

A. *JAFFE Database*

This database contains 213 images in total. There are 10 subjects and 7 facial expressions for each subject. Each subject has about twenty images and each expression includes two to three images. The seven expressions are angry, happy, disgust, sadness, surprise, fear and neutral respectively. Fig.2 shows the seven expressions from one subject.

In this experiment, images have size of 256×256. After acquiring the face region from the face image, we adjust the size to 156 × 156. And then we apply the techniques mentioned above to detect and extract the facial components and adjust them to the same size. In our experiments, size of the eye-brows is 52×106, the dimensionality of the corresponding HOG encoded feature is 1×2160. Size of the nose-mouth is 78×104, and the dimensionality of the corresponding HOG encoded feature is 1×3456. We concatenate the two feature vectors into a single one. The final feature is a 1 ×5616 vector.

We adopt the leave-one-sample-out strategy to test our method and compare with the other methods. There are 10 subjects in this database. Each subject has a few images. From each group, we randomly select two or three images as the test data set and the remaining images as the training set. At last, there are 23 images in the test set and 190 images constitute the training set. The results are shown in Table Ⅰ.

TABLE I. Classification Results Of Four Methods On The Jaffe Datase.

Method	Classification Rate
Gabor+FSLP [19]	91.0%
LBP [9]	89.1%
Patch-based Gabor [8]	92.3%
Our method	94.3%

In [9], they applied the Local Binary Pattern (LBP) descriptors to represent the facial expression and used the Adaboost to select the optimal features. The average classification rate was about 89.1%. In [19], 18 Gabor filters were convolved with the face images to get the filtered images, and only the amplitudes of selected fiducial points were used as feature vectors. They tested different classifiers and the best performance was about 91%. In [8], Zhang et al. adopted the salient distance features to do the facial expression recognition. They extracted the 3-D Gabor features, selected the "salient" patches, and matched the patches to obtain the salient distance features. The classification rate that they obtained was about 92.3%. From the results, we could find that our method outperforms the other three methods tested.

B. The Extended Cohn-Kanade Dataset

The dataset has 123 subjects and 593 sequences. There are seven expressions and neutral in this dataset. The seven expressions are angry, happy, sad, surprise, contempt, fear, and disgust. Fig.3 shows the 8 expressions with each from a different subject. Among 593 sequences, only 327 sequences have expression labels. We used the peak frame of each labeled sequences as the sample image. The frequency of each expression is shown in Table Ⅱ.

TABLE II. THE FREQUENCY OF EACH EXPRESSION IN THE EXTENDED COHN-KANADE DATASET.

Expression	Frequency
Angry	45
Contempt	18
Disgust	59
Happy	25
Surprise	69
Sad	28
Fear	83

Note that the neural expression is excluded from the experiments. We follow the similar procedure applied in the JAFFE dataset experiments. The original size of the image is 640 ×490. We detect the face first, and adjust the size of face to 256×256. Once we obtained the face region, we could detect the eyes and extract the facial components. The final size of the eye-brow is 74×150 and the nose-mouth 130×128. Down sampling is used for the extracted facial components before applying the HOG to reduce the dimensionality. At last, the HOG encoded features of the eye-brow component are a 1× 864 vector and the HOG encoded features of the nose-mouth component are a 1 × 1764 vector. The final feature is a 1×2628 vector.

In this experiment, we divide the images into two sets. One is the training set and the other is the test set. About one-fifth images of each group are randomly selected for the test set. The remaining images form the training set. At last, there are 59 samples for the test and 268 samples for the training. In order to eliminate the influence of the randomness, we repeated the process 10 times and compute the average classification rate. We achieved an average of 88.7 with a variance of ±2.3% classification rate at last.

In order to compute the classification rate of each expression, we follow the baseline method and adopt the leave-one-subject out strategy. This strategy promises each subject can be evaluated once. There are 118 subjects. Each time, the expression images of one subject are picked out for the test and the images of the other subjects are used for training. We repeat 118 times and compute the average. The results are shown in table Ⅲ. The diagonal values are the hit rates. We could find that the expression "contempt" has the lowest hit rates. This may be this expression is easy to be mixed with the other expressions. The "surprise", "disgust" and "happy" expressions get high hit rates. The three kinds of expressions are more distinct than the other expressions.

We also compare our method with three baseline methods with the results shown in Table Ⅳ. The baseline methods use different features: SPTS, CAPP and SPTS+CAPP, respectively. From Table Ⅳ, we can see that the performance of our method is much better than SPTS and CAPP, especially for the expression "contempt". The hit rate can be improved nearly by 40%. As for the combination of SPTS and CAPP, the hit rate of the expression "contempt" is higher than our method, However, the hit rates of the "anger" and "fear" expressions are lower than our method. Compare with the baseline methods, our method achieves a good performance under a more strict condition. Note that the neutral faces are not used as the reference in our system.

TABLE III. THE CONFUSION MATRIX OF THE EXPRESSIONS.

	AN	**CO**	**DI**	**FE**	**HA**	**SA**	**SU**
AN	0.84	0.04	0.07	0.00	0.02	0.00	0.02
CO	0.06	0.61	0.00	0.11	0.11	0.11	0.00
DI	0.02	0.00	0.95	0.00	0.03	0.00	0.00
FE	0.08	0.04	0.00	0.72	0.12	0.00	0.04
HA	0.01	0.03	0.00	0.00	0.96	0.00	0.00
SA	0.07	0.04	0.00	0.04	0.00	0.82	0.04
SU	0.00	0.01	0.00	0.00	0.00	0.00	0.99

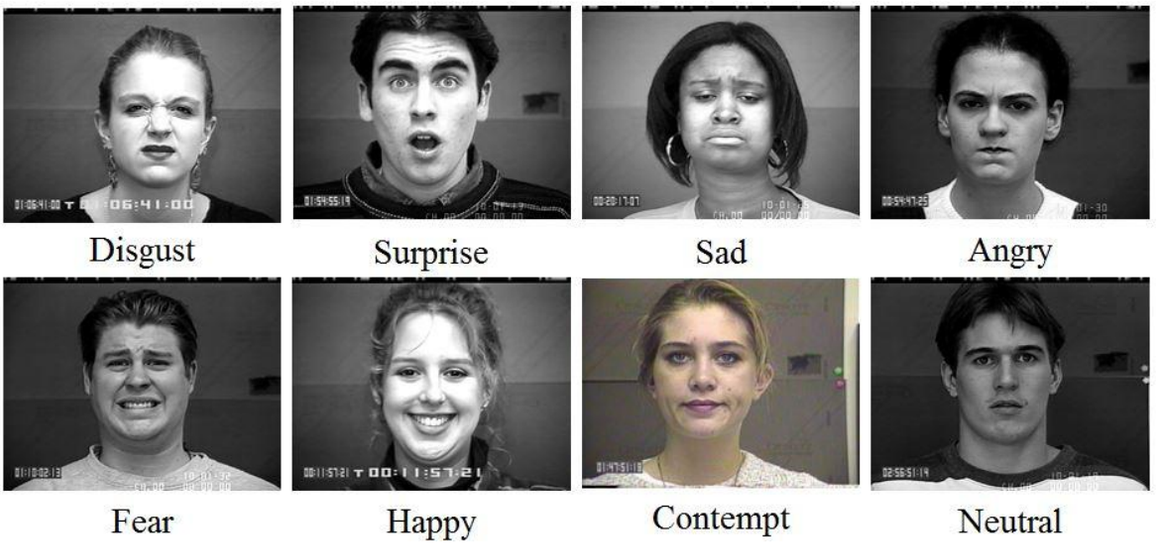


Fig. 3. The eight expressions with different subjects.

TABLE IV. THE CLASSIFICATION RATES OF EACH EXPRESSION WITH DIFFERENT METHODS

	Our method	SPTS [18]	CAPP [18]	SPTS+CAPP [18]
AN	0.84	0.35	0.70	0.75
CO	0.61	0.25	0.22	0.84
DI	0.95	0.68	0.95	0.95
FE	0.72	0.22	0.22	0.65
HA	0.96	0.98	1.0	1.0
SA	0.80	0.28	0.60	0.68
SU	0.99	1.0	0.99	0.96

IV. CONCLUSION

In this paper, we propose an effective method to handle the facial expression recognition problem. Instead of using the whole face, we detect and extract the facial components from the face image. Facial expressions are caused by facial muscle movements and these movements or subtle changes can be described by the HOG features, which are sensitive to the object shapes. The encoded features are used to train a linear SVM. Experiment results on two databases, JAFFE and the extended Cohn-Kanade dataset, show that our proposed method can achieve a good performance. The classification rates of our method on the two datasets are 94.3% and 88.7±2.3%, respectively. Facial expression recognition is a very challenging problem. More efforts should be made to improve the classification performance for important applications. Our future work will focus on improving the performance of the method in the wild environment and on the more subtle expressions such as “contempt”.

ACKNOWLEDGMENT

This work was partially supported by a research grant from The Hong Kong Polytechnic University (Project Code: G-YJ87).

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods Audio, Visual, and Spontaneous Expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 39-58, 2009.

[2] P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Consulting Psychologists Press,* 1978.

[3] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial Action Coding System: The Manual on CD ROM. A Human Face," 2002.

[4] S. Z. Li and A. K. Jain, *Handbook of face recognition*: springer, 2011.

[5] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Automatic Face and Gesture Recognition Proceedings. Third IEEE International Conference on*, 1998, pp. 454-459.

[6] X. Feng, A. Hadid, and M. Pietikäinen, "A coarse-to-fine classification scheme for facial expression recognition," in *Image Analysis and Recognition*, ed: Springer, 2004, pp. 668-675.

[7] S. Khandait, R. C. Thool, and P. Khandait, "Automatic facial feature extraction and expression recognition based on neural network," *International Journal of Advanced Computer Science and Applications,* vol. 2, pp. 113-118, 2012.

[8] L. Zhang and D. Tjondronegoro, "Facial expression recognition using facial movement features," *Affective Computing, IEEE Transactions on,* vol. 2, pp. 219-229, 2011.

[9] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing,* vol. 27, pp. 803-816, 2009.

[10] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, 2008, pp. 1-8.

[11] P. Viola and M. Jones, "Robust Real-Time Face Detection," *International journal of computer vision,* vol. 57, pp. 137-154, 2004.

[12] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International journal of computer vision,* vol. 60, pp. 91-110, 2004.

[13] H. G. Feichtinger and T. Strohmer, *Gabor analysis and algorithms: Theory and applications*: Springer, 1998.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 971-987, 2002.

[15] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Computer Vision and Pattern Recognition, 2005. IEEE Conference on*, 2005, pp. 886-893.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, 2011.

[17] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," in *Automatic Face and Gesture Recognition, Proceedings. Third IEEE International Conference on*, 1998, pp. 200-205.

[18] 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," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94-101.

[19] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: face expression recognition," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 35, pp. 477-488, 2005.

Simulation for the Motion of Nano-Robots in Human Blood Stream Environment

S. Y. Ahmed
Scientific Computing Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
syserry@hotmail.com

S. Amin
Scientific Computing Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
ahmed_andeel76@hotmail.com

T. El-Arif
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
taha_elarif@hotmail.com

***Abstract*—this paper is concerned with the problem of demonstrating the motion of a swarm of nano-robots until reaching the unspecified target areas in the human blood stream environment in which the red and white cells worked as obstacles for the nano-robots. Thus, each nano-robot must make predication on the movement of each obstacle as it moves in unknown dynamic environment. Additionally, we have to study the effects of the fluid flow of the blood on the motion of each nano-robot. This problem is solved by combining the modified obstacle avoidance and improved PSO algorithms. In which, the collision avoidance of the blood cells is achieved by modified obstacle avoidance algorithm. Furthermore, the communication, coordination, navigation, and definition of the target areas problems for the nano-robots are carried out in an organized way by applying improved Particle Swarm Optimization (PSO) algorithm to enable nano-robots to make decisions. This research includes a simulation of the performance and behavior of nano-robots. This simulation was designed, implemented and evaluated against the original requirement considerations. The results demonstrating that swarm nano-robots navigation system can be effectively simulated by utilizing the proposed algorithms.**

Keywords—nano-robot; swarm intelligence; blood vessel; Reylonds Number; collision time; coverage;

I. INTRODUCTION

Nanomedicine has developed into one of the most capable sides of nanotechnology. [1] Adapted health care and drug design, in addition to targeted delivery of healing agents are just some of the advantages that nanomedicine may supply. [2] As such, the use of nano-robots in biomedical applications is presently a main focus of several research groups. Nano-robotics is a promising area of scientific and technological prospect. It is a new and rapidly growing interdisciplinary field dealing with the assembly, construction and use of molecular devices based on nanoscale principles and/or dimensions. The size-related challenge is the ability to measure, manipulate, and assemble matter with features on the scale of 1–100 nm.

A nano-robot is a controllable machine at the nanometer or molecular scale that is composed of nano-scale components and algorithmically responds to input forces and information. Bionano-robotics, namely biomolecular robots, represents a specific class of nano-robots where proteins and DNA could be used as motors, mechanical joints, transmission elements, or sensors. If all of these different components were assembled together they can form bio-nano-robots with multi-degree-of-freedom, able to apply forces and manipulate objects in the nanoscale world.

These bio-components seem to be a very logical choice for designing nano-robots. In addition since some of the core applications of nano-robots are in the medical field, using bio-components for these applications seems to be a good choice as they both offer efficiency and variety of functionality [3]. This idea is clearly inspired by nature's construction of complex organisms which are capable of movement, sensing and organized control.

The design of nano-robotic systems requires the use of information from an enormous variety of sciences ranging from quantum molecular dynamics to kinematic analysis. Until now, there does not exist any particular guideline or a prescribed manner that details the methodology of designing a bio-nano-robot [4]. As research development is at the interface of physical sciences and biology it requires multi-skilled teams. To achieve this it is essential for the future molecular roboticians to be able to design and prototype the bio-nanomechanics, to develop dynamic and kinematic models, to study their dynamic performances and to optimize their structure.

Although many of the described technologies have been developed into more or less mature products for robots acting in the macro-world, the nano-size of the molecular robots posestreme challenges and requires a complete rethinking of the design and prototyping methodologies.

Nano-robots are visualized to be tremendously connected to medicine, by being coded to perform particular biological mission. The principle aspects such nano-robots must have incorporate [5]: Biocompatibility: The immune system works as an obstacle for any foreign body presented in the living being. So nano-robots must be programmed to avoid it;

Communication: Nano-robots would need to convey with one another and with their external environment; Navigation: To attain focused on activities the nano-robots should incorporate some type of path planner; Coordination: The nano-robots must have the capacity to arrange their activities in a decentralized way.

In this paper, to enable the nano-robots to avoid such immune system we modify the obstacle avoidance algorithm which is presented in [6]. Furthermore, to enable the nano-robots to communicate, navigate, and coordinate cooperatively, the swarm intelligence algorithm called particle swarm optimization (PSO) is improved and utilized.

The paper is organized as follows. Section II provides a brief theoretical background on swarm intelligence algorithms also; swarm nano-robots and PSO algorithm are discussed. Section III presents the modified obstacle avoidance algorithm in details. A description of the complete algorithm that will combine both proposed algorithms follows in Section IV. In Section V the simulation results and subsequent analysis is introduced. Finally, Section VI discusses some brief conclusions and a short discussion for further work.

II. SWARM INTELLIGENCE

The term swarm intelligence got more being used throughout the most recent 15 years. “Swarm” is naturally valid to huge random systems that perform something motivating. The idea of swarm seems more to be nearly connected with frameworks prepared to do completing helpful events as well as "intelligent" tasks. From the robotic side, swarms self-organized into particles. The creation of organized particles is a feature of intelligence. Another is the recognition and analysis of particles, which swarms do when they optimize function. So, when swarms doing something intelligent this will lead to “Swarm Intelligence”. [7]

Another definition of “Swarm intelligence” in which it handles the developing of the collaborative behaviors of simple agents cooperates with themselves, and their environment. These models are motivated by the collective behavior of insects and other animals [9]. Swarm intelligence handles the developing of the collaborative behaviors of simple agents cooperate with themselves, and their environment [10]. These models are motivated by the collective behavior of insects and other animals [9]. From the algorithmic view point, swarm intelligence models are computing algorithms that are effective for carrying out distributed optimization problems.

The principle of swarm intelligence focuses on probabilistic-based search algorithms. All swarm intelligence models exhibit a number of general properties [11]. There are three commonly used swarm intelligence models which are ant colony optimization, particle swarm optimization, and bee colony optimization.

A. Swarm Nano-robots

Swarm robotics has added a set of “standard” problems. One set of problems is based on pattern arrangement: aggregation, self-organization into a network, deployment of distributed arrays of sensors, covering of areas, mapping of the environment.

A second set of problems focuses on some specific unit in the environment: goal searching, finding the source of a chemical plume, foraging, etc. And another set of problems handle more complex group behavior: collective transport, grouping, etc. other specific robotic tasks, such as obstacle avoidance and all environment navigation, apply to swarms in addition. [8]

A brief review of these problems, as well as their relation to swarming in general, is given in the paper by Passino [12] whose work has focused on swarm robotics control.

For nano-robots, it is possible that each of which encompasses of incomplete capabilities because of small size. This analysis proved that swarm intelligence from social interactions among agents could deal with the limited capabilities that would be inevitable in future nano robots.

The swarm nano-robots system must be designed so that nanoparticles in the swarm cooperate and together organize themselves into structures. Additionally, they are able to repair the structure when damaged as long as they are in the environment. In this research our system aims to demonstrate that a swarm system of nano-robots with some essential characteristics can be communicated, coordinated, and navigated in cooperative way using swarm intelligence algorithm PSO.

B. Particle Swarm Optimization (PSO)

PSO is a swarm intelligence technique derived from the flocking behavior of birds [13]. In PSO, each particle searches through the problem space, improves its knowledge, modifies its velocity based on the information that it gathers, and updates its position.

The position and velocity of each particle are randomly produced. Afterward, using the fitness evaluation function, compare the *pbest* (best found position for a particle) fitness of each particle to its current fitness. If the current fitness is better, so update the *pbest* fitness to the current fitness, and assign the current position to replace the *pbest* position. Likewise, compare the fitness of each particle with the *gbest* (best found position for all particles) fitness. If the recent fitness is better than *gbest*, replace *gbest* with the current position, and assign the current fitness as the *gbest*’s fitness. The associated position and velocity for each particle are updated based on equation (1), and equation (2) respectively.

$$\mathrm{Vel_m} = w * vel_{m-1} + c_1 * rand\ () * (pbest - Pos_{m-1}) +$$

$$c_2 * rand\ () * (gbest - Pos_{m-1}) \quad (1)$$

$$Pos = Pos_{m-1} + vel_m \quad (2)$$

Where vel and Pos refer to the velocity and position of the particle in the search space, m is the iteration number, w is the inertia weight, c1 is contant factor called cognitive parameter and c2 is constant factor called social parameter, and rand() is a randomly generated value between 0 and 1. Table I, presents the pseudo-code of the PSO algorithm.

TABLE I. PARTICLE SWARM OPTIMIZATION ALGORITHM PSEUDO-CODE

Begin
(1) Iteration number = 0
(2) Initialize a population of particles = n
(3) **While** (termination condition is not reached) **do**
(4) **For** each particle
(5) Calculate fitness value;
(6) **If** the fitness value is better than the *pbest*
(7) Set current value as the new *pbest*
(8) **End If**
(9) **End For**
(10) Choose the particle with the best fitness value of all the particles as the *gbest*
(11) **For** each particle
(12) Calculate particle velocity according equation (1)
(13) Update particle position according equation (2)
(14) **End For**
(15) **End While**
End

C. Optimizing swarm nano-robots using PSO

The PSO is improved to enable the nano-robots to communicate, coordinate, and navigate to reach the target area in the human blood stream environment. So, each particle becomes a nano-robot. Moreover, the inertial force (f), viscosity (η) of the blood for a nano-robot is influence the new positions. As a result, the Reynolds number R (the ratio of inertial to viscous forces for a nano-robot moving in a blood) which is measured by equation (3) must be added to equation (1) to adjust the velocity of the nano-robot, see equation (4).

$$R = \delta f / \eta \quad (3)$$

(Where δ is the blood density)

$$Vel_m = R + w * vel_{m-1} + c_1 * rand\ () * (pbest - Pos_{m-1}) +$$

$$c_2 * rand\ () * (gbest - Pos_{m-1}) \quad (4)$$

In PSO, the path is relied on the fitness value. For swarm nano-robots, the fitness value indicates how well each nano-robot is directed to the target area. So, it will be evaluated for each nano-robot over its neighborhoods based on the coverage of the target area.

III. OBSTACLE AVOIDANCE ALGORITHM

Every nano-robot inside the human body environment will encounter red and white cells as obstacles during moving within blood stream. Therefore, nano-robot must apply algorithm for avoiding the colliding of such obstacles. We modify the obstacle avoidance algorithm which is presented in [6].

Assume that the obstacles are represented as circles so the configuration of each obstacle is described by its center, the radium and the velocity. Note that obstacles and nano-robots have the same fluid velocity.

When nano-robot moves on a new location, and verifies that there are moving obstacle in the target area. Then, the modified obstacle avoidance algorithm is started which is shown in Table II.

Based on the distance between the nano-robot and the obstacle, nano-robot is headed to the collision free direction θ , and using the value of θ we can calculate the new velocity v consequently, calculate the collision free position p based on the value of v .

TABLE II. MODIFIED OBSTACLE AVOIDANCE ALGORITHM PSEUDO-CODE

Begin
(1) Calculate d the distance between nano-robot and encountered obstacle

(2) Calculate θ the angle between the centroid of nano-robot and the centroid of obstacle

(3) **If** the distance d less than the threshold value

(4) Nano-robot steps on the opposite direction of obstacle (counter clock wise) and θ is calculated by:

$$(\theta = \theta + 180^\circ)$$

(5) **If** the distance d greater than threshold value and the target area in the positive y-axis direction

(6) Nano-robot steps on the perpendicular direction on the obstacle and θ is calculated by:

$$(\theta = \theta + 90^\circ)$$

(7) **If** the distance d greater than threshold value and the target area in the negative y-axis direction

(8) Nano-robot steps on the perpendicular direction of the obstacle and θ is calculated by:

$$(\theta = \theta - 90^\circ)$$

(9) Else

(10) Randomly calculate θ by adding or subtracting 90°

(11) Calculate new velocity for nano-robot using the value of θ where

$$v\ x = \cos(\theta) \text{ and } v\ y = \sin(\theta)$$

(12) Calculate the new position of nano-robot using the value of the velocity where pos = pos + v;

End

IV. CONTROL ALGORITHMS FOR SWARM NANO-ROBOTS

In this section the nano-robot must use control algorithms to enable them to move around the human blood stream environment without colliding any obstacles while travelling to find the target areas, and in an efficient way.

The control algorithms mainly are the combination of the improved PSO algorithm and the modified obstacle avoidance algorithm which are discussed in details in the previous sections.

The control algorithm will proceed as follows, in the case of a nano-robot not detect an obstacle, it moves according to the improved PSO algorithm. Where while a nano-robot moves randomly, it can find the target area, and then it shares the best value with its neighbor (so it will move to the best location based on this best value). Otherwise, if an obstacle is encountered, a nano-robot follows the obstacle avoidance algorithm. After the nano-robot avoids the obstacle, a nano-robot moves again according to the PSO algorithm. Table III shows how to combine improved PSO algorithm, with obstacle avoidance algorithm.

TABLE III. CONTRO ALGORITHMS FOR NANO-ROBOT PSEUDO-CODE

Begin
(1) **For** each nano-robot s_i
(2) **If** an Obstacle is encountered
(3) Run ***modified obstacle avoidance algorithm***
(4) **Else If** nano-robot find target
(5) Update gbest value
(6) **End if**
(7) **For** nano-robot s_i and its neighbor nano-robot s_j
(8) Update gbest of nano-robot s_i and nano-robot s_j with max value
(9) **End for**
(10) **If** time < T_{free}
(11) Move randomly
(12) **Else**
(13) Calculate new velocity v using gbest position
(14) Calculate new position using v
(15) **End if**
(16) **End for**
End

V. SIMULATION RESULTS AND ANALYSIS

This section incorporates the simulation results of the behavior of swarm nano-robots moving in the blood stream environment until finding the target area. The environment will include red and white cells as obstacles that must be avoided by the nano-robots.

The movement plan composes of swap short movements with random changes in direction. Each time obstacles are recognized, the nano-robots move to the new obstacle free positions based on the obstacle avoidance algorithm.

After obstacle is avoided, each nano-robot will change its position based on the PSO algorithm so that the new position information is transmitted to the others. To investigate the efficiency of control algorithms, several simulation parameters shown in table IV are used.

TABLE IV. SIMULATION PARAMETERS

Parameter	Value
Radius of Nano-robot	3μm
Radius of Red Cell	7μm
Radius of White Cell	12 μm
T_{free}	20s
Fluid Velocity	100 μm /s
Density	1 g/cm3
Viscosity	10^{-2} g/cm.s

A. *Interface validation*

The simulator was implemented in a C programming environment. The validation of the interface was done in many steps with tests of increasing complexity in different environments. The interface composes of main functions for plotting the workspace, initiating the movements of the obstacles and the nano-robots, invoking the control algorithms to allow the swarm nano-robots to reach the target areas without colliding any obstacle, controlling and showing the velocity of each nano-robot, displaying the number of nano-robots that arrived to the target area at a specific time, and demonstrating the time spend by each nano-robot until reach the target area.

B. *Simulation Experiments*

Some results of the tests with the simulator are presented in Fig. 1. The blue, red, white, and orange circles are corresponding to nano-robots, red cells, white cells, and target areas respectively. Fig. 1 describes the scenario of obstacle avoidance of 25 nano-robots moving according to the inertia and viscosity forces of the blood. Efficiently, all the nano-robots arrived to the target areas after 44 seconds. During motion, if obstacles are detected, the nano-robot can rotate its heading rapidly to avoid the collision by invoking the obstacle avoidance algorithm. After passing the obstacle, the nano-robots locate their positions based on the coverage of the target environment which is based on the PSO algorithm. In this scenario, we can evidently observe that whenever the obstacles are detected, the new obstacle free positions can be located efficiently, and all the nano-robots successes to arrive to the target regions effectively.

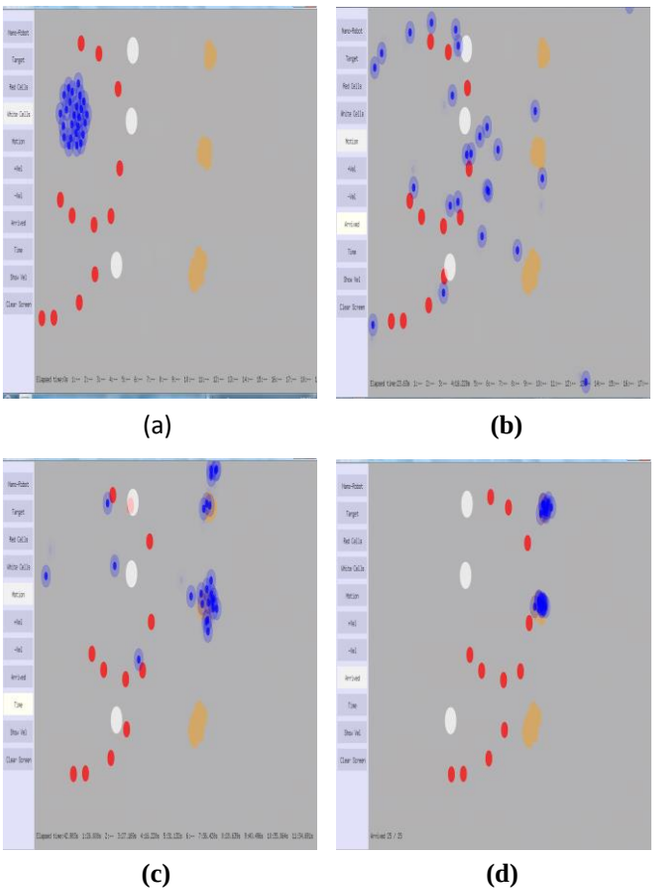


Fig. 1. Movement control for 25 nano-robots navigating in human blood stream environment

VI. CONCLUSIONS

It is believed that nano-robots can greatly contribute to the evolution of modern medical approaches and practices. In this research, the modified obstacle avoidance algorithm and PSO swarm intelligent algorithm are combined efficiently to allow the nano-robots to investigate within human body until reach the target regions. The simulation results have supported that the proposed representation effectively constructs an obstacle free self-organized path. Our future work would include improvement on the design of the simulator. This could include simulating the drug delivery to the unhealthy cells by set of collaborating nano-robots in 3D virtual reality environment at the nanoscale level.

References

[1] J. Panyam and V. Labhasetwar, "Biodegradable Nanoparticles for Drug and Gene Delivery to Cells and Tissue". Advanced Drug Delivery Reviews, 2003, 55: p. 329-347.

[2] Aristides A. G. Requicha, "Nanorobots, NEMS, and nanoassembly," Proceedings of the IEEE, 2003.

[3] Insoluble Photosensitizing Anticancer Drugs: A Novel Drug-Carrier System for Photodynamic Therapy". Journal o f the American Chemical Society, 2003, 125: p. 7860-7865 .

[4] Adriano Cavalcanti, Bijan Shirinzadeh, Robert A Freitas Jr., and Tad Hogg, "Nanorobot architecture for medical target identification," Nanotechnology, 2008.

[5] M. Foldvari, M. Bagonluri, "Carbon nanotubes as functional excipients for nanomedicines: II. Drug delivery and biocompatibility issues", Nanomed.: Nanotechnol. Biol. Med. 4(3),2008, 183–200

[6] Khin Haymar Saw Hla, YoungSik Choi, Jong Sou Park, Obstacle Avoidance Algorithm for Collective Movement in Nanorobots, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[7] G. Al-Hudhud, "On Swarming Medical Nanorobots," International Journal of Bio-Science and Bio-Technology, vol. 4, no. 1, 2012, pp. 75–90.

[8] T. Nantapat, B. Kaewkamnerdpong, T. Achalakul, and B. Sirinaovakul, "Best-so-far ABC based nanorobot swarm," in Proceedings of the Third International Conference on Intelligent Human-Machine Systems and Cybernetics - Volume 01, ser. IHMSC '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 226–229.

[9] Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kaufmann Publisher, San Francisco ,2001.

[10] Venayagamoorthy, G.K., Harley, R.G.: Swarm Intelligence for Transmission System Control. In: IEEE Power Engineering Society General Meeting, 2007, pp. 1–4.

[11] Bai, H., Zhao, B.: A Survey on Application of Swarm Intelligence Computation to Electric Power System. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, vol. 2,2006, pp. 7587–7591.

[12] Beni, G., Liang, P.: Pattern Reconfiguration in Swarms‐Convergence of a Distributed Asynchronous and Bounded Iterative Algorithm. IEEE Trans. Robotics and Autom.,1996, pp. 485‐490.

[13] Baykasoglu, A., Özbakır, L., Tapkan, P.: Artificial Bee Colony Algorithm and Its Application to Generalized Assignment Problem. Intelligence. In: Chan, F.T.S., Tiwari, M.K. (eds.) Swarm Intelligence: Focus on Ant and Particle Swarm Optimization, 2007, pp. 532–564.

Probabilistic Models for 2D Active Shape Recognition using Fourier Descriptors and Mutual Information

Natasha Govender
MIAS (CSIR)
South Africa
ngovender@csir.co.za

Jonathan Warrell
MIAS (CSIR)
South Africa
jwarrell@csir.co.za

Philip Torr
University of Oxford
United Kingdom
philip.torr@eng.ox.ac.uk

Fred Nicolls
UCT
South Africa
fred.nicolls@uct.ac.za

I. ABSTRACT

Shape recognition is essential for robots to perform tasks in both human and industrial environments. Many algorithms have been developed for shape recognition with varying results. However, few of the proposed methods actively look for additional information to improve the initial shape recognition results. We propose an initial system which performs shape recognition using the euclidean distances of Fourier descriptors. To improve upon these results we build multinomial and Gaussian probabilistic models using the extracted Fourier descriptors and show how actively looking for cues using mutual information can improve the overall results. These probabilistic models achieves excellent results while significantly improving on the initial system.

II. INTRODUCTION

The use of robots in industrial and household environments is steadily on the increase. A huge part of robots functioning in these environments is recognising objects. Textural and feature-based approaches are often not appropriate for these types of applications because parts may contain little or no distinctive features other than boundary shape. Environments may not have consistent lighting conditions which can adversely affect these approaches. We use Fourier descriptors to extract boundary information to perform shape recognition. Polar co-ordinates are selected as our shape signature for the descriptors.

Two shape recognition systems are proposed in this paper. The first system uses the euclidean distance between the descriptors to determine the class of each shape. We chose the toy problem of a child's shape puzzle because the shapes were relatively arbitrary and certain shapes were also similar. This was selected to determine the robustness of the system. The second system aims to improve on the shape recognition system that uses just euclidean distance. For this system close-up images of the shapes were captured as input to the system. Here we propose using the Fourier descriptors extracted in a probabilistic manner. Multinomial and Gaussian distributions are built using the Fourier descriptors. We then include an active vision component in the form of mutual information. When the system is determining the correct object sequence, mutual information provides the system with the ability to select the position in the sequence which it is most unsure about.

In our experiments we show that using the probabilistic models with mutual information outperforms both using just the Fourier descriptors as well as the probabilistic models without mutual information.

The structure of the paper is as follows: The next section describes related work. Section IV elaborates on the problem and Section V discusses the Fourier descriptors. Section VI describes the polar co-ordinates and the results for the first system. A complete description of the probability models is presented in Sections VII. Section VIII presents further experimental results and discussion. The conclusion is discussed in Sections IX.

III. RELATED WORK

Various shape representation methods, or shape descriptors, exist in the literature. These methods can be classified into two categories: region based versus contour based. In region based techniques, all the pixels within a shape are taken into account to obtain the shape representation [19],[18]. Contour based shape representation exploits shape boundary information.

Fourier descriptors are contour based and capture global shape features in the first few low frequency terms, while higher frequency terms capture finer features of the shape. Wavelet descriptors can also be used to model shape and have an advantage over Fourier descriptors in that they maintain the ability to localise a specific artifact in the frequency and spatial domains [20]. However, wavelet descriptors are impractical for higher dimensional feature matching [22]. The Fourier Descriptor method can also be easily normalized.

Fourier descriptors are a widely used, all purpose shape description and recognition technique. They have been used in a variety of fields over the years, including commerce, medical, space exploration, and technical sectors. In the field of computer vision, Fourier descriptors have been used for human silhouette recognition [12] for surveillance systems, content based image retrieval [24],[13], shape analysis [23],[14], character recognition [15],[16] and shape classification [11]. In these methods, different shape signatures have been exploited to obtain the Fourier descriptors. These include central distance, complex coordinates, curvature function, and

Fig. 1. The board with the shapes removed

Fig. 2. The shapes that need to be recognised

cumulative angles [7]. Most systems use complex coordinates to model the shape boundary [21],[12] but we use polar coordinates because in our experiments this method produces more accurate results.

There have been a number of probabilistic based models for shape recognition proposed such as using Procrustean models[10], probability density functions [1], geometric features [5] and generative models [3] to name a few. None of these methods use Fourier descriptors as their input parameter to the shape recognition system. In addition none of these methods use active vision by incorporating mutual information to improve their initial results. Mutual information was introduced by [9] as a viewpoint selection mechanism for active vision, which has been subsequently used/proposed by [6][2][17]. As noted, this can be expensive to calculate, and requires the collection of extensive statistics at training time, although as [9] discussed, it provides the optimal strategy provided the underlying models are correct. Using mutual information also makes it easy to incorporate probabilistic assumptions to assist with active information selection. Our framework follows that of [9] and [4] in terms of the general Bayesian form of our updates and we use a sampling scheme to make the mutual information calculations tractable.

IV. The problem

A board containing cut out cartoons of different animals was used in the experiments. The shapes were removed from the board and placed on table. Figure 1 and Figure 2 display the board and the shapes used in the experiments. For each shape 20 close-up images were also captured. Information from these images are used in the probabilistic models.

V. Fourier Descriptors

These images are initially converted to grayscale and then into binary images. The method presented in [8] is used to detect and label the various objects boundaries. Each shape is then segmented from the image and stored. The same procedure is followed for all captured images.

The set of (x,y) boundary coordinates for each shape is converted to polar coordinates. The Fast Fourier Transform (FFT) is then taken for each set of values. The formula used is described in equation 1. Rotation and changes in the starting point only affect the phase of the descriptor. All the phase information can be removed by taking the absolute values of the descriptor elements. It has been shown that the low frequency components of the Fourier Transform are sufficient for shape recognition[12][24] and thus the entire transform does not need to be used. We found that using the first 15 Fourier co-efficients (excluding the very first component F(0)) provided sufficient discriminatory information to model a shape. F(0) is the lowest frequency term and is the only component in the Fourier descriptor that is dependent on the actual location of the shape. By ignoring the first component, it becomes translation invariant. F(0) tells us nothing about the shape; only mean position. The Fourier Descriptor is then normalized to remove any scaling effects. The FFT of the shape is described as:

$$\mathcal{F}(i) = \mathrm{FFT}\{\mathbf{r}\}_i, \tag{1}$$

where

$$r(s) = \sqrt{(x(s) - x_c)^2 + (y(s) - y_c)^2}, \tag{2}$$

x_c, y_c are the shape centroids, x(s), y(s) are the boundary coordinates of the s'th point, $\mathbf{r}$ is the vector of radii, and $\mathrm{FFT}\{.\}$ denotes the discrete fast fourier transform.

VI. Recognition using Polar co-ordinates

Each shape boundary is then matched to the shapes extracted from the board using euclidean distance. Since the energy in the Fourier components decreases sequentially, we artificially boost the contribution of each component. We have found that as the number of components increases, increasing the value used to boost the components works best when calculating the euclidean distance. The shape on the board

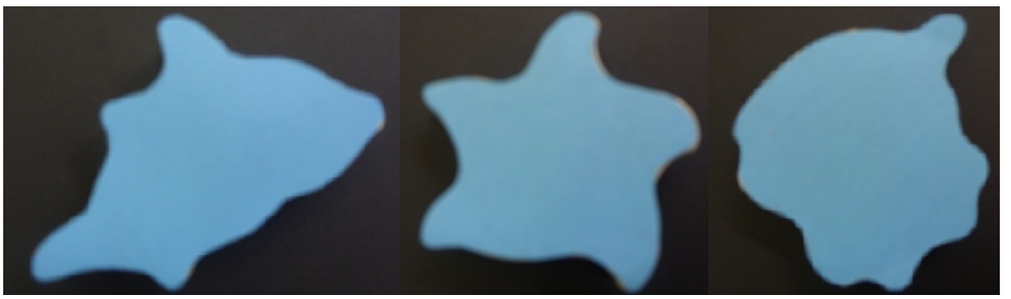


Fig. 3. Close-up images of the shapes

TABLE I
RECOGNITION RESULTS

Shape	Complex Coordinate Method	Polar Coordinate Method
Whale	yes	yes
Seal	no	yes
Fish	no	no
Crab	no	yes
Dolphin	yes	yes
Mussel	no	yes
Snail	no	yes
Octopus	no	yes
Star Fish	yes	yes
Tortoise	no	no

with the smallest euclidean distance to the shape on the table is considered to be the match.

A. Results

The shapes used in the experiments are in the form of animals which include a tortoise, whale, seal, dolphin, fish, crab and so on as seen in Figures 1-3. There are ten shapes in total. Many shape recognition systems use complex coordinates to model the shape boundary but we opted to use polar coordinates because in our experiments this method produces more accurate results. Table I shows the results obtained from both methods.

The complex coordinate system recognises four shapes correctly while the polar coordinate system correctly recognises eight out of the ten shapes. It incorrectly identifies the fish and the tortoise shapes. Figure 5 displays the first fifteen dimensions of the Fourier descriptors extracted for the fish shape from the board and the tortoise shape and the fish shape extracted from the cut out pieces. The system incorrectly recognises the fish shape as the tortoise as this produces the smallest euclidean distance. Looking at the Fourier descriptor components, we can see that the tortoise and fish descriptors are fairly similar (produces similar peaks) and could be easily confused. The fish shape actually has the second smallest euclidean distance. A similar situation occurs when trying to recognise the tortoise shape.

VII. PROBABILITY MODELS

A. Multinomial distribution

For the multinomial distribution we extracted the Fourier Descriptors from the dataset containing the close-up images of the shapes. The 20 close-up images for each shape were split into two sets containing 14 images for training and 6 images for testing. The training set was further split into two sets containing 7 images each. One was used for training and the other as a validation set. This was done to determine a quasi-ground truth histogram distribution which can be used for testing. The euclidean distance was calculated between every image in the training and validation set. This process was carried out 10 times. The minimum distance value was then determined which identified which object class the system thought each image belonged to. A distribution histogram for each image class was then calculated. A bias was placed at the correct class to provide the system with a reliable ground truth distribution.

Let N be the number of shapes. D the dimension of the Fourier transforms (in this case we used 15 descriptors). Let x represent as possible permutation x $\in \mathcal{P} \subset \{1..N\}^N$. Here, $\mathcal{P}$ denotes all permutations of N objects, hence is the subset of $\{1..N\}^N$ which contains no repetitions. Observation O for the close-up shapes takes the form $O = [O_1, O_2, ...O_N]$, where $O_n \in \mathcal{V} \subset (\mathbb{Z}^+)^N$ which are the counts in the histogram for test images in class n derived above. Let $\theta = [\theta_1, \theta_2...\theta_N]$ represent the parameters of the distributions for each of the object classes. For the multinomial model we use $\theta_n = \alpha_n$, where α_n is the multinomial mean vector set using the counts from the training images. For initialisation a noisy prior is selected for the board. This is done to incorporate the effects/noise that may occur due to illumination changes and the camera or lens used. The prior for the board can be represented by $\pi(x)$. The probability of a permutation given all observations is described as:

$$P(x|O) = \frac{P(O|x) \cdot \pi(x)}{P(O)} \propto \pi(x) \prod_n P'(O_n|\theta_{x_n}), \qquad (3)$$

where $P'(O_n|\theta_{x_n}) = \text{Mult}(O_n|\alpha_{x_n}) = (M!/\prod_m O_{nm}!) \prod_n \alpha_{x_n}^{O_{nm}}$ for the multinomial likelihood, where m ranges across the histogram bins, and M is the number of test images per class.

Bayes theorem can be used to update the probability after each new individual observation. This is given by:

$$\begin{aligned} P_0(x) &= \pi(x) \\ P_t(x|O_1..O_t) &\propto P'(O_{n(t)}|\theta_{x_{n(t)}}) P_{t-1}(x|O_1, ..O_{t-1}), \end{aligned} \qquad (4)$$

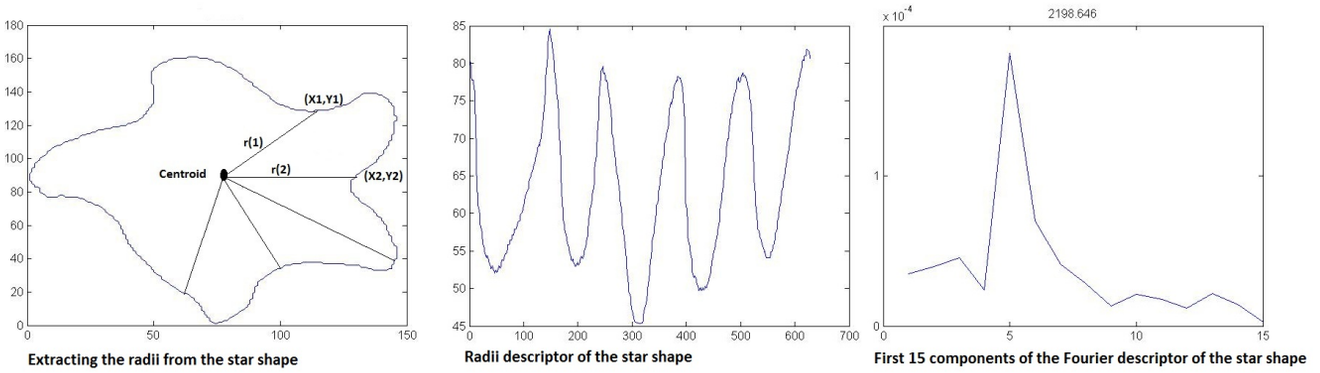


Fig. 4. Boundary of a shape depicting the Fourier descriptors converted to polar co-ordinates

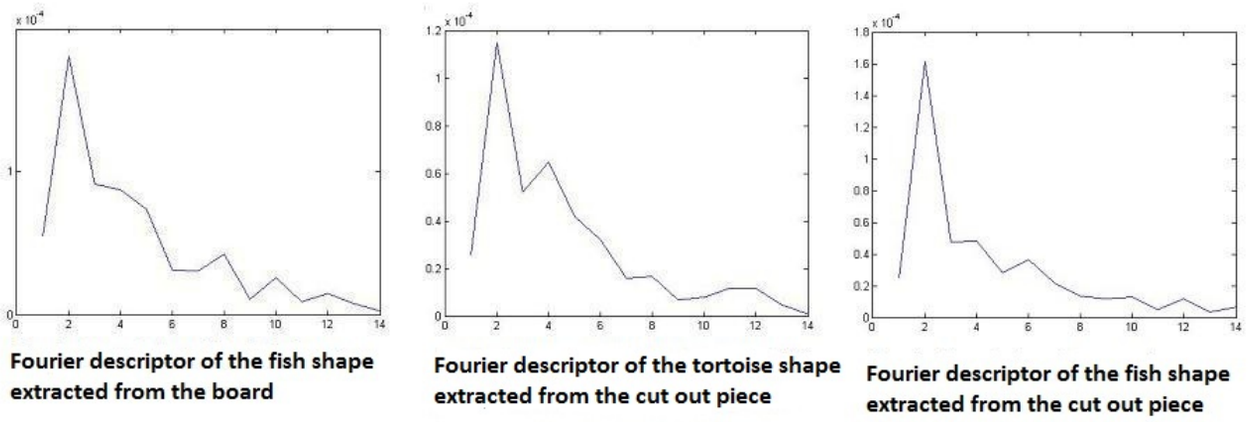


Fig. 5. Fourier Descriptors

where n(t) is the index of the observation seen at time t.

Mutual information(MI) assists in the selection of the position to look at since there can be no repetitions. Once the system is fairly certain of the position of a class in the permutation, mutual information can assist in deciding which position to look at next i.e. which is the most uncertain. Randomly selecting the next position to look at does not not take this information into account.The Mutual Information selection rule is as follows:

$$n(t+1) = \text{argmax}_{n \neq n(1)..n(t)} \, \text{MI}(O_n; x). \tag{5}$$

Mutual information values increases with uncertainty. In this equation we want to select the position in the permutation with the most uncertainty for a given observation.

We can rewrite the above equation in terms of the conditional entropy as follows:

$$\text{MI}(O_n; x) = H(x) - H(x|O_n), \tag{6}$$

where H(.) represents the Shannon entropy and $H(.|.)$ represents the conditional entropy. We need to minimise the conditional entropy. This is described as:

$$n(t+1) = \text{argmin}_{n \neq n(1)..n(t)} \, H(x|O_n). \tag{7}$$

The conditional entropy can be written as:

$$\begin{aligned} H(x|O_n) \quad = \quad & -\sum_{O_n \in \mathcal{V}} P_t(O_n)[\sum_{x' \in \mathcal{P}} P(x'|O_n, O_{n(1)}, ...O_{n(t)}) \cdot \\ & \log(P(x'|O_n, O_{n(1)}, ...O_{n(t)}))]. \end{aligned} \tag{8}$$

To evaluate $P_t(O_n)$, we introduce mixing coefficients β

$$\beta_m = \sum_{(x|x_n=m)} P_t(x), \tag{9}$$

for m = 1..N, which weight the likelihoods for each class. This gives us

$$P_t(O_n) = \sum_m \beta_m \cdot P'(O_n|\theta_m). \tag{10}$$

To avoid exhaustively summing across $\mathcal{V}$ in equation 8, we can consider the conditional entropy as the expectation across $P_t(O_n)$ and approximately evaluate the sum by sampling from

this distribution.

$$\begin{aligned} H(x|O_n) &= E_{o\sim P_t(O_n)}[H(x|o)] \\ &\approx \frac{1}{n}\sum_{o_i} H(x|o_i) \\ &= -\frac{1}{n}\sum_{o_i}[\sum_{x'\in\mathcal{P}} P(x'|o_i, O_{n(1)}, ...O_{n(t)}) \cdot \\ &\quad \log(P(x'|o_i, O_{n(1)}, ...O_{n(t)}))], \end{aligned} \tag{11}$$

where E denotes the expectation. o_i in equation 10 represents the samples drawn from the mixture distribution described in equation 10 where i ranges from 1 to n number of samples.

B. Gaussian Distribution

The image set was treated in the same manner as used in the multinomial distribution. The training images were used to learn a Gaussian distribution for each class, $\theta_n = (\mu_n, \sigma_n)$. For σ_n we used a diagonal covariance matrix. For each observation O_n we included all test images $O_n = [O_{n1}...O_{nM}]$, where M is the number of test images per class. The feature space is $\mathcal{V} = (\mathbb{R}^+)^{DM}$ since we have one D dimensional Fourier descriptor for each image. For the likelihood in equation 10 , we used joint likelihood of these observations.

$$P'(O_n|\theta) = P'(O_n|\mu, \sigma) = \prod_{i=1..M} \mathcal{N}(O_{ni}|\mu, \sigma), \tag{12}$$

where $\mathcal{N}$ represents the Gaussian distribution and O_{ni} is the i'th descriptor of observation n.

Since feature space is now continuous the summation in equation 8 changes to an integral. We can use the sampling technique to approximate this integral as in equation 11.

$$\begin{aligned} H(x|O_n) &= \int_{\mathcal{V}} H(x|o)P_t(o)do \\ &= E_{o\sim P_t(O_n)}[H(x|o)] \\ &\approx -\frac{1}{n}\sum_{o_i}[\sum_{x'\in\mathcal{P}} P(x'|o_i, O_{n(1)}, ...O_{n(t)}) \cdot \\ &\quad \log(P(x'|o_i, O_{n(1)}, ...O_{n(t)}))]. \end{aligned} \tag{13}$$

The sampling distribution used here is the same as described in equation 10, with $P'(O_n|\theta)$ as in equation 12.

VIII. Experiments

The sequence on the board was used as the ground truth. Noise was added to the initial models to take into account possible illumination changes and noise introduced by the camera. Each simulation was run 100 times with a different split of the training and the testing images each time. For both models, we want to identify the correct sequence. Once the system is fairly certain about the object at a specific position mutual information allows us to select the next position to look at which the system is most unsure about. In the random case this position is randomly selected. For the shape puzzle the initial model for board was very good so we introduced an artificial flipping method where two object positions would be flipped at random for 20% of the objects. The reason for doing this was we wanted to demonstrate the effectiveness of using mutual information when the initial guesses are not very accurate. We ran simulations with restricted numbers of objects ranging from 4 to 10, and display the results when only 7 and 10 objects are used. Since we found it was computationally expensive to go through all the possible combinations when using 9 or 10 objects, we introduced the sampling method as discussed to reduce this complexity.

TABLE II
Timings for single position update

Method	# Objects	Random (s)	MI (s)
Multinomial	7	0.003	0.074
	10	0.155	33.916
Guassian	7	0.013	0.090
	10	0.173	33.651

The multinomial distribution is initialised using the class histogram calculated at the start.

The probability after looking at 7 and 10 view are 92% and 94% respectively. This provides better results than just using the euclidean distance of the Fourier descriptors as show in Table 1. Also as shown in Figure 6, MI information outperforms randomly selecting the next position to visit. The figure shows how the probability of the correct permutation changes with the number of views seen, and the percentage of correct objects when the permutation with the highest probability is chosen at each number of viewpoints.

In the Gaussian case, the class histogram is not used. Instead we use a covariance matrix to calculate the likelihoods as explained in Section VII-B. The probability after looking at 7 and 10 view are 99.2% and 99.8% respectively. This provides much better results than just using the euclidean distance of the Fourier descriptors and also outperforms the multinomial distribution.

In Table II the average timings are given for choosing the position to view next and updating the current distribution after making an observation. For the mutual information, we used 20 samples per position. The timings are similar for both the multinomial and Guassian distributions. As shown, the mutual information increases the time taken over random selection, although this could be reduced by using fewer samples while trading off accuracy.

IX. Conclusion

We have presented a system which extracts Fourier descriptors from ten different animal shapes to be used for shape recognition. Initially recognition was performed using the euclidean distance between the shapes. This resulted is an accuracy of 80% with the system confusing the fish and the tortoise shapes. We then set about using the Fourier descriptors from the shapes in probability models. We showed that using mutual information to actively select the next most uncertain position in the sequence provides better results than randomly selecting the next position. Both the models correctly identify all the objects. This paper has shown that using probability models for shape recognition, incorporating information about

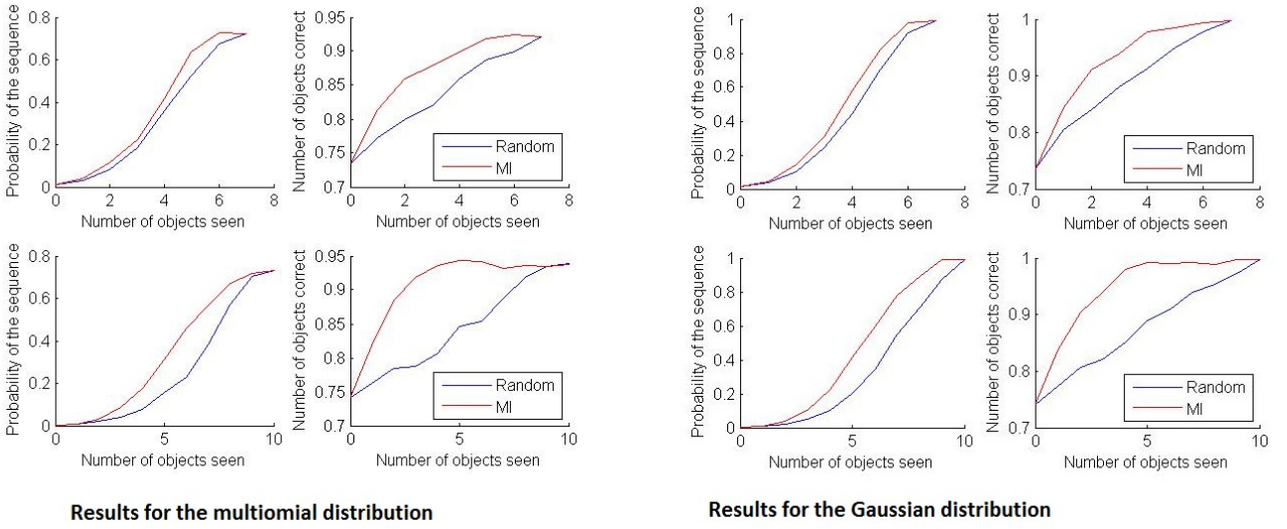


Fig. 6. Results for the multinomial and Guassian distributions using sequences of 7 and 10 objects averaged over 100 splits of data

the current state of the system (MI) and actively selecting which uncertain position to look at produces excellent results.

REFERENCES

[1] C.B Akgul, B.Sankur, Y. Yemez, and F.Schmitt. 3d model retrieval using probability density-based shape descriptors. In *IEEE Transactions of Pattern Analysis and Machine Intelligence*, volume 31, 2009.

[2] A.Singh A.Krause and C.Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 2008.

[3] B.Krishnapuram, C.M. Bishop, and M.Szummer. Generative models and bayesian model comparison for shape recognition. *Ninth International Workshop on Frontiers of Handwriting Recognition*, 2004.

[4] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Active object recognition in parametric eigenspace. In *British Machine Vision Conference (BMVC)*, pages 629–638, 1998.

[5] D.Macrini, C.Whiten, R.Laganiere, and M.Greenspan. Probabilistic shape parsing for view-based object recognition. In *21st International Conference on Pattern Recognition*, 2012.

[6] E.Sommerlade and I.Reid. Information-theoretic active scene exploration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2002.

[8] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*. Addision-Wesley Longman Publishing, 1992.

[9] J.Denzler and C.M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. In *IEEE Transactions on PAMI*, 2002.

[10] J.M.Glover. Probabilistic procrustean models for shape recognition with an application to robotic grasping. Master's thesis, MIT, 2008.

[11] H. Kauppinen, T. Seppanen, and M. Pietikainen. An experimental comparison of autoregressive and fourier-based descriptors in 2d shape classification. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, volume 17, pages 201–207, 1995.

[12] R. D. D. Leon and L. E. Sucar. Human silhouette recognition with fourier descriptors. In *15th International Conference on Pattern Recognition*, pages 709–712, 2000.

[13] G. Lu and A. Sajjanahr. Region-based shape representation and similarity measure suitable for content-based image retrieval. In *IEEE Transactions on Pattern Analysis and Machine Learning*, pages 164–174, 1999.

[14] P. J. Van Otterloo. *A Contour oriented Approach to Shape Analysis*. Prentice Hall, 1991.

[15] E. Persoon and K. Fu. Shape discrimination using fourier descriptors. In *IEEE Transaction On Systems, Man and Cybernetics*, pages 170–179, 1977.

[16] T. W. Rauber. Two-dimensional shape description. Technical report, University Nova de Lisboa, Portugal, 1994.

[17] R.Rosales S.Yu, B.Krishnapuram and R.Rao. Active sensing. In *IEEE International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 639 – 646, 2009.

[18] G. Taubin. *Recognition and Positionning of Rigid Objects using Algebraic and Moment Invariants*. PhD thesis, Brown University, December 1990.

[19] C. H. Teh and R. T. Chin. On image analysis by the methods of moments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 10, 1988.

[20] Q. M. Tieng and W. W. Boles. Recognition of 2d object contours using wavelet transform zero crossing representation. In *IEEE Transaction on Pattern Analysis and Machine Learning*, 1997.

[21] F. J. Janse van Rensburg, J. Treurnicht, and C. J. Fourie. The use of fourier descriptors for object recogntion in robotic assembly. In *5th CIRP International Seminar on Intelligent Computation in Manufacturing Engineering*, 2006.

[22] H. S. Yang, S. U. Lee, and K. M. Lee. Recognition of 2d contours using starting-point-independent wavelet coefficient matching. In *Journal of Visual Communication and Image Representation*, volume 9, pages 171–181, 1998.

[23] C. T. Zahn and R. Z. Roskies. Fourier descriptors for plane closed curves. In *IEEE Transaction on Computer*, volume 21, pages 269–281, 1972.

[24] D. Zhang and G. Lu. A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *Victoria*, volume 14, pages 1–9, 2001.

Vehicle Logo Recognition using Image Matching and Textural Features

Nacer Farajzadeh
Faculty of IT and Computer Engineering
Azarbaijan Shahid Madani University
Tabriz, Iran
n.farajzadeh@azaruniv.edu

Negin S. Rezaei
Department of Mechatronics
Islamic Azad University, Ahar Branch
Ahar, Iran
negin.saberrezayi@yahoo.com

***Abstract*—In recent years, automatic recognition of vehicle logos has become one of the important issues in modern cities. This is due to the unlimited increase of cars and transportation systems that make it impossible to be fully managed and monitored by human. In this research, an automatic real-time logo recognition system for moving cars is introduced based on histogram manipulation. In the proposed system, after locating the area that contains the logo, image matching technique and textural features are utilized separately for vehicle logo recognition. Experimental results show that these two methods are able to recognize four types of logo (Peugeot, Renault, Samand and Mazda) with an acceptable performance, 96% and 90% on average for image matching and textural features extraction methods, respectively.**

***Keywords*—Vehicle, logo recognition, textural features, image matching.**

I. Introduction

The vehicle logo is one of the fundamental signs of the vehicles. Automatic vehicle logo recognition plays an important role in intelligent transportation systems in modern cities. Some vehicle logo recognition applications include vehicle tracking, policing and security [1, 2]. Due to the wide variety in the appearance of vehicles of the same vehicle manufacturer, it is difficult to categorize vehicles using simple methods such as morphological functions and so on. In recent years, several studies have been carried out for vehicle-type classification and vehicle manufacturer recognition [3-5].

In [6], the authors used Scale Invariant Feature Transform (SIFT) for vehicle logo recognition. These features are invariant to scale, rotation and partially invariant to illumination differences [7]. In their study, images taken from the rear view of vehicles were used and they obtained 89.5% recognition accuracy. However, it was reported that this system did not have real-time performance and the speed of recognition process were not mentioned in this article. In another work [8], a new approach for vehicle logo recognition from frontal view was presented with 93% recognition accuracy. In this research, the vehicle manufacture and model were treated as a single class and recognized simultaneously and no results for recognition speed were reported. In [9], a car detection system is presented based on color segmentation and labeling, which performs color recognition. Author of [3] used textural features such as contrast, homogeneity, entropy and momentum for frontal view of vehicle images. The classification accuracy of their work was reported 94% using a three-layered neural network. In [3], the processing times are also not reported.

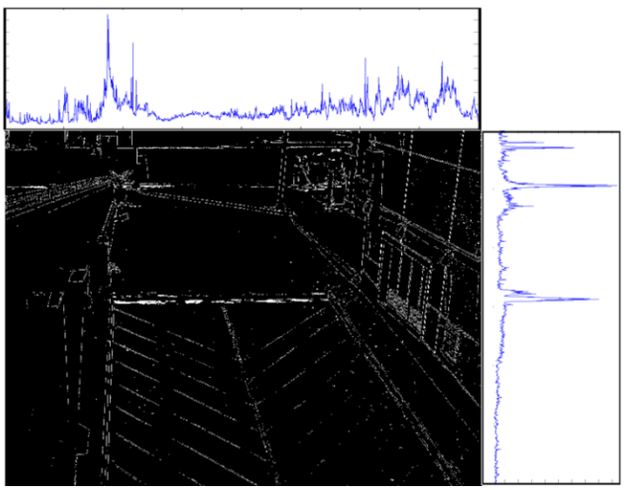
In this research, an autonomous system that aims at obtaining reliable real-time vehicle model recognition for moving vehicles is presented; first by locating the license plate in a vehicle frontal view image and detecting the region of interest over the vehicle, including logo area. Then, the vehicle logo is identified with two different methods: image matching and textural features extraction. This system is flexible and can be used in any situation with an acceptable real-time recognition rate compared to the existing methodologies.

The rest of this paper is organized as follows. In the next section, the proposed system is introduced. In Section 3, the image matching technique and textural features extraction, which are used in the proposed method, are described. Section 4 provides experimental results and Section 5 concludes our study.

II. Proposed Method

The proposed system for vehicle logo recognition is composed of two phases. The first phase is to build a

(a)



(b)

Fig. 1: (a) A parking entrance with no vehicle (main image) (b) Detected edges with vertical and horizontal projections of the pixels lying on the edges.

prototype of the main image. The main image is the image that is taken from an empty scene of interest, i.e. a scene that includes no vehicle or any other moving objects (Fig. 1a).

The prototype of an image is defined as the vertical and horizontal projections (histograms) of its edges. In the proposed system, the red band is used to convert a given image into gray-scale image and Sobel edge detection is used to detect edges. In this phase, we also count the number of pixels, Tm, laying on the edges for the further processing in the next phase. Fig. 1b shows an example of vertical and horizontal histograms of the main image that is taken from an empty parking entrance.

The second phase is the recognition phase where the logo of a vehicle in the given image is identified. This phase consists of three steps. The first step is to investigate whether there is a vehicle in the image or not (vehicle detection step). To this end, the proposed system converts the test image into a gray-scale image and detects the edges on the converted image. Then, the total number of the pixels laying on the edges of the test image, T_t, is calculated. If $500 \leq |T_m - T_t| \leq 5000$, we conclude that there is a vehicle in the scene. Furthermore, if , we may conclude that there are more than one moving cars in the scene. These boundaries were obtained according to our empirical experiments.

In the second step of this phase, the prototype of the test image is built and compared with the prototype of the main image. An example of this comparison is shown in Fig. 2. As this figure shows, there are two ranges on vertical and horizontal projections that the histograms of test image have different values, say over 20%, with the histograms of the main image. Therefore, we define a rectangle with the widths and the heights equal to the widths and the heights of the measured ranges in the vertical and horizontal projections respectively.

As it is seen in Fig. 2c, one sixth of the obtained rectangle is selected to reduce the number of calculations. This area of interest more likely contains the license plate and the logo. The size of the area of interest depends on the width and the length of the obtained rectangle; the bigger the moving object, the bigger the size of the rectangle. We should note that in case that there are two or more vehicles in the image, the size of the obtained rectangle will be larger accordingly.

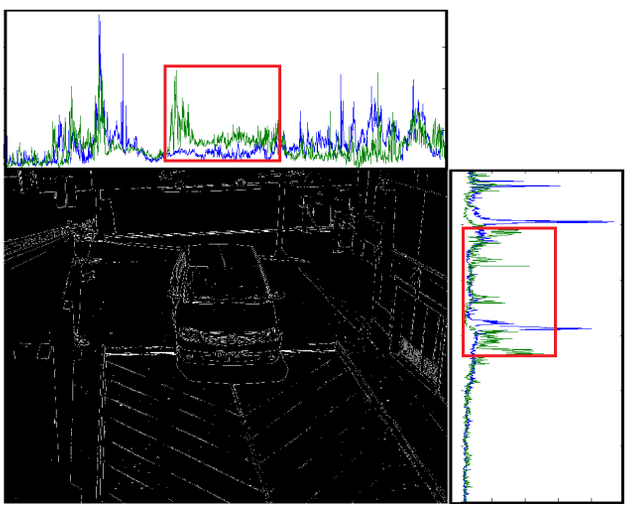
The third step, the last step in this phase, is to recognize the logo in the area of interest. In this step we use image matching and textural features extraction methods. These methods are described in the next section for immediate reference. The block diagram of the proposed system is demonstrated in Fig. 3.

III. Recognition Techniques

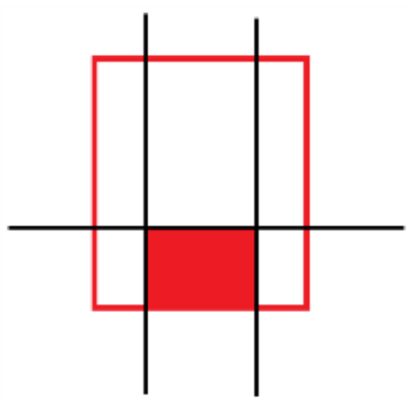
A. Image Matching

he Normalized Cross-Correlation (NCC) is one of the most popular methods for image matching. This method is one of the basic statistical approaches for image registration. It is used for template matching or

(a)



(b)



(c)

Fig. 2: (a) A parking entrance with one vehicle (b) Prototype comparison of the main image and the test image (c) the extracted rectangle that contains license plate and logo.

pattern recognition. Template can be considered a sub-image from the reference image, and the image can be considered as a sensed image. In [10], the authors have proposed a method of medical image registration by template matching based on NCC.

In [11], a fast pattern matching algorithm is proposed based on the NCC criterion by combining adaptive multi-level partition with the winner update scheme to achieve an efficient search. In [12], the author has proposed a combined approach to enhance the performance of template matching system using image pyramid in conjunction with Sum of Absolute Difference (SAD) similarity measure. Based on experimental results, it was found that the capabilities provided by the proposed method in [12] significantly improved the accuracy and execution time of template matching system. From the review of literature, it is observed that the template matching algorithm based on NCC is one of the best approaches for matching the template with same image accurately. In this paper, we use NCC as a model to recognize vehicle logos. The 2D NCC is calculated using Eq. 1.

$$\gamma(u,v) = \frac{\sum_{x,y} \left[(x,y) - \overline{f}_{u,v}\right] \left[t(x-u,y-v) - \overline{t}\right]}{\left(\sum_{x,y} \left[(x,y) - \overline{f}_{u,v}\right]^2 \left[t(x-u,y-v) - \overline{t}\right]^2\right)^{0.5}} \tag{1}$$

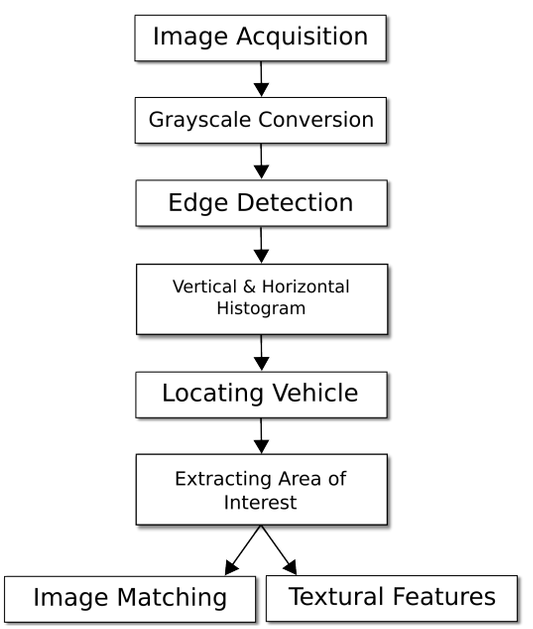


Fig. 3: Block diagram of the proposed system.

In this equation, f and t are the input vehicle image and the logo (template) images, respectively, $\bar{t}$ are the mean gray-level value of the template, $\overline{f}_{u,v}$ is the mean of $f(x,y)$ in the region under the template, (x,y) stands for position on the main image and (u,v) stands for position on the template.

In the proposed method, the template images are scaled up and down in order the proposed system become scale invariant. The template images are scale down until the width of the template image is 4 times smaller than the width of the sub-image cropped from the test image (area of interest), and scale up until the width of the template is equal to the width of the area of interest.

B. Textural Features Extraction

For each retrieved original image, Gray Level Co-occurrence Matrix (GLCM) [13] is used to capture the spatial dependence of gray-level values for different angles of pixel relativity (0° , 45°, 90°, and 135°). Each matrix is run through probability-density functions to calculate different textural parameters. After analyzing the color features of the focused image, the textural features are extracted. In one review, 21 textural parameters were identified [14]. However, another report indicated that only three textural parameters were useful in identifying logo recognition; contrast, homogeneity, entropy and momentum [9]. In this research, three textural parameters are used in identifying image characteristics: entropy, energy, and homogeneity; defined as below [15]:

$$Entropy = -\sum_i \sum_j P_d(i,j) \log_d P(i,j) \quad (2)$$

$$Energy = \sum_i \sum_j P_{(1,0)}(i,j)^2 \quad (3)$$

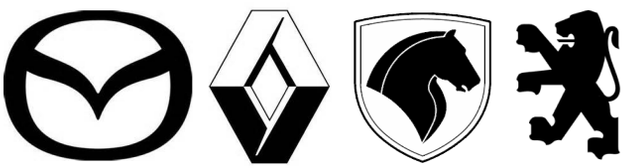
$$Homogeneity = \sum_i \sum_j \frac{P_{(1,0)}(i,j)}{1+(i-j)^2} \quad (4)$$

where d is the distance between two neighboring resolution cells; q is the angle between two neighboring cells; $P_{(1,0)}(i,j)$ is joint probability density function at $d=1$ and $q=0$.

IV. Experimental Results

The automatic and real-time vehicle logo recognition of moving cars faces many challenges. Therefore, preparing a proper dataset is essential to enhance the input data and making it more suitable for the next processing steps. 210 images from the vehicles were captured in noon with natural ambient lightening in a public parking of a mall (Fig. 4). A CCD digital camera (G12 Powershot, Canon) was used to capture images between 12:00 AM and 13:00 AM with 5 seconds interval in June 2013. Obtained Images are 1600×1200 pixels. In the experiments we use four conventional Iranian vehicle logos as shown in Fig. 5. To extract the textural features, we measured textual parameters for 24 rectangles (six images for each type of logo i.e. Peugeot, Renault, Samand and Mazda). Table I shows the value of textural parameters (in pixel) for each type of the vehicle manufacturer.

Fig. 4: Snapshot of a public parking lot (Hyperstar, Tehran, Iran) entrance.



(a) (b) (c) (d)

Fig. 5: Logos of vehicle manufacturers: (a) Mazda (b) Renault (c) Samand (d) Peugeot.

Table II shows the performance of the proposed system with two methods for the recognition of the logos. As it is seen, image matching technique has more precision compared to the textural features. However, it is more time consuming. One of the important advantages of these methods is that their results are not dependent to the color of the logos.

TABLE I: Textural parameters for four types of vehicle logos.

Manufacturer	Entropy (± 10%)	Energy (± 20%)	Homogeneity (± 15%)
Peugeot	7455	2456	7675
Renault	4546	4657	5656
Samand	5565	5568	8854
Mazda	3125	5446	3435

TABLE II: Performance of two proposed methods for vehicle logo recognition.

	Image Matching		Textural Features	
	Precision(%)	Speed(s)	Precision(%)	Speed(s)
Peugeot	98.1	3.7	91.4	2.1
Renault	97.5	3.7	92.9	2.1
Samand	93.4	3.7	86.7	2.1
Mazda	96.0	3.7	89.1	2.1
Average	96.2	3.7	90.0	2.1

V. CONCLUSION

In this study, we proposed an automatic system for vehicle logo recognition. We used two methods to recognize the logos of interest; image matching and textural features. Experimental results showed that these two methods are capable to recognize four types of logo with an acceptable performance, 96% and 90% on average for image matching and textural features extraction methods, respectively. However, the textural features was less accurate than the image matching, it was about 80% faster than it. These two methods can be used for FPGA based programmable boards for increasing the speed of processes. The proposed system that presented in this article can be used as a commercial system for traffic monitoring, tracking stolen cars, managing parking toll, red-light violation enforcement, border and customs checkpoints, etc.

REFERENCES

[1] T. Kato, Y. Ninomiya and I. Masaki, "Preceding vehicle recognition based on learning from sample images", *IEEE Transaction on Intelligent Transportation Systems*, vol. 3, no. 4, pp. 252-260, 2002.

[2] A.H.S. Lai and N.H.C. Yung, "Vehicle-type identification through automated virtual loop assignment and block-based direction-biased motion estimation", *IEEE Transaction on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 86-97, 2000.

[3] H.J. Lee, "Neural network approach to identify model of vehicles", *Lecture notes in computer science*, vol. 3973, pp. 66-72, 2006.

[4] C.N. Anagnostopoulos, I. Anagnostopoulos, V. Loumos and E. Kayafas, "A license plate recognition algorithm for intelligent transportation system applications", *IEEE Transaction on Intelligent Transportation Systems*, vol. 7, no. 3, pp. 377-392, 2006.

[5] A.H.S. Lai, G.S.K. Fung and N.H.C. Yung, "Vehicle type classification from visual-based dimension estimation", *Proceedings of IEEE Intelligent Transportation Systems*, pp. 201-206, 2001.

[6] L. Dlagnekov and S. Belongie, Recognizing cars, University of California, San Diego, 2005.

[7] A.P. Psyllos, C.E. Anagnostopoulos and E. Kayafas, "Vehicle logo recognition using a SIFT-based enhanced matching scheme", *IEEE Transaction on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 322-328, 2010.

[8] V. Petrovic and T. Cootes, "Analysis of features for rigid structure vehicle type recognition", *Proceedings of the British Machine Vision Conference*, pp. 587-596, 2004.

[9] M. Merler, Car color and logo recognition, CSE 190A Projects in Vision and Learning, University of California, 2006.

[10] J. N Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross correlation", *Proceedings of the International Conference on Advances Computing, Control, Telecommunication Technologies*, pp. 819-822, 2009.

[11] S.D. Wei and S.H. Lai, "Fast template matching algorithm based on normalized cross correlation with adaptive multilevel winner update", *IEEE Transation on Image Processing*, vol. 17, no. 11, pp. 2227-2235, 2008.

[12] F. Alsaade, "Fast and accurate template matching algorithm based on image pyramid and sum of absolute difference similarity measure", *Research Journal of information Technology*, vol. 4, no.4, pp.204-211, 2012.

[13] R. Jain, R. Kasturi and B.G. Schunck, Machine Vision, McGraw-Hill, 1995.

[14] C. Zheng, D.W. Sun and L. Zheng, "Recent applications of image texture for evaluation of food qualities: a review", *Trends on Food Science and Technology*, vol. 17, pp. 113-128, 2006.

[15] R.M. Haralick, K. Shanmugam and I. Dinstein, "Textural features for image classification", *IEEE Transaction on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610-621, 1973.

Detecting and Tracking Moving Objects in Video Sequences Using Moving Edge Features

Aziz Karamiani
Faculty of IT and Computer Engineering
Azarbaijan Shahid Madani University
Tabriz, Iran
a.karamiani@azaruniv.edu

Nacer Farajzadeh
Faculty of IT and Computer Engineering
Azarbaijan Shahid Madani University
Tabriz, Iran
n.farajzadeh@azaruniv.edu

***Abstract*—Detecting and tracking moving objects in a sequence of video images is an important application in the field of computer vision. This topic has many applications in surveillance systems, human-computer interaction, robotics, etc. Since the these systems require real-time processing, providing an efficient method with lower computational complexity is a challenge. In this paper, a fast and robust method for detecting and tracking moving objects is presented. This method is based on following mobility edge through fixed edges. The results show that the proposed method, further to its efficiency, is able to overcome challenges such as brightness variations and background changes over time.**

***Keywords*—Object tracking, edge detection, moving object.**

I. Introduction

Detection and tracking moving objects has many applications in the field of machine vision such as: video compression, monitoring systems, industrial control and gesture-based computer interaction. Yilmaz et al. evaluated and classified moving object tracking methods [1]. According to their classification, tracking methods have been divided into three categories: point-based tracking, kernel-based tracking and Silhouette tracking. Point-based method is further divided into two groups: deterministic and statistical. Kernel-based method is also divided into two groups, which are pattern matching and classifier-use. The last one, Silhouette method, uses the shape of objects and evaluate object contour methods.

According to Yilmaz et al., moving object tracking methods in various areas are faced with problems such as overlapping moving objects, change in brightness, little background motion, lack of motion stability in background and camera moving. To fix any of these problems, we should seek appropriate solutions [1].

Slim et al. [2] proposed a method for tracking pedestrians with a mobile camera using a color histogram. The problem of pedestrians overlapping was eliminated with considering histogram of people's head on the overlapping region. Lee et al. [3] introduced a method for tracking a mobile robot by another mobile robot; in their method, for tracking the target robot by the tracker, they set up a view angle of mobile tracking camera for tracking desired position and target robot by means of position information and motion information about both robots. However, in their method, tracking may face problems such as barriers in front of the tracker robot camera, overlapping target robot and its disappearance. Yokoyama [4] used contour-based object method for tracking and gradient feature method for detecting and tracking objects based on optical flow and edge. Zhan et al. [5] used background difference method for moving object based on update background model.

A noticeable improvement for background models is the use of statistical methods for pixel colors. For example, Stauffer and Grison [6] used a Gaussian matrix for pixel color. In this method, a pixel in a healthy frame was compared against the background model of Gaussian. If an adaptation accorded, then average and variance is updated; otherwise, Gaussian average is set to value of pixel color and initial variance. Jepson et al. [7] presented an object tracker that is a combination of three components including object appearance features, unstable features and noise process. The fixed component detects more reliable appearance for estimating object motion where regions of object do not change rapidly over time. Unstable components find rapid changes in pixels and, finally, control outliers noise point are created by noise.

In this paper, a fast and robust yet simple method

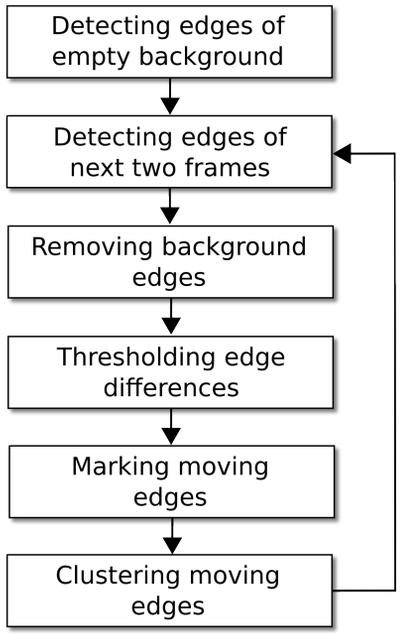


Fig. 1: Block diagram of the proposed system.

for detecting and tracking moving objects is presented. This method is based on following mobility edge through fixed edges. The results show that the proposed method, further to its efficiency, is able to overcome challenges such as brightness variations and background changes over time.

The rest of this paper is organized as follows. The proposed method is explained in section 2. Sections 3 presents experimental results, and Section 4 concludes our work.

II. Proposed Method

In this paper, for tracking moving objects in video sequence, we use edge features. The reason for using edges is that they are less sensitive to light changes. Thus, the system is able to act properly in different environmental conditions. The block diagram of the system is shown in Fig. 1.

A. Excluding the background

In this step, an image of the background is created in the desired range without the presence of moving objects. The image obtained in this step is used to remove fixed edges throughout the entire frames. Fig. 2 shows a sample background image without moving objects.

Fig. 2: Background image without moving objects.

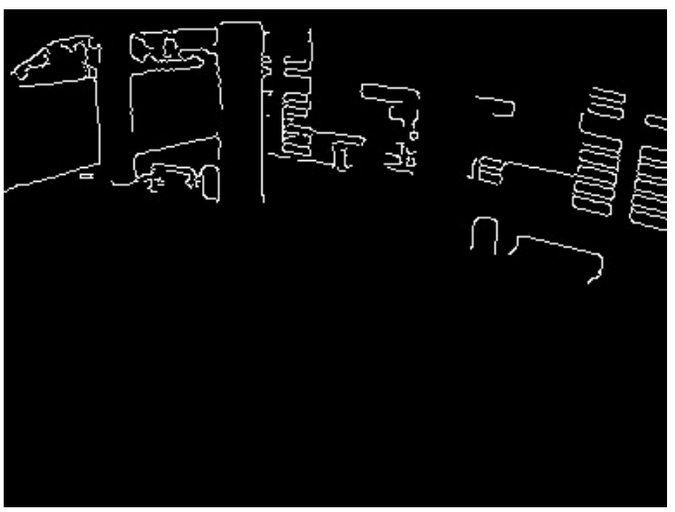


Fig. 3: Canny edge detection for background image.

B. Edge detection in background image

The edges of the background image without the presence of moving objects are detected using Canny algorithm. Canny algorithm for detecting edges is one of the common methods that are used in many applications. Before detecting edges, we smooth the image using Gaussian method. Fig. 3 shows the result of Canny algorithm for Fig. 2.

C. Processing new frames

In this step, the proposed method takes 2 successive frames of the incoming frames (t and $t+1$) , for the next stage. Here, we again employ Gaussian smoothing

Fig. 4: Canny algorithm result on frames 376 (left) and 377 (right).

Fig. 5: Removing background fixed edges from frames t (left) and $t+1$ (right).

and Canny edge detection algorithm on two new frames. Therefore, all the edges of the objects in the scene, which can include stationary objects and moving objects in the background, can be extracted. For example, Fig. 4 shows the resulting edges for frames 376 and 377.

D. Removing the background fixed edges

In this step, the edges of the background image without moving objects in frames t and $t+1$ subtracted from each other to eliminate fixed edges. This can be seen in Fig 5.

Fig. 6: Moving edges marked for frames 94 (left), 202 (middle), and 250 (right).

E. Thresholding

In this step, the proposed method eliminates minimal movement as background noise, e.g. slight movement including moving leaves or someone making a permanent or temporary stop in the scene. To this end, we use difference between successive frames. This step results in removing edges that have no movement between consecutive frames (Eq. 1).

$$f(x,y) = \begin{cases} 255 & \text{if } f_t(x,y) \neq f_{t+1}(x,y) \\ 0 & \text{otherwise} \end{cases}, \qquad (1)$$

where $f(x,y)$ represents the pixel intensity at (x,y).

F. Marking the moving edges

In this step of the proposed method, the remaining edges of the previous frame is marked. The marked edges are obtained from the elimination of repetitive edges in the consecutive frames and background edges. These marked edges are used in the next step to identify moving objects. The results for frames 94, 202 and 250 are shown in Fig. 6. Obtained moving pixels are displayed in red.

G. Tracking moving objects

The final step in the proposed method is to cluster moving edges as moving objects which are attributable to a particular object. To this end, we use an initiative method as follows. We scan the image marked in the previous step from top-left to bottom-right. When a pixel marked as a moving edge is met, we search within a window of size 80×160 pixels, where the marked pixel (the red dot in Fig. 7) is located in the middle of upper side of the mentioned window. In this window, we will count the number of marked pixels. If this number is greater than a threshold, then the target object will be assumed within this window. And a rectangular box with a size of 80×40 pixels is considered as the moving object accordingly. Then, the counted pixels in the gray area,

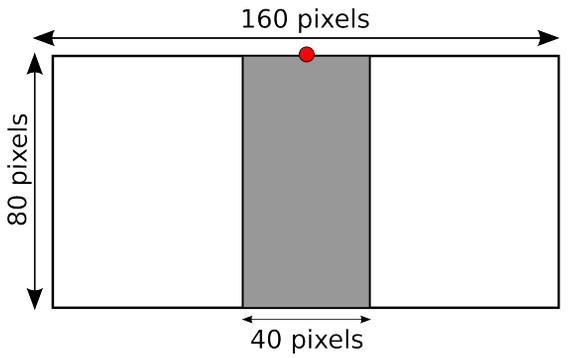


Fig. 7: Particular rectangular area for the moving human-like objects.

as is shown in Fig. 7, is removed to avoid recounting other objects. Note that, this particular rectangular area is assumed for human-like objects and may be changed according to the shape of the object being tracked.

III. Experimental Results

In the testing phase, we consider a video taken from a college [8]. This video, which is a low resolution video (288×384 pixels), contains a scene depicting a passer-by. It captures a scene of human movement that includes various types of motions, changes in global lighting conditions, and occlusion of people due to other people or fixed objects in the scene [8].

To implement the proposed method, EMGUCV library is used. The implemented method is ran on Lenovo B590 with 4GB of memory, video card NVIDIA GEFORCE 1 GB and an Intel (R) Core (TM) i3-3120M CPU@2.50GHz processor. Tracking results are shown in Fig. 8 for the different frame sequences.

The total number of moving objects in the video is counted manually which is 400. The number of correct detections by the proposed method is 374 which results in 93.5% of accuracy. We also use average run time of 25 experiments to evaluate the processing time of the proposed algorithm. The average processing time for the test video with 2294 frames is equal to 60,020 ms. Therefore, the processing time per frame is 26 ms. As a result, 38 frames per second can be processed. As real-time processing requires to process 30 frames per second [9], the proposed algorithm has a desired running time.

IV. Conclusion

Detecting and tracking moving objects in videos is an important application in the field of computer vision.

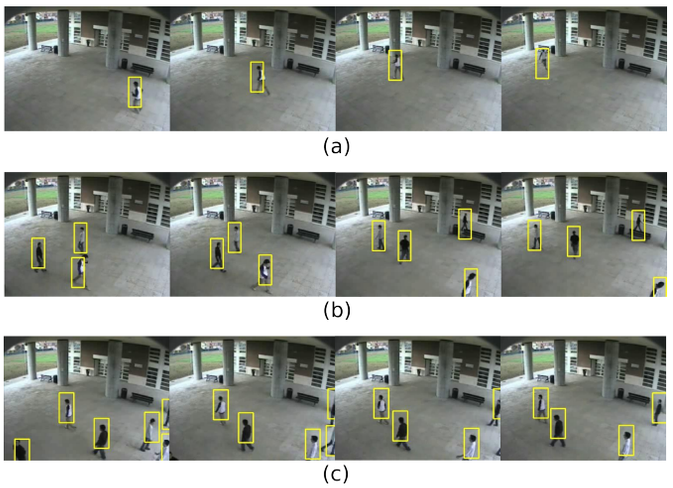


Fig. 8: Samples of tracking results by the proposed method (left to right): (a) frames 94 to 97, (b) frames 203 to 206, (c) frames 334 to 337.

In recent years, many methods have been proposed in the literature. Most of the existing methods are sensitive to changes in brightness and background. In this paper we proposed a method based on edge features, background subtraction, and frame difference for detecting and tracking moving objects. The results showed that the proposed method has a comparable performance with the competitors with a desirable computational complexity, and is robust to changes in illumination and background.

References

[1] A. Yilmaz, O. Javed and M. Shah "Object tracking: a survey", *ACM Computing Surveys*, vol. 38, no. 4, pp. 1-45, 2006.

[2] J.S. Lim and W.H. Kim, "Detection and tracking multiple pedestrians from a moving camera", *International Symposium on Visual Computing*, pp. 527-534, 2005.

[3] C. Lee, "Vision tracking of a moving robot from a second moving robot using both relative and absolute position referencing methods", *37th Annual Conference on IEEE Industrial Electronics Society*, pp. 325-330, 2011.

[4] M. Yokoyama, "A contour-based moving object detection and tracking", *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 271-276, 2005.

[5] R. Zhang "Object tracking and detecting based on adaptive background subtraction", *International Workshop on Information and Electronics Engineering*, pp.1351-1355, 2012.

[6] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real time tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747-757, 2000.

[7] A.D. Jepson, D.J. Fleet and T. Andelmaraghi, "Robust online appearance models for visual tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1296-1311, 2003.

[8] G. Doretto, T. Sebastian, P.H. Tu and J. Rittscher, “Gait-based identification of people in urban surveillance video”, *Journal of Ambient Intelligence and Humanized Computing*, vol.2, no.2, pp. 127-151, 2011.

[9] “Video surveillance trade-offs, a question of balance: finding the right combination of image quality, frame rate and bandwidth ” (Available: http://www.motorolasolutions.com/web/Business/_Documents/staticfiles/VideoSurveillance_WP_3_keywords.pdf).

Vision based mobile Gas-Meter Reading

Machine Learning method and application on real cases

Mehdi Chouiten
WASSA
5 Rue de l'Eglise, 92100, Boulogne-Billancourt,
Region of Paris, France
mehdi.chouiten@wassa.fr

Peter Schaeffer
WASSA
5 Rue de l'Eglise, 92100, Boulogne-Billancourt,
Region of Paris, France
peter.schaeffer@wassa.fr

***Abstract*— The constant increase of smartphones computation capabilities has allowed a growing number of applications. This, combined with the improvements of sensors quality and to third generation (3G) and fourth generation (4G) telecommunication network coverage, made possible the development of robust and reliable computer vision applications exchanging significant amount of data.**

In the gas distribution industry, the consumption reporting is a very important issue. In France, the major gas provider (GDF Suez) plans to deploy 11 million smart meters within 2022. In the meantime, employees of GDF are periodically sent to manually collect data from customers.

In this paper, we present a solution developed for GDF – Suez to solve this problem using mobile technologies and computer vision algorithms.

Keywords; Mobile, Computer Vision, Segmentation, Gas-Meter, OCR, machine learning

I. INTRODUCTION

Smartphones are becoming the prefered plateform for developing innovative applications in several domains like computer vision or augmented reality.

First of all, they are now widely used, a recent study [1] show that more than half of U.S people has one. It's obvious that smartphones are now mainstream, and their use will continue to grow in the near future.

Secondly, they are becoming more and more powerfull, with improved computing and imaging capabilities, it's now possible to embedd in a mobile applications more computer intensive task that were even not possible a few years earlier on desktop computers. For example, the CPU is 40 times more efficient in the latest iPhone comparing to the initial one released in 2007.

Lastly, they are permently connected to fast and reliable cellular networks allowing cloud based mobile applications, for example for data storage or computing of tasks that are still to heavy to be performed on mobile devices.

The challenge we are going to tackle is to find an easy and affordable way to invoice gas customer by allowing the client to automatically read his gas meter through a smartphone application. This task is currently performed by employees sent to the client home to read the counter. Currently, this procedure is very expensive for energy providers and they are searching new ways to lower the cost of this procedure.

Our Solution consist of an automated read and send of the gas-meter current value through a smartphone camera.

II. AIM AND MOTIVATION

Capitalizing on the growth of mobile usage and the inherent increase in performances related to computation capabilities and to network speed, the challenge we are going to tackle is to find an easy and affordable way to bill gas customers by allowing the customers to automatically read their gas-meter through a smartphone application. This task is currently performed by human operators sent to the client's home to read the gas-meter.

Currently, this procedure can be very costly for energy providers and they are searching new ways to lower the cost of this procedure. Among the solutions is to have smart meters connected to the network. Though this solution is effective, the governements and providers plans to replace the existing meters takes from 5 to 10 years depending on countries and the cost of this replacement is valued to bilions of dollars.

Our goal is to provide a solution that consists of an automated gas-meter reading application that works on most smartphones

using only the embedded camera and that is able to record and send the data to the provider.

III. Previous work

Though it is not widespread, the challenge of being able to recognize consumption digits of meters (gas-meters, electricity meters...) is not new. Some previous works have been initiated, and some of them include the use of computer vision. Though most of these systems are not optimized for a mobile device, we will present thei main features and used approaches. We will then have a discussion on their accuracy and introduce the need of a new approach.

In 2011, Cai and al. [2] has initiated research on electric meter recognition using computer vision. The electric meter is considered as a Region Of Interest (ROI) characterized by its color. This detection is then followed by a post processing step aiming to achieve a finer detection taking in consideration the format of the display of the consumption digits.

The character segmentation is done after threshold of the detected zone, some morphologic operations and character segmentation. The numbers classification is based on the number of white pixels.

The presented results are superior to 90% success. These results are based on very specific meters and limited database.

One of the most recent works on the subject is the research done by Vanetti and al. [3]. Using a set of supervised neural models able to detect an object in a cognitive manner. Each node us trained on a set of different points extracted from the training database. This is used for both the counter detection and for digit segmentation.

Within the counter detection step, the segmentation presented some lack of precision in the boundaries. This resulted in missing some parts of the image and sometimes having unnecessary background included in the region detected. To avoid this problem, the team used fast watersheds algorithm.

For the classification of the digits, an SVM with radial function based kernel is used as a discriminative model. The final accuracy of this system was from 45% to 90% depending on the number of noisy pixels.

In the same field, Grafmüller and Bayerer [4] have worked on the improvement of the performance of character recognition algorithms for industrial applications. They use prior knowledge to help the system to take advantage from this information and make a better decision.

The image captured is not binarized since all gray levels are kept. Also, the information of lines and skew is taken in consideration in all the process.
A study comparing different combinations is performed and the result is that the best system is the one using prior knowledge, using DCT (Discrete Cosine Transform) as a feature (instead of PCA) and having SVM as a classifier for individual characters.

The two first systems are quite close to our application. However, their generalization to a heterogeneous set of gas-meters seems complicated and their invariance to light and color variations are not robust because of the chosen features. The naïve OCR used in [2] counting the number of pixels is also a minus.
Finally, our need is to have a mobile application that performs on very heterogeneous devices (including low computation capabilities devices). That is why we introduce our new approach.

IV. Context and Constraints

The application has been designed to fulfill the previously stated objective. However, while studying the needs of the users, we identified some constraints to respect.

First, the application should be able to address most of the users. Today, the market of smartphones is dominated by Apple iPhones based on iOS and Android system based smartphones. Combined, these 2 OS cover more than 96% [5] of the smartphones market. This is why we decided to target the 2 platforms.

Second constraint is about the computation time. Though we want to maximize the performance, it doesn't make sense for the user to wait longer than typing the digits manually. That is why we defined a 3 seconds maximum computation time on any device. The algorithm adapts the processing depending on the device and can work in degraded mode. Also, some gas-meters being in cellars and places without network connection, it was necessary to perform the computation offline.

Finally, another constraint is to allow users to see their history and allow GDF to make statistics of the usage of the application and eventual effects on the user consumption.

V. Solution

To answer the need respecting the previously stated constraints. We proposed a solution packaged as an SDK embeddable on android as well as iOS platforms. The SDK is coded in C++ but also provides all necessary wrappers to be easily included in an java or objective-c application. The solution includes a history of all reports and captured images. It also involves a back-office destined to GDF.

Algorithmically speaking, the solution follows the steps described in figure 1.

First of all, the user has to log into the application with his credentials. Then he simply chooses to add a new report. The camera is then launched. A visual assistant helps the user to target the right zone (the region containing the digits). After having some frames captured, the capture stops and the processing is performed. The solution could process at the same time it makes the capture but keeping in mind that we want the solution to run on most smartphones including old ones, we noticed that some low computation capabilities phones couldn't handle both and multithreading was not really useful for such phones.

The next step is then to detect the ROI (Region Of Interest). It is done using a Haar cascade (Viola and Jones method). The obtained ROI is then converted to HSV (Hue, Saturation, Value) format. We then normalize the obtained cropped image on V channel. After that, we divide the zone in 2 sub-images.One includes the useful consumption part (black part that represents the consumption in cubic meters) and the other is the decimal part in red.

On each of these 2 images, a thresholding is done using Otsu method [6]. Some morphological operations are performed on these two images before blob detection. The transformations include erosion/dilation depending on the result of thresholding. The obtained blobs are then filtered. Too small blobs are deleted. Blobs that don't respect aspect ratios of digits are also eliminated. As shown in figure 2, the obtained blobs represent the digits. These blobs are then aligned. Depending on the result of the alignment, some morphologic operations are performed (to avoid blobs fusion for example). Our blobs are then ready to be passed to the OCR (Optical Character Recognition).

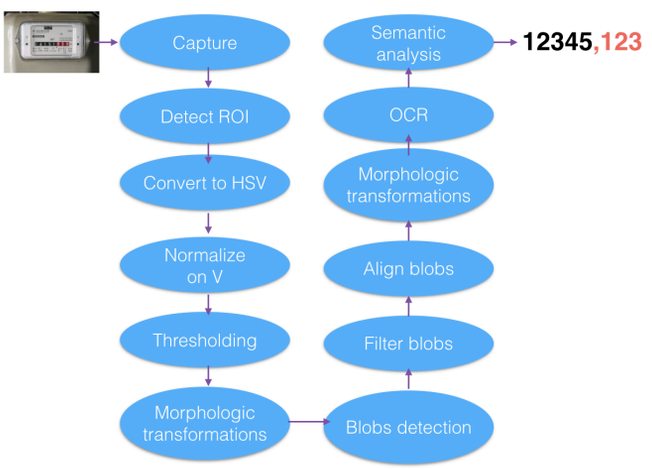


Figure 1. Main steps of the solution

We used GOCR open source OCR in our project because it's very lightweight and because the result of our pre-processing are good enough to produce easy to recognize image. The last step is a semantic analysis verifying the coherence of the consumption depending on history and consumption estimations. If the consumption is valid (eg. not inferior to previous one or too large comparing to estimates…) then the report is stored and sent to GDF back-office.

Figure 2. An example of the output of the algorithm after the "filter blob" step (top: original image, bottom: result).

If the user is not satisfied with the detection. He still can modify the wrong characters. The algorithm takes in consideration this modification and includes it in a learning database for results improvement.

Figure 3. Result interface (allowing the user to change before submit)

VI. TESTS

Tests of the solution have been performed. We performed tests on 20 devices (16 android-based and 4 iOS-based phones). The tests have been performed on heterogeneous gas-meters representing the variety of GDF customers' gas-meters. Tests have been made on 5 different models of gas-meters.

The obtained results are listed below (Table 1). Performance on android phones is lower due to the heterogeneous phones including some low-budget phones having lower computation capabilities and lower resolution camera sensors.
The test phase is divided in 2 different situations. We tested the vision algorithm alone and the full solution including the semantic analysis (comparing to previous consumptions and future consumption estimates).

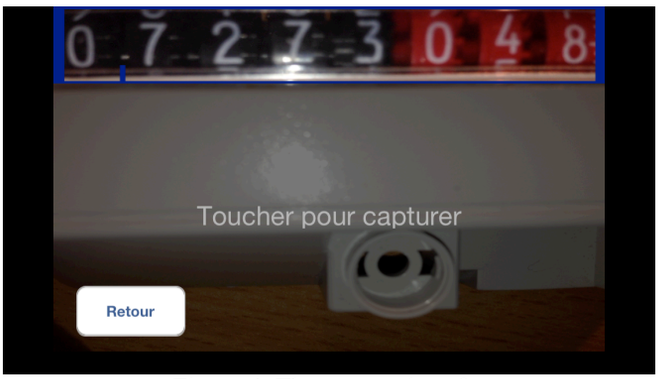


Figure 4. The capture interface

	Pure Vision	Vision + Semantic
Android phones	87 %	92%
iOS phones	92 %	> 99%
Average	89%	93 %

Table 1. Test results (success rate).

The tests include particular situations like very dark places, light reflections, unreadable characters, unclear choices (see Figure 5). The success rates given on Table1 are based on individual digits and not full consumption. This means that if we have on one test 7 digits correct on a total of 8. We will consider 87.5% correct on this test and not consider the whole test as wrong.

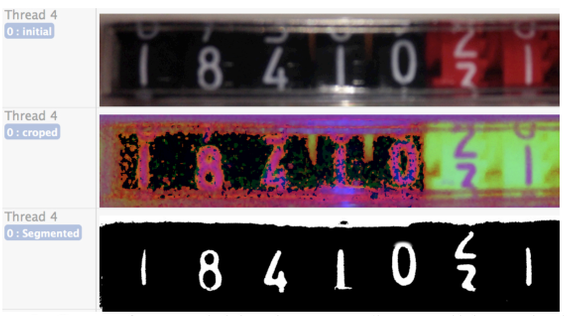


Figure 5. Case of unreadable character (impossible to decide 2 or 3 on 7th character)

VII. Conclusion and Future Work

In this paper, we proposed an innovative solution that avoids a huge amount of money spending and time wasting to send human operators to collect information directly from the customers. Our solution answers the needs and constraints specified in the paper working in a reasonable time on the 2 main mobile Operation Systems and supporting low computation capabilities and offline processing. The machine learning techniques involved and the algorithm itself proved to be quite efficient especially when combined with prior information (semantic analysis).

It is possible to improve the results by having a larger set of training and we can also make a more adaptable algorithm that maximizes performance depending on the device power. We can also have a better management of situations were we can't decide. We currently simply select one of the available choices and in case we don't have any, we select the same digit as what we have in history.

VIII. Aknowledgements

We express our gratitude to GDF-Suez for contributing to this research effort. First, financially. And also for making so many gas-meters models images available and for involving some of their users for the test phase.

IX. References

[1] Aaron Smith. Smartphone ownership 2013. Technical report, pewinternet, 2013.

[2] Zemin Cai, Chuliang Wei, Ye Yuan, An Efficient Method for Electric Meter Readings Automatic Location and Recognition, Procedia Engineering, Volume 23, 2011, Pages 565-571, ISSN 1877-7058.

[3] Marco Vanetti, Ignazio Gallo, and Angelo Nodari. 2013. GAS meter reading from real world images using a multi-net system. *Pattern Recogn. Lett.* 34, 5 (April 2013), 519-526.

[4] Martin Grafmüller, Jürgen Beyerer, Performance improvement of character recognition in industrial applications using prior knowledge for more reliable segmentation, Expert Systems with Applications, Volume 40, Issue 17, 1 December 2013, Pages 6955-6963, ISSN 0957-4174.

[5] IDC, Smartphone OS Market Share, Q1 2014

[6] Nobuyuki Otsu, « A threshold selection method from gray-level histograms », *IEEE Trans. Sys., Man., Cyber.*, vol. 9, 1979, p. 62–66

BAYESIAN BLIND DECONVOLUTION OF IMAGES COMPARING JMAP, EM AND BVA WITH A STUDENT-T A PRIORI MODEL

A. Mohammad-Djafari

Laboratoire des signaux et systèmes (L2S)
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
plateau de Moulon, 3 rue Joliot-Curie, 91192 GIF-SUR-YVETTE Cedex, France

ABSTRACT

Blind image deconvolution consists in restoring a blurred and noisy image when the point spread function of the blurring system is not known a priori. This inverse problem is ill-posed and need prior information to obtain a satisfactory solution. Regularization methods, well known, for simple image deconvolution is not enough. Bayesian inference approach with appropriate priors on the image as well as on the PSF has been used successfully, in particular with a Gaussian prior on the PSF and a sparsity enforcing prior on the image. Joint Maximum A posteriori (JMAP), Expectation-Maximization (EM) algorithm for marginalized MAP and Variational Bayesian Approximation (VBA) are the methods which have been considered recently with some advantages for the last one. In this paper, first we review these methods and give some original insights by comparing them, in particular for their respective properties, advantages and drawbacks and their computational complexity. Then we propose to look at these methods in two cases: A simple one which is using Gaussian priors for both the PSF and the image and a more appropriate case which is a Student-t prior for the image to enhance the sharpness (sparsity) of the image while keeping Gaussian prior for the PSF. We take advantages of the Infinite Gaussian Mixture (IGM) property of the Student-t to consider a hierarchical Gaussian-Inverse Gamma prior model for the image. We give detailed comparison of these three methods for this case.

Keywords

Blind Deconvolution; Bayesian JMAP; Expectation-Maximization (EM); Variational Bayesian Approximation (BVA); Student-t prior models; Blind Image restoration.

1. INTRODUCTION

A blurred image $g(x, y)$ can be modeled as the convolution of the original sharp image $f(x, y)$ with a point spread function (pdf) $h(x, y)$:

$$g(x, y) = f(x, y) * h(x, y) + \epsilon(x, y), \qquad (1)$$

where * represents the convolution operation and $\epsilon(x, y)$ the errors. The inverse problem of the deconvolution consists in estimating $f(x, y)$ from the blurred and noisy image $g(x, y)$ when the Point Spread Function (PSF) $h(x, y)$ of the blurring system is known a priori. This inverse problem is ill-posed and needs prior information on the original image. Regularization theory and the Bayesian inversion have been successful for this task. See for example [1, 2] and [3, 4, 5, 6, 7, 8, 9, 10].

Blind Deconvolution consists in restoring the blurred and noisy image $g(x, y)$ when the PSF $h(x, y)$ is not known a priori. This inverse problem is still more ill-posed and need strong prior information to obtain a satisfactory solution. Regularization theory and simple Bayesian inversion, well known, for simple deconvolution are no more enough [11, 12]. Bayesian inference approach with appropriate priors on the image as well as on the PSF has been used successfully [2, 13, 14, 15, 16]. In particular, a Gaussian prior on the PSF and a sparsity enforcing prior on the image has been used successfully [17, 12, 18].

Joint Maximum A posteriori (JMAP) estimation of the image $f(x, y)$ and the PSF $h(x, y)$, Expectation-Maximization algorithm for marginal MAP and the Variational Bayesian Approximation (BVA) are three main methods which have been considered recently with some advantages for the last one [19, 20, 21, 15, 22, 23].

In this paper, first we review the basic ideas of these methods and give some original insights by comparing these three methods and their associated algorithms. Then, we discuss more in detail their properties as well as their computational costs and complexities for two cases: one for the case of Gaussian priors for both image and the PSF and the second for the case where we still keep a Gaussian prior for the PSF but we propose to use the Student-t prior for the image. The Student-t model has the advantage of sparsity enforcing property and its Infinite Gaussian Mixture property gives the possibility of proposing a hierarchical structure generative graphical model for the output data. Finally, we give details of the three estimation methods of JMAP, BEM and VBA for this prior model and discuss more in detail their properties as well as their computational costs and complexities.

2. BACKGROUND ON BAYESIAN APPROACH FOR BLIND DECONVOLUTION

Assuming a forward convolution model, additive noise, and discretized model, we have:

$$\boldsymbol{g} = \boldsymbol{h} * \boldsymbol{f} + \boldsymbol{\epsilon} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon} = \boldsymbol{F}\boldsymbol{h} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{f}$ represents the unknown sharp image, $\boldsymbol{h}$ the unknown PSF, $\boldsymbol{\epsilon}$ the errors, $\boldsymbol{H}$ the 2D convolution matrix (Toeplitz-Bloc-Toeplitz) obtained from the PSF $\boldsymbol{h}$ and $\boldsymbol{F}$ the 2D convolution matrix obtained from the image $\boldsymbol{f}$ [24, 25, 26].

Using this forward model and assigning the forward $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h})$ and the prior laws $p(\boldsymbol{f})$ and $p(\boldsymbol{h})$, the Bayesian approach starts with the expression of the joint posterior law

$$p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g}) = \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h})\, p(\boldsymbol{f})\, p(\boldsymbol{h})}{p(\boldsymbol{g})}. \tag{3}$$

From here, basically, two approaches have been proposed to estimate both $\boldsymbol{f}$ and $\boldsymbol{h}$:

- JMAP:

$$(\widehat{\boldsymbol{f}},\widehat{\boldsymbol{h}}) = \arg\max_{(\boldsymbol{f},\boldsymbol{h})} \{p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\} \tag{4}$$

  and

- Marginal likelihood estimate of $\boldsymbol{h}$:

$$\widehat{\boldsymbol{h}} = \arg\max_{\boldsymbol{h}} \{p(\boldsymbol{h}|\boldsymbol{g})\} \tag{5}$$

  followed by the marginal MAP estimate of $\boldsymbol{f}$:

$$\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}|\widehat{\boldsymbol{h}},\boldsymbol{g})\right\}. \tag{6}$$

The first one is easily understood and linked to the classical regularization theory, if we note that:

$$(\widehat{\boldsymbol{f}},\widehat{\boldsymbol{h}}) = \arg\max_{(\boldsymbol{f},\boldsymbol{h})} \{p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\} = \arg\min_{(\boldsymbol{f},\boldsymbol{h})} \left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\right\}$$

$$\text{with} \quad J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h}) = -\ln p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h}) - \ln p(\boldsymbol{f}) - \ln p(\boldsymbol{h}) \tag{7}$$

which, with the following Gaussian priors: $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|0, v_\epsilon \boldsymbol{I})$, $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|0, v_f \boldsymbol{I})$ and $p(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{h}|0, v_h(\boldsymbol{C}_h'\boldsymbol{C}_h)^{-1})$ becomes:

$$J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h}) = \frac{1}{v_\epsilon}\|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|_2^2) + \frac{1}{v_f}\|\boldsymbol{f}\|_2^2 + \frac{1}{v_h}\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2. \tag{8}$$

2.1. Joint MAP estimation:

Noting that $\|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|_2^2 = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 = \|\boldsymbol{g} - \boldsymbol{F}\boldsymbol{h}\|_2^2$, the JMAP Criterion (8) can be written as:

$$\begin{aligned} J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h}) &= \tfrac{1}{v_\epsilon}\|\boldsymbol{g} - \boldsymbol{H} * \boldsymbol{f}\|_2^2) + \tfrac{1}{v_f}\|\boldsymbol{f}\|_2^2 + \tfrac{1}{v_h}\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2 \\ &= \tfrac{1}{v_\epsilon}\|\boldsymbol{g} - \boldsymbol{F} * \boldsymbol{h}\|_2^2) + \tfrac{1}{v_f}\|\boldsymbol{f}\|_2^2 + \tfrac{1}{v_h}\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2. \end{aligned} \tag{9}$$

So, its alternate optimization with respect to $\boldsymbol{f}$ (with fixed $\boldsymbol{h}$) and $\boldsymbol{h}$ (with fixed $\boldsymbol{f}$) result to the following iterative algorithm:

JMAP Algorithm:

Initialization:

$$\boldsymbol{h}^{(0)} = \boldsymbol{h}_0, \quad \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(0)})$$

Iterations:

$$\begin{aligned} &\boldsymbol{f}^{(k)} = \arg\min_{\boldsymbol{f}} \left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\right\} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g} \\ &\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}^{(k-1)}) \\ &\boldsymbol{h}^{(k)} = \arg\min_{\boldsymbol{h}} \left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\right\} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}\boldsymbol{F}'\boldsymbol{g} \\ &\boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(k-1)}) \end{aligned} \tag{10}$$

where $\lambda_f = \frac{v_f}{v_\epsilon}$ and $\lambda_h = \frac{v_h}{v_\epsilon}$. This algorithm can be visualized as in the following figure:

$$\begin{array}{ccccc} \widehat{\boldsymbol{h}}^{(0)} & \rightarrow \widehat{\boldsymbol{h}} \rightarrow & \boxed{\begin{array}{l}\boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}) \\ \widehat{\boldsymbol{f}} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g}\end{array}} & \rightarrow \widehat{\boldsymbol{f}} \\ & \uparrow & & \downarrow \\ & \widehat{\boldsymbol{h}} \leftarrow & \boxed{\begin{array}{l}\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}) \\ \widehat{\boldsymbol{h}} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\end{array}} & \leftarrow \widehat{\boldsymbol{f}} \end{array}$$

2.2. Bayesian Expectation-Maximization (BEM)

The second method, needs first the integration (marginalization):

$$p(\boldsymbol{h}|\boldsymbol{g}) = \int p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\, \mathrm{d}\boldsymbol{f} \tag{11}$$

which can not often be done analytically and needs approximation methods to obtain the solution. The Expectation-Maximization (EM) and its Bayesian version (BEM) try to find this solution by alternate maximizing of some lower bound $p^*(\boldsymbol{h}|\boldsymbol{g})$ to it. In summary, the BEM algorithm can be written as a two step iterative algorithm:

- E step:Compute the expected value:

$$Q(\boldsymbol{h},\boldsymbol{h}^{(k-1)}) = \left\langle \ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\rangle_{p(\boldsymbol{f}|\boldsymbol{h}^{(k-1)},\boldsymbol{g})} \tag{12}$$

- M step:

$$\boldsymbol{h}^{(k)} = \arg\max_{\boldsymbol{h}} \left\{Q(\boldsymbol{h},\boldsymbol{h}^{(k-1)})\right\} \tag{13}$$

For the Gaussian case, noting that

$$\begin{aligned}-\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g}) &= c+\tfrac{1}{2}J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\\ &= c+\tfrac{1}{2}\left[\tfrac{1}{v_\epsilon}\|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|_2^2)+\tfrac{1}{v_f}\|\boldsymbol{f}\|_2^2+\tfrac{1}{v_h}\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\right]\end{aligned} \tag{14}$$

where c is a constant which will be eliminated since after, and that

$$\begin{aligned}&<-\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})>\\ &=<\|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|_2^2)>+\lambda_f<\|\boldsymbol{f}\|_2^2)>+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\tfrac{1}{v_\epsilon}\Big[\|\boldsymbol{g}\|^2-2\boldsymbol{g}'<\boldsymbol{F}>\boldsymbol{h}+\|<\boldsymbol{F}>\boldsymbol{h}\|^2+\\ &\text{Tr}\left\{\boldsymbol{H}\text{Cov}\left[\boldsymbol{f}\right]\boldsymbol{H}'\right\}\Big]+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\left[\|\boldsymbol{g}-<\boldsymbol{F}>\boldsymbol{h}\|^2+\|\boldsymbol{D}_f\boldsymbol{h}\|^2\right]+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\end{aligned} \tag{15}$$

where we assumed that $\text{Tr}\left\{\boldsymbol{H}\text{Cov}\left[\boldsymbol{f}\right]\boldsymbol{H}'\right\}$ can be written as $\|\boldsymbol{D}_f\boldsymbol{h}\|^2$ which is possible. Then, with this relation, it is easy to write down the Bayesian EM algorithm as follows:

Bayesian EM Algorithm:

Initialization:

$\boldsymbol{h}^{(0)}=\boldsymbol{h}_0, \quad \boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(0)})$

Iterations:

$\boldsymbol{\Sigma}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}$

$\boldsymbol{f}^{(k)}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g}$

$\boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f}^{(k-1)})$

$\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_f\boldsymbol{h}\|_2^2$

$\boldsymbol{h}^{(k)}=(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\boldsymbol{F}'\boldsymbol{g}$

$\boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(k-1)})$

(16)

This algorithm can be visualized as follows:

$$\widehat{\boldsymbol{h}}^{(0)}\rightarrow\widehat{\boldsymbol{h}}\rightarrow\boxed{\begin{array}{l}\boldsymbol{H}=\texttt{Convmtx}(\widehat{\boldsymbol{h}})\\ \widehat{\boldsymbol{f}}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g}\\ \widehat{\boldsymbol{\Sigma}}=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\end{array}}\Rightarrow\begin{array}{l}\widehat{\boldsymbol{f}}\\ \widehat{\boldsymbol{\Sigma}}\end{array}$$

$$\uparrow\qquad\qquad\qquad\qquad\Downarrow$$

$$\widehat{\boldsymbol{h}}\leftarrow\boxed{\begin{array}{l}\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_f\boldsymbol{h}\|_2^2\\ \boldsymbol{F}=\texttt{Convmtx}(\widehat{\boldsymbol{f}})\\ \widehat{\boldsymbol{h}}=(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\end{array}}\Leftarrow\begin{array}{l}\widehat{\boldsymbol{f}}\\ \widehat{\boldsymbol{\Sigma}}\end{array}$$

2.3. Variational Bayesian Approximation (VBA)

The third approach which, in some way, generalizes BEM, is the VBA method which consists in approximating the joint posterior law $p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})$ by a separable one $q(\boldsymbol{f},\boldsymbol{h})=q_1(\boldsymbol{f}|\boldsymbol{h})\,q_2(\boldsymbol{h}|\boldsymbol{f})$ by minimizing the Kullback-Leibler KL($q:p$). It is easily shown that the alternate optimization of this criterion results to the following iterative algorithm:

- E step:Compute the expected values $\left\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\rangle_{q_1}$ and $\left\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\rangle_{q_2}$ and deduce:

$$\left\{\begin{array}{l}q_1(\boldsymbol{f}|\boldsymbol{h}^{(k)})\propto\exp\left\{\left\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\rangle_{q_2(\boldsymbol{h}|\boldsymbol{f}^{(k-1)})}\right\}\\ q_2(\boldsymbol{h}|\boldsymbol{f}^{(k)})\propto\exp\left\{\left\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\rangle_{q_1(\boldsymbol{f}|\boldsymbol{h}^{(k-1)})}\right\}\end{array}\right. \tag{17}$$

- M step:

$$\left\{\begin{array}{l}\boldsymbol{f}^{(k+1)}=\arg\max_{\boldsymbol{f}}\left\{q_2(\boldsymbol{f}|\boldsymbol{h}^{(k)})\right\}\\ \boldsymbol{h}^{(k+1)}=\arg\max_{\boldsymbol{h}}\left\{q_2(\boldsymbol{h}|\boldsymbol{f}^{(k)})\right\}\end{array}\right. \tag{18}$$

Here too, it can be shown that with the Gaussian priors, we obtain the following algorithm:

VBA Algorithm :

Initialization:

$\boldsymbol{h}^{(0)}=\boldsymbol{h}_0; \quad \boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(0)})$

$\boldsymbol{\Sigma}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}$

$\boldsymbol{f}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g}$

$\boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f})$

$\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_f\boldsymbol{h}\|_2^2$

Iterations:

$\boldsymbol{\Sigma}_h=v_\epsilon(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}$

$\boldsymbol{h}^{(k)}=(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+v_\epsilon\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\boldsymbol{F}'\boldsymbol{g}$

$\boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(k-1)})$

$\text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\}=\|\boldsymbol{D}_h\boldsymbol{f}\|_2^2$

$\boldsymbol{\Sigma}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I}+v_\epsilon\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}=\|\boldsymbol{D}_h\boldsymbol{f}\|^2$

$\boldsymbol{f}^{(k)}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I}+v_\epsilon\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{H}'\boldsymbol{g}$

$\boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f}^{(k-1)})$

$\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_f\boldsymbol{h}\|_2^2$

(19)

$$\begin{array}{l}\widehat{\boldsymbol{h}}^{(0)}\\ \widehat{\boldsymbol{\Sigma}}_f^{(0)}\end{array}\Rightarrow\begin{array}{l}\widehat{\boldsymbol{h}}\\ \widehat{\boldsymbol{\Sigma}}_f\end{array}\Rightarrow\boxed{\begin{array}{l}\boldsymbol{H}=\texttt{Convmtx}(\widehat{\boldsymbol{h}})\\ \text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\}=\|\boldsymbol{D}_f\boldsymbol{f}\|_2^2\\ \widehat{\boldsymbol{\Sigma}}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I}+\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\\ \widehat{\boldsymbol{f}}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I}+\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\boldsymbol{H}'\boldsymbol{g}\end{array}}\Rightarrow\begin{array}{l}\widehat{\boldsymbol{f}}\\ \widehat{\boldsymbol{\Sigma}}_f\end{array}$$

$$\Uparrow\qquad\qquad\qquad\qquad\Downarrow$$

$$\begin{array}{l}\widehat{\boldsymbol{h}}\\ \widehat{\boldsymbol{\Sigma}}_h\end{array}\Leftarrow\boxed{\begin{array}{l}\boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f})\\ \text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_h\boldsymbol{h}\|_2^2\\ \widehat{\boldsymbol{h}}=(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\\ \widehat{\boldsymbol{\Sigma}}_h=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\end{array}}\Leftarrow\begin{array}{l}\widehat{\boldsymbol{f}}\\ \widehat{\boldsymbol{\Sigma}}_f\end{array}$$

2.4. Comparison JMAP1, BEM1 and VBA1

Comparing the three algorithms JMAP (10), BEM (16) and VBA (16), we can make the following remarks:

- In JMAP, there is no need to matrix inversion. At each step, we can find $\boldsymbol{f}^{(k)}$ and $\boldsymbol{h}^{(k)}$ using an optimization algorithm.
- In BEM, at each step, wee need to compute $\boldsymbol{\Sigma}_f$ and do the matrix decomposition $\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}'_f\boldsymbol{h}\|^2$. This is a very costly operation due to the size of the matrices $\boldsymbol{H}'$ and $\boldsymbol{\Sigma}_f$.
- In VBA, at each step, wee need to compute $\boldsymbol{\Sigma}_f$ and do the matrix decomposition $\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}'_f\boldsymbol{h}\|^2$ and also to compute $\boldsymbol{\Sigma}_f$ and do the matrix decomposition $\text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\} = \|\boldsymbol{D}'_h\boldsymbol{f}\|^2$. There are two very costly operations.

For practical applications, we have to write specialized algorithm taking account of the particular structures of the matrix operators $\boldsymbol{H}$ and $\boldsymbol{F}$. In particular, in Blind deconvolution, these matrices are Toeplitz (or Block-Toeplitz) and we can approximate them with appropriate circulant (or Bloc-circulant) matrices and use the Fast Fourier Transform (FFT) to write appropriate algorithms.

3. JMAP, BEM AND VBA WITH A STUDENT-T PRIOR

As we are, in general, looking for a sharp image, a Gaussian prior is not very appropriate. We may use any sparsity enforcing priors. Between those prior law, one is very interesting, the Student-t prior:

$$\mathcal{T}(f_j|\nu,\mu_j,v_f) = \int_0^\infty \mathcal{N}(f_j|\mu_j, z_j^{-1}v_f)\,\mathcal{G}(z_j|\nu/2,\nu/2)\,\mathrm{d}z_j$$

where

$$\mathcal{N}(f_j|\mu_j, z_j^{-1}v_f) = |2\pi v_f/z_j|^{-1/2}\exp\left\{-\frac{1}{2v_f}z_j(x_j-\mu_j)^2\right\}$$

and

$$\mathcal{G}(z_j|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}z_j^{\alpha-1}\exp\left\{-\beta z_j\right\}.$$

Now, using the forward model (2) and the following priors

$$\left\|\begin{array}{l} p(\boldsymbol{\epsilon}|v_\epsilon) = \mathcal{N}(\boldsymbol{\epsilon}|0, v_\epsilon\boldsymbol{I}) \rightarrow p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h},v_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{h}*\boldsymbol{f}, v_\epsilon\boldsymbol{I}),\\ p(\boldsymbol{h}|v_h) = \mathcal{N}(\boldsymbol{h}|0, v_h(\boldsymbol{C}'_h\boldsymbol{C}_h)^{-1})\\ p(\boldsymbol{f}|\boldsymbol{z},v_f) = \mathcal{N}(\boldsymbol{f}|0, v_f\boldsymbol{Z}^{-1}) \text{ with } \boldsymbol{Z} = \text{Diag}[z_1,\cdots,z_N]\\ p(\boldsymbol{z}|\alpha,\beta) = \prod_{j=1}^N \mathcal{G}(z_j|\alpha,\beta)\end{array}\right. \tag{20}$$

we have

$$\begin{aligned} p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h}|\boldsymbol{g},) &\propto p(\boldsymbol{g}|\boldsymbol{h},\boldsymbol{f})\,p(\boldsymbol{h}|v_h)\,p(\boldsymbol{f}|\boldsymbol{z},v_f)\,p(\boldsymbol{z}|\alpha,\beta)\\ &\propto \mathcal{N}(\boldsymbol{g}|\boldsymbol{h}*\boldsymbol{f}, v_\epsilon\boldsymbol{I})\mathcal{N}(\boldsymbol{h}|0, v_h(\boldsymbol{C}'_h\boldsymbol{C}_h)^{-1})\\ &\quad \mathcal{N}(\boldsymbol{f}|0, v_f\boldsymbol{Z}^{-1})\prod_{j=1}^N\mathcal{G}(z_j|\alpha,\beta)\\ &\propto \exp\left\{-\tfrac{1}{v_\epsilon}J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h})\right\}\end{aligned} \tag{21}$$

which results to:

$$\begin{aligned} J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h}) &= \|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2 + \lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|^2\\ &+\lambda_f\|\boldsymbol{Z}^{1/2}\boldsymbol{f}\|^2 + 2v_\epsilon\left[\sum_{j=1}^N(\alpha-1)\ln z_j + \beta z_j\right]\end{aligned} \tag{22}$$

Using this expression, we can obtain easily the necessary developments to describe the the algorithms JMAP, BEM and VBA with this prior model.

JMAP Blind Deconvolution Algorithm with Student-t prior:

Initialization:

$$\boldsymbol{h}^{(0)} = \boldsymbol{h}_0, \quad \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(0)}), \quad \boldsymbol{z}^{(0)} = 1$$

Iterations:

$$\begin{array}{l}\boldsymbol{f}^{(k)} = \arg\min_{\boldsymbol{f}}\left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h})\right\} = (\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1}\boldsymbol{H}'\boldsymbol{g}\\ \boldsymbol{z}^{(k)} = \arg\min_{\boldsymbol{z}}\left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h})\right\} \rightarrow \widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j}\\ \quad \text{with} \quad \widehat{\alpha}_j = \frac{1}{2}+\alpha \text{ and } \widehat{\beta}_j = \beta + \frac{1}{2} + \|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2\\ \boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}^{(k-1)})\\ \boldsymbol{h}^{(k)} = \arg\min_{\boldsymbol{h}}\left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h})\right\} = (\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}'_h\boldsymbol{C}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\\ \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(k-1)})\end{array} \tag{23}$$

where $\lambda_f = \frac{v_f}{v_\epsilon}$ and $\lambda_h = \frac{v_h}{v_\epsilon}$. This is illustrated in the following:

$$\begin{array}{ccc} \widehat{\boldsymbol{h}}^{(0)} & & \\ \widehat{\boldsymbol{z}}^{(0)} & \Rightarrow \widehat{\boldsymbol{h}} \Rightarrow \widehat{\boldsymbol{z}} & \boxed{\begin{array}{l}\boldsymbol{Z} = \text{diag}\left[\widehat{\boldsymbol{z}}\right], \quad \boldsymbol{H} = \texttt{Convmtx}(\widehat{\boldsymbol{h}})\\ \widehat{\boldsymbol{f}} = (\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1}\boldsymbol{H}'\boldsymbol{g}\end{array}} \rightarrow \widehat{\boldsymbol{f}}\\ & \Uparrow & \downarrow \\ \widehat{\boldsymbol{h}} & \Leftarrow & \boxed{\begin{array}{l}\widehat{\alpha}_j = \frac{1}{2}+\alpha, \widehat{\beta}_j = \beta+\frac{1}{2}+\|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2\\ \widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j}\\ \boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f})\\ \widehat{\boldsymbol{h}} = (\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}'_h\boldsymbol{C}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\end{array}} \leftarrow \widehat{\boldsymbol{f}} \\ \boldsymbol{z} & & \end{array}$$

Following the same approach, for BEM we obtain:

BEM Blind Deconvolution Algorithm with Studet-t prior:

Initialization:

$$\boldsymbol{h}^{(0)} = \boldsymbol{h}_0, \quad \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(0)}), \quad \boldsymbol{z}^{(0)} = 1$$

Iterations:

$$\begin{array}{l}\boldsymbol{\Sigma}_f = v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1}\\ \boldsymbol{f}^{(k)} = (\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1}\boldsymbol{H}'\boldsymbol{g}\\ \boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}^{(k-1)})\\ \text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}_f\boldsymbol{h}\|^2\\ \widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j}\\ \quad \text{with } \widehat{\alpha}_j = \frac{1}{2}+\alpha \text{ and } \widehat{\beta}_j = \beta+\frac{1}{2}+\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2\\ \boldsymbol{h}^{(k)} = (\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}'_h\boldsymbol{C}_h+\boldsymbol{D}'_f\boldsymbol{D}_f)^{-1}\boldsymbol{F}'\boldsymbol{g}\\ \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(k-1)})\end{array} \tag{24}$$

The flow diagram of this algorithm is shown in the following:

$$
\begin{array}{l}
\widehat{\boldsymbol{h}}^{(0)} \\
\widehat{\boldsymbol{z}}^{(0)}
\end{array}
\Rightarrow
\begin{array}{l}
\widehat{\boldsymbol{h}} \\
\widehat{\boldsymbol{z}}
\end{array}
\Rightarrow
\boxed{
\begin{array}{l}
\boldsymbol{Z} = \text{diag}\left[\widehat{\boldsymbol{z}}\right], \quad \boldsymbol{H} = \texttt{Convmtx}(\widehat{\boldsymbol{h}}) \\
\widehat{\boldsymbol{\Sigma}}_f = v_\epsilon(\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{Z})^{-1} \\
\widehat{\boldsymbol{f}} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{Z})^{-1}\boldsymbol{H}'\boldsymbol{g}
\end{array}
}
\Rightarrow
\begin{array}{l}
\widehat{\boldsymbol{f}} \\
\widehat{\boldsymbol{\Sigma}}_f
\end{array}
$$

$$
\Uparrow \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Downarrow
$$

$$
\begin{array}{l}
\widehat{\boldsymbol{h}} \\
\widehat{\boldsymbol{z}}
\end{array}
\Leftarrow
\boxed{
\begin{array}{l}
\widehat{\alpha}_j = \frac{1}{2} + \alpha, \widehat{\beta}_j = \beta + \frac{1}{2} + \|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|^2 \\
\widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j} \\
\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}) \\
\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}_h\boldsymbol{h}\|^2 \\
\widehat{\boldsymbol{h}} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h = \boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}
\end{array}
}
\Leftarrow
\begin{array}{l}
\widehat{\boldsymbol{f}} \\
\widehat{\boldsymbol{\Sigma}}_f
\end{array}
$$

Again, following the same steps, we obtain for VBA:

VBA Blind Deconvolution Algorithm with Studet-t prior:

Initialization:

$$
\begin{array}{l}
\boldsymbol{h}^{(0)} = \boldsymbol{h}_0; \quad \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(0)}), \quad \boldsymbol{z}^{(0)} = 1; \\
\boldsymbol{\Sigma}_f = v_\epsilon(\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1} \\
\boldsymbol{f} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g} \\
\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}) \\
\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}_f\boldsymbol{h}\|^2
\end{array}
$$

Iterations:

$$
\begin{array}{l}
\boldsymbol{\Sigma}_h = v_\epsilon(\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h + \boldsymbol{D}_h'\boldsymbol{D}_h)^{-1} \\
\boldsymbol{h}^{(k)} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h + v_\epsilon\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g} \\
\boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(k-1)}) \\
\text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\} = \|\boldsymbol{D}_h\boldsymbol{f}\|^2 \\
\boldsymbol{\Sigma}_f = v_\epsilon(\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I} + v_\epsilon\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1} = \|\boldsymbol{D}_h\boldsymbol{f}\|^2 \\
\boldsymbol{f}^{(k)} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I} + v_\epsilon\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\boldsymbol{H}'\boldsymbol{g} \\
\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}^{(k-1)}) \\
\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}_f\boldsymbol{h}\|^2 \\
\widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j} \\
\quad \text{with } \widehat{\alpha}_j = \frac{1}{2} + \alpha \text{ and } \widehat{\beta}_j = \beta + \frac{1}{2} + \|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|^2
\end{array}
\tag{25}
$$

The flow diagram of this algorithm is shown in the following:

$$
\begin{array}{l}
\widehat{\boldsymbol{h}}^{(0)} \\
\boldsymbol{\Sigma}_h^{(0)} \\
\widehat{\boldsymbol{z}}^{(0)}
\end{array}
\Rightarrow
\begin{array}{l}
\widehat{\boldsymbol{h}} \\
\boldsymbol{\Sigma}_h \\
\widehat{\boldsymbol{z}}
\end{array}
\Rightarrow
\boxed{
\begin{array}{l}
\boldsymbol{Z} = \text{diag}\left[\widehat{\boldsymbol{z}}\right], \quad \boldsymbol{H} = \texttt{Convmtx}(\widehat{\boldsymbol{h}}) \\
\text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\} = \|\boldsymbol{D}_h\boldsymbol{f}\|^2 \\
\widehat{\boldsymbol{\Sigma}}_f = v_\epsilon(\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{Z} + \boldsymbol{D}_f'\boldsymbol{D}_f)^{-1} \\
\widehat{\boldsymbol{f}} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{Z} + \boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\boldsymbol{H}'\boldsymbol{g}
\end{array}
}
\Rightarrow
\begin{array}{l}
\widehat{\boldsymbol{f}} \\
\widehat{\boldsymbol{\Sigma}}_f
\end{array}
$$

$$
\Uparrow \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Downarrow
$$

$$
\begin{array}{l}
\widehat{\boldsymbol{h}} \\
\widehat{\boldsymbol{\Sigma}}_h \\
\widehat{\boldsymbol{z}}
\end{array}
\Leftarrow
\boxed{
\begin{array}{l}
\widehat{\alpha}_j = \frac{1}{2} + \alpha, \widehat{\beta}_j = \beta + \frac{1}{2} + \|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|^2 \\
\widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j} \\
\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}) \\
\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}_h\boldsymbol{h}\|^2 \\
\widehat{\boldsymbol{h}} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h + \boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g} \\
\widehat{\boldsymbol{\Sigma}}_h = v_\epsilon(\boldsymbol{H}'\boldsymbol{H} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}
\end{array}
}
\Leftarrow
\begin{array}{l}
\widehat{\boldsymbol{f}} \\
\widehat{\boldsymbol{\Sigma}}_f
\end{array}
$$

4. CONCLUSIONS

In this paper, we considered the blind deconvolution problem in a Bayesian framework. Then, first, using Gaussian priors, we compared three main algorithms based on JMAP, BEM and VBA giving some insight and comparison between these methods. Then, using a Gaussian prior for the PSF but a Student-t prior for the images to enhance the sparsity, again, we compared those three methods.

The main conclusion can be summarized as follows:

- In JMAP, at each iteration, only the value of the previous estimate of h and f are transferred to the next step. So, in one hand, the computational cost of this approach is low because there is no need for matrix inversion. At the other hand, we do not know a lot about the convergence and the properties of the obtained solution.

- In EM, the value of the IRF h is transferred, but for $\boldsymbol{f}$, its expected value and its uncertainty (covariance matrix) are transferred for the next iteration computation of $\boldsymbol{h}$. So, in one hand, the computational cost of this approach is higher than JMAP because here we need the computation of $\boldsymbol{\Sigma}_f$ which needs a huge matrix inversion. At the other hand, we know a little more about the convergence (to local maximum of the marginal likelihood) and the properties of the obtained solution.

- In VBA, at each step, not only the values of the estimates, but also theirs uncertainties (in fact the whole approximated marginal laws) are transferred. So, in one hand, the computational cost of this approach is still higher than BEM because here we need the computation of $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_h$ which needs two huge dimensional matrix inversion. At the other hand, not only we get the estimates of $\boldsymbol{f}$ and $\boldsymbol{h}$ but also their approximated marginals $q_1(\boldsymbol{f})$ and $q_2(\boldsymbol{h})$, from which, we can compute any statistical properties of these estimates.

5. REFERENCES

[1] G. Demoment and R. Reynaud, "Fast minimum-variance deconvolution," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-33, pp. 1324–1326, 1985.

[2] T. F. Chan and C. K. Wong, "Total variation blind deconvolution.," *IEEE Trans Image Process*, vol. 7, no. 3, pp. 370–375, 1998.

[3] Simone Fiori, "Fast fixed-point neural blind-deconvolution algorithm.," *IEEE Trans Neural Netw*, vol. 15, no. 2, pp. 455–459, Mar 2004.

[4] Liang Wei, Lei Hua-ming, and Que Pei-wen, "Sparsity enhancement for blind deconvolution of ultrasonic

signals in nondestructive testing application.," *Rev Sci Instrum*, vol. 79, no. 1, pp. 014901, Jan 2008.

[5] Haiyong Liao and Michael K Ng, "Blind deconvolution using generalized cross-validation approach to regularization parameter estimation.," *IEEE Trans Image Process*, vol. 20, no. 3, pp. 670–680, Mar 2011.

[6] Filip Sroubek and Peyman Milanfar, "Robust multichannel blind deconvolution via fast alternating minimization.," *IEEE Trans Image Process*, vol. 21, no. 4, pp. 1687–1700, Apr 2012.

[7] Yu-Wing Tai, Xiaogang Chen, Sunyeong Kim, Seon Joo Kim, Feng Li, Jie Yang, Jingyi Yu, Yasuyuki Matsushita, and Michael S Brown, "Nonlinear camera response functions and image deblurring: theoretical analysis and practice.," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 10, pp. 2498–2512, Oct 2013.

[8] Xiang Zhu and Peyman Milanfar, "Removing atmospheric turbulence via space-invariant deconvolution.," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 1, pp. 157–170, Jan 2013.

[9] Hanjie Pan and Thierry Blu, "An iterative linear expansion of thresholds for -based image restoration.," *IEEE Trans Image Process*, vol. 22, no. 9, pp. 3715–3728, Sep 2013.

[10] Thibault Lelore and Frdric Bouchara, "Fair: a fast algorithm for document image restoration.," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 8, pp. 2039–2048, Aug 2013.

[11] Jianlin Zhang, Qiheng Zhang, and Guangming He, "Blind deconvolution of a noisy degraded image.," *Appl Opt*, vol. 48, no. 12, pp. 2350–2355, Apr 2009.

[12] Joao Oliveira, Mario Figueiredo, and Jose Bioucas-Dias, "Parametric blur estimation for blind restoration of natural images: Linear uniform motion and out-of-focus.," *IEEE Trans Image Process*, Oct 2013.

[13] J. Idier and Y. Goussard, "Markov modeling for bayesian multi-channel deconvolution," *Proceedings of IEEE ICASSP*, p. 2, 1990.

[14] J. Zhang, "The mean field theory in em procedures for blind markov random field image restoration.," *IEEE Trans Image Process*, vol. 2, no. 1, pp. 27–40, 1993.

[15] Sevket Babacan, Jingnan Wang, Rafael Molina, and Aggelos Katsaggelos, "Bayesian blind deconvolution from differently exposed image pairs.," *IEEE Trans Image Process*, vol. 19, no. 11, Nov 2010.

[16] Hacheme Ayasso and Ali Mohammad-Djafari, "Joint NDT image restoration and segmentation using Gauss–Markov–Potts prior models and variational bayesian computation," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2265–2277, 2010.

[17] A Mohammad-Djafari, "Bayesian approach with prior models which enforce sparsity in signal and image processing," *EURASIP Journal on Advances in Signal Processing*, vol. Special issue on Sparse Signal Processing, pp. 2012:52, 2012.

[18] Esteban Vera, Miguel Vega, Rafael Molina, and Aggelos K Katsaggelos, "Iterative image restoration using nonstationary priors.," *Appl Opt*, vol. 52, no. 10, pp. D102–D110, Apr 2013.

[19] Rafael Molina, Javier Mateos, and Aggelos K Katsaggelos, "Blind deconvolution using a variational approach to parameter, image, and blur estimation.," *IEEE Trans Image Process*, vol. 15, no. 12, pp. 3715–3727, Dec 2006.

[20] Se Un Park, Nicolas Dobigeon, and Alfred O Hero, "Semi-blind sparse image reconstruction with application to mrfm.," *IEEE Trans Image Process*, vol. 21, no. 9, pp. 3838–3849, Sep 2012.

[21] Zhimin Xu and Edmund Y Lam, "Maximum a posteriori blind image deconvolution with huber-markov random-field regularization.," *Opt Lett*, vol. 34, no. 9, pp. 1453–1455, May 2009.

[22] L. Blanco and L. M. Mugnier, "Marginal blind deconvolution of adaptive optics retinal images.," *Opt Express*, vol. 19, no. 23, pp. 23227–23239, Nov 2011.

[23] Siamak Yousefi, Nasser Kehtarnavaz, and Yan Cao, "Computationally tractable stochastic image modeling based on symmetric markov mesh random fields.," *IEEE Trans Image Process*, vol. 22, no. 6, pp. 2192–2206, Jun 2013.

[24] N. N. Abdelmalek, T. Kasvand, and J. P. Croteau, "Image restoration for space invariant pointspread functions.," *Appl Opt*, vol. 19, no. 7, pp. 1184–1189, Apr 1980.

[25] Wided Souidene, Karim Abed-Meraim, and Azeddine Beghdadi, "A new look to multichannel blind image deconvolution.," *IEEE Trans Image Process*, vol. 18, no. 7, pp. 1487–1500, Jul 2009.

[26] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman, "Understanding blind deconvolution algorithms.," *IEEE Trans Pattern Anal Mach Intell*, Jul 2011.

Modular Multilevel Converter Based HVDC for Grid Voltage Stability

Ngoc-Thinh Quach, Eel-Hwan Kim, Ho-Chan Kim
Dept. of Electrical Engineering
Jeju National University
Jeju City, Korea
{ngoct1984, ehkim, hckim}@jejunu.ac.kr

Min-Jae Kang
Dept. of Electronic Engineering
Jeju National University
Jeju City, Korea
minjk@jejunu.ac.kr

***Abstract*—This paper proposes a control method of the modular multilevel converter based high voltage direct current (MMC-HVDC) system for the grid voltage stability instead of using a static synchronous compensator (Statcom). The d-axis current component is used to control the grid voltage via the reactive power. Meanwhile, the q-axis current component is employed to control the active power according to its application. With this control method, the power system will operate stably with the maximum efficiency. Moreover, it can save the cost of installing Statcom. The simulation results are performed in the PSCAD/EMTDC software environment to confirm the effectiveness of the proposed control method.**

Keywords—MMC-HVDC, Statcom, Grid voltage stability

I. INTRODUCTION

The increasing of electricity demand requires the development of the power systems with high quality, reliability and stability. However, the receiving voltage at the load side is often unstable at the nominal value because of losses in the transmission line and transformer or the random change of load. To solve this problem, the static synchronous compensator (Statcom) has been used [1], [2]. It is a shunt connected device at the point of common coupling (PCC) to stabilize the grid voltage. Nevertheless, the use of Statcom will increase cost. In the power system, which the modular multilevel converter based high voltage direct current (MMC-HVDC) system is connected to, the use of Statcom is not necessary. The function of voltage control will belong to the MMC-HVDC system. A MMC-HVDC system is a new type of voltage source converter for the medium or high voltage applications. Its operation under various conditions has been researched by many authors over the world [3]-[5]. It can control the active and reactive powers independently. The main function of the MMC-HVDC system is to transfer the active power between two power sources. In almost applications, the reactive power is set to zero. This wastes the ability of using the device because the MMC-HVDC system can absorb or generate the reactive power from or to the power system. Thus, this paper proposes a control method of the MMC-HVDC system for the grid voltage stability instead of using Statcom. With this control method, the device is used with the maximum efficiency, meanwhile the power system is operating stably. Moreover, it can save the cost of installing Statcom.

II. COMPENSATION TECHNIQUE OF STATCOM

Fig. 1 describes the single line and vector diagrams of the compensation technique of Statcom. The load requires the reactive power from the source. The load current, I_l, contains two components, the active and reactive currents. Without compensation, the source current, I_s, is high because of the reactive current as shown in Fig. 1(a). However, the source current will be decreased in case of compensation because the reactive current from the source is compensated by Statcom as

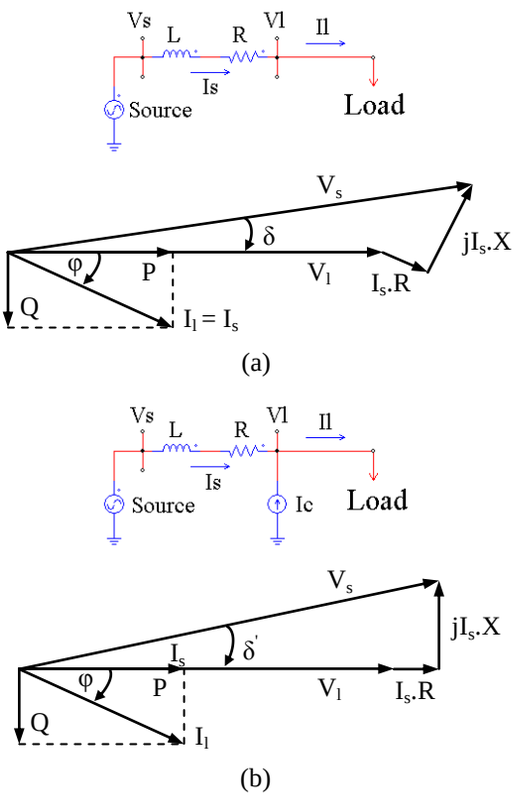


Fig. 1. Single line and vector diagrams of Statcom. (a) Without compensation, (b) With compensation.

expressed in Fig. 1(b). Thus, the losses will reduce significantly. As a result, the voltage received at the load side will be improved. If the Statcom is applied for regulating the grid voltage stability, the amount of the reactive power supplied from Statcom will depend on the reference value of the grid voltage.

III. Proposed Control Method of the MMC-HVDC System

The configuration of the MMC-HVDC system is shown in Fig. 2. The power system is created by the power generation source, transformer, transmission line and load. The source 2 will support for the power system via a MMC-HVDC system.

In this study, the controller of the MMC-HVDC system is designed by combining the compensation technique of Statcom and the conventional control method of MMC-HVDC system. The MMC-1 is employed to control the active power and the grid voltage. The MMC-2 is used to control the dc-link voltage and the reactive power ($Q_2 = 0$).

The control diagram of the MMC-1 for regulating the grid voltage is shown in Fig. 3. Because the MMC-HVDC system is a type of the voltage source converter, thus the d-axis and q-axis current components are controlled independently. The grid voltage is adjusted by using the d-axis current component via the reactive power control. The controller of grid voltage is the proportional-integral (PI) controller. The output signal of the grid voltage controller is the reference reactive power for the MMC-1. The active power is controlled by using the q-axis current component. The output signals of the current control loop are the reference voltages for the pulse-width modulation (PWM) method. Then, the gating signals of the IGBT will be generated from the capacitor voltage balancing method.

IV. Simulation Results

The 250 MW MMC-HVDC system and the power system in Fig. 2 are modeled by using the PSCAD/EMTDC simulation program. The grid voltage is 154 kV. The parameter of the power system is shown in Table 1. The simulation results are set up in two cases.

A. *Independent control between the active and reactive powers*

The simulation results are shown in Fig. 4. The initial conditions are $P_1^* = 100$ MW, $V_{dc} = 100$ kV, $Q_1^* = Q_2^* = 0$. At t = 3 s, the active power is ramped up to 250 MW. Then, the reactive power will be ramped up from 0 to 100 MVar at t = 5 s. Fig. 4(a) is the dq-axis current components of the MMC-1. The active and reactive powers of the MMC-1 are shown in Fig. 4(b). It can be seen that the active and reactive powers do not depend on each other while one of them is changed. The dc-link voltage is always kept at its reference value as depicted in Fig. 4(c). Besides, the reactive power of the MMC-2 is controlled to zero in this case as illustrated in Fig. 4(d). The capacitor voltages of the MMC-1 and MMC-2 are also shown in Fig. 4(e), (f). It is controlled around its nominal value when the active and reactive powers of the MMC-1 are changed.

B. *Grid Voltage Stability with the MMC-HVDC system*

Fig. 5 shows the operation of the MMC-HVDC system with the proposed control method. Because of losses on transmission lines and the transformers, the grid voltage at the PCC will be dropped below the nominal value. Moreover, the source 2 is an offshore wind farm in this study. Thus, the active power which is transferred by the MMC-HVDC system

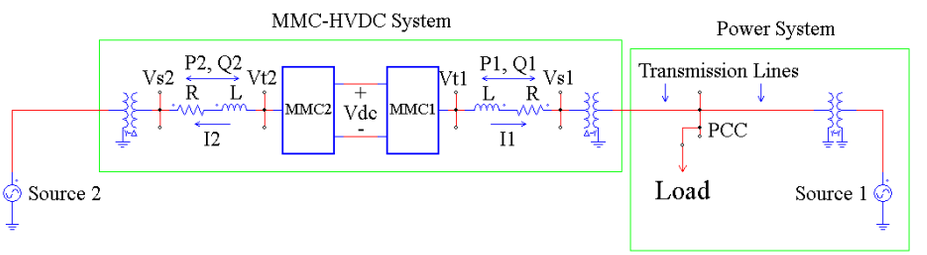


Fig. 2. The configuration of the MMC-HVDC system in the power system

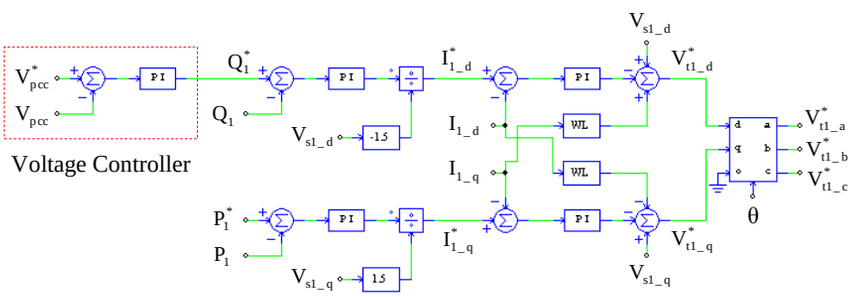
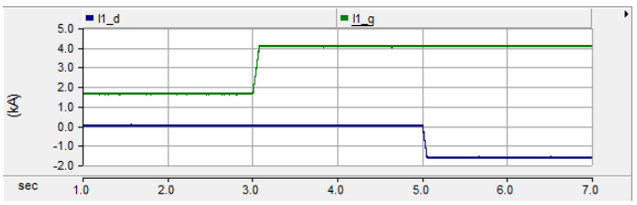
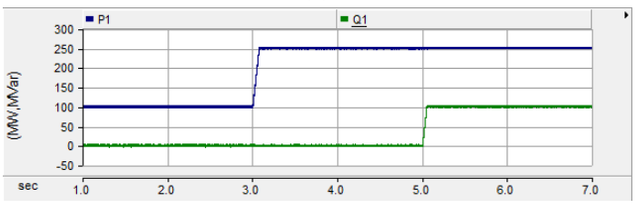


Fig. 3. Control diagram of MMC-1 for regulating the grid voltage

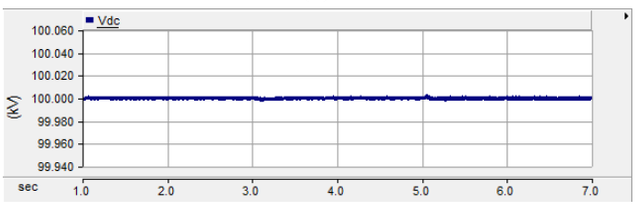
is variable as described in Fig. 5(a). This can cause an oscillation on the grid voltage. At t = 8 s, the voltage controller is activated. The MMC-1 will supply the reactive power to the grid. Therefore, the grid voltage is regulated to the nominal value of 154 kV as expressed in Fig. 5(b). Finally, Fig. 5(c) shows the active and reactive powers of loads.



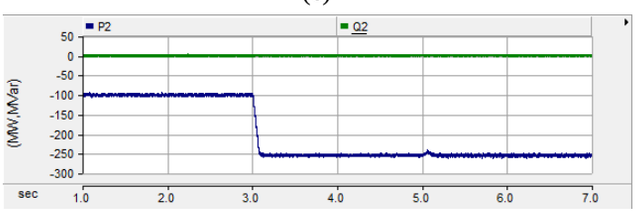
(a)



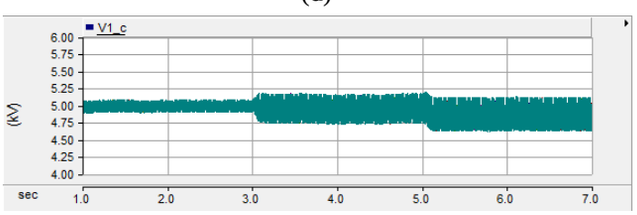
(b)



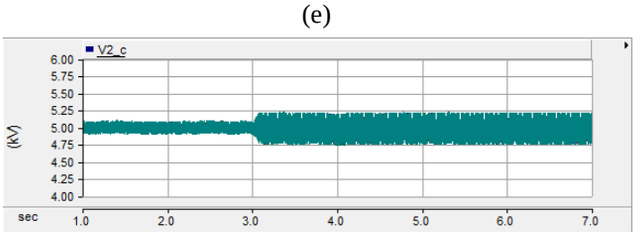
(c)



(d)

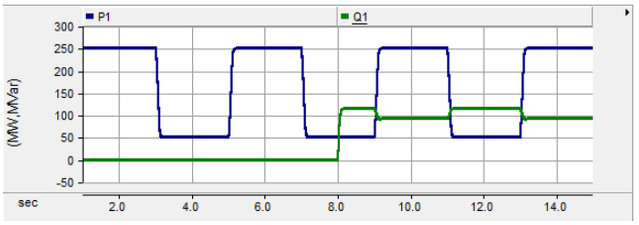


(e)

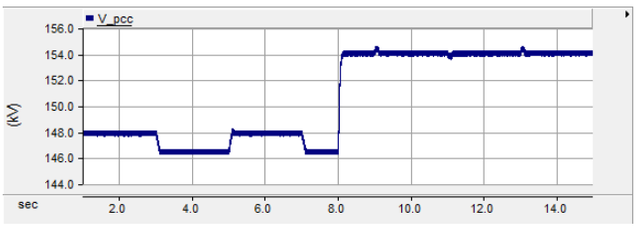


(f)

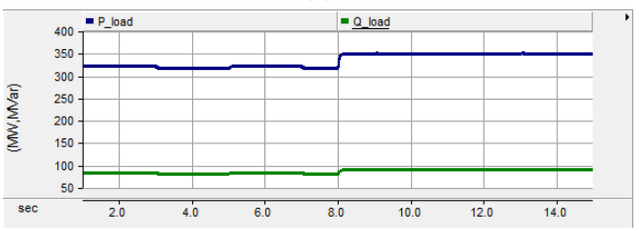
Fig. 4 Independent control



(a)



(b)



(c)

Fig. 5. Grid voltage stability

TABLE I. SIMULATION PARAMETER

		Value
MMC-HVDC system	Active power	250 MW
	Ac voltage	154 kV
	Nominal frequency	60 Hz
	Transformer ratio	154 kV/ 50 kV
	Dc-link voltage	100 kV
	Sub-module capacitor	9700 μF
Grid	Load	350 MW
	Power factor of power sytem	0.97

V. CONCLUSION

This paper has proposed a control method of the MMC-HVDC system for the grid voltage stability instead of using Statcom. The simulation results have demonstrated that the MMC-HVDC system can control the active and reactive powers independently. Moreover, the grid voltage is almost stable at the nominal value without depending on the losses and the power variation from the source. With the proposed control method, the using efficiency of the MMC-HVDC system is maximum and the use of Statcom is not necessary. Thus, it can save the cost.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0025438, 2011-0012202).

REFERENCES

[1] P. Rao, M.L. Crow, and Z. Yang, "Statcom control for power system voltage control applications," IEEE Trans. on Power Del., vol. 15, pp. 1311-1317, Oct. 2000.

[2] A.H. Norouzi and A.M. Sharaf, "Two control schemes to enhance the dynamic performance of the Statcom and SSSC," IEEE Trans. on Power Del., vol. 20, pp. 435-442, Jan. 2005.

[3] J. Qin and M. Saeedifard, "Predictive control of a modular multilevel converter for a back-to-back HVDC system," IEEE Trans. on Power Del., vol. 27, pp. 1538-1547, July 2012.

[4] Q. Tu, Z. Xu, and L. Xu, "Reduced switching-frequency modulation and circulating current suppression for modular multilevel converters," IEEE Trans. on Power Del., vol. 26, pp. 2009-2017, July 2011.

[5] Q. Tu, Z. Xu, and Y. Chang, and L. Guan, "Suppressing DC voltage ripples of MMC-HVDC under unbalanced grid conditions," IEEE Trans. on Power Del., vol. 27, pp. 1332-1338, July 2012.

DIAGNOSIS OF ORAL CANCERS USING IMPLANTED ANTENNAS

Omar K. Hammouda
Department of Information Engineering and Technology
German University in Cairo
omar.abdel-ghany@student.guc.edu.eg

New Cairo City, Cairo, Egypt

Prof. Dr. A. M. M. Allam
Department of Information Engineering and Technology
German University in Cairo
abdelmegid.allam@guc.edu.eg

New Cairo City, Cairo, Egypt

Abstract -- **The main objective of this article is to design an antenna and measure the return-loss behavior of healthy mouth tissues at the ISM frequency band (2.4 - 2.5 GHz) and compare it with those of the mouth tissues infected with malignant cancerous tumors. One can observe the change of the resonant frequency of the antenna in case of infected tissues and thus can detect the existence of malignant tumors in the mouth. Spiral PIFA id utilized. A human mouth model and the antenna are designed using the CST software. The antenna is fabricated on the Roger4350 substrate material which has a thickness of 1.524 mm and a relative permittivity of ε_r = 3.66. The antenna is tested on the face in three locations, left and right cheeks and under the chin. Testing on malignant tumors is only simulated on the CST. There is a noticeable resonance frequency shift of the return-loss when the malignant tissues are considered. There is a good matching between the measured and simulated results.**

Index Terms — Implanted Antenna, Mouth Cancer, Spiral PIFA, ISM band.

I. INTRODUCTION

Cancer in the simplest terms is the abnormal growth of cells somewhere in the body. Each year, more than a million people receive a cancer diagnosis, and the most common types of cancer include breast cancer, prostate cancer, lung cancer, and colorectal cancer. In addition to the three major types of cancer treatment — surgery, chemotherapy and radiation therapy — researchers are working to find new and more effective ways of fighting cancer [1]. Although some cancers cannot be prevented even by living a healthy life style, almost all types of cancers can be totally cured if they were diagnosed in an early stage [1]. As technology continues to change, so too does the practice of medicine. A new device that is in the works is the implantable antennas. It would use new wireless technologies and allow for patient care to occur outside of the doctor's office by monitoring many vital statistics in the human body, including the mutation of cells into malignant tumors.

The challenges in the way of implantable antennas are power loss in the biological tissues, the effect of the surroundings on the antenna impedance and antenna efficiency, size constraints and the difficulties of having actual measurements with the live tissues [2,3,4]. Antennas used to elevate cancer tissues temperature are positioned inside or outside the patient's body. The shapes of antennas depend on their locations. Indeed there are antennas implanted internally and others implanted externally [2,5]. The electrical properties of the body tissues are frequency dependant and should be identified for the frequency of interest. The biological tissues are extremely lossy and this makes it difficult to get a reasonable level of power out of the body. In addition, it is required as impedance matching of the antenna inside the human tissue [3,6].

Safety issues should be taken into consideration when implanting antennas inside human body referring to IEEE C95.1 [7] standard definitions. The Specific Absorption Rate (SAR) is a measure of the rate at which energy is absorbed by the body when exposed to a radio frequency (RF) electromagnetic field. For human body the average SAR is to be below 0.8W/Kg and spatial peak SAR, averaged over any 1-g of tissue, is to be less than 1.6W/Kg across the body. The SAR value depends on the exact location, on the geometry of the part of the body that is exposed to the RF energy and geometry of the RF source [2,3,5].

The main objective of the paper is to design an antenna to be placed on different location of the mouth to measure its return-loss characteristics. Using the fact that malignant tissues and healthy tissues have different electrical

properties and observing the difference in antenna performance could help in early detection of oral cancer. The paper contains four main sections in addition to the conclusion. Section II presents the design of the human mouth model on the. Depicted in section III is the design and fabrication of the antenna. Section IV provides the simulation and measurement results carried out on the CST model. The antenna is tested on two people's face for the healthy tissues and only the CST model for malignant tissues. Section V concludes the paper.

II. DESIGN OF THE HUMAN MOUTH MODEL

An abstracted model is shown in figure 1. The dimensions of the model was based on the dimensions of the CST Voxel Family Katja model shown in figure 2. The model includes skin, muscle, fat, bone, teeth, mucosa and tongue. Other mouth components are excluded due to the lack of information about their electrical properties, that is why it is an abstracted model. Table 1 shows the electrical properties of the previously mentioned mouth parts at 2.45 GHz frequency.

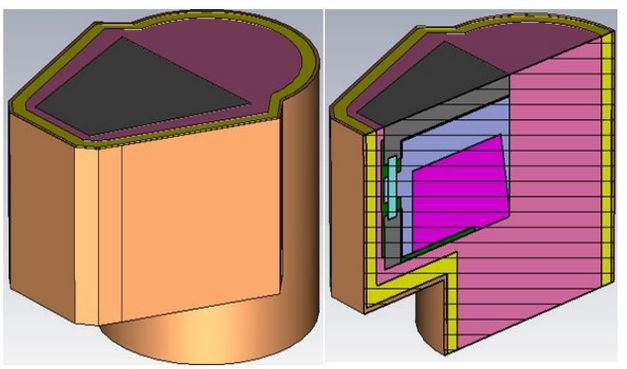


Figure 1. Abstracted mouth model, side view and cut side view

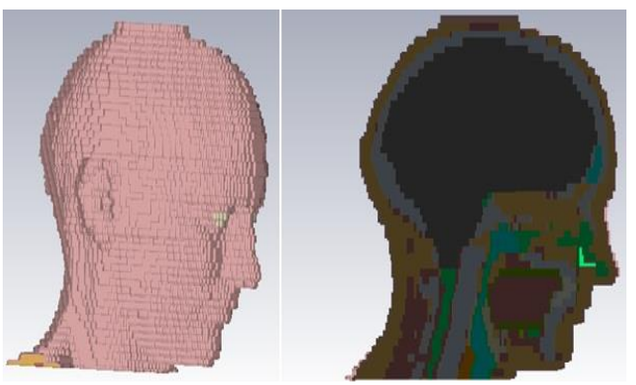


Figure 2. CST Voxel Family Katja human model, side view and cut side view

Table 1. Electrical properties of mouth tissues at 2.45 GHz frequency [6]

Mouth tissue	Conductivity [S/m]	Relative permittivity
Skin	1.464	38.007
Fat	0.10452	5.2801
Muscle	1.7388	52.729
Bone	0.80517	18.548
Tongue	1.8026	52.628
Teeth	0.39431	11.381
Mouth Cavity	0	1
Mucosa	0.5232	65.696

III. DESIGN AND FABRICATION OF SPIRAL PIFA

Planar inverted-F antennas (PIFAs) are miniature designs that offer great versatility for both mobile and wireless applications. Such antennas offer multiband, broadband operation, omni-directional radiation patterns, high efficiency and small size. The spiral PIFA is designed as shown in figure 3. The antenna dimensions are 30x30 mm^2, using the material Roger4350 substrate material whish has a thickness of 1.524 mm and a relative permittivity of $\varepsilon r = 3.66$. The thickness of the ground and the top spiral strip is 0.015 mm.

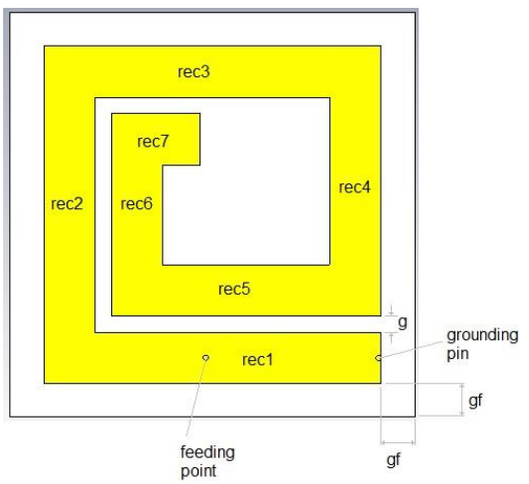


Figure 3. Top view of the spiral PIFA

The distance between the copper spiral strip and the edge of the substrate is *gf* = 2.5 mm, the inner gap of the copper spiral is *g* = 1.2 mm. The grounding pin has a diameter of 0.4 mm and the feeding pin's diameter is 1 mm. the feeding pin midpoint is 12.8 mm away from the grounding pin midpoint. All rectangles have a width of 3.8 mm. Table 2 shows the lengths of the rectangles.

Table 2. Length of the spiral strip rectangles

Rectangle number	1	2	3	4	5	6	7
Length [mm]	25	25	25	20	20	15	6.6

Figure 4 illustrates the manufactured spiral PIFA. The return-loss is measured in air using the network analyzer shown in figure 5. The measured result is compared to the simulated result as seen in figure 6.

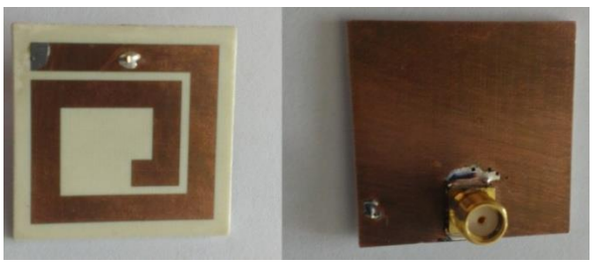


Figure 4. Top and bottom view of the manufactured antenna

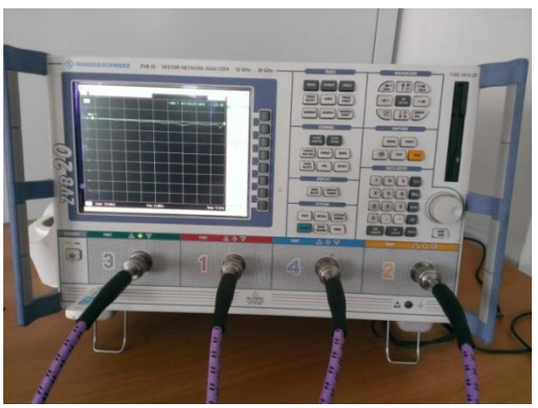


Figure 5. Network Analyzer

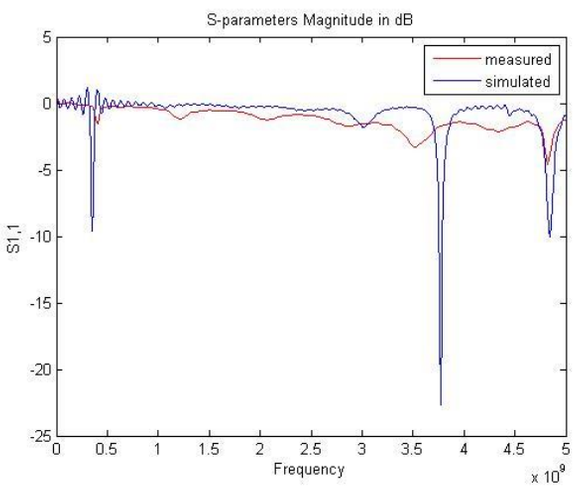


Figure 6. Return-loss of spiral PIFA in air

IV. SIMULATION AND MEASUREMENT OF THE SPIRAL PIFA ON THE MOUTH

The antenna is tested on healthy tissues of two different people. Three main areas of the face are tested on; left cheek, right cheek and under the chin. The results are compared with those from testing the antenna on the mouth model. Figures 7, 8, 9 show the locations where the antenna is tested and figures 10, 11, 12 show the measured return-loss compared with the simulated ones.

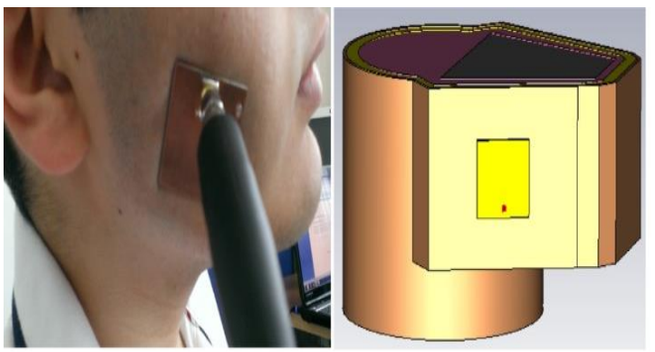


Figure 7. Antenna placed on the right cheek

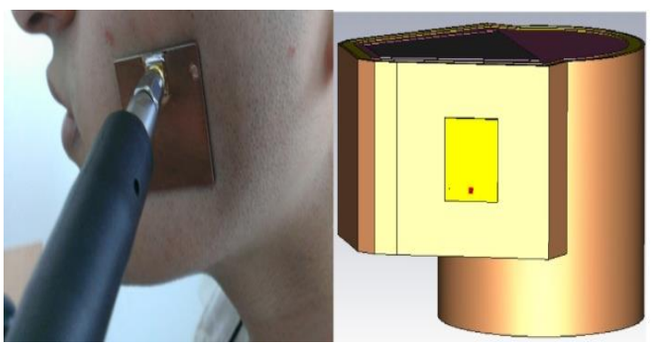


Figure 8. Antenna placed on the left cheek

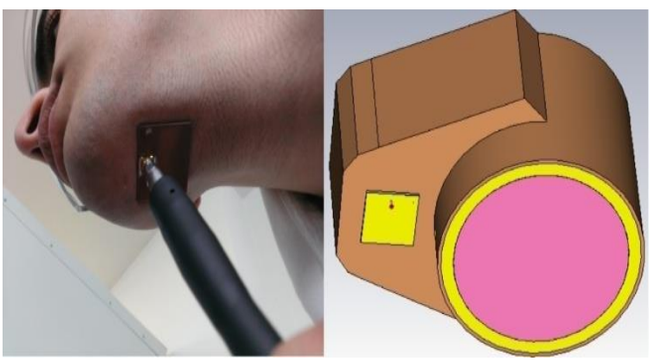


Figure 9. Antenna placed under the chin

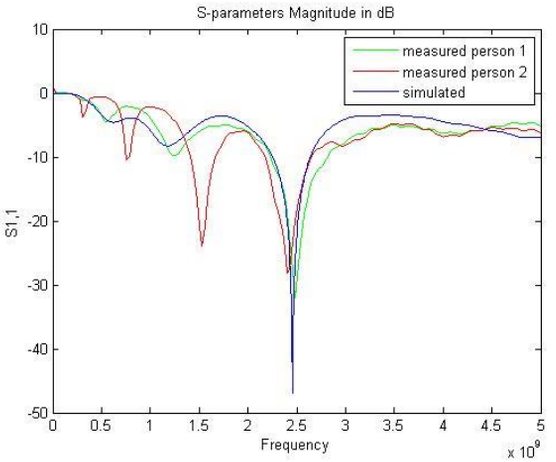


Figure 10. Return-loss of spiral PIFA on right cheek

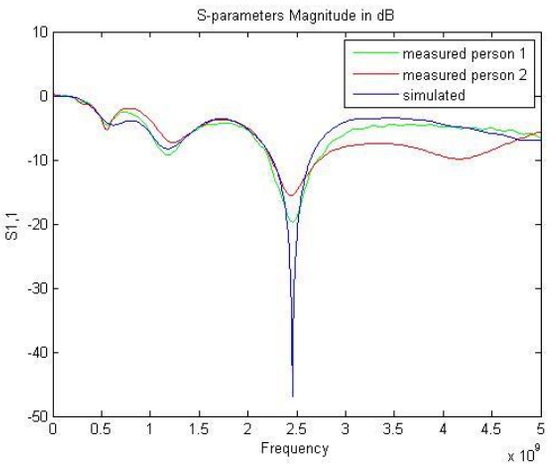


Figure 11. Return-loss of spiral PIFA on left cheek

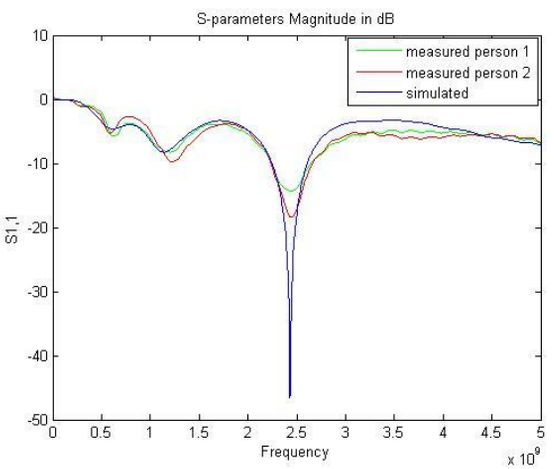


Figure 12. Return-loss of spiral PIFA under the chin

Due to the lack of any information about the electrical properties of malignant mouth tumors, an assumption is made by using the breast cancer cells electrical properties to be able to visualize the diagnosis of cancer infections in mouth. These tissues have a relative permittivity of $\varepsilon_r = 35$ and conductivity of 4.9 S/m. The two main locations where malignant tumors are usually found [1] is either around the tongue or on the inner walls of the cheeks and the lips, as shown in figures 13,14. For the cells surrounding the tongue, they are chosen to have a diameter of 2 cm while the cells located on the inner wall of the cheeks and lips are made of 0.5 cm diameter. Figures 15 - 22 show the simulated results from the healthy mouth tissues compared with the simulated results from the defected tissues. It should be pointed out that the following results show the cancer cells covering either the right side only of the model or both sides. Also, the antenna placed under the chin did not sense the cancer cells located anywhere on the mouth model, that is why its results are excluded from this paper.

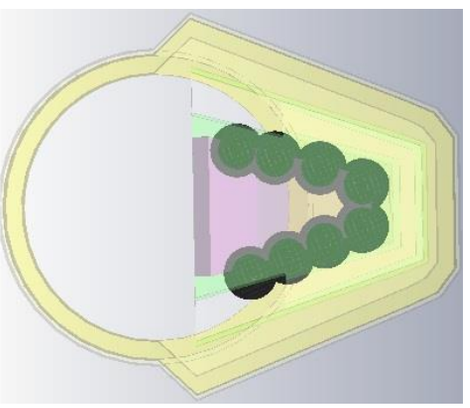


Figure 13. Top view of mouth model, cancer cells surround the tongue

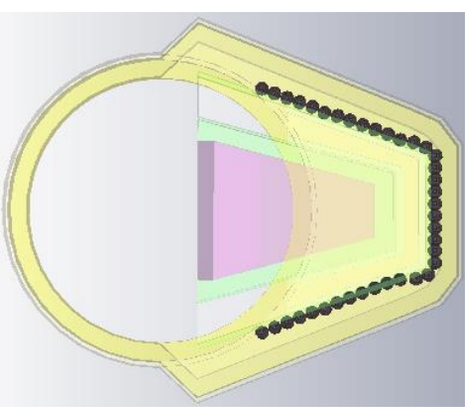


Figure 14. Top view of mouth model, cancer cells on inner cheeks and lips walls

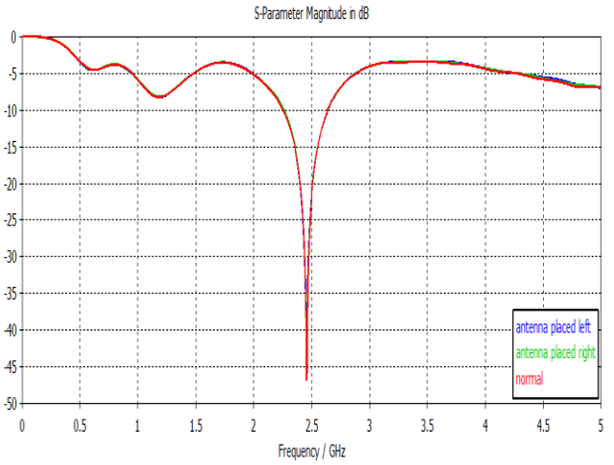


Figure 15. Return-loss of the cancer cells covering the right side of the tongue compared with those of normal tissues

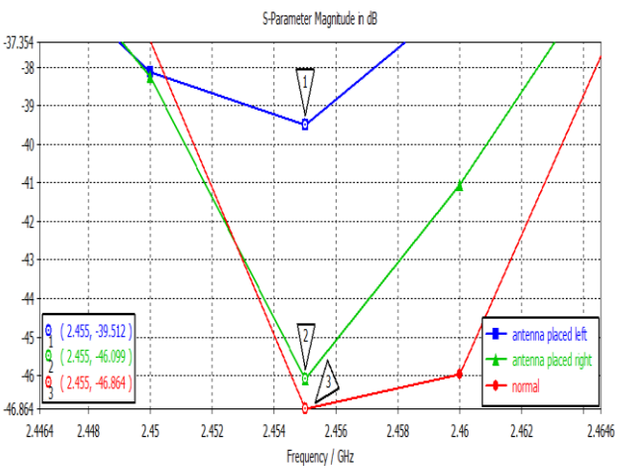


Figure 16. Zoom on figure 15 with markers

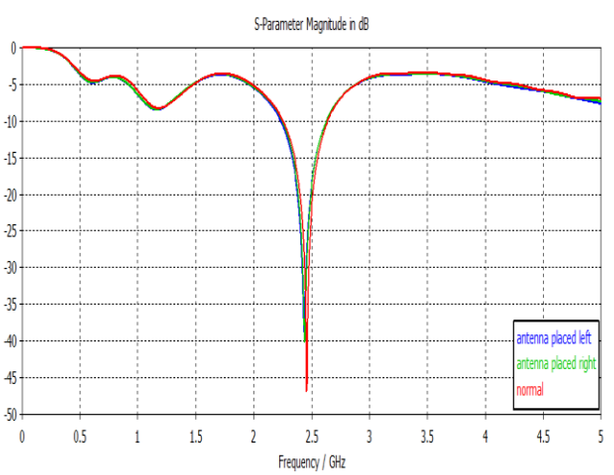


Figure 17. Return-loss of the cancer cells surrounding the tongue compared with those of normal tissues

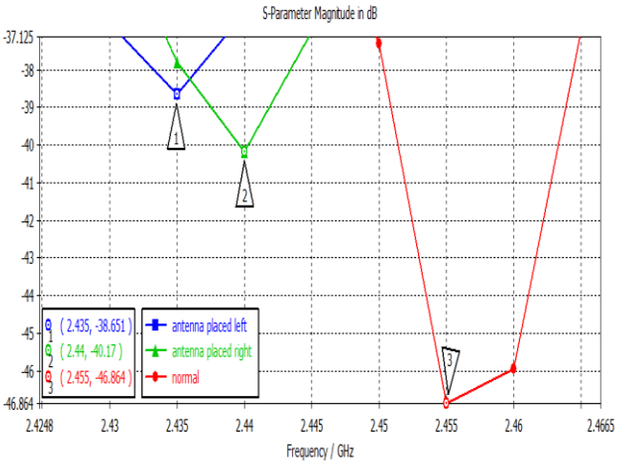


Figure 18. Zoom on figure 17 with markers

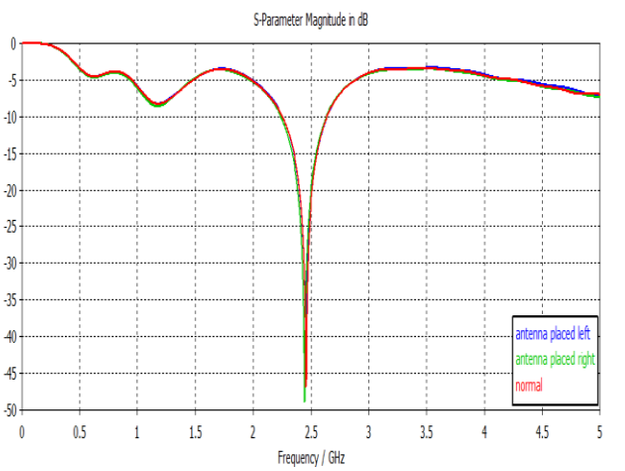


Figure 19. Return-loss of the cancer cell covering the right inner cheeks and lips walls compared with those of normal tissues

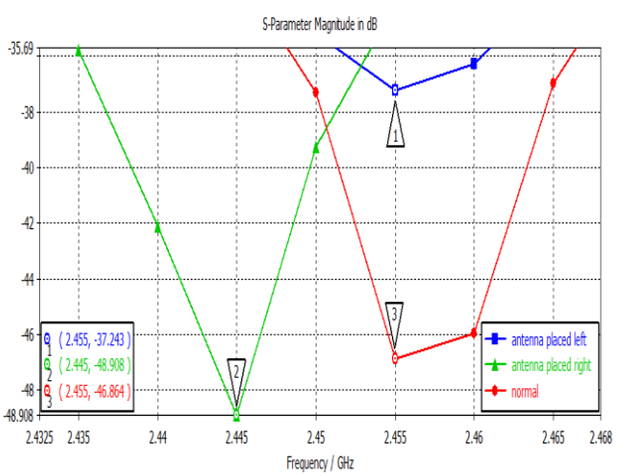


Figure 20. Zoom on figure with markers

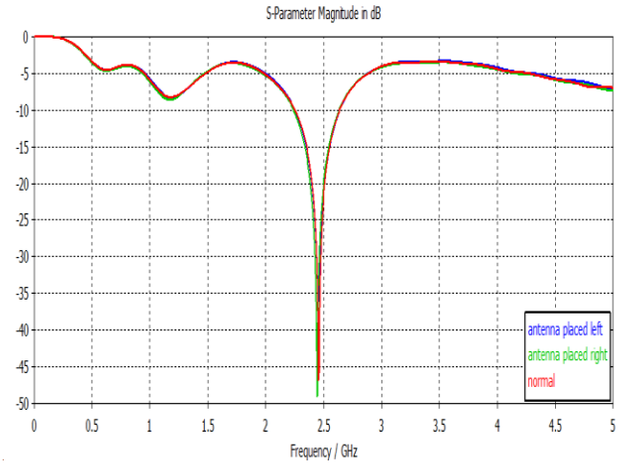


Figure 21. Return-loss of the cancer covering all the inner cheeks and lips walls compared with those of the normal tissues

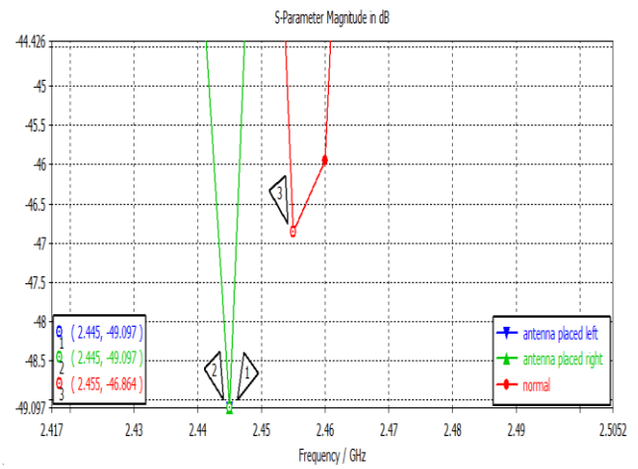


Figure 22. Zoom on figure 21 with markers

One notices that there is a resonance shift in the range of 10000 - 20000 kHz which can achieve cancer diagnosis.

V. CONCLUSION AND FUTURE WORK

The main objective of this project is to design an antenna to be placed on the human mouth. Due to the emphasis on the role of early detection for oral cancer tumors on its treatment, studies are focused on finding ways for early detection of the tumor. The variation of electrical properties of the normal oral tissues and the malignant tissues are used to observe the change in the antenna performance. Thus by comparing both cases, one can detect according to the outcome whether there exists a tumor or not. A spiral PIFA antenna working in the ISM band is designed. It is simulated on my built CST mouth model to get an idea of what to expect when testing on people. The antenna is manufactured and tested on two different persons using the network analyzer. Due to the lack of phantoms simulating the malignant tumors in the mouth, the antenna is only tested on the mouth model in which the cancer cells are added. The simulated and measured results show a good agreement, and there is a noted difference between the healthy simulated tissues and the defected ones.

REFERENCES

[1] National Health Information Society, "Oral Cancer Facts", Office of the Assistant Secretary for Health, Office of the Secretary, U.S. Department of Health and Human Services, USA

[2] Y.Rahmat-Samii and J.Kim," Implanted Antennas in Medical Wireless Communications," A Publication in the Morgan and Claypool Publishers' series, 1st edition, vol.01, 2006.

[3] S.M.Abdelsayed, N.K.Nikolova and M.J.Deen," Radiation Characteristics of Loop Antennas for Biomedical Implants," The National Sciences and Engineering Research Council (NSERC) of Canada and Research in Motion (RIM).

[4] A.Khaleghi and I.Balasingham, " On the Ultra Wideband Propagation Channel Characterizations of the Biomedical Implants,"IEEE 978-1-4244-2517-4, 2009.

[5] J.Kim and Y.Rahmat-Samii," Implanted Antennas Inside a Human Body: Simulations, Designs, and Characterizations," IEEE Transactions on Microwave Theory and Techniques, vol.52, no.8, pp.1934-1943, August2004.

[6] Gabriel, C. and S. Gabriel, "Compilation of the dielectric Improving in-body ultra wideband communication 13 properties of body tissues at RF and microwave frequencies," Brooks Air force Tech. Rep AL/OE-TR-1996-0037, 1996.

[7] IEEE Standard for Safety Levels with Respect to human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz, IEEE Standard C95.1-1999,1999.

BAYESIAN BLIND DECONVOLUTION USING A STUDENT-T PRIOR MODEL AND VARIATIONAL BAYESIAN APPROXIMATION

A. Mohammad-Djafari

Laboratoire des signaux et systèmes (L2S)
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
plateau de Moulon, 3 rue Joliot-Curie, 91192 GIF-SUR-YVETTE Cedex, France

ABSTRACT

Deconvolution consists in estimating the input of a linear and invariant system from its output knowing its Impulse Response Function (IRF). When the IRF of the system is unknown, we are face to Blind Deconvolution. This inverse problem is ill-posed and needs prior information to obtain a satisfactory solution. Regularization theory, well known for simple deconvolution, is no more enough to obtain a satisfactory solution. Bayesian inference approach with appropriate priors on the unknown input as well as on the IRF has been used successfully, in particular with a Gaussian prior on the IRF and a sparsity enforcing prior on the input. Joint Maximum A posteriori (JMAP), Expectation-Maximization (EM) algorithm for marginalized MAP and Variational Bayesian Approximation (VBA) are the methods which have been considered recently with some advantages for the last one. In this paper, first we review these methods and give some original insights by comparing them, in particular for their respective properties, advantages and drawbacks and their computational complexity. Then, we propose to use a Student-t prior law for the unknown input which has the property of sparsity enforcing and which gives the possibility to give a hierarchical graphical structure for the generating model of the observations. Finally, we present detailed algorithms of JMAP, EM and VBA for the joint estimation of the input, the IRF and the hidden variables of the infinite Gaussian mixture model of the Student-t probability law.

Keywords
Blind Deconvolution, Bayesian JMAP, Expectation Maximization (EM), Variational Bayesian Approximation (VBA), Gaussian, Mixture of Gaussians (MoG) and Student-t prior models

I. INTRODUCTION

In a Linear and Invariant System (LIS), the output $g(t)$ can be modeled as the convolution of the input $f(t)$ with the impulse response function (IRF) $h(t)$:

$$g(t) = f(t) * h(t) + \epsilon(t), \tag{1}$$

where * represents the convolution operation and $\epsilon(t)$ the errors. The inverse problem of the deconvolution consists in estimating $f(t)$ from the output $g(t)$ when the IRF $h(t)$ of the blurring system is known a priori. This inverse problem is ill-posed and needs prior information on the input signal $f(t)$. Regularization theory and the Bayesian inversion have been successfully for this task [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19].

Blind Deconvolution consists in estimating both the input $f(t)$ and the IRF $h(t)$ from the output $g(t)$. This inverse problem is still more ill-posed and need strong prior information to obtain a satisfactory solution. Regularization theory and simple Bayesian inversion, well known, for simple deconvolution are no more enough. Bayesian inference approach with appropriate priors on the input as well as on the IRF has been used successfully, in particular with a Gaussian prior on the IRF and a sparsity enforcing prior on the input[5], [20], [21], [22].

Joint estimation of the input $f(t)$ and the IRF $h(t)$ can be done by Joint Maximum A posteriori (JMAP), Bayesian Expectation-Maximization (BEM) or the Variational Bayesian Approximation (VBA) which are three main methods which have been considered recently with some advantages for the last one [23], [24], [25], [26], [27], [28].

In this paper, first we review these methods in general and give some original insights by comparing them for the Gaussian priors model. Then, we propose to keep the Gaussian model for the IRF $h(t)$ but to use a Student-t prior model for the input $f(t)$. The Student-t model has the advantage of sparsity enforcing property and its Infinite Gaussian Mixture property gives the possibility of proposing a hierarchical structure generative graphical model for the output data. Finally, we give details of the three estimation methods of JMAP, BEM and VBA for this prior model and discuss more in detail their properties as well as their computational costs and complexities.

Even if we have applied this prior model and some of these algorithms in different 1D signal deconvolution and 2D image restoration, in this paper no particular result is shown, but the focus will be more on the comparison of these

algorithms and in particular on their computational costs.

II. BASICS OF THE BAYESIAN APPROACH FOR BLIND DECONVOLUTION

Assuming a forward convolution model $g(t) = f(t) * h(t) + \epsilon(t)$ with additive noise, and discretized model, we have:

$$\boldsymbol{g} = \boldsymbol{h} * \boldsymbol{f} + \boldsymbol{\epsilon} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon} = \boldsymbol{F}\boldsymbol{h} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{f}$ is the vector of the unknown input samples, $\boldsymbol{h}$ is the vector of the unknown IRF samples, $\boldsymbol{\epsilon}$ is the vector of errors, $\boldsymbol{H}$ is a Toeplitz convolution matrix obtained from the IRF $\boldsymbol{h}$ and $\boldsymbol{F}$ is a Toeplitz convolution matrix obtained from the samples of the input $\boldsymbol{f}$.

Using this forward model and assigning the forward $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h})$ and the prior laws $p(\boldsymbol{f})$ and $p(\boldsymbol{h})$, the Bayesian approach starts with the expression of the joint posterior law

$$p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g}) = \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h})\, p(\boldsymbol{f})\, p(\boldsymbol{h})}{p(\boldsymbol{g})}. \tag{3}$$

From here, as we will see in the next sections, basically three approaches have been proposed to estimate both $\boldsymbol{f}$ and $\boldsymbol{h}$. The first one is Joint Maximum A Posteriori (JMAP):

$$(\widehat{\boldsymbol{f}},\widehat{\boldsymbol{h}}) = \arg\max_{(\boldsymbol{f},\boldsymbol{h})} \left\{p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\}. \tag{4}$$

The second one is based first on the computation of the Marginal likelihood:

$$\widehat{\boldsymbol{h}} = \arg\max_{\boldsymbol{h}} \left\{p(\boldsymbol{h}|\boldsymbol{g})\right\} \tag{5}$$

followed by the MAP estimate

$$\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}|\widehat{\boldsymbol{h}},\boldsymbol{g})\right\}. \tag{6}$$

But as we will see the first step needs a marginalization which is hard to do. However, the Bayesian Expectation-Maximization (BEM) method tries to find a solution to this method through an iterative algorithm which converges to a local maximum of the marginal likelihood. The third one is based on the approximation of this joint posterior in such a way that its use can be done more easily. These methods are summarized in the next three subsections.

II-A. Joint MAP

The first one is easily understood and linked to the classical regularization, if we note that:

$$\begin{aligned}(\widehat{\boldsymbol{f}},\widehat{\boldsymbol{h}}) &= \arg\max_{(\boldsymbol{f},\boldsymbol{h})} \left\{p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\right\} \\ &= \arg\min_{(\boldsymbol{f},\boldsymbol{h})} \left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\right\}\end{aligned} \tag{7}$$

with

$$J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h}) = -\ln p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h}) - \ln p(\boldsymbol{f}) - \ln p(\boldsymbol{h}) \tag{8}$$

which, with the following Gaussian priors :

$$\left\{\begin{array}{l} p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|0, v_\epsilon \boldsymbol{I}), \\ p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|0, v_f \boldsymbol{I}), \\ p(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{h}|0, v_h (\boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}) \end{array}\right. \tag{9}$$

becomes:

$$\begin{aligned} &J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h}) \\ &= \tfrac{1}{v_\epsilon}\|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|_2^2) + \tfrac{1}{v_f}\|\boldsymbol{f}\|_2^2 + \tfrac{1}{v_h}\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2 \\ &= \tfrac{1}{v_\epsilon}\left[\|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|_2^2) + \lambda_f\|\boldsymbol{f}\|_2^2 + \lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\right]. \end{aligned} \tag{10}$$

where $\lambda_f = \frac{v_f}{v_\epsilon}$ and $\lambda_h = \frac{v_h}{v_\epsilon}$.

Noting that:

$$\|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|_2^2 = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 = \|\boldsymbol{g} - \boldsymbol{F}\boldsymbol{h}\|_2^2,$$

its alternate optimization with respect to $\boldsymbol{f}$ (with fixed $\boldsymbol{h}$) and $\boldsymbol{h}$ (with fixed $\boldsymbol{f}$) result to:

$$\left\{\begin{array}{l} \textbf{Algorithm JMAP:} \\ \\ \textbf{Initialization:} \\ \\ \boldsymbol{h}^{(0)} = \boldsymbol{h}_0, \quad \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(0)}) \\ \\ \textbf{Iterations:} \\ \\ \boldsymbol{f}^{(k)} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g} \\ \boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}^{(k-1)}) \\ \\ \boldsymbol{h}^{(k)} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}\boldsymbol{F}'\boldsymbol{g} \\ \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(k-1)}) \end{array}\right. \tag{11}$$

We may note that the computation of $\boldsymbol{f}^{(k)}$ and $\boldsymbol{h}^{(k)}$ can be done via gradient based optimization algorithms:

$$\left\{\begin{array}{l} \boldsymbol{f}^{(k)} = \arg\min_{\boldsymbol{f}} \left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\right\} \\ \boldsymbol{h}^{(k)} = \arg\min_{\boldsymbol{h}} \left\{J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\right\} \end{array}\right. \tag{12}$$

II-B. Algorithm BEM

The second method, needs first the integration (marginalization):

$$p(\boldsymbol{h}|\boldsymbol{g}) = \int p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\, \mathrm{d}\boldsymbol{f} \tag{13}$$

which can not often be done analytically and needs approximation methods to obtain the solution. The Expectation-Maximization (EM) and its Bayesian version try to find this solution by alternate maximizing of some lower bound $p^*(\boldsymbol{h}|\boldsymbol{g})$ to it. In summary, the BEM algorithm can be written as a two step iterative algorithm:

- E step: Compute the Expected value:

$$Q(\boldsymbol{h},\boldsymbol{h}^{(k-1)}) = \langle \ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle_{p(\boldsymbol{f}|\boldsymbol{h}^{(k-1)},\boldsymbol{g})} \tag{14}$$

  which is now considered as a function of $\boldsymbol{h}$.

- M step (Maximization):

$$\boldsymbol{h}^{(k)} = \arg\max_{\boldsymbol{h}} \left\{Q(\boldsymbol{h},\boldsymbol{h}^{(k-1)})\right\} \tag{15}$$

It is shown that, subject to some mild conditions, this algorithm converges to a local maximum of the marginal likelihood.

For the Gaussian case, noting that

$$\begin{aligned}-\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g}) &= c+\tfrac{1}{2v_\epsilon}J_{\text{MAP}}(\boldsymbol{f},\boldsymbol{h})\\ &= c+\tfrac{1}{2v_\epsilon}\left[\|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|_2^2)+\lambda_f\|\boldsymbol{f}\|_2^2+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\right]\end{aligned} \tag{16}$$

where c is a constant which will be eliminated since after. Now, looking at the expression of $\langle -\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle_{p(\boldsymbol{f}|\boldsymbol{h}^{(k-1)},\boldsymbol{g})}$ and keeping only the terms depending on $\boldsymbol{h}$ we obtain:

$$\begin{aligned}&<-\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})>\\ &=<\|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|_2^2)>+\lambda_f<\|\boldsymbol{f}\|_2^2)>+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\left[\|\boldsymbol{g}\|^2-2\boldsymbol{g}'\boldsymbol{H}*<\boldsymbol{f}>+<\|\boldsymbol{H}\boldsymbol{f}\|_2^2>\right]\\ &+\lambda_f<\|\boldsymbol{f}\|_2^2)>+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\left[\|\boldsymbol{g}\|^2-2\boldsymbol{g}'\boldsymbol{H}<\boldsymbol{f}>+\text{Tr}\left\{\boldsymbol{H}<\boldsymbol{f}\boldsymbol{f}'>\boldsymbol{H}'\right\}\right]\\ &+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\Bigg[\|\boldsymbol{g}\|^2-2\boldsymbol{g}'\boldsymbol{H}<\boldsymbol{f}>\\ &+\text{Tr}\left\{\boldsymbol{H}(\text{Cov}\left[\boldsymbol{f}\right]+<\boldsymbol{f}><\boldsymbol{f}'>)\boldsymbol{H}'\right\}\Bigg]+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\tfrac{1}{v_\epsilon}\Bigg[\|\boldsymbol{g}\|^2-2\boldsymbol{g}'<\boldsymbol{F}>\boldsymbol{h}+\|<\boldsymbol{F}>\boldsymbol{h}\|^2+\\ &\text{Tr}\left\{\boldsymbol{H}\text{Cov}\left[\boldsymbol{f}\right]\boldsymbol{H}'\right\}\Bigg]+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\\ &=\left[\|\boldsymbol{g}-<\boldsymbol{F}>\boldsymbol{h}\|^2+\|\boldsymbol{D}_f\boldsymbol{h}\|^2\right]+\lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|_2^2\end{aligned} \tag{17}$$

where we assumed that $\text{Tr}\left\{\boldsymbol{H}\text{Cov}\left[\boldsymbol{f}\right]\boldsymbol{H}'\right\}$ can be written as $\|\boldsymbol{D}_f\boldsymbol{h}\|^2$ which is possible. Then, with this relation, it is easy to write down the Bayesian EM algorithm as follows:

$$\left\{\begin{array}{l}\textbf{Algorithm BEM:}\\ \textbf{Initialization:}\\ \boldsymbol{h}^{(0)}=\boldsymbol{h}_0,\quad \boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(0)})\\ \textbf{Iterations:}\\ \boldsymbol{\Sigma}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\\ \boldsymbol{f}^{(k)}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g}\\ \boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f}^{(k-1)})\\ \text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_h'\boldsymbol{h}\|_2^2\\ \boldsymbol{h}^{(k)}=(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\\ \boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(k-1)})\end{array}\right. \tag{18}$$

II-C. Algorithm VBA

The third approach which, in some way, generalizes BEM, is the VBA method which consists in approximating the joint posterior law $p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})$ by a separable one $q(\boldsymbol{f},\boldsymbol{h})=q_1(\boldsymbol{f})\,q_2(\boldsymbol{h})$ by minimizing the Kullback-Leibler $\text{KL}(q:p)$. It is easily shown that the alternate optimization of this criterion results to the following iterative algorithm:

- E step: Compute the expected values $\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle_{q_1}$ and $\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle_{q_2}$ and deduce:

$$\left\{\begin{array}{l}q_1(\boldsymbol{f}|\boldsymbol{h}^{(k)})\propto\exp\left\{\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle_{q_2(\boldsymbol{h}|\boldsymbol{f}^{(k-1)})}\right\}\\ q_2(\boldsymbol{h}|\boldsymbol{f}^{(k)})\propto\exp\left\{\langle\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle_{q_1(\boldsymbol{f}|\boldsymbol{h}^{(k-1)})}\right\}\end{array}\right. \tag{19}$$

- M step:

$$\left\{\begin{array}{l}\boldsymbol{f}^{(k+1)}=\arg\max_{\boldsymbol{f}}\left\{q_2(\boldsymbol{h}|\boldsymbol{h}^{(k)})\right\}\\ \boldsymbol{h}^{(k+1)}=\arg\max_{\boldsymbol{h}}\left\{q_2(\boldsymbol{h}|\boldsymbol{f}^{(k)})\right\}\end{array}\right. \tag{20}$$

Here too, it can be shown that with the Gaussian priors, we obtain the following algorithm:

$$\left\{\begin{array}{l}\textbf{Algorithm VBA:}\\ \textbf{Initialization:}\\ \boldsymbol{h}^{(0)}=\boldsymbol{h}_0;\quad \boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(0)})\\ \boldsymbol{\Sigma}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\\ \boldsymbol{f}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g}\\ \boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f})\\ \text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_h'\boldsymbol{h}\|_2^2\\ \textbf{Iterations:}\\ \boldsymbol{\Sigma}_h=v_\epsilon(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\\ \boldsymbol{h}^{(k)}=(\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+v_\epsilon\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g}\\ \boldsymbol{H}=\texttt{Convmtx}(\boldsymbol{h}^{(k-1)})\\ \text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\}=\|\boldsymbol{D}_f'\boldsymbol{f}\|_2^2\\ \boldsymbol{\Sigma}_f=v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I}+v_\epsilon\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}=\|\boldsymbol{D}_f'\boldsymbol{f}_h\|_2^2\\ \boldsymbol{f}^{(k)}=(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{I}+v_\epsilon\boldsymbol{D}_f'\boldsymbol{D}_f)^{-1}\boldsymbol{H}'\boldsymbol{g}\\ \boldsymbol{F}=\texttt{Convmtx}(\boldsymbol{f}^{(k-1)})\\ \text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_h'\boldsymbol{f}\|_2^2\end{array}\right. \tag{21}$$

II-D. Comparison between JMAP, BEM and VBA

Comparing the three algorithms JMAP (11), BEM (18) and VBA (21), we can make the following remarks:

- In Joint MAP, there is no need to matrix inversion. At each step, we can find $\boldsymbol{f}^{(k)}$ and $\boldsymbol{h}^{(k)}$ using an optimization algorithm.
- In BEM, at each step, wee need to compute $\boldsymbol{\Sigma}_f$ and do the matrix decomposition $\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_h'\boldsymbol{h}_h\|_2^2$. This is a very costly operation due to the size of the matrix $\boldsymbol{H}'$ and the matrix $\boldsymbol{\Sigma}_f$.
- In VBA, at each step, wee need to compute $\boldsymbol{\Sigma}_f$ and do the matrix decomposition $\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\}=\|\boldsymbol{D}_h'\boldsymbol{h}_h\|_2^2$ and also to compute $\boldsymbol{\Sigma}_f$ and do the

matrix decomposition $\mathrm{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h\boldsymbol{F}'\right\} = \|\boldsymbol{D}_f'\boldsymbol{f}_h\|_2^2$. There are two very costly operations.

For practical applications, we have to write specialized algorithm taking account of the particular structures of the matrix operators $\boldsymbol{H}$ and $\boldsymbol{F}$. In particular, in Blind deconvolution, these matrices are Toeplitz (or Block-Toeplitz) and we can approximate them with appropriate circulant (or Bloc-circulant) matrices and use the Fast Fourier Transform (FFT) to write appropriate algorithms.

III. JMAP, BEM AND VBA WITH A STUDENT-T PRIOR

As we are, in general, looking for a sharp input signal, a Gaussian prior is not very appropriate. We may use any sparsity enforcing priors. Between those prior law, one is very interesting, the Student-t prior with its Infinite Gaussian Mixture (IGM) property:

$$\mathcal{T}(f_j|\nu,\mu_j,v_f) = \int_0^\infty \mathcal{N}(f_j|\mu_j,z_j^{-1}v_f)\,\mathcal{G}(z_j|\nu/2,\nu/2)\,\mathrm{d}z_j \tag{22}$$

where

$$\mathcal{N}(f_j|\mu_j,z_j^{-1}v_f) = (\tfrac{2\pi v_f}{z_j})^{-1/2}\exp\left\{-\tfrac{1}{2v_f}z_j(f_j-\mu_j)^2\right\} \tag{23}$$

and

$$\mathcal{G}(z_j|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}z_j^{\alpha-1}\exp\{-\beta z_j\}. \tag{24}$$

This property can be used to propose a hierarchical prior structure which can be used to propose a hierarchical graphical generating model for the observed signal $\boldsymbol{g}$ which can be summarized as follows:

$$\left\{\begin{array}{l} p(\boldsymbol{\epsilon}|v_\epsilon) = \mathcal{N}(\boldsymbol{\epsilon}|0,v_\epsilon\boldsymbol{I}) \\ p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{h},v_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{h}*\boldsymbol{f},v_\epsilon\boldsymbol{I}), \\ p(\boldsymbol{h}|v_h) = \mathcal{N}(\boldsymbol{h}|0,v_h(\boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}) \\ p(\boldsymbol{f}|\boldsymbol{z},v_f) = \mathcal{N}(\boldsymbol{f}|0,v_f\boldsymbol{Z}^{-1}) \\ \quad \text{with } \boldsymbol{Z} = \mathrm{Diag}[z_1,\cdots,z_N] \\ p(\boldsymbol{z}|\alpha,\beta) = \prod_{j=1}^N \mathcal{G}(z_j|\alpha,\beta) \end{array}\right. \tag{25}$$

Then, from the expression of the joint posterior:

$$\begin{array}{l} p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h}|\boldsymbol{g}) \\ \propto p(\boldsymbol{g}|\boldsymbol{h},\boldsymbol{f})\,p(\boldsymbol{h}|v_h)\,p(\boldsymbol{f}|\boldsymbol{z},v_f)\,p(\boldsymbol{z}|\alpha,\beta) \\ \propto \mathcal{N}(\boldsymbol{g}|\boldsymbol{h}*\boldsymbol{f},v_\epsilon\boldsymbol{I})\,\mathcal{N}(\boldsymbol{h}|0,v_h(\boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}) \\ \quad \mathcal{N}(\boldsymbol{f}|0,v_f\boldsymbol{Z}^{-1})\prod_{j=1}^N\mathcal{G}(z_j|\alpha,\beta) \\ \propto \exp\left\{-\frac{1}{v_\epsilon}J_{\mathrm{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h})\right\} \end{array} \tag{26}$$

we can deduce the JMAP criterion:

$$\begin{array}{l} J_{\mathrm{MAP}}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{h}) = \|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2 + \lambda_h\|\boldsymbol{C}_h\boldsymbol{h}\|^2 \\ \quad +\lambda_f\|\boldsymbol{Z}^{1/2}\boldsymbol{f}\|^2 + 2v_\epsilon\left[\sum_{j=1}^N(\alpha-1)\ln z_j + \beta z_j\right] \end{array} \tag{27}$$

Using this expression, we can obtain easily the necessary developments to describe the the algorithms JMAP, EM and VBA for which, we added the appendix 2 to distinguish them from the Gaussian models of the last section.

$$\left\{\begin{array}{l} \textbf{Algorithm JMAP2:} \\ \textbf{Initialization:} \\ \boldsymbol{h}^{(0)} = \boldsymbol{h}_0, \quad \boldsymbol{H} = \mathtt{Convmtx}(\boldsymbol{h}^{(0)}), \quad \boldsymbol{z}^{(0)} = 1 \\ \textbf{Iterations:} \\ \boldsymbol{f}^{(k)} = (\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1}\boldsymbol{H}'\boldsymbol{g} \\ \widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j} \\ \quad \text{with} \quad \widehat{\alpha}_j = \frac{1}{2}+\alpha \quad \text{and} \\ \qquad \widehat{\beta}_j = \beta + \frac{1}{2} + \|\boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2 \\ \boldsymbol{F} = \mathtt{Convmtx}(\boldsymbol{f}^{(k-1)}) \\ \boldsymbol{h}^{(k)} = (\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h)^{-1}\boldsymbol{F}'\boldsymbol{g} \\ \boldsymbol{H} = \mathtt{Convmtx}(\boldsymbol{h}^{(k-1)}) \end{array}\right. \tag{28}$$

where $\lambda_f = \frac{v_f}{v_\epsilon}$ and $\lambda_h = \frac{v_h}{v_\epsilon}$.

Following the same approach and finding the expressions of $\langle -\ln p(\boldsymbol{f},\boldsymbol{h}|\boldsymbol{g})\rangle$ with respect to the marginals $p(\boldsymbol{f}|\boldsymbol{h},\boldsymbol{g})$ and $p(\boldsymbol{h}|\boldsymbol{f},\boldsymbol{g})$, we obtain the necessary relations for BEM and VBA with the proposed hierarchical IGM prior. These two algorithms are summarized in the following:

$$\left\{\begin{array}{l} \textbf{Algorithm BEM2:} \\ \textbf{Initialization:} \\ \boldsymbol{h}^{(0)} = \boldsymbol{h}_0, \quad \boldsymbol{H} = \mathtt{Convmtx}(\boldsymbol{h}^{(0)}), \quad \boldsymbol{z}^{(0)} = 1 \\ \textbf{Iterations:} \\ \boldsymbol{\Sigma}_f = v_\epsilon(\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1} \\ \boldsymbol{f}^{(k)} = (\boldsymbol{H}'\boldsymbol{H}+\lambda_f\boldsymbol{Z})^{-1}\boldsymbol{H}'\boldsymbol{g} \\ \boldsymbol{F} = \mathtt{Convmtx}(\boldsymbol{f}^{(k-1)}) \\ \mathrm{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f\boldsymbol{H}'\right\} = \|\boldsymbol{D}_h'\boldsymbol{h}_h\|_2^2 \\ \widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j} \\ \quad \text{with} \quad \widehat{\alpha}_j = \frac{1}{2}+\alpha \quad \text{and} \\ \qquad \widehat{\beta}_j = \beta + \frac{1}{2} + \boldsymbol{g}-\boldsymbol{h}*\boldsymbol{f}\|^2 \\ \boldsymbol{h}^{(k)} = (\boldsymbol{F}'\boldsymbol{F}+\lambda_h\boldsymbol{C}_h'\boldsymbol{C}_h+\boldsymbol{D}_h'\boldsymbol{D}_h)^{-1}\boldsymbol{F}'\boldsymbol{g} \\ \boldsymbol{H} = \mathtt{Convmtx}(\boldsymbol{h}^{(k-1)}) \end{array}\right. \tag{29}$$

$$
\left\{
\begin{array}{l}
\textbf{Initialization:} \\
\boldsymbol{h}^{(0)} = \boldsymbol{h}_0; \quad \boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(0)}), \quad \boldsymbol{z}^{(0)} = 1; \\
\boldsymbol{\Sigma}_f = v_\epsilon (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1} \\
\boldsymbol{f} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I})^{-1} \boldsymbol{H}'\boldsymbol{g} \\
\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}) \\
\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f \boldsymbol{H}'\right\} = \|\boldsymbol{D}_h' \boldsymbol{h}_h\|_2^2 \\
\textbf{Iterations:} \\
\boldsymbol{\Sigma}_h = v_\epsilon (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h' \boldsymbol{C}_h + \boldsymbol{D}_h' \boldsymbol{D}_h)^{-1} \\
\boldsymbol{h}^{(k)} = (\boldsymbol{F}'\boldsymbol{F} + \lambda_h \boldsymbol{C}_h' \boldsymbol{C}_h + v_\epsilon \boldsymbol{D}_h' \boldsymbol{D}_h)^{-1} \boldsymbol{F}'\boldsymbol{g} \\
\boldsymbol{H} = \texttt{Convmtx}(\boldsymbol{h}^{(k-1)}) \\
\text{Tr}\left\{\boldsymbol{F}\boldsymbol{\Sigma}_h \boldsymbol{F}'\right\} = \|\boldsymbol{D}_f' \boldsymbol{f}_h\|_2^2 \\
\boldsymbol{\Sigma}_f = v_\epsilon (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I} + v_\epsilon \boldsymbol{D}_h' \boldsymbol{D}_h)^{-1} = \|\boldsymbol{D}_f' \boldsymbol{f}_h\|_2^2 \\
\boldsymbol{f}^{(k)} = (\boldsymbol{H}'\boldsymbol{H} + \lambda_f \boldsymbol{I} + v_\epsilon \boldsymbol{D}_f' \boldsymbol{D}_f)^{-1} \boldsymbol{H}'\boldsymbol{g} \\
\boldsymbol{F} = \texttt{Convmtx}(\boldsymbol{f}^{(k-1)}) \\
\text{Tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}_f \boldsymbol{H}'\right\} = \|\boldsymbol{D}_h' \boldsymbol{h}_h\|_2^2 \\
\widehat{z}_j^{(k)} = \frac{\widehat{\beta}_j}{\widehat{\alpha}_j} \\
\quad \text{with} \quad \widehat{\alpha}_j = \frac{1}{2} + \alpha \quad \text{and} \\
\qquad \widehat{\beta}_j = \beta + \frac{1}{2} + \|\boldsymbol{g} - \boldsymbol{h} * \boldsymbol{f}\|^2
\end{array}
\right. \tag{30}
$$

As we can see the complexity of these algorithms with the Student-t prior compared to the equivalent Gaussian cases are not too much different. However, in real applications, still we need to do simplification. One way to go ahead is to choose a full separable approximation.

IV. CONCLUSIONS

In this paper, we considered the blind deconvolution problem in a Bayesian framework. Then, first, using Gaussian priors, we compared three main algorithms based on JMAP, BEM and VBA giving some insight and comparison between these methods using a Gaussian prior for both IRF $\boldsymbol{h}$ and the input signal $\boldsymbol{f}$. Then, using a Gaussian prior for the IRF but a Student-t prior for the images to enhance or to account for the sparsity structure of the input signal, again, we compared those three methods and their corresponding algorithms. The main conclusion can be summarized as follows:

- In JMAP, at each iteration, only the value of the previous estimate of h and f are transferred to the next step. So, in one hand, the computational cost of this approach is low because there is no need for matrix inversion. At the other hand, we do not know a lot about the convergence and the properties of the obtained solution.

- In EM, the value of the IRF h is transferred, but for $\boldsymbol{f}$, its expected value and its uncertainty (covariance matrix) are transferred for the next iteration computation of $\boldsymbol{h}$. So, in one hand, the computational cost of this approach is higher than JMAP because here we need the computation of $\boldsymbol{\Sigma}_f$ which needs a huge matrix inversion. At the other hand, we know a little more about the convergence (to local maximum of the marginal likelihood) and the properties of the obtained solution.

- In VBA, at each step, not only the values of the estimates, but also theirs uncertainties (in fact the whole approximated marginal laws) are transferred. So, in one hand, the computational cost of this approach is still higher than BEM because here we need the computation of $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_h$ which needs two huge dimensional matrix inversion. At the other hand, not only we get the estimates of $\boldsymbol{f}$ and $\boldsymbol{h}$ but also their approximated marginals $q_1(\boldsymbol{f})$ and $q_2(\boldsymbol{h})$, from which, we can compute any statistical properties of these estimates.

V. REFERENCES

[1] T.J. Deeming, "Deconvolution and reflection coefficient estimation using a generalized minimum entropy principle," *Proc. 51st Ann. Meeting of the Soc. of Exploration Geophysicists*, p. 1, 1981.

[2] D.L. Donoho, "On minimum entropy deconvolution," *Applied Time Series Analysis II*, p. 1, 1981.

[3] G. Wahba, "Constrained regularization for ill-posed linear operator equations with applications in meteorology and medicine," *in Statistical Decision Theory and Related Topics III New-York: Academic*, pp. 383–417, 1982.

[4] G. Demoment and R. Reynaud, "Fast minimum-variance deconvolution," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-33, pp. 1324–1326, 1985.

[5] T. F. Chan and C. K. Wong, "Total variation blind deconvolution.," *IEEE Trans Image Process*, vol. 7, no. 3, pp. 370–375, 1998.

[6] T. J. Holmes, "Blind deconvolution of quantum-limited incoherent imagery: maximum-likelihood approach.," *J Opt Soc Am A*, vol. 9, no. 7, pp. 1052–1061, Jul 1992.

[7] S. U. Pillai and B. Liang, "Blind image deconvolution using a robust gcd approach.," *IEEE Trans Image Process*, vol. 8, no. 2, pp. 295–301, 1999.

[8] M. K. Ng, R. J. Plemmons, and S. Qiao, "Regularization of rif blind image deconvolution.," *IEEE Trans Image Process*, vol. 9, no. 6, pp. 1130–1134, 2000.

[9] Stuart Jefferies, Kathy Schulze, Charles Matson, Kurt Stoltenberg, and E. Keith Hege, "Blind deconvolution in optical diffusion tomography.," *Opt Express*, vol. 10, no. 1, pp. 46–53, Jan 2002.

[10] Simone Fiori, "Fast fixed-point neural blind-deconvolution algorithm.," *IEEE Trans Neural Netw*, vol. 15, no. 2, pp. 455–459, Mar 2004.

[11] Liang Wei, Lei Hua-ming, and Que Pei-wen, "Sparsity enhancement for blind deconvolution of ultrasonic signals in nondestructive testing application.," *Rev Sci Instrum*, vol. 79, no. 1, pp. 014901, Jan 2008.

[12] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman, "Understanding blind deconvolution algo-

rithms.," *IEEE Trans Pattern Anal Mach Intell*, Jul 2011.

[13] Haiyong Liao and Michael K Ng, "Blind deconvolution using generalized cross-validation approach to regularization parameter estimation.," *IEEE Trans Image Process*, vol. 20, no. 3, pp. 670–680, Mar 2011.

[14] Andrs G Marrugo, Michal Sorel, Filip Sroubek, and Mara S Milln, "Retinal image restoration by means of blind deconvolution.," *J Biomed Opt*, vol. 16, no. 11, pp. 116016, Nov 2011.

[15] Mohammad Rostami, Oleg Michailovich, and Zhou Wang, "Image deblurring using derivative compressed sensing for optical imaging application.," *IEEE Trans Image Process*, vol. 21, no. 7, pp. 3139–3149, Jul 2012.

[16] Filip Sroubek, Gabriel Cristbal, and Jan Flusser, "A unified approach to superresolution and multichannel blind deconvolution.," *IEEE Trans Image Process*, vol. 16, no. 9, pp. 2322–2332, Sep 2007.

[17] Filip Sroubek and Peyman Milanfar, "Robust multichannel blind deconvolution via fast alternating minimization.," *IEEE Trans Image Process*, vol. 21, no. 4, pp. 1687–1700, Apr 2012.

[18] Luxin Yan, Houzhang Fang, and Sheng Zhong, "Blind image deconvolution with spatially adaptive total variation regularization.," *Opt Lett*, vol. 37, no. 14, pp. 2778–2780, Jul 2012.

[19] Xiang Zhu and Peyman Milanfar, "Removing atmospheric turbulence via space-invariant deconvolution.," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 1, pp. 157–170, Jan 2013.

[20] J. Idier and Y. Goussard, "Markov modeling for bayesian multi-channel deconvolution," *Proceedings of IEEE ICASSP*, p. 2, 1990.

[21] Sevket Babacan, Jingnan Wang, Rafael Molina, and Aggelos Katsaggelos, "Bayesian blind deconvolution from differently exposed image pairs.," *IEEE Trans Image Process*, vol. 19, no. 11, Nov 2010.

[22] A Mohammad-Djafari, "Bayesian approach with prior models which enforce sparsity in signal and image processing," *EURASIP Journal on Advances in Signal Processing*, vol. Special issue on Sparse Signal Processing, pp. 2012:52, 2012.

[23] S. Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos, "Variational bayesian blind deconvolution using a total variation prior.," *IEEE Trans Image Process*, vol. 18, no. 1, pp. 12–26, Jan 2009.

[24] Michael M Bronstein, Alexander M Bronstein, Michael Zibulevsky, and Yehoshua Y Zeevi, "Blind deconvolution of images using optimal sparse representations.," *IEEE Trans Image Process*, vol. 14, no. 6, pp. 726–736, Jun 2005.

[25] Rafael Molina, Javier Mateos, and Aggelos K Katsaggelos, "Blind deconvolution using a variational approach to parameter, image, and blur estimation.," *IEEE Trans Image Process*, vol. 15, no. 12, pp. 3715–3727, Dec 2006.

[26] Se Un Park, Nicolas Dobigeon, and Alfred O Hero, "Semi-blind sparse image reconstruction with application to mrfm.," *IEEE Trans Image Process*, vol. 21, no. 9, pp. 3838–3849, Sep 2012.

[27] Zhimin Xu and Edmund Y Lam, "Maximum a posteriori blind image deconvolution with huber-markov random-field regularization.," *Opt Lett*, vol. 34, no. 9, pp. 1453–1455, May 2009.

[28] L. Blanco and L. M. Mugnier, "Marginal blind deconvolution of adaptive optics retinal images.," *Opt Express*, vol. 19, no. 23, pp. 23227–23239, Nov 2011.

m-Cardiac System for Real-time ECG Monitoring Using an RN-XV WiFly Module

Nazrul Anuar Nayan, Susamraine A/L Yi Lak, Nur Sabrina Risman
Department of Electrical, Electronic and Systems Engineering
Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia
Bangi Selangor, Malaysia
nazrul@ukm.edu.my

***Abstract*—Cardiovascular diseases are the leading cause of mortality worldwide. Research has shown that close monitoring can help improve the health of cardiovascular patients. Real-time monitoring of electrocardiograph (ECG) data can be performed with the advancement of wireless technology. This paper discusses the development and testing of a low-cost ECG monitoring system that uses a smartphone application. This system utilizes an ECG sensor connected to an Arduino UNO microcontroller and an ultra-low power RN-XV WiFly module for data communication. Real-time ECG signals are displayed on a smartphone and can also be stored in cloud storage to provide references for doctors. The system has a simple architecture and is easy to set up for ECG monitoring.**

Keywords—cardiovascular; wireless technology; ECG monitoring system; RN-XV WiFly; Arduino microcontroller

I. INTRODUCTION

The prevalence of cardiovascular diseases has recently increased to the point where they have become the leading cause of death worldwide. According to the 2012 World Health Organization statistics report [1], cardiovascular diseases account for the largest proportion of deaths from non-communicable diseases (48%). The most common reason for critical delays in medical treatment is the lack of early warning and patient unawareness. The electrocardiograph (ECG) sensor has become one of the most commonly used diagnostic tests for monitoring heart activities. The accuracy of ECG depends on the condition being tested. In normal practice, ECG leads are attached to the body while the patient lies flat on a bed. For high-risk cardiac patients, ECG signals are the obvious data that should be collected continuously and given priority over other sensor data. Storing ECG signals for further analysis by cardiologists is also important [2]. What doctors actually prefer is to constantly monitor these parameters such that data regarding patient history and daily changes in condition are always available. When such findings and data points are accessible, early intervention can be made available to patients [3].

Smartphones can be used to find information, purchase items, or make video calls through wireless networks. Many applications can now be run in smartphones. Reference [4] concluded that long term-Evolution (LTE) and LTE-A are good candidates for delivering biomedical data from the smartphones down to the recipients.

This study presents a low-cost ECG monitoring system that uses a smartphone application. The system is intended for patients with a known cardiovascular disease who require round-the-clock monitoring. The proposed system is a portable device that is easy to use on patients. In addition, patients can upload ECG data to a cloud database, which can be used by doctors for future references. Such data can help improve diagnoses, save time for doctors, and save the lives of patients.

II. PREVIOUS DEVELOPMENT ON WIRELESS ECG MONITORING

Observational studies conducted in [5] suggest that telemonitoring (either used alone or as part of a multidisciplinary approach) may decrease hospitalizations and readmission rates among patients with heart failure. Reference [6] conducted a randomized controlled trial to test the effect of 3 months of patient care via home monitoring. This trial collected 12-lead ECG data during video consultations with clinic staff. The authors reported improved patient outcomes. The ECG signal can be transmitted to smartphones through Bluetooth IEEE 802.15.1 [7]. The system can efficiently detect and transmit high-quality ECG waves. This application can run on smartphones wherein ECG signals are plotted with body temperature and blood pressure. This system can also track patient location. The functions of the software can be improved by adding some algorithms to propagate diagnostic ECG waves. The disadvantage of this system is its high power consumption, which is mainly attributed to the type of microcontroller used and to Bluetooth.

A Wearable Mobile Electrocardiogram Monitoring System (WMES) mainly consists of a wearable ECG acquisition device, a mobile phone with global positioning system, and healthcare server [8]. With the wireless communication technique, WMEMS can monitor patient's heart rate continuously anywhere in the globe if they are under GSM's coverage cellular network. Therefore, the WMEMS provides a good system prototype for ECG telemedicine applications.

Reference [9] stated that the iPhone, iPod Touch, and iPad have been accepted as target media for mobile health (m-health). Many developers of m-health applications have chosen iOS devices to provide convenient tools to consumers. This situation has been proven by the increasing number of m-health applications. Such applications exhibit great potential in

public health care and health education. Advancements in mobile and wireless health care solutions contribute to different aspects of our lives ranging from diagnosis to treatment of various health problems such as cardiovascular diseases in [9]. If any abnormalities are found, the patient will be notified through an audible alarm and first aid techniques will be shown to the patient in the phone's display [10]. Android applications are also part of the diverse groups of products that can provide health care solutions. These applications have also been adapted as references for developing Bluetooth applications in the Android platform. Reference [11] proposed that telemedicine can be applied to a greater extent in cardiology wherein ECG serves as a primary tool. Patient vital signs, such as ECG, heart rate, respiratory rate, temperature, and peripheral capillary oxygen saturation values, are captured and entered into the database. Then, the data are uploaded to a web-based server, which sends them to doctor phones with Android technology. Clifton et al. explained that several technologies that promise to significantly improve patient care are currently available or are being developed [12]. Vital sign recordings can be enhanced by automated transmission of the measured parameters to an electronic patient record. Therefore, these technologies should be carefully executed because poor-quality deployment can lead to bad patient care.

III. METHODOLOGY

Fig. 1. Complete m-Cardiac system architecture

The proposed m-cardiac system is shown in Fig. 1. The system consists of ECG electrodes with an embedded Arduino microcontroller placed in a belt under the chest and belly of a patient. The RN-XV WiFly module is used as the communication medium to transmit ECG data to mobile phones. Real-time ECG data are displayed by using a mobile application. The application then sends the data via WLAN or GPRS to a patient medical profile (PMP), i.e., a personalized cloud-based health data center. The huge amount of health data is processed by using a specific algorithm tool. This tool performs real-time classification of vital signs based on data mining techniques.

We work with a three-lead ECG sensor in our prototype. Noise, interference, and non-rest conditions of the patient can contaminate signals. This condition implies that focus should be placed on extreme ECG signals. We used Lead II (Fig. 2) in the proposed ECG monitoring system because the voltage from the right arm to the left leg provides the strongest signal as it moves across the heart. Electrode placement on the human body is shown in Fig. 3.

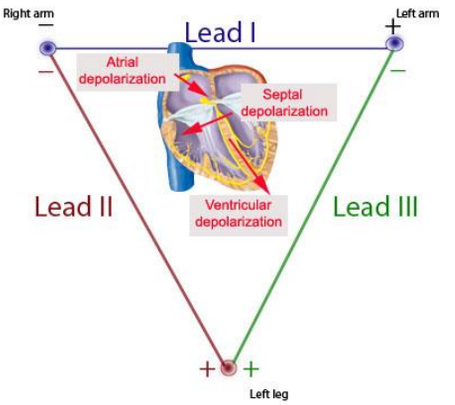


Fig. 2. ECG leads as explained in McGill Physiology Virtual Laboratory [10]

The block diagram of the proposed wireless ECG monitoring system is shown in Fig. 4. The system consists of the following: i) ECG electrodes, ii) microcontroller Arduino, iii) e-health kit, iv) the RN-XV WiFly module, v) smartphone, and vi) database.

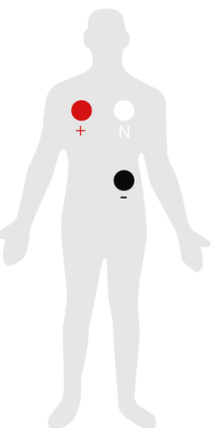


Fig. 3. Electrode placement as suggested in Libelium Comunicaciones Distribuidas [11]

A. *Hardware System*

The hardware system has a significant role in the operation of the proposed ECG monitoring system. We will provide a brief introduction to the sensor and the microcontroller in this section.

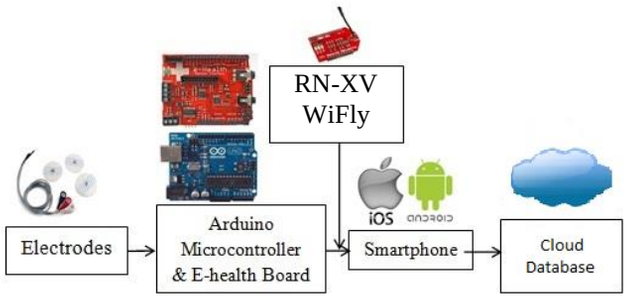


Fig. 4. Architecture of the proposed system

Arduino UNO is the main component used in the proposed system. It processes signals sent by the biosensors. This microcontroller board is based on ATMega328 and consists of 14 digital pin entries (input) and 6 analog productions (output), a 16 MHz ceramic resonator, a USB connection, a power jack, an In-Circuit Serial Programming header, and a reset button. Arduino UNO possesses the necessary features required to support the microcontroller by connecting it to a computer through a USB cable. The RN-XV WiFly is used to establish communication between the microcontroller and the smartphone. This module functions as a Wi-Fi antenna for transmitting data from the Arduino to the smartphone, as shown in Fig. 8.

B. Software System

For the software system, we use Arduino for programming and Tera Term to configure the RN-XV WiFly module. The connection properties of Tera Term are shown in Fig. 7. Before connecting the smartphone to the RN-XV WiFly module (Fig. 8), the module should be configured first. Tera Term is used to set up the IP address of the module to perform this process.

An Xbee USB adapter is connected to the RN-XV WiFly module to produce a serial connection during configuration. Configurations for Android (Fig. 4) and iOS (Fig. 5) are performed independently.

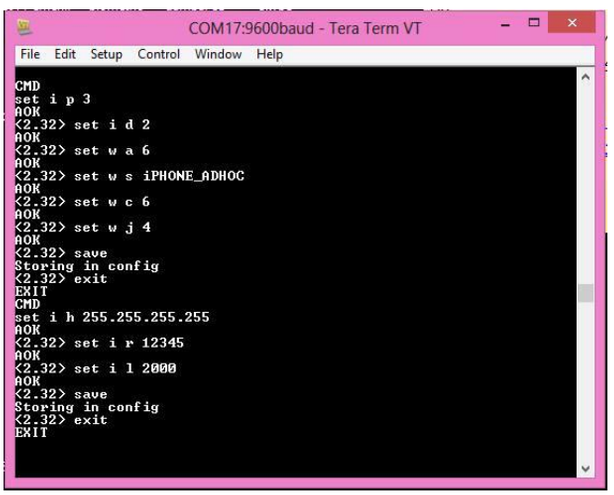


Fig. 5. iOS configuration on the RN-XV WiFly module

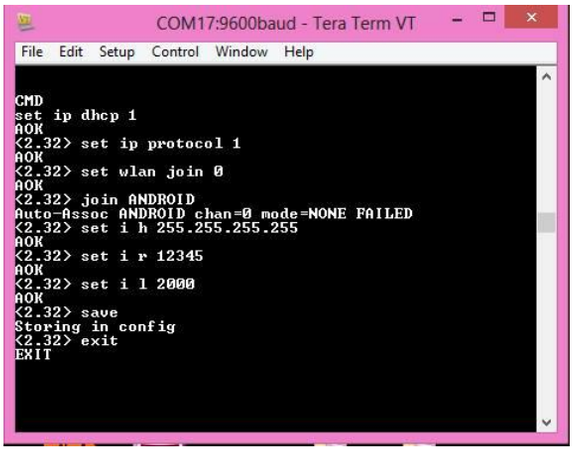


Fig. 6. Android configuration on the RN-XV WiFly module

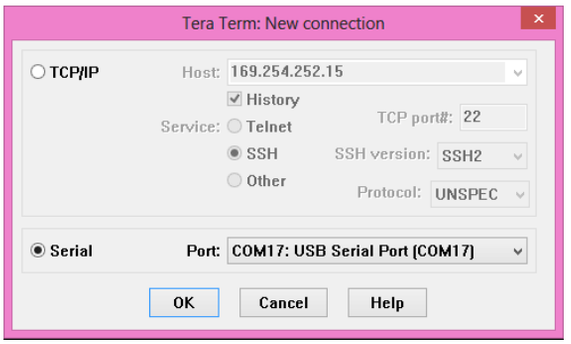


Fig. 7. Tera Term connection properties

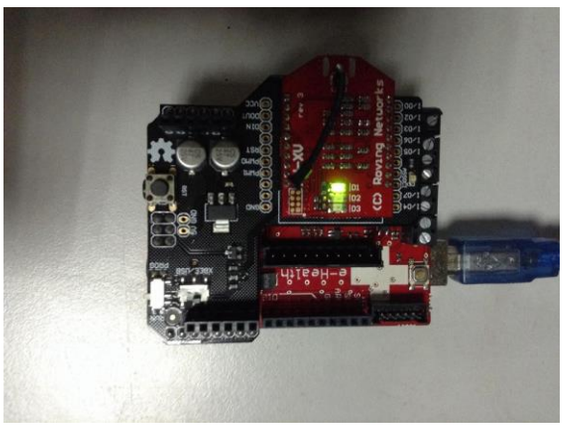


Fig. 8. The RN-XV WiFly module connected to a smartphone

IV. Results and Discussion

This chapter, which presents the results, is divided into two parts: (1) the serial monitor for the desktop and (2) data transmitted to smartphones. The output is initially tested on the serial monitor (Fig. 9) to ensure that the sensor is fully operational before being displayed on a smartphone. The smartphone processes sensor data. Raw data are then converted into ECG signals via Gaussian process regression to

eliminate noise. This algorithm will be further improved to generate ECG signal more accurately. The final ECG signal is shown on the smartphone based on Fig. 10 and Fig. 11. In Fig. 10, due to low sampling rate, the ECG signal appears inaccurately.

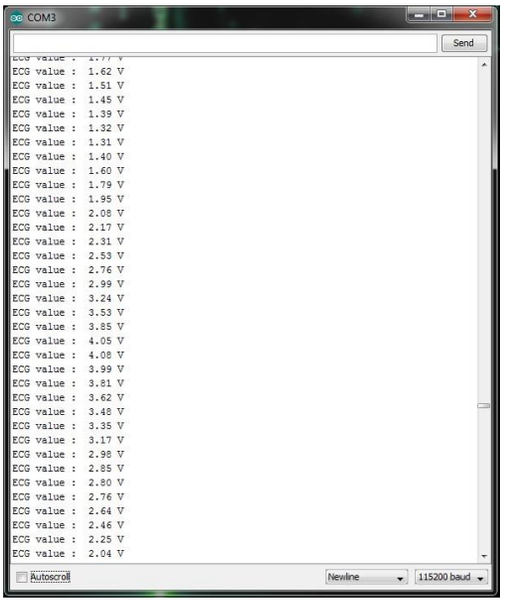


Fig. 9. ECG signals verified by using the serial monitor Arduino

Fig. 10. Output on an iOS smartphone

Fig. 11. Output on an Android smartphone

Users also have the option to store data in text form in PMP, as shown Fig. 12. Doctors can replot ECG data for a particular period that requires attention.

ECG value :	2.44 V	ECG value :	2.20 V
ECG value :	2.02 V	ECG value :	2.64 V
ECG value :	1.65 V	ECG value :	2.87 V
ECG value :	1.43 V	ECG value :	2.80 V
ECG value :	1.37 V	ECG value :	2.46 V
ECG value :	1.50 V	ECG value :	2.06 V
ECG value :	1.78 V	ECG value :	1.66 V
ECG value :	2.17 V	ECG value :	1.41 V
ECG value :	2.60 V	ECG value :	1.33 V
ECG value :	2.88 V	ECG value :	1.44 V
ECG value :	2.86 V	ECG value :	1.73 V
ECG value :	2.55 V	ECG value :	2.13 V
ECG value :	2.13 V	ECG value :	2.58 V
ECG value :	1.75 V	ECG value :	2.89 V
ECG value :	1.47 V	ECG value :	2.88 V
ECG value :	1.36 V	ECG value :	2.57 V
ECG value :	1.44 V	ECG value :	2.16 V
ECG value :	1.69 V	ECG value :	1.75 V
ECG value :	2.07 V	ECG value :	1.48 V
ECG value :	2.52 V	ECG value :	1.37 V
ECG value :	2.86 V	ECG value :	1.45 V
ECG value :	2.91 V	ECG value :	1.74 V
ECG value :	2.66 V	ECG value :	2.11 V
ECG value :	2.24 V	ECG value :	2.57 V
ECG value :	1.83 V	ECG value :	2.92 V
ECG value :	1.52 V	ECG value :	2.97 V
ECG value :	1.39 V	ECG value :	2.72 V
ECG value :	1.43 V	ECG value :	2.29 V

Fig. 12. ECG data stored in a cloud database

V. CONCLUSION

An m-cardiac wireless ECG monitoring system that uses smartphones has been developed. We have constructed a working prototype that focuses on ECG sensors. In addition, the features of the ultra-low RN-XV WiFly module as a communication medium between microcontrollers and smartphones have been described. This system helps reduce the number of times that lithium batteries should be recharged. ECG data that have been saved in the database can be retrieved by doctors for future reference. The target users for our application are patients who have suffered from a heart attack or at high risk of suffering from a heart attack. We have learned from discussions with cardiologists that these patients are worried that a heart attack may occur/reoccur, and thus, are willing to wear a monitoring device that can help reassure them of their safety. Intrusiveness is not an issue for these highly motivated patients.

ACKNOWLEDGMENT

This work is supported by Universiti Kebangsaan Malaysia ETP-2013-078 research grant.

REFERENCES

[1] Fadéla Chaib. "New data highlight increases in hypertension, diabetes incidence", World Health Organization. http://www.who.int/mediacentre/news/releases/2012/world_health_statistics_20120516/en/ 04 July 2014

[2] P. Leijdekkers, V. Gay, "Personal Heart Monitoring System Using Smart Phones To Detect Life Threatening Arrhythmias," Proc. of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06) 2006.

[3] P. Sundaram, "Patient Monitoring System Using Android Technology," International Journal of Computer Science and Mobile Computing. IJSMC, vol. 2, pp. 191-201, May 2013.

[4] Adibi S., Mabasher A., Tofigh T., "LTE networking:extending the reach for sensors in mHealth applications," Transactions on Emerging Telecommunications Technologies, vol.25, pp. 692-206, 2013.

[5] A.A.Louis, Turner T, Gretton M, "A systematic review of telemonitoring for the management of heart failure," Eur. J. Heart Failure, vol. 5, pp. 583–90, 2003.

[6] Woodend AK, Sherrard H, Fraser M, et al. "Telehome monitoring in patients with cardiac disease who are at high risk of readmission," Heart Lung J. Acute Crit. Care, vol.37, pp.36–45,2008.

[7] Noureddine Belgacem, Fethi Bereksi-Reguig, "Bluetooth Portable Device for ECG and Patient Motion Monitoring,". Nature & Technology,pp.19-23,2011.

[8] I.J. Wang, L. D. Liao., Y. T. Wang, C.Y. Chen, B.S. Lin, S. W. Lu, C.T. Lin, "A Wearable Mobile Electrocardiogram measurement device with novel dry polymer-based electrodes," TENCON 2010-2010 IEEE Region 10 Conference, pp. 379-378, 2010.

[9] C. Liu, Q. Zhu, K. A. Holroyd, E. K. Seng, "Status and Trends of Mobile-Health Applications For iOS Devices: A Developer's Perspective," The Journal of Systems and Software 84, pp. 2022-2033, 2011.

[10] J. Vijay, Sathisha M.S, Shivakumar K.M, "Android Based Portable ECG Monitor," International Journal of Engineering and Computer Science (IJECS), vol.2, pp. 1560-1567, May 2013.

[11] A. Akram, R. Javed, A. Ahmad, "Android Based ECG Monitoring System," International Journal of Science and Research (IJSR) ISSN (Online), pp. 2319-7064 Paper ID : 02013402, November 2013.

[12] D. A. Clifton, T. Bonnice, L. Tarassenko and P. Watkinson., "The Digital Patient," Clinical Medicine 2013, vol. 13, No 3 pp. 252-7, 2013.

[13] http://www.medicine.mcgill.ca/physio/vlab/cardio/ecgbasics.htm

[14] http://internetmedicine.com/2013/12/20/e-health-sensor-platform-10-apps-in-one-wow/

A Context-sensitive E-Learning Tool for Back-Propagation Neural Network

G. N. Reddy, Gurpreet Singh, and Vishnudev Vasanthan
Drayer Department of Electrical Engineering, Lamar University
Beaumont, TX, USA, gnreddy@lamar.edu
International Workshop on Artificial Intelligence IWAI-2014, Istanbul, Turkey; August 22-23, 2014: Update 2.1, 7/5/2014

Abstract -- This paper presents an e-learning tool for exploring the back-propagation neural network architecture. It is developed using MS Visual Studio / Visual C++ 2012. It is context sensitive meaning that it displays systems dynamic equations according to the context. Functionality of the developed package include: 1. At its highest-level, the software has two modes of operation: the training mode and the recall mode. 2. While in training mode, it has two sub-modes: one is for the learning and the other is for application-development. 3. In learning mode, the software generates text-output traces corresponding to each step in the top-down design of the neural network architecture. 4. The generated numeric traces have dual-usage: they can be used either for learning purposes or for generating class room tests. 5. While training at application-level, software displays only the input-output relations -- before and after the training values, as the goal here is to develop an application. 6. In training mode the software generates a cumulative error-index to monitor the progress of the network training both at the application-level as well as at the individual measurement data pair-level. 7. It enables you to include the network training termination criteria. 8. At the end of the network training, it is stores the trained network into a text-file. 9. In test or recall mode, the trained network is retrieved from the stored-file, and then it generates the network response due to a test input. 10. The software is essentially tailored for class room teaching and for creating class-room tests.

Key Words: BP Neural Network; e-Learing tool; Context-sensitive help; Modeling and simulation; Educational software.

I. INTRODUCTION

Back propagation neural network architecture is complex and it requires a good e-learning tool to master its understanding [1, 2]. Some of the commercial and open-source BP-NNA software packages include: 1. MathWorks Neural Network Toolbox [3]; 2. Back-propagation neural network software from soft112.com [4]; 3. BP-C#-program by McCaffrey [4]; 4. Neural network C#-libraries from codeproject [5]; 5. Griiffith's-trace [7]; 6. Wikipedia has excellent review on neural network software [8]. Each of the above tools have their own merits, the tool presented in this paper excels for learning and teaching. Our output trace generated is similar to the one in Griiffith's-trace [7] with exception of embedded system dynamic equations. The following sections describe the BP e-Learing tool.

II. BACK-PROPAGATION NEURAL NET ARCHITECTURE

Fig. 1 shows the architecture of the back-propagation neural network architecture with one hidden layer. One can have any number of hidden layers in the back-propagation network [1]. Typically there are two hidden layers, but one is sufficient for majority of the applications. In this software we have used one hidden layer to simplify overall network complexity. As shown, it is a multi-layer, fully-connected, feed-forward network. The three layers are: the input layer (in), the hidden-layer (hl), and the output-layer (ol).

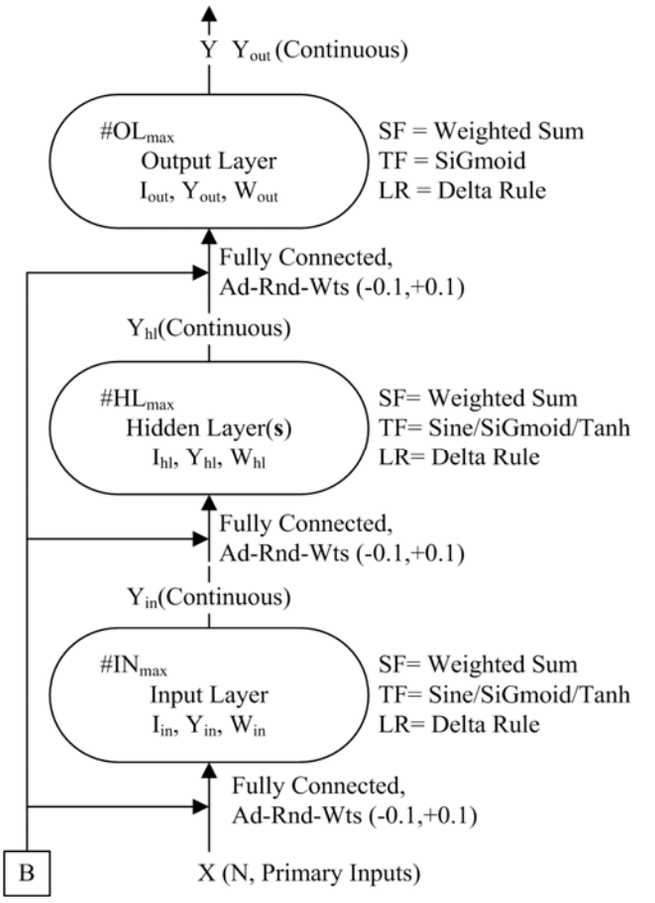


Figure 1. Back-propagation Neural Network Architecture.

The weight-matrices of the input-layers, hidden-layer, and the output-layer are correspondingly denoted as Win, Whl, and Wout. Initially all its weight matrices are initialized with small adaptive random weights (Ad-Rnd-Wts) between ± 0.1. A bias-elements B (the 0^{th}-element) is added to the input and the hidden-layer. How information is processed within each of the neural elements is specified by their neurodynamics. The neurodynamics is a combination of a summation function SF followed by a transfer function TF. All of the layers use the weighted summation function (weighted-sum). The transfer function, however, can be different for each layer. Input and the hidden-layers can have sine or sigmoid or tanh transfer functions (S/G/T). Output layer can also have above three transfer functions; however, we have fixed it as sigmoid to simplify the complexity of the network training algorithm. One can have any number of elements in each layer (INmax, HLmax, OLmax). This educational version of the software number of elements is limited to 25. For each layer the internal activations are denoted by I or sum (Iin, Ihl, Iout) and the corresponding output activations denoted as Y or act (Yin, Yhl, Yout).

III. BP NETWORK TRAINING MODE

The training of BP-network involves the following steps: *1a. Initialize the network weights* with small random weights:

$$W_{ji}^{k} = \text{rand}(-0.1, +0.1) \tag{1}$$

where, k is the layer number: k = 0, is the primary input layer PI; k = 1, is the input-layer IL; k = 2, is the hidden layer HL; and k = 3, is the output layer OL. W_{ji}^{k}, is the weight from i-th-element in (k-1)-th layer to the j-th-element in k-th layer.

1b. Set initial values: Initialize the training cycle number to zero: n = 0; and tolerable error-level to a desirable value: TssTh = typically 0.1.

2. Set initial values for each epoch of training cycle: Initialize pattern number to zero: p = 0; Global and local error-flags to zero: flagG = 0, flagL = 0. An error occurs when the computed value is different from the desired value.

3. Do a forward-pass: apply the primary input vector Xp to the network and compute the corresponding output vector Yout. The generalized equations to compute internal activations Is and corresponding output activations Ys for any layer is given by:

$$I_j^k = \sum_{i=0}^{N_k} W_{ji}^k * X_i^{k-1} \tag{2}$$

$$Y_j^k = TF_k * I_j^k \tag{3}$$

Here, X0 represents primary input vector X; TF3 is the TF for the output layer which is fixed as sigmoid in this software; and TF1 and TF2 are the TFs for the input and the hidden layer TFs which can be any one of sine or sigmoid or tanh. Individual layer internal activations corresponding outputs are given by the following sets of equations. Input layer sums Is and acts Ys are given by:

$$Iin_j = \sum_{i=0}^{PImax} Win_{ji} * X_{pi} \tag{4}$$

$$Yin_j = TF_1 * Iin_j \tag{5}$$

With TF_1 = sigmoid, Yin is given by:

$$Yin_j = \frac{1}{1+e^{-(Iinj*G)}} \tag{6}$$

where G is the gain factor and is usually between 1 and 10. Hidden layer sums and acts are given by:

$$Ihl_j = \sum_{i=0}^{Ilmax} Whl_{ji} * Yin_i \tag{7}$$

$$Yhl_j = TF_2 * Ihl_j \tag{8}$$

Output layer sums and acts are given by:

$$Iol_j = \sum_{i=0}^{Hlmax} Wol_{ji} * Yhl_i \tag{9}$$

$$Yol_j = TF_3 * Iol_j \tag{10}$$

In this software TF_3 is sigmoid.

4. Compute Tssp: find the mean square error of the current pattern:

$$Tss_p = \sqrt{\frac{1}{N_3}\sum_{j=0}^{N3}(D_{pj}^3 - Y_j^3)^2} \tag{11}$$

Here, Dp and Y are desired and the correspondingly computed values at the output layer.

If $(D_{pj}^3 - Y_j^3) > TssTh$ for any j = 1,..N_3;

then set flagL = flagG = 1; else Go to step 6.

5a. Find error functions, weight-changes; and new weights
If flagL = 1, compute error functions δs for each element; the weight-changes DWs; and the new weights W's. The error functions are needed to find weight-changes to the network. Error functions at the output-layer, with sigmoid TF, are given by:

$$\delta_j^3 = Y_j^3(1 - Y_j^3) * (D_j^3 - Y_j^3) \tag{12}$$

Here, 3 is the output-layer number. The error function is a product of gradient * error. In (12) the gradient is Y(1 – Y), and the error is (D - Y). The gradient for different TFs is different, for sigmoid it is Y(1 –Y) [1]. The error in known at the output-layer, as the desired value D and correspondingly computed value Y are known. For other lower-level layers Ys are known but not the desired values Ds.
The generic error functions for the other-layers are computed as:

$$\delta_j^k = Y_j^k\,(1-Y_j^k) * \sum_{m=1}^{N_{k+1}} W_{jm}^{k+1}\,\delta_m^{k+1} \tag{13}$$

That is, error-functions of the lower-level layers are computed from the upper-level layers. The error functions for the hidden layer elements are computed as:

$$\delta hl_j = Yhl_j\,(1-Yhl_j) * \sum_{m=1}^{OLmax} Wol_{jm}\;\delta ol_m \tag{14}$$

Here, the error of a hidden layer element is computed as the weighted summation of the output-layer error-functions. For each value of j, m varies from 1 to OLmax. The error functions for the input layer elements are computed as:

$$\delta in_j = Yin_j\,(1-Yin_j) * \sum_{m=1}^{HLmax} Whl_{jm}\;\delta hl_m \tag{15}$$

Here, the error of an input layer element is computed as the weighted summation of the hidden-layer error-functions.

5b. Find weight changes:
Find weight changes from primary inputs to the input-layer elements, DWin:

$$DWin_{ji} = \alpha * \delta in_j * X_{pi} \tag{16}$$

Here, α is the training coefficient ranging from 0.1 to 1.0; i = 0,..,PImax; and j = 1,..ILmax. Find weight changes from input-layer-elements to the hidden-layer elements, DWhl:

$$DWhl_{ji} = \alpha * \delta hl_j * Yin_i \tag{17}$$

Here, i = 0,..,ILmax; and j = 1,..HLmax.
Find weight changes from hidden-layer-elements to the output-layer elements, DWol:

$$DWol_{ji} = \alpha * \delta ol_j * Yhl_i \tag{18}$$

Here, i = 0,..,HLmax; and j = 1,..OLmax.

5c. Find new weights W(n+1):

New weights W(n+1) are computed as the old weights W(n) plus the weight-changes DW(n), n being the previous cycle and n+1 is the current cycle:

$$Win(n+1)_{ji} = Win(n)_{ji} + DWin(n)_{ji} \tag{19}$$

$$Whl(n+1)_{ji} = Whl(n)_{ji} + DWhl(n)_{ji} \tag{20}$$

$$Wol(n+1)_{ji} = Wol(n)_{ji} + DWol(n)_{ji} \tag{21}$$

6. Go to next pattern to train:
Set p = p + 1; if (p < pmax) go to Step 3.

7a. Compute TssC:
Normalized cumulative error, in cycle n, of all patterns is given as:

$$TssC(n) = \frac{1}{pmax} * \sum_{i=0}^{pmax-1} Tssp_i \tag{22}$$

7b. Go to next epoch-training:

If flagG = 0; then Go to Step 8.
Else Set n = n + 1; then Go to Step 2.

That is, repeat Steps 2 through 7 until all patterns are trained with acceptable error.

8. Write trained network to a file; Write cumulative network error TssC to a file; End network training.

IV. BP RECALL MODE

In recall or test mode, for given test input X, the network estimates Yout by successively computing activation vectors Yin, Yhl, and Yout. The estimated Yout-response will be the nearest output-match corresponding to the entered input. You can enter into the recall mode only after the network is trained. In this mode, first the trained network is read from bp-ckt.txt file which is generated at the end of training mode. For a test input X it finds the corresponding output Yout. The activation vectors Yin, Yhl, and Yout are computed as:

$$Yin_j = Y_j^1 = TF_1 * Iin_j = TF_1 * \sum_{i=0}^{PImax} Win_{ji} * X_i \tag{23}$$

$$Yhl_j = Y_j^2 = TF_2 * Ihl_j = TF_2 * \sum_{i=0}^{ILMax} Whl_{ji} * Yin_i \tag{24}$$

$$Yout_j = Y_j^3 = TF_3 * Iout_j = TF_3 * \sum_{i=0}^{HLmax} Wout_{ji} * Yhl_i \tag{25}$$

In (23), j = 1,..ILmax; in (24), j = 1,..JLmax; and in (25), j = 1,..OLmax;

V. THE BP-SOFTWARE ARCHITECTURE

Fig. 2 shows the overall architecture of the BP-software.

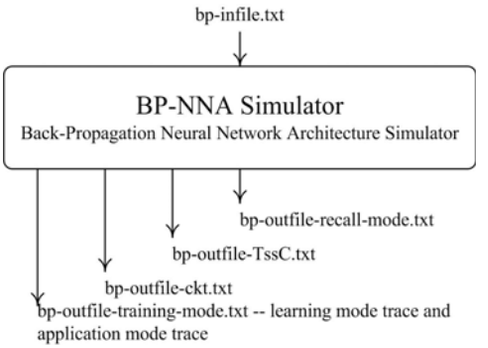


Figure 2. The BP-Software Architecture.

In Fig. 2, bp-infile.txt is the input data file; bp-outfile-training-mode.txt is the output trace generated during the network training; bp-outfile-ckt.txt is the output file that contains the trained bp-network; bp-outfile-recall-mode.txt is the output simulation trace generated during the network recall.

VI. THE BP-SIMULATOR: OUTPUT SIMULATION TRACE: TRAINING MODE

Table 1is the input data file to run the network in learning mode. The tables 2 through 7 are the generated output files in different modes of BP-simulation. The table 1 input data file contains network specification – number of elements in each layer of the BP-network; their transfer functions; percenatage of weight changes to make in each successive cycle of training; training termination criteria; and the patterns to train. Table 2 is the output simulation trace while network is in learning-training-mode; this trace is useful for learning about the BP-network; it can also be used for test generation – formulation of numeric problems on BP-NNA. Major phases of training include: 1. doing forward-pass – where activations of each element of the network are computed for a given input vector X; 2. finding the error functions for each element of the network; 3. finding weight-changes to the network weights; 4. finding new weights of the network; and finally 5. finding cumulative network error TssC. Table 3 contains the trained BP-network -- the number elements in each layer; each-layer's transfer functions; and the trained network weights. Table 4 gives the cumulative RMS-error in successive cycles of training. This is also shown in a chart-form in Fig. 3. The network is continues to be trained until the network's cumulative error TssC is less than the set threshold error TssTh. For the trained network shown in Tables 2; it took 61 cycles to train with an initial error of 0.51. Table 5 contains an input-data file to train BP-network in application-development mode with multiple-patterns. Table 6 is the corresponding output simulation trace. Here the details of training are disabled; the emphasis is placed on the application development. Table 7 contains the output simulation while BP is in recall mode or test mode. Various phases of network recall include: 1. reading and printing the trained network; 2. for a given input vector X, doing forward-pass to find Yout; and 3. prompting how to terminate the recall session.

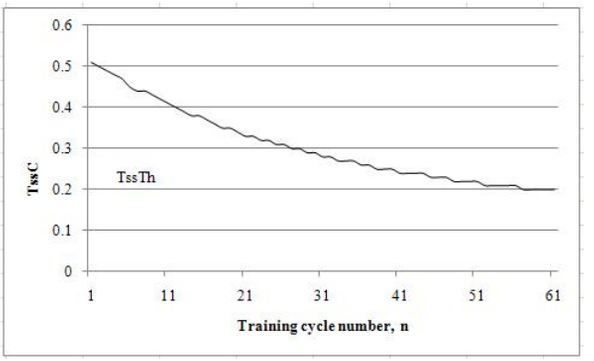


Figure 3. Cumulative network RMS-error in successive training cycles: TssC

VII. CONCLUSIONS

This paper presents Back-Propagation Neural Network Architecture software for educational use. Back-propagation is a complex neural network; e-learning tool such as the one presented in this paper will be an invaluable tool for mastery of the BP-network.

REFERENCES

[1] G. N. Reddy, A book manuscript to be polished, "Artificial Neural Networks: Theory, Computer Simulation Programs, and Applications", Department of Electrical Engineering, Lamar University, Beaumont, Texas, 2013.

[2] Stephen T. Welstead, Neural Network and Fuzzy Logic Applications in C/C++, Wiley, 1994.

[3] MathWorks.com, Neural Network Toolbox, Novi, Michigan, 2013

[4] Soft112.com, Backpropagation neural network, 2013.

[5] James McCaffrey, Neural Network Back-Propagation Using C#, 2013

[6] Codeproject.com, Andrew Kirillov, Neural Networks on C#, November 2006.

[7] Niall Griffith, Backpropagation algorthm, MIT Computer Scienec and Artificial Intelligence, CIS, Tutorial 10, 2013

[8] Wikipedia.org, Neural network software, July 2013.

Table 1. Input Data File (bp-infile.txt)

```
bp-infile.txt
4           : PImax, Number of PIs
3           : ILmax, Nuber of elemnets in IL
3           : HLmax, Nuber of elemnets in HL
3           : OLmax, Nuber of elemnets in OL
0.1         : alpha, Training coefficient
G           : iltf & hltf: S/T/G: sime/tan/sigmoid
G           : oltf, TF for OL
2.0         : Gain for all TFs: I' = I * Gain
0.2         : Tssth, Threshold error: 0.3 => 30%
1           : Pmax & the Pattern associations: Yi, Xi
1 0 0 1
1 1 0
```

Table 2. BP-Simulator Output simulation trace – in learning mode (bp-outfile-traning-mode).

```
************************************************
** Back-Propagation Neural Network Simulator    **
** Traning mode: Output Simulation trcae        **
************************************************
Reading input data from: bp-infile.txt
BP NETWORK - TRAINING MODE:
# of PEs in PI/IN/HL/OL:          4  3  3  3
LR for IN, HL, and OL:            Delta Rule
Training Coefficient (alpha):     0.10
TF Input/Hidden Layers (hltf):    Sigmoid
TF of Output Layer (oltf):        Sigmoid
Gain factor for the TFs:          2.00
Error Threshold (TssTh):          0.20
# of pattern associations (pmax): 1
             0         1         2         3         4
X[ 0]      1.00      1.00      0.00      0.00      1.00
D[ 0]                1.00      1.00      0.00
Network training starts here:
CYCLE:  1
Weight Matrices of the Network:
          PI: 0     PI: 1     PI: 2     PI: 3     PI: 4
IL: 1 -0.018000  0.034000 -0.032000 -0.100000  0.038000
IL: 2 -0.052000  0.056000  0.016000  0.024000  0.028000
IL: 3 -0.090000 -0.010000  0.062000 -0.046000  0.022000

          IL: 0     IL: 1     IL: 2     IL: 3
HL: 1  0.082000  0.090000 -0.016000 -0.046000
HL: 2 -0.028000  0.082000 -0.092000 -0.096000
HL: 3  0.006000  0.084000  0.064000 -0.058000

          HL: 0     HL: 1     HL: 2     HL: 3
OL: 1 -0.068000 -0.064000  0.090000 -0.006000
OL: 2 -0.048000  0.042000 -0.024000  0.038000
OL: 3 -0.076000  0.034000  0.098000 -0.030000

Forward-pass: Activations of each PE in the Network:
           IL: 0     IL: 1     IL: 2     IL: 3
sum        0.000     0.054     0.032    -0.078
act        1.000     0.527     0.516     0.461

           HL: 0     HL: 1     HL: 2     HL: 3
sum        0.000     0.100    -0.077     0.057
act        1.000     0.550     0.462     0.528

           OL: 1     OL: 2     OL: 3
sum       -0.065    -0.016    -0.028
act        0.468     0.492     0.486
Tssp: rms errors: p0...pmax:  0.51
TssC[cycle]: Cumulative-Tssp rms error(before)-: 0.51

CYCLE:  1 Pattern: 0
Delta-Fn for each PE in the Network:
        IL: 1         IL: 2         IL: 3
   -6.31e-005    2.20e-004    4.40e-005
        HL: 1         HL: 2         HL: 3
   -1.80e-003   -7.50e-004    1.91e-003
        OL: 1         OL: 2         OL: 3
    1.33e-001    1.27e-001   -1.21e-001
Delta Weight Matrices of the Network:
          PI: 0     PI: 1     PI: 2     PI: 3     PI: 4
IL: 1 -0.000006 -0.000006 -0.000000 -0.000000 -0.000006
IL: 2  0.000022  0.000022  0.000000  0.000000  0.000022
IL: 3  0.000004  0.000004  0.000000  0.000000  0.000004

          IL: 0     IL: 1     IL: 2     IL: 3
HL: 1 -0.000180 -0.000095 -0.000093 -0.000083
HL: 2 -0.000075 -0.000040 -0.000039 -0.000035
HL: 3  0.000191  0.000101  0.000099  0.000088

          HL: 0     HL: 1     HL: 2     HL: 3
OL: 1  0.013253  0.007287  0.006120  0.007001
OL: 2  0.012696  0.006980  0.005863  0.006706
OL: 3 -0.012142 -0.006676 -0.005607 -0.006414
CYCLE:  2

Weight Matrices of the Network:
          PI: 0     PI: 1     PI: 2     PI: 3     PI: 4
IL: 1 -0.018006  0.033994 -0.032000 -0.100000  0.037994
IL: 2 -0.051978  0.056022  0.016000  0.024000  0.028022
IL: 3 -0.089996 -0.009996  0.062000 -0.046000  0.022004

          IL: 0     IL: 1     IL: 2     IL: 3
HL: 1  0.081820  0.089905 -0.016093 -0.046083
HL: 2 -0.028075  0.081960 -0.092039 -0.096035
HL: 3  0.006191  0.084101  0.064099 -0.057912

          HL: 0     HL: 1     HL: 2     HL: 3
OL: 1 -0.054747 -0.056713  0.096120  0.001001
OL: 2 -0.035304  0.048980 -0.018137  0.044706
OL: 3 -0.088142  0.027324  0.092393 -0.036414

Forward-pass: Activations of each PE in the Network:
           IL: 0     IL: 1     IL: 2     IL: 3
sum        0.000     0.054     0.032    -0.078
act        1.000     0.527     0.516     0.461

           HL: 0     HL: 1     HL: 2     HL: 3
sum        0.000     0.100    -0.077     0.057
act        1.000     0.550     0.462     0.528

           OL: 1     OL: 2     OL: 3
sum       -0.041     0.007    -0.050
act        0.480     0.503     0.475
Tssp: rms errors: p0...pmax:  0.50
TssC[cycle]: Cumulative-Tssp rms error(before)-: 0.50
…
CYCLE: 61
Tssp: rms errors: p0...pmax:  0.20
TssC[cycle]: Cumulative-Tssp rms error(before)-: 0.20
```

Table 3. Trained network (bp-outfile-ckt.txt)

```
4  3  3  3  G  G 2

 -0.010052  0.041949 -0.032000 -0.100000  0.045949
 -0.051681  0.056319  0.016000  0.024000  0.028319
 -0.095006 -0.015006  0.062000 -0.046000  0.016994

  0.107394  0.103536 -0.002890 -0.034394
 -0.003866  0.094858 -0.079541 -0.084966
  0.042027  0.103176  0.082598 -0.041517

  0.344648  0.165523  0.283310  0.216568
  0.344183  0.260153  0.159735  0.249546
 -0.450879 -0.174549 -0.077649 -0.232241
```

Table 4. Cumulative error TssC (bp-outfile-TssC.txt)

```
Tssc: Cumulative error at each cycle
Cycle:  TssC:
1       0.51
2       0.50
3       0.49
...
61      0.20
```

Table 5. Input Data File (bp-infile.txt): application mode

```
bp-infile.txt
4          : PImax, Number of PIs
3          : ILmax, Nuber of elemnets in IL
3          : HLmax, Nuber of elemnets in HL
3          : OLmax, Nuber of elemnets in OL
0.6        : alpha, Training coefficient
G          : iltf & heltf: S/T/G: sime/tan/sigmoid
G          : oltf, TF for OL
5.0        : Gain for the TF: I' = I * Gain
0.25       : Tssth, Threshold error: 0.3 => 30%
4          : Pmax & the Pattern associations: Yi, Xi
1 0 0 1
0 1 0
0 1 1 0
1 0 1
1 1 1 0
1 0 0
0 1 1 1
0 0 1
```

```
CYCLE:574
YOL[ 0]-:   0.10   0.87   0.00
YOL[ 1]-:   0.81   0.00   1.00
YOL[ 2]-:   0.79   0.07   0.05
YOL[ 3]-:   0.23   0.01   0.84
Tssp: rms errors: p0...pmax:  0.09 0.11 0.13 0.16
TssC[cycle]: Cumulative-Tssp rms error(before)-: 0.12

Network Training Complete: Cycles: 574

Writing Network into the File bp-ckt.txt...

Writing TssC into the File bp-outfile-TssC.txt...

END BACK-PROPAGATION SIMULATION: TRAINING SESSION
```

Table 6. BP-Simulator Output simulation trace - in application mode (bp-outfile-training-mode.txt)

```
************************************************
** Back-Propagation Neural Network Simulator    **
** Traning mode: Output Simulation trcae        **
************************************************
Reading input data from: bp-infile.txt
BP NETWORK - TRAINING MODE:
# of PEs in PI/IN/HL/OL:          4  3  3  3
LR for IN, HL, and OL:            Delta Rule
Training Coefficient (alpha):     0.60
TF Input/Hidden Layers (hltf):    Sigmoid
TF of Output Layer (oltf):        Sigmoid
Gain factor for the TFs:          5.00
Error Threshold (TssTh):          0.25
# of pattern associations (pmax): 4
               0         1         2         3         4
X[ 0]      1.00      1.00      0.00      0.00      1.00
D[ 0]                0.00      1.00      0.00
X[ 1]      1.00      0.00      1.00      1.00      0.00
D[ 1]                1.00      0.00      1.00
X[ 2]      1.00      1.00      1.00      1.00      0.00
D[ 2]                1.00      0.00      0.00
X[ 3]      1.00      0.00      1.00      1.00      1.00
D[ 3]                0.00      0.00      1.00
Network training starts here:
Weight Matrices of the Network:
          PI: 0     PI: 1     PI: 2     PI: 3     PI: 4
IL: 1 -0.018000  0.034000 -0.032000 -0.100000  0.038000
IL: 2 -0.052000  0.056000  0.016000  0.024000  0.028000
IL: 3 -0.090000 -0.010000  0.062000 -0.046000  0.022000
          IL: 0     IL: 1     IL: 2     IL: 3
HL: 1  0.082000  0.090000 -0.016000 -0.046000
HL: 2 -0.028000  0.082000 -0.092000 -0.096000
HL: 3  0.006000  0.084000  0.064000 -0.058000
          HL: 0     HL: 1     HL: 2     HL: 3
OL: 1 -0.068000 -0.064000  0.090000 -0.006000
OL: 2 -0.048000  0.042000 -0.024000  0.038000
OL: 3 -0.076000  0.034000  0.098000 -0.030000
CYCLE:  1
YOL[ 0]-:   0.41   0.49   0.46
YOL[ 0]+:   0.28   0.66   0.31
DOL[ 0] :   0.00   1.00   0.00
YOL[ 1]-:   0.28   0.66   0.31
YOL[ 1]+:   0.47   0.46   0.50
DOL[ 1] :   1.00   0.00   1.00
YOL[ 2]-:   0.47   0.46   0.50
YOL[ 2]+:   0.64   0.31   0.34
DOL[ 2] :   1.00   0.00   0.00
YOL[ 3]-:   0.64   0.31   0.34
YOL[ 3]+:   0.45   0.24   0.53
DOL[ 3] :   0.00   0.00   1.00
Tssp: rms errors: p0...pmax:  0.46 0.69 0.50 0.56
TssC[cycle]: Cumulative-Tssp rms error(before)-: 0.55
```

Table 7. BP-Simulator Output simulation trace - in Recall mode (bp-outfile-recall-mode.txt)

```
BACK-PROPAGATION NETWORK - RECALL MODE:
Reading Network Weights from: bp-outfile-ckt.txt
Trained BP Network:
Number of Elements: PI/IL/HL/OL:  4  3  3  3
Transfer Function for the IN, HL:      G
Transfer Function for the OL:          G
Gain factor for Sine/siGmiod/Tanh TFs:  5.0

Weight Matrix WIL:
            0         1         2         3         4
  1 -0.117929  0.373708 -0.169350 -0.237350 -0.004736
  2 -0.290731  1.331819 -0.518565 -0.510565  0.278290
  3 -0.305431  0.594064 -0.425380 -0.533380  1.221188
Weight Matrix WHL:
            0         1         2         3
  1  0.095929 -0.120610 -0.397326 -0.594480
  2  0.319776 -0.298146 -0.848796 -0.394399
  3 -0.021513  0.066980  0.139034 -0.572369
Weight Matrix WOL:
            0         1         2         3
  1 -0.588324  0.765280 -0.185429  1.184759
  2  0.436281 -1.127587 -1.576824 -0.380790
  3 -1.586459  1.415305  2.336450  0.023212

BACK-PROPAGATION NETWORK: Reacll Mode:
Back-Propagation Network: Results of Testing:
Enter 9 9 9 9 9 to Terminate Testing:
Enter a Test Input: x1..x4
          0      1      2      3      4
 PI:  1.000  1.000  0.000  0.000  1.000
IIL:          0.251  1.319  1.510
YIL:  1.000  0.778  0.999  0.999
IHL:         -0.989 -1.154 -0.403
YHL:  1.000  0.007  0.003  0.118
IOL:         -0.444  0.379 -1.566
YOL:          0.098  0.869  0.000

Enter a Test Input: x1..x4
          0      1      2      3      4
 PI:  1.000  0.000  1.000  1.000  0.000
IIL:         -0.525 -1.320 -1.264
YIL:  1.000  0.068  0.001  0.002
IHL:          0.086  0.298 -0.018
YHL:  1.000  0.606  0.816  0.478
IOL:          0.290 -1.716  1.189
YOL:          0.810  0.000  0.997
…
Enter a Test Input: x1..x4
END BACK-PROPAGATION SIMULATION: RECALL SESSION
```

A Modular Fuzzy Logic Expert System for Autonomous Mobile Robots

G. N. Reddy, Vishnudev Vasanthan, Gurpreet Singh, and Sreelatha Maila
Drayer Department of Electrical Engineering, Lamar University
Beaumont, TX, USA, gnreddy@lamar.edu
International Workshop on Artificial Intelligence IWAI-2014, Istanbul, Turkey; August 22-23, 2014: Update 2.1, 7/520/14

Abstract -- This paper presents modular fuzzy logic expert system software, written in C++, for building real-time autonomous mobile robots. The source can be used with any microcontroller that supports C++ such as Arduino-microcontrollers. The software is a fuzzy logic expert system FLES simulator. It can be used to simulate an FLES off-line; or can be downloaded into a microcontroller to operate in real-time. The software is modularized so that it can be used with any controller or estimation type of applications. The embeddable-software has four, three-input and one-output, modules. The three input modules represent FLES sub-sections in the form its: rule-based knowledge base; input/output variables with their corresponding membership functions; and the sensed input variable values. The output module containing the generated output signals. It also generates text-based output simulation trace illustrating the detailed sequential steps executed by the inference engine. While operating off-line, the three input modules are represented by three input data files and the output module by a generated output file. While operating in real-time, the knowledge-base and the i/o membership functions are merged into the FLES C++ code; the input signals and the outputs signals are mapped to the microcontroller's i/o-ports. In real-time mode the output simulation trace can be disabled. We have embedded this software into Arduino-Uno-Due microcontrollers to built two mobile-RT-FLES: One an FLES for precise estimation of battery state of charge SoC; and the other an FLES to implement the classical controller for balancing an inverted pendulum.

Key Words: Fuzzy Logic Expert Systems; Fuzzy Logic Controllers; Embeddable Software; Micro-controllers; Autonomous mobile robots.

I. INTRODUCTION

Fuzzy logic expert systems have been used for variety applications [1, 2]. This paper is on using FLES to solve control [3, 4] and estimation problems [5]. Fuzzy logic controllers FLCs are attractive because they are robust, multiple in and multiple out, and simpler to implement. FLCs are suitable to run on PCs main-frame computers and on microcontrollers. To make them mobile one should run them on microcontrollers. If microcontrollers are used, the commercial FLES-software packages are difficult use. One needs to write their own FLES source code. The software presented is in this paper is one such code to run on the microcontrollers. It is modularly-structured so it can used with any application. The following sections describe this microcontroller-embeddable FLES software.

II. EMBEDDABLE FLES SOFTWARE FOR AUTONOMOUS MOBILE ROBOTS

Fig. 1 shows the five basic elements of the FLES: Knowledge base KB, Inference engine IE, User interface UI, Input-fuzzifier, and output-defuzzifier. The Knowledge base KB is a set of empirical rules by which the overall system behavior is summarized. In an empirical rule the input/output values are in linguistic-form such as positive-small ps, negative-large nl, and positive-medium pm. Inference engine IE is the kernel of the FLC that executes all of the sequential steps involved in the overall execution of the FLC, starting from reading the inputs till the control outputs are generated. These steps are described in the following section. Input-fuzzifier converts absolute values into linguistic values. Here the absolute input variable values are expressed as a function of the membership functions of the input variables IMFs. Output-defuzzifier converts fuzzy or linguistic values into absolute values. The computing system is the one on which the FLC-kernel is executed. For off-line applications it is usually a personal computer. For on-line autonomous mobile applications, it should be a microcontroller.

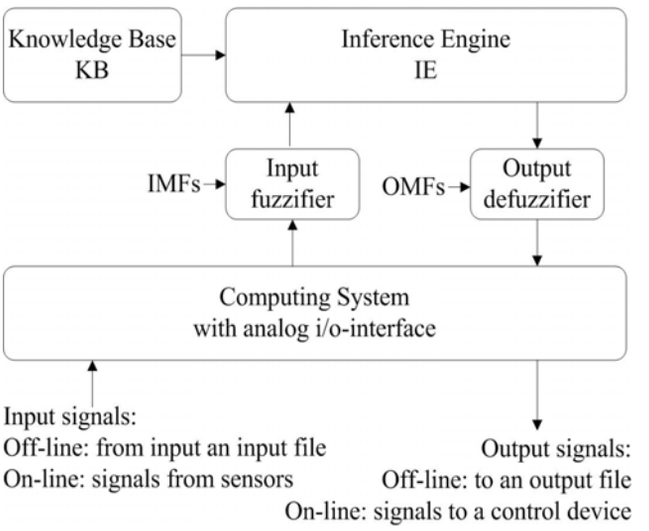


Figure 1. Basic elements of an FLES for control applications.

III. THE FLES SOFTWARE ARCHITECTURE

Fig.2 shows the overall architecture of the embeddable software. It has four, three-input and one-output, modules. The first input module reads sensor values in absolute form. The second input module has membership functions of the input and the output variables. The third input module has the knowledge base of the system in the form of a rule-set. The output-module generates defuzzified output signals.

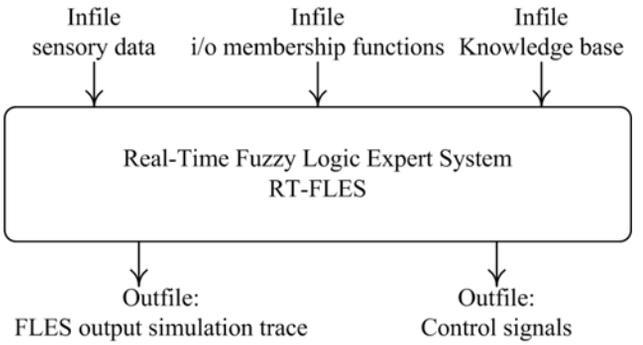


Figure 2. Embeddable FLES-software architecture.

While running off-line (in simulation-mode) on a personal computer, the inputs are read from three input files and the output is stored into an output file. While running on a real-time autonomous robot's micro-controller, the input knowledge base and the membership functions are merged into the FLES-source code. Sensory inputs are mapped to the analog-input-ports of the controller; and the outputs are mapped to a controller's analog output-ports. Overall execution sequence of the inference engine, to implement FLC, is detailed in the following section.

IV. FLC-KERNEL: EXECUTION SEQUENCE OF THE INFERENCE ENGINE

Step-by-step sequences of tasks executed by the inference engine are as follows:

IE1: Fuzzify input variables
Express input variables as a percentage of the input membership functions IMFs. Translate inputs into linguistic form.

IE2: Find the activated rule-set Ra
Find the rule set Ra in the KB, where their input-variable-value requirements match with the current input-values in their linguistic form.

IE3: Find effective input membership function eimf-Ri for each of the activated rule-Rai.
Example: assuming there are two input variables in1 and in2 and the rules are logically ANDed such as:
if (in1 is ns) AND (in2 is pm) then y is nm.

eimf-Rai = min(in1v-Rai, in2v-Rai),
eimf-Raj = min(in1v-Raj,in2v-Raj),

where eimf-Rai is the effective input membership function of the active rule Rai. in1v-Rai and in2v-Rai are the fuzzified values of in1 and in2 in Rai. Similarly, eimf-Raj is the effective input membership function of the rule Raj; in1v-Raj and in2v-Rj are fuzzified values of in1 and in2 in Raj. The fuzzified variable values in1v and in2v being expressed as a percentage of the input membership functions. To find eimf, select minimum of in1v and in2v.

IE4: Find the final rule to execute with its eimf and eomf
among the active-rules identify the rule to execute:
Rule to fire Rf = R with: max(eimf-Rai, eimf-Raj, …)
a. Select the rule Rf corresponding to the maximum of the eimfs
b. eimf-Rf = min(in1v-Rf, in2v-Rf)
c. eomf-Rf = omf-Rf (with one output variable) with multiple output variables eomf should be computed similar to the eimf.

IE5: Find defuzzified control output y
output = eimf * eomf
y = eimf-Rf * eomf-Rf

The embeddable software presented in this paper executes these five steps in sequence. It is also generates an output simulation trace corresponding to each of these steps. This output trace is illustrated in the following section for solving a specific control problem.

V. TEST RESULTS: THE FLES OUTPUT SIMULATION TRACE

Test results of the software are presented in the form of solving a specific control problem. The control problem of this case study is to balance the inverted pendulum IP using a cart driven by dc-motors as shown in Fig. 3. The function of the controller is to keep the IP straight-up. When the pendulum tilts away from the center by θ (angle a), degrees, at a rate of dθ/dt (derivative of the angle da); then the controller must generate a control signal to move cart by x-units in the right direction at right rate of dx/dt. The movement of the cart is proportional to dc-motor current mc. If the motor is controlled at fixed interval then dx/dt is not needed. The problem now has two input variables a and da and one output variable mc. The FLES configuration of this problem is shown in Fig. 4. The input membership functions IMFs for the input variables a and da are shown in Fig. 5. The triangular MFs are represented by three vertices in the output simulation trace as shown in Table 4. For example, imf-nm has vertices p0, p1, p2: -54, -36, -18. The fuzzified or linguistic value of the variable is "nm". Input variables a and da have min/max values of: -54/+54; and units of degrees and degrees/sec respectively. The range is from -54 to +54 with 7 membership functions each. Fig. 6 shows the output membership function OMF for the output variable mc. It also has 7 membership functions but they are singleton. The knowledge base used for the IP-problem shown is shown in Fig. 7. The final goal now is for a given sensed values of a and da the FLC must generate the appropriate motor current mc.

In the software: the sensed values of a and da are specified through an input file “infile1-SenData.txt” as shown in Table 1. The input/output membership functions are specified through the input file “infile2-IOMFs.txt”, refer to Table 2. The knowledge base is specified through another input file “infile3-KB.txt”, this is shown in Table 3. The generated output is stored into an output file “fles-outfile.txt”, this file is shown in Table 4.

While operating in real-time the knowledge base and the MFs are merged into the FLES-source code. Sensed input values and the generated outputs are mapped to i/o-ports of the microcontroller being used. Depending on the problem the input, the output, and the knowledge base modules can be specified accordingly. It includes the name and units of input and output variables; the IMFs and the number of IMFs, OMFs and the number OMFs, and finally the KB.

The detailed results of execution tasks IE1 through IE5 of the inference engine, specified in section IV, are shown in Table 4. This can be disabled in on-line mode. In this table, the initial sections A, B, and C will display input data to the FLC. This includes the sensor data of the input variables; the input/output membership functions; and the knowledge base of the problem. The task here is to build an FLC to balance an inverted pendulum problem. It is a two-input and one-output problem. The input variable values for the FLC are: the pendulum’s angular displacement (angle a) is -12.0 degrees away from the center, and rate of angular displacement (derivative of angle da) is -3.0 degrees/sec. Input variables in1 and in2 both have 7 membership functions. Minimum and maximum values are -54 to +54 with 18-units between the vertices with units of degrees for in1 and degrees/sec for in2. The output variable mc has 7-MFs with minimum and maximum of -18 ma and +18 ma with 6 ma separation. The mc linguistic values ranging from negative-large (nl) to positive-large (pl). The knowledge base of the IP-problem is specified by 13-rules denoted as R0-to-R12. These are logically ANDed-rules.

The 5-step procedure by which the inference engine generates the control signal from the sensor input data is shown in section-D of Table 4. The five steps are denoted as: IE1 through IE5.

```
The IE1-Trace: In IE1 the input variables
are fuzzified: for in1 = -12 degrees, the
fuzzified values are: (12/18)-ns or
(6/18)-zr; and for in2 = -3 degrees/sec,
the fuzzified values are: (3/18)-ns or
(15/18)-zr.

The IE2-trace: In IE2 active rule set Ra
is determined: the activated rules are R7
and R8. Ra = {R7, R8}.

The IE3-trace: In IE3 effective input
membership functions eIMF-Rai for each
active rule Rai is determined:
eIMF-R7 = (12/18)-ns;
eIMF-R8 = (6/18)-zr;

The IE4-trace: In IE4 Rule to be fired is
determined:
Rule fired = max(eIMFs) = eIMF-R7 =
(12/18);
Rule fired is R7

The IE5-trace: In IE5 find output – motor
current mc:
output = eIMF * eOMF = (12/18) * ps
       = (12/18) * 6.0 = 4.0 mille-amps
```

VI. CONCLUSIONS

This paper presents embeddable FLC-software-code that can be downloaded into microcontrollers to build autonomous mobile-robots. We have used this software to implement two fuzzy logic systems, one is an estimation problems and the other a control problem. The estimation problem is an application one that is used for precise estimation of the battery’s state of charge SoC [5]. For this we have used Handyboard that supports interactive C. This required partial modification C++ into interactive-C. Recently we have implemented the same application on Arduino-Due microcontroller. The other is a control application that we have implemented is to balance an inverted pendulum. This is again implemented on a Arduino-Due microcontroller. This is an autonomous robot. The current microcontrollers are very powerful yet very inexpensive (under $50). With your own embeddable source-code, such as the one presented in this paper, one can build extremely complex yet inexpensive mobile-FLC.

REFERENCES

[1] Kim C. Ng, A Neuro-Fuzzy Controller for Mobile Robot Navigation and Multirobot Convoying, IEEE Transactions on Systems, Man, and Cybernetics – part B, VOL. 28, NO. 6, Pg. 829-840, 1998.

[2] Dimiter Driankov and Alessandro S affiotti, Fuzzy Logic Techniques for Autonomous Vehicle Navigation, A Springer-Verlag, 2001.

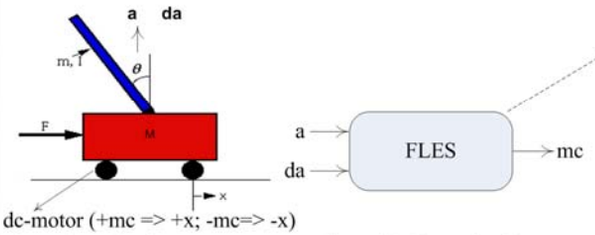
[3] Chuen Lee, "Fuzzy Logic in Control Systems: Fuzzy logic Controller – Part I", IEEE International Conference on Systems, Man, and Cybernetics, 1990, Vol: 20, No. 2, Pages: 404-418.

[4] Chuen Lee, "Fuzzy Logic in Control Systems: Fuzzy logic Controller – Part Part II", IEEE International Conference on Systems, Man, and Cybernetics, 1990, Vol: 20, No. 2, Pages: 419-435

[5] Maila, Sreelatha (G. N. Reddy), “Embedded-System Controlled Fuzzy Logic Expert System to Estimate Battery-SOC, Summer II, Electrical Engineering, Lamar University, 2008.

Fuzzy-Logic Expert Systems FLES: Inference Engine IE: Case Study

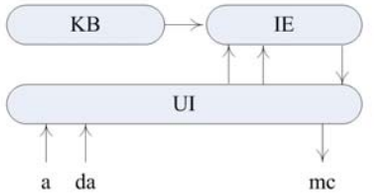
Overall operations by the Inference Engine IE in the RT-FLES



dc-motor (+mc => +x; -mc=> -x)

Figure 3. Balancing an Inverted Pendulum.

Figure 4a. Fuzzy Logic Expert System FLES with i/o.



FLES: Fuzzy Logic Expert System; KB: Knowledge Base; IE: Inference Engine; UI: User Interface

Figure 4b. Fuzzy Logic Expert System FLES with i/o.

Note: da = dθ/dt is computed: "θ" is sampled into the FLES at a constant interval of around 1000-samples/sec (or at a sampling fs frequency of 1 kHz) from a photo-sensor. That is, dt = 1/fs = 1 msec. $d\theta_n/dt = (\theta_n - \theta_{n-1}) / 1$ msec; For $\theta_n = 25.02$ degrees, $\theta_{n-1} = 25.00$ degrees; dθ = (25.02 – 25.0*) = 0.02 degrees; da = dθ/dt = 0.02/1 ms = 20 deg/sec.

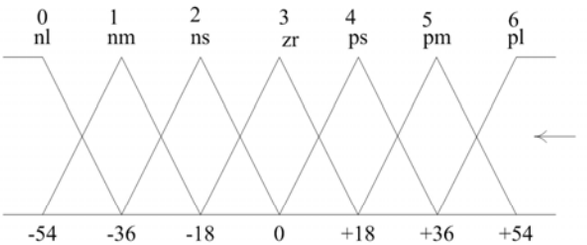


Figure 5. Fuzzy Logic Input Membership Functions IMFs for the input variables: a and da.

Input variable 1: a; Units: degrees; Min / Max: -54 / +54

Input variable 2: da; Units: degrees/sec; Min / Max: -54 / +54

Table 1. Sensor Input Data.

```
// infile1-SenData.txt
// Input values
-12.0    : degrees
-3.0     : degrees/sec
```

Output variable: mc

Units: milli_amps

Min/Max: -18/18

Member Ship Functions:

0 nl	-18
1 nm	-12
2 ns	-6
3 zero	0
4 ps	6
5 pm	12
6 pl	18

Note::
a => angle, θ, theta, angular displacement;
da => derivative of the angle, dθ/dt, θ'

Figure 6. Output Membership Functions OMFs for the output variable mc.

Table 2. Input/Output Membership Functions IOMFs.

```
// infile2-IOMFs.txt
// Input Membership Functions IMFs: min, max, n
-54.0  : in1min a: degrees
+54.0  : in1max a: degrees
 7     : in1num
-54.0  : in2min da: degrees/sec
+54.0  : in2max da: degrees/sec
 7     : in2num
-18.0  : in3min mc: milli-amps
+18.0  : in3max mc: milli-amps
 7     : in3num
```

FLES Knowledge-Base: Rule-Set

```
R0: IF a IS zr AND da IS nm THEN mc IS pm
R1: IF a IS ps AND da IS ns THEN mc IS ps
R2: IF a IS zr AND da IS ps THEN mc IS ns
R3: IF a IS zr AND da IS pm THEN mc IS nm
R4: IF a IS zr AND da IS pl THEN mc IS nl
R5: IF a IS nl AND da IS zr THEN mc IS pl
R6: IF a IS nm AND da IS zr THEN mc IS ps
R7: IF a IS ns AND da IS zr THEN mc IS ps
R8: IF a IS zr AND da IS zr THEN mc IS zr
R9: IF a IS ps AND da IS zr THEN mc IS ns
R10:IF a IS pm AND da IS zr THEN mc IS nm
R11:IF a IS pl AND da IS zr THEN mc IS nl
R12:IF a IS ns AND da IS ps THEN mc IS ps
```

Figure 7. The FLES Knowledge Base KB for the IP-Problem

Table 3. The Knowledge Base for the IP-Problem.

```
// infile3-KB.txt
// Knowledge base for the fuzzy logic expert system
// Example: infile3KB.txt for: 13 rules, 2-inputs, 1-output
// MFs: 0-nl; 1-nm; 2-ns; 3-zr; 4-ps; 5-pm; 6-pl
// Rule: 0  1  2  3  4  5  6  7  8  9  10 11 12
//       ---------------------------------------------
         zr ps zr zr zr nl nm ns zr ps pm pl ns :kbIn1 MFs
         nm ns ps pm pl zr zr zr zr zr zr zr ps :kbIn2 MFs
         pm ps ns nm nl pl ps ps zr ns nm nl ps :kbOut MFs
```

Table 4. FLES Output Simulation Trace: FLC Kernel Execution Steps IE1 through IE5 of the Inference Engine

```
FUZZY LOGIC EXPERT SYSTEM FLES OUTPUT SIMULTATION TRACE:
Dr. G. N. Reddy, LUEE, Summer II, 2013
Update 2.3: 8/15/2013; 10 am

Input/Output Files:
infile1-SenData.txt: Sensor inputs: angle & dangle
infile2-IOMFs.txt:   Input/Output MFs: Ranges & NumMFs
infile3-KB.txt:      Knowledge Base: Rule Set
fles-outfile.txt:    FLES Output Simulation Trace

A. Reading input sensor values from: infile1SenData.txt
INPUT DATA: infile1SenData.txt:
in1, degrees     = -12.0
in2, degrees/sec = -3.0

B. Reading input/Output MFs from: infile2_IOMFs.txt
B1. Input/Output Variables:
    names, units, min, max, nummf, deltamf
1. in1: name, units, min, max, nummf, deltamf
   a degrees -54.0 54.0  7.0 18.0

2. in2: name, units, min, max, nummf, deltamf
   a degrees/sec -54.0 54.0  7.0 18.0

3. out: name, units, min, max, nummf, deltamf
   a milli-amps -18.0 18.0  7.0  6.0

B2: Corresponding Vertices of the MFs:
mf, po, p1, p2:
in1MF[0]:  nl,  -54.0,  -54.0,  -36.0
in1MF[1]:  nm,  -54.0,  -36.0,  -18.0
in1MF[2]:  ns,  -36.0,  -18.0,    0.0
in1MF[3]:  zr,  -18.0,    0.0,   18.0
in1MF[4]:  ps,    0.0,   18.0,   36.0
in1MF[5]:  pm,   18.0,   36.0,   54.0
in1MF[6]:  pl,   36.0,   54.0,   54.0

in2MF[0]:  nl,  -54.0,  -54.0,  -36.0
in2MF[1]:  nm,  -54.0,  -36.0,  -18.0
in2MF[2]:  ns,  -36.0,  -18.0,    0.0
in2MF[3]:  zr,  -18.0,    0.0,   18.0
in2MF[4]:  ps,    0.0,   18.0,   36.0
in2MF[5]:  pm,   18.0,   36.0,   54.0
in2MF[6]:  pl,   36.0,   54.0,   54.0

outMF[0]:   nl,    -18.0
outMF[1]:   nm,    -12.0
outMF[2]:   ns,     -6.0
outMF[3]:   zr,     +0.0
outMF[4]:   ps,     +6.0
outMF[5]:   pm,    +12.0
outMF[6]:   pl,    +18.0


C. Reading the knowledge base from: infile3-KB.txt
C1. The Knowledge Base is:
Rule:          0  1  2  3  4  5  6  7  8  9  10 11 12
               --------------------------------------
kbin1MF:  a:   zr ps zr zr zr nl nm ns zr ps pm pl ns
kbin2MFs: da: nm ns ps pm pl zr zr zr zr zr zr zr ps
kbOMFstr: mc: pm ps ns nm nl pl ps ps zr ns nm nl ps

c2. Knowledge Base KB (Rule-Set) is:
R0 : If a is zr AND da is nm Then mc is pm;
R1 : If a is ps AND da is ns Then mc is ps;
R2 : If a is zr AND da is ps Then mc is ns;
R3 : If a is zr AND da is pm Then mc is nm;
R4 : If a is zr AND da is pl Then mc is nl;
R5 : If a is nl AND da is zr Then mc is pl;
R6 : If a is nm AND da is zr Then mc is ps;
R7 : If a is ns AND da is zr Then mc is ps;
R8 : If a is zr AND da is zr Then mc is zr;
R9 : If a is ps AND da is zr Then mc is ns;
R10: If a is pm AND da is zr Then mc is nm;
R11: If a is pl AND da is zr Then mc is nl;
R12: If a is ns AND da is ps Then mc is ps;

Note: Membership functions:
0-nl; 1-nm; 2-ns; 3-zr; 4-ps; 5-pm; 6-pl
```

Table 4. FLES Output Simulation Trace: Contd...

```
D. Inference Engine: Compute output using the Inputs & KB
IE1: Compute membership functions
IE2: Find activated rules
IE3: Find effective input membership function eimf
IE4: Find the rule to fire and corresponding omf
IE5: Compute output = motor current = mc = eimf * eomf



IE1: Compute membership functions
     Find fuzzified or linguistic values of the inputs

in1: Expressed as function of IMF1
in1 = -12.0 degrees
in1mf_1: mf, V11/deltamf: ns, +12.0/18.0
in1mf_2: mf, V12/deltamf: zr, +6.0/18.0

in2: Expressed as a function of IMF2
in2 = -3.0 degrees/sec
in2mf_1: mf, V21/deltamf: ns, +3.0/18.0
in2mf_2: mf, V22/deltamf: zr, +15.0/18.0



IE2: Find the activated rule-set
The activated rule-set is:
7, 8,
Number of activated rules = 2



IE3: Find eIMF for each the activated rule
Generate 10-value vector for each rule:
Active rules & the corresponding 10-value vectors:

Rule; i1mf:n,str,v; i2mf:n,str,v; eimf: n,str,v
7     2 ns 12.00    3 zr 15.00    2 ns 12.00
8     3 zr 6.00     3 zr 15.00    3 zr 6.00

IE4: Find the rule to fire:
Find max(eimf_Ri), Ri-fired, eOMF_Ri

Rule fired:        7
eIMF = max_eIMFs:  12.00 / 18.00



IE5: Find Output mc:
mc = eIMF * eOMF;
eIMF = 12.00
Rule fired is: 7
eOMFstr:       ps
OMF number, eOMFn: 4
OMF value, eOMFv:  6.00

Final output: mc = eIMFv * eOMFv
mc, in ma = 4.00

End of the output simulation trace: LUEE-FLES
```

Detection of Fluorescent Bacteria Using VPNP Phototransistors Arrays Integrated on Multi-Labs-On-A-Chip System (MLoC)

Abdullah Tashtoush
Biomedical systems and Informatics Engineering Department
Yarmouk University
Irbid, Jordan
Abdullah.t@yu.edu.jo

***Abstract*—Recently, the need of a low-cost, rapid, selective and sensitive method for detecting bacterial pathogens in medical diagnosis, and food-safety inspection has become a prevalent demand. Traditional methods, such as polymerase chain reaction and cell culture techniques take several hours to days to give accurate results, and require bulky, expensive equipment. In this paper, the MLoC system, including optical approach that has been used to create a compact and portable immunosensor for sensitive and rapid detection of bacteria, will be introduced as a novel technique. However, MLoC system requires laser-induced fluorescence and is neither low-cost nor quantitative, therefore making it suboptimal for point-of-care diagnostic purposes. MLoC technology includes, amongst other techniques, optical biosensing, which uses bacteria and bacteriophages (phages) as biological detecting elements, along with red laser 650 nm as an excitation source to identify bacteria rapidly and safely.**

Keywords— Fluorescent; VPNP; MLoC; Bactria; Bacteriophages; Microfluidic channel

I. INTRODUCTION

In this paper, a new generation of multibiosensors named MLoC that includes optical biosensor where a fluorescence spectroscopy as a sensing mechanism has been introduced [1][2], employing bacteriophage or phageorganisms as recognition elements to detect deadly bacteria such as *E*-Coli and Salmonella. Fluorescence spectroscopy can be applied to a wide range of problems in the chemical and biological sciences. The measurements can provide information on a wide range of molecular processes, including the interactions of solvent molecules with fluorophores, rotational diffusion of biomolecules, binding interactions, conformational changes, and distances between sites on biomolecules. Advances in technology for cellular imaging and single-molecule detection are expanding the usefulness of fluorescence. These advances in fluorescence technology are fast response times, easy implementation, stand-off detection, and decreasing the cost and complexity of previously complex instruments. Fluorescence spectroscopy will continue to contribute to rapid advances in biology, biotechnology, medical diagnostics, DNA sequencing, forensics, genetic analysis and nanotechnology [3]. Fluorescence detection is highly sensitive, and there is no longer the need for the expense and difficulties of handling radioactive tracers for most biochemical measurements. There has been dramatic growth in the use of fluorescence for cellular and molecular imaging. Fluorescence imaging can reveal the localization and measurements of intracellular molecules, sometimes at the level of single-molecule detection.

Typically, fluorescence based sensors excite optically active recognition elements that are selective to particular media (i.e. LB). Emission from the bacteriophage, at wavelengths longer than the excitation wavelength, is monitored and provides information regarding the concentration of multiple bacteria or viruses in real-time. Thus, in this technique, it is important to immobilize the bacteriophage at the sensor surface and maximize the contact surface area to maximize interaction of the media with the recognition element during the sensor operation [4][5].

II. SURFACE ACTIVATION AND BACTERIO-PHAGE IMMOBILIZATION

A. Selecting a Template (Heading 2)

Phages, which are bacterial viruses, have recently been postulated as promising recognition elements in bacterial biosensors. They are ubiquitous in nature, highly specific to bacteria and hence harmless to humans, much cheaper to produce than antibodies and more stable than the latter. In addition, they can also be immobilized on transducing devices in pretty much the same manner as antibodies or DNA probes. The potential use of phages in the development of novel and unique diagnostics and therapeutics for life-threatening bacteria is virtually limitless. This is the reason why the present project aims to develop a method for rapid pathogen detection based on phages [6][7]. A phageorganisms (phage) is a type of virus that infects bacteria.

III. EXPERIMENTAL PROCEDURES

A. Multibiosensors Instrument

- The multibiosensors system called as multi-labs-on-a chip; (MLoC); in this work measures 2.23 mm x 3.04 mm was fabricated using the CMOSP35 process available through CMC; as shown in Fig. 1. The optical detector consists of vertical phototransistor array (VPNP), current mirror, current-to-voltage converter, amplifier, band-pass filter, and phase detector [8][7].

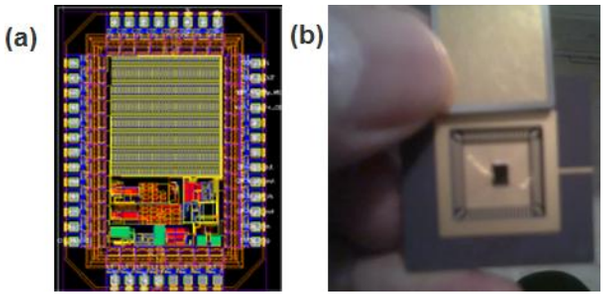


Fig. 1. Multi-Labs-On-a chip, (a) layout of MLOC (b) the die mounted on PGA 68 sockets (CMC).

- The optical detector was fabricated based on a single VPNP phototransistor pixel that measures 40.0 µm x 41.775 µm, whilst the VPNP phototransistor 32 x 32 array measures 1210.8 µm x 1318.4 µm was fabricated using CMOSP35 process available through CMC. The vertical phototransistor is formed by the p-active (emitter)/n-well (base)/p-substrate (collector), and has one of the highest responsivities of the photodetectors available in this standard CMOS process. The advantages of the CMOS-based system include its operation using low supply voltages and low cost fabrication[9][12].
- A National Instruments DAQ516 PCMCIA card installed in a laptop computer provided digital I/O lines and an analog-to- digital conversion channel so that the CMOS microchip detection elements were individually accessed and read out.
- A custom written software interface constructed with Labview program controlled the data acquisition process for MLoC system.

B. Phage immobilization on gold surface and protocol

- Clean the chips three times with ethanol and dry with air.
- Immerse the chips in 1mM Mercaptoundecanoic acid in ethanol for 24 hours for 20 ml of ethanol we need 0.0044g of Mercaptoundecanoic acid:

$$\frac{1\ \text{mml}}{1000\ \text{ml}} \times \frac{1\ \text{mol}}{1000\ \text{mml}} \times \frac{218.36\ \text{g}}{1\ \text{mol}} \times 20\ \text{ml} = 0.0044\ \text{g}$$

- Wash the chips five times with ethanol and then dry with air.
- Immerse the chip with EDC/NHS 1.15 g EDC in 15 ml high purity water, 200 mg NHS in 15 ml high quality water. Prepare EDC/NHS as required only as it decomposes very quickly. Required: For 10 ml of DI water, 0.77 g required of EDC, and 133.33 mg of NHS, immersed for 1 hour.
- Wash the chips 5 times with water and then 3 times with phosphate buffer.
- Incubate the phage with the activated surface for 3 hours with shaking.
- Wash 5-7 times with phosphate buffer.

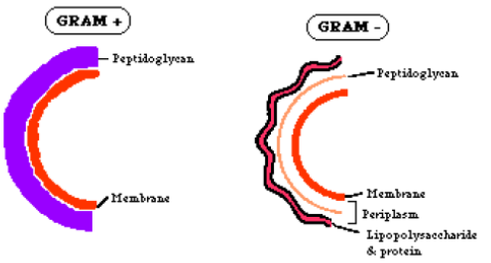


Fig. 2. The structure of Gram-positive bacteria, and Gram-negative bacteria.

- Then block the surface with 1 mg/ml (Bovine Serum Albumin) BSA in phosphate buffer. BSA is used to block the surface to potentially improve non-specific adsorption of the bacteria to the 1-hexadecanethiol.
- The reference will be exactly the same procedure as above except replace BSA instead of phage T4 [7].

C. Bacteria Preparation Protocol

Escherichia coli ATCC 11303 and wild-type T4 bacteriophages were prepared at Biophage Pharma Inc. (Montreal, Canada). Escherichia coli is a Gram negative bacterium that is commonly found in the lower intestine of warm-blooded organisms (endotherms). Most E. coli strains are harmless, but some, such as serotype O157:H7, can cause serious food poisoning in humans, and are occasionally responsible for costly product recalls [9]. The gram reaction is based on the structure of the bacterial cell wall, and it is sited in two categories; as shown in Fig. 2. In Gram-positive bacteria, the layer of peptidoglycan, which forms the outer layer of the cell, traps the purple crystal violet stain. In Gram-negative bacteria, the outer membrane prevents the stain from reaching the peptidoglycan layer in the periplasm. The outer membrane is then permeabilized by acetone treatment, and the pink safranin counterstain is trapped by the peptidoglycan layer [9].

D. Design and Fabrication Microfluidic Channels (MFC)

Microfluidic was designed and fabricated using soft photolithography technique in different styles; that is coated fully or partially by gold; as shown in Fig. 3:

1. Plastic channels; with one inlet and outlet.
2. Microfluidic channels fabricated using PDMS with one inlet and outlet, and two inlets and outlet. The outcome was good results.

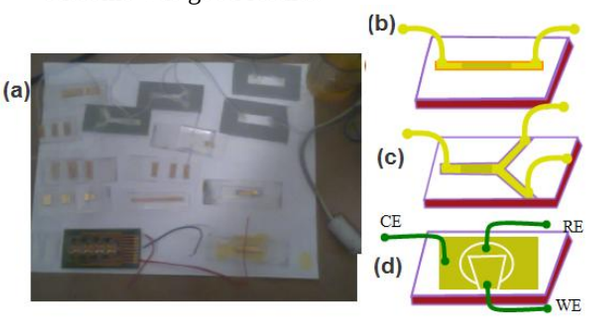


Fig. 3. Microfluidic channels (MFC).

3. Microfluidic channels fabricated by etching glass using a mixture of phosphoric acid (H_3PO_4), and HF is one possible solution, then coated by gold; The outcome of this technique was not good because the roughness on the surface.
4. Microfluidic channels fabricated by etching Si using TMAH, then coated by gold. The outcome was good result.

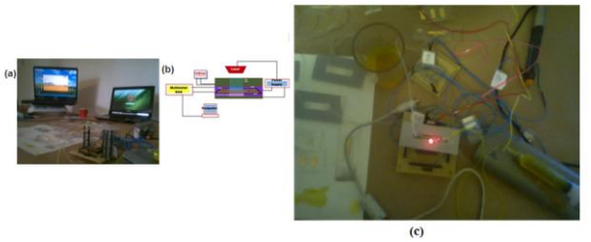


Fig. 4. Experimental setup.

IV. Result and Discussion

The present work performed using IC consists of high-gain phototransistors 32 x 32 array, the structure of phototransistor pixel was unfolded in previous work [10]-[14], and E. coli 12 as a target to detect by this experiment. Fig. 1 presents a view of the integrated MLoC system. A red laser 650 nm is used as excitation source, and the fluorescence is detected by optical chip using a phototransistor 32 x 32 array on-chip, as shown in Fig. 4.

A pinhole was used to eliminate extraneous light from the laser. The laser beam was focused onto the optical chip array using a capillary holder that's adjusted using a translational stage so that the laser beam passed through the center of the microfluidic channel (MFC). Fluorescence from the MFC was detected with the CMOS microchip that was lay on it and collected the data using a 2700 Multimeter/Data Acquisition System and National Instruments DAQ516 PCMCIA card installed in a laptop computer. A band pass optical filter (cut-off position: 510 nm, Edmund Industrial Optics) was attached on MLoC to eliminate the laser scattering.

A. *Before Encapsulation the Chip*

Fig. 5 shows the response of phototransistors detection elements. This figure shows the fluorescence detection of E-coli obtained using MLoC microchip system after the MFC was filled with high concentration of bacteria 10^9 cfu/mL by means of microfluidic pump or syringe. The laser beam irradiation onto MFC was well-controlled.

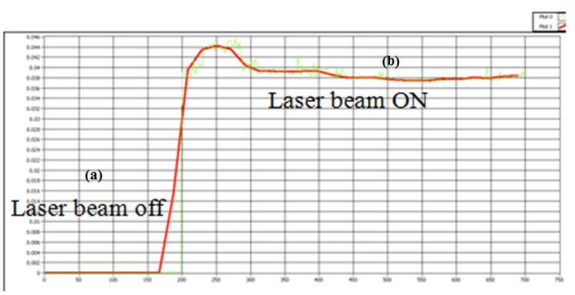


Fig. 5. Optical response of phototransistors array, (a) When the laser beams OFF, (b) ON.

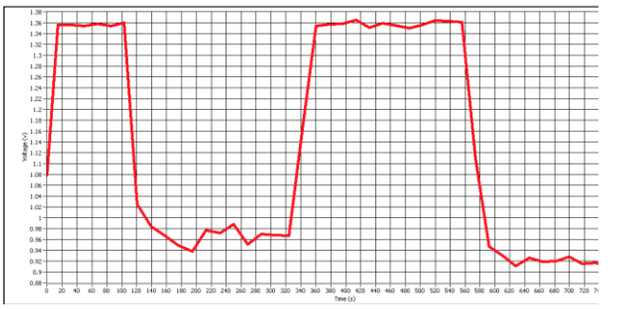


Fig. 6. The whole cycle profile for DI water- high concentration of bacteria.

When the laser beam was irradiating MFC, the phototransistors showed obvious optical response. The dark current signal was obtained when there was no laser beam irradiation onto MFC. The fluorescence intensities of E-coli 12 decreased while DI water pumping through MFC. The bacteria detection was performed after this optical adjustment for the detection of the highest fluorescence signal.

A full cycle was recorded, started with the dark current signal; when the laser beam OFF and DI water through this step pumping followed by high concentration of bacteria whilst the laser switch ON. The signal was observed a little bit higher than the previous one because of a residual bacteria presence along with pumping high concentration of bacteria. Thereafter; the laser beam switch OFF and DI water injected as well, Fig. 6 shows the whole of this cycle.

MLoC system was showing remarkable results in regarding of reproducibility of the experiment that makes sure about the response behavior and performance of phototransistors array. Fig. 6 shows the profiles as a result of DI water and high concentration of bacteria, where observed the reproducibility of the experiment, ultimately.

B. *After encapsulation the chip*

The bacteria; E-coli 12 was prepared in two categories Gram-positive (PB) and Gram-negative (NB) and pumping through MFC, by recording the response of MLoC system, a significant detection voltage was observed.

Fig. 7 shows the profile because of flowing DI water through MFC then bacteria in LB media was pumping with

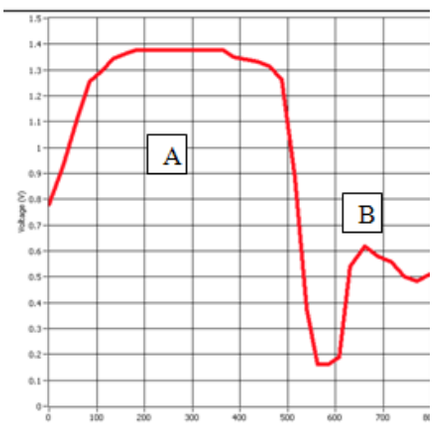


Fig. 7. The profile of high concentration of bacteria in both types PB and NB.

high concentration (10^9 cfu/mL; Colony Forming Units/mL); both types of bacteria; PB and NB; were used in this case at the same concentration; a voltage around 1.39V for the PB and 0.18 V for NB were detected.

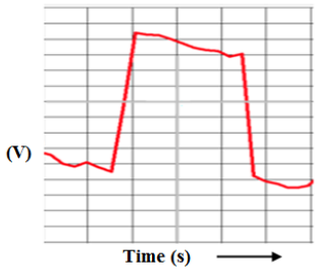


Fig. 8. The profile of fluorescent material detection after encapsulating the optical chip.

The experiment was repeated after the optical chip was encapsulated using Norland Optical Adhesives; NOA60 are clear, colorless, one part adhesives that contain no solvents. When exposed to ultraviolet light, they gel in seconds and full cure in minutes to give a tough, resilient bond. These adhesives are designed for fast, precision bonding where low strain and optical clarity are required; the following experiments have been done. A profile was recorded as a result of flowing DI water through MFC and then fluorescent material; respectively. The profile result as shown in Fig. 8 it is recording no significant difference from the previous experiment results. This means that encapsulating is significant to protect MLoC system without any influence.

V. CONCLUSIONS

Attributable to its compact design and multiplex capability, the CMOS microchip system combined with phototransistors provided high-gain and throughput analysis as tool for the detection of bacteria in medical diagnosis, food-safety inspection and bacterial pathogens DNA microarray analysis. Based on its compactness, low cost, multiplex capability, selective and sensitive method, the integrated CMOS microchip system as a detector is expected to be compatible with conventional micro-fabricated devices. This technique is allowing more rapid and high throughput analysis.

REFERENCES

[1] J.R. Lakowicz, “Enhanced oxygen detection using porous polymeric gratings with integrated recognition elements”, Principles of Fluorescence Spectroscopy, Kluwer Academic/ Plenum Publishers, New York, 1999.

[2] J.R. Lakowicz, Principles of Fluorescence Spectroscopy, Kluwer Academic/ Plenum Publishers, New York, 1999.

[3] Magdalena Gabig-Ciminska, Marcin Los, Anders Holmgren, J€org Albers, Agata Czyz, Reiner Hintsche, Grzegorz Wegrzyn, and Sven-Olof Enfors, “Detection of bacterial pathogen DNA using an integrated complementary metal oxide semiconductor microchip system with capillary array electrophoresis”, Analytical Biochemistry 324, 84–91, 2004.

[4] Lei Yao, Mohamad Hajj Hassan, Vamsy Chodavarapu, Arghavan Shabani, Beatrice Allain, Mohammed Zourob, Rosemonde Mandeville, “CMOS Imager Microsystem for Multi-Bacteria Detection”, IEEE 2006.

[5] N. Nikkhoo C. Man, K. Maxwell P. G. Gulak, “A 0.18µm CMOS Integrated Sensor for the Rapid Identification of Bacteria”, IEEE International solid-state Circuit Conference, pp. 636-617, 2008.

[6] Turgut Sefket Aytur, “A CMOS Biosensor for Infectious Disease Detection”, Electrical Engineering and Computer Sciences, University of California at Berkeley.

[7] L. Gervais, M. Gela, B. Allain, M. Tolba, L. Brovko, M. Zourob, R. Mandeville, M. Griffiths, S. Evoy, “ Immobilization of biotinylated bacteriophages on biosensor surfaces”, Sensors and Actuators B 125 (2007) 615–621.

[8] R. J. Baker; “CMOS Circuit Design, Layout and Simulation”, 2nd ed., New York: Wiley-IEEE Press, 2008.

[9] Escherichia coli O157:H7". CDC Division of Bacterial and Mycotic Diseases. http://www.cdc.gov/ncidod/dbmd/diseaseinfo/escherichiacoli_g.htm. Retrieved on 2007-01-25.

[10] Abdullah Tashtoush, Adel Omar Dahmane, “Multi-Labs-On- a Chip (MLoC) For Atto-molar Cancer Markers Concentration Using VPNP Phototransistor Detection”, Manuscript submitted to be published.

[11] Abdullah Tashtoush, Adel Omar Dahmane, “A new generation for multibiosensors”, Manuscript submitted to be published.

[12] L. Yao, A. Tashtoush, E. Ghafer-Zadeh, R. Mandeville, V. Chodavarapu, "CMOS Imaging of Biological and Chemical Sensor Microarrays", Presented at the Canadian Institute for Photonics Innovation (CIPI) Annual Meeting, Quebec city, May 2009.

[13] Vamsy P. Chodavarapu, Daniil O. Shubin, Rachel M. Bukowski, Albert H. Titus, Alexander N. Cartwright, and Frank V. Bright, “CMOS-based phase fluorometric oxygen sensor system”, IEEE Transactions On Circuits And Systems—I: Regular Papers, Vol. 54, No. 1, January 2007.

[14] Lei Yao, Rifat Khan, Vamsy P. Chodavarapu, Vijay S. Tripathi, and Frank V. Bright, “Sensitivity-Enhanced CMOS Phase Luminometry System Using Xerogel-Based Sensors”, IEEE Transactions On Biomedical Circuits And Systems, PP. 1-8.

Fully Label-Free Impedimetric Immunosensor Chip Based on Interdigitated Microelectrodes for a Thyroid Hormones Detection Portable System

Abdullah Tashtoush
Biomedical systems and Informatics Engineering Department
Yarmouk University
Irbid, Jordan
abdullah.t@yu.edu.jo

Abstract*— Recently, many researchers have reported integrated analysis systems called lab-on-a-chip or micro total analysis systems that are small, light, and capable of integrating all sample-handling steps in the microfluidic channels (MFC). Immunoassay scheme can be sorted based on the simplicity and rapidity in performing clinical measurements. This paper will introduce a novel technique based on CMOS electrochemical impedance spectroscopy incorporating with interdigitated microelectrodes arrays (IDMA) embedded on microfluidic channel (MFC) as a new generation of multibiosensors named as multi-labs-on- a single chip system (MLoC) that is capable to detect low level concentration and be high qualify thyroid hormones markers as a response of frequency variations.

Keywords*— EIS, IDMA, MFC, HSA, HPLC, Thyroid hormones

I. INTRODUCTION

Thyroid-stimulating hormone (TSH) control the thyroid grand hormones, the hypothalamus produces thyrotropin-releasing hormone (TRH) that controls anterior pituitary, which forms TSH. TSH spurs the thyroid to excrete the hormones thyroxine (T4) and triiodothyronine (T3). The EIS detection is much preferred to other methods including amperometric, voltammetric, and/or potentiometric methods. In recent studies, Park et al. [1]-[5] demonstrated that EIS measurements can provide significantly more sensitive responses than cyclic voltammograms (CVs), beside; electrochemical detection is advantageous because it is inexpensive and easy to miniaturize while maintaining relatively high sensitivity.

The interdigitated microelectrodes will remarkably show reduction in the response current because of the technology that the system made of, which is CMOSP35 that provides by CMC. Fig. 1 shows the MLoC layout and design, respectively.

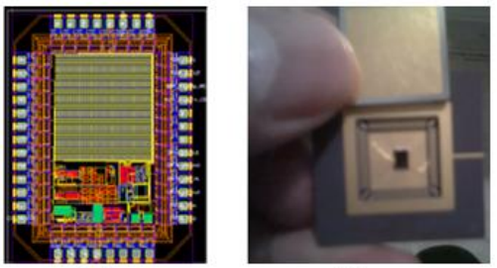


Fig. 1. MLoC system (a) layout (b) the chip.

In biochemistry and analytical chemistry high performance liquid chromatography (HPLC) is used for separate, identify, and quantify compounds based on their idiosyncratic polarities and interactions with the column's stationary phased due to HPLC used to purify individual chemical compounds from mixture of compounds such as drugs and endogenous compounds; such as Thyroxine T3, T4 which are foremost thyroid hormones. In case of hormones as type of drugs and endogenous compounds; they have the capability to jump to or released from proteins in plasma such as human serum albumin (HSA) where it is the most plentiful protein found in human plasma. An extensive assortment of drugs and endogenous compounds including T4 and T3 are readily bound to human serum albumin [6]. Such situation is playing an important role in their circulations in the human body. The variations and activities causing by releasing the hormones is a clinical concern can be monitored premeditated using HPLC incorporated with electrochemical impedance measurements accordingly. Contrast to traditional techniques such as mass spectroscopy involves problems [7]-[10], such as adsorption and leakage of the bound form through the membrane leading to change the concentration of the target sample thus yielding unpredicted results. For that reason; an electrochemical impedemetric technique due to its high sensitivity and selectivity it is introduced to detect

the hormones pharmacological and biological activities thus the variations measured by the impedance as a result of unbound concentration changing and so significant to prognosis the effect of the illness at early stages [11]-[12].

Thyroid hormones organize a sort of biological processes, including growth and development. The variation level occurs in T3, T4 and rT3 profiles are analytically problem-solving of harshness of non-thyroidal diseases consequently prognosis these levels is so important for human life [13]-[17]. The thyroxine T4 can be transformed into two forms either active from T3 or inactive forms rT3 [18]. The unbound concentrations of thyroid hormones are measurable function via impedance measurements. In view of the fact that T3 the active form of the thyronine the low level of serum T3 almost certainly being required for continued existence in many general diseases [19].

The mechanism of the thyroid grand relies on the level of thyroid hormones (T3 and T4). The pituitary release of TSH is inversely affected by the concentration of these hormones in the blood that ends up to create a regulatory negative feedback loop, which means TSH is increased for low T3, T4.

II. ELECTROCHEMICAL METHODOLOGY AND BEHAVIOR

Basically; the electrochemical technique can be performed through measuring the generated electrical signals either the current, voltage, charge of the labeled species or the resulting complex. Otherwise; this technique has the potential to be performed by measuring the impedance, capacitance or admittance [20]. Practically; electrochemical immunosensors merge immunoglobulin; antibody, antigen interaction with electrochemical impedemetric measurements through utilizing a sort of developed techniques such as impedance, capacitive, amperometric, potentiometric, and conductometric. The electrodes where the interaction is taking place play an essential role in the performance of electrochemical biosensors. Therefore, biosensors performance; i.e. sensitivity, will be effect on the light of electrodes material, geometry, size, and its modification. Therefore, the smallest sensing surface area the more surface-to-volume ratio of biomolecular samples the higher capture efficiency the higher electrical properties response to the applied potentials. For that reasons the sensitivity of the transducer definitely increased [21]. Electrochemical detection approach regularly has need of a reference electrode (RE); i.e. Ag/AgCl, is reserved at a distance from the sensing surface so that specific potential stays stable throughout the experiment. Secondly; a counter electrode (CE); i.e. Pt, and a working electrode (WE); i.e. Au; noble materials selecting specific material rely on the analyte nature, where the later is the sensing or redox electrode [22], as shown in Fig. 2. Despite the fact that the counter electrode ascertains a link to the analyte; a current can be applied for the working electrode (WE); accordingly. Detail characteristics of electrochemical biosensors were unfolded in details by Mehrvar et al. [6].

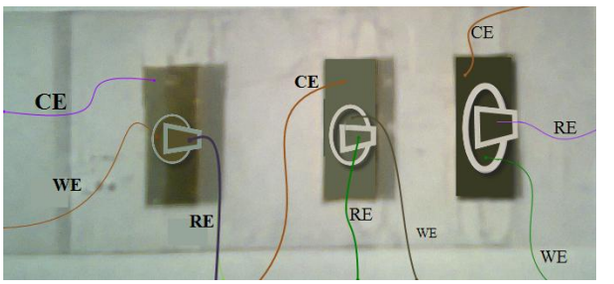


Fig. 2. Microfluidic channel (MFC).

The Nyquist Plot as shown in Fig. 3 is demonstrated is characteristic of a single "time constant"; one resistor and capacitor in parallel. The Electrochemical analytical process can be more significant if the impedance measurements run with a range of frequencies "$10^{-3} < f < 10^{5}\ Hz$". In such condition they are mostly controlled by the interfacial properties of the modified electrodes. The applied frequency is main controller over the impedance measurements. For low frequencies less than 10-3 Hz this results the impedance value to be calculated by the DC-conductivity of the electrolyte solution. While for frequencies greater than 105 Hz more parameters get significant and affect the impedance spectrum as well [23]. Theoretically, the microfluidic cell with three-electrode system can be treated as mentioned before; using electrical equivalent circuit that mainly constructs from the electrolyte solution resistance (Rsol), Faradaic impedance (Zf), and double layer capacitance (Cdl). If there is no electrochemical reaction on the electrode surface the Faradaic path Zf is inactive, and only non-Faradaic impedance is operative. In such case the equivalent circuit can be simplified as a serial combination of the electrolyte solution resistance and the double layer capacitance, which will results the total impedance (ZT) of the system as shown in Fig. 3.

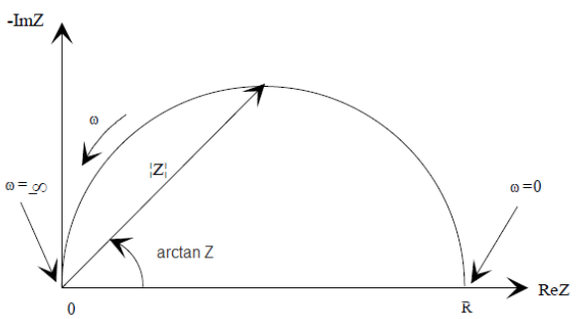


Fig. 3. The Nyquist Plot characteristics.

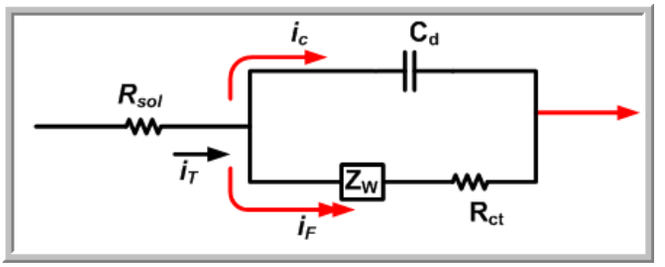


Fig. 4. Randles equivalent electrical circuit model.

So far; the total impedance measurement somehow still in standard form; for real and more accurate system the situation is more complex and actual plot of impedance in the complex plane will unite the two limiting features; mass-transfer and kinetic control regions. Nyquist plot is so essential for electrochemical system impedance where the two limiting features regions are defined at low and high frequencies, mass-transfer and kinetic control regions; respectively. Among the entire parameters of the total impedance, the key that control the process is the charge-transfer resistance, R_{ct} incorporating with Warburg impedance as function of "σ". The semicircular region does not look perfect due the entire behavior of the system where the mass transfer is a significant factor. The charge-transfer resistance is dominant as long as the chemical system is kinetically sluggish. On the other hand; the charge-transfer resistance, R_{ct} might be slightly small by comparison to the solution resistance and the Warburg impedance over nearly the whole available range of "σ". Then the system is so kinetically shallow that mass transfer always plays a role, Fig. 4 illustrates the behavior of the electrochemical system [16].

III. MICROFLUIDIC CHANNEL (MFC) MICROFABRICATION

In electrochemical impedance immunoassay the microfluidic channel and the sensing surface play a key role in biosensors characteristics. Each of them has been fabricated separately afterward integrated together as shown in Fig. 5. The IDMA as an electrochemical sensor was surface-mounted using PVA TePla Microwave Asher; oxygen bonding technique on a microfluidic channel that makes of microfluidic channels with three inlets; analyte, washing and micelles, and outlet fabricated using PDMS bonding on substrate either polymer or glass using PVA bonding technique. Fig. 5 illustrates the basic microfluidic channel (MFC) with three inlets, analyte, washing and micelles, and outlet. Fig. 6 shows the microfluidic channel (MFC) embedded with interdigitated microelectrode arrays (IDMA) as a sensing surface referring to working electrode in the basic electrochemical cell structure was introduced [24].

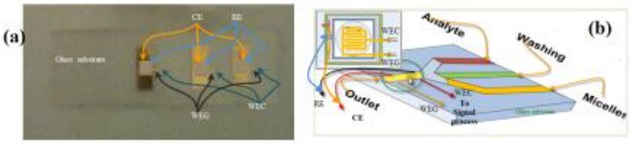


Fig. 5. MCF embedded with IDMA (a) Polymeric lab on a chip on top of CMOS chips (b) with embedded interdigitated microelectrodes array.

All microchannels are 2.54 cm wide, 5.8 mm length and 1 mm depth. The basic structural concept of the microfluidic channel interconnection is shown in Fig. 2. The reaction and sensing chamber along with three inlets and outlet so that allows the three different solutions; analyte, washing and micelles which be full of the required analyte for an immunoassay to be one after the other inserted into the chamber with the MFC and the sensing surface. To increase detection and analysis sensitivity of the transducer the MFC was designed to reduce interactions between biochemical reagents [25].

Microfluidic channel was fabricated using soft photolithography technique in a cleanroom protocols. The microfabrication process of the microfluidic channel can be describing simply by preparing mold and master where the mold can be preparing using in different ways such etching silicon using TMAH [26], etching glass, or soft photolithography technique and master can be preparing by coating polymer such as PDMS on the mold. The soft photolithography technique will be infolded here for simplicity and reasonably priced technique. The protocol as follows: firstly; preparing the mask on a thick glass board; depends on the available photoresist either positive or negative. Secondly; lay the slice on the spinner and pour few drops of SU-8 (SU-8, Microchem, MA) then spinning it for 1 min at 5000 rpm. Afterward place the slice in the oven for 15min for hardening the structure. Thirdly; place the slice on the aligner; Top and Bottom Side lithography EVG620, to pattern the required shape of MFC, afterward developed the photoresist and dry it with dry air or nitrogen gas. Fourthly; a polydimethylsiloxane (PDMS) is well preparing using cast-coating and cure protocols then pouring it over the mold and leaving it over night in the oven at 95 oC. Fifthly; the structure of MFC is almost ready for drilling the holes for the inlets and outlet. Sixthly; bring the final PDMS after cutting in proper way fitting the desired MCF shape to substrate where sensing surface; IDMA present then performing wafer bonding using O2 Asher protocol to seal the microfluidic channel. Finally; the substrate should make possible for electrical interconnections such as connections for RE, CE, WEC, and WEG.

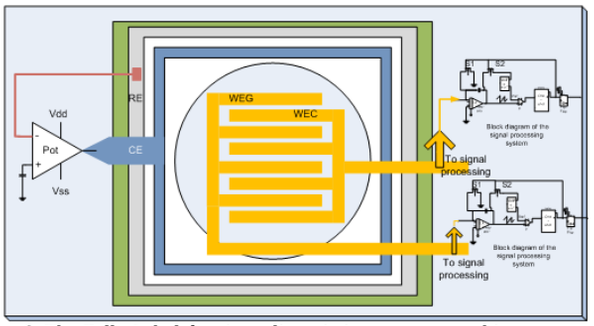


Fig. 6. The Fully Label-free impedimetric immunosensor chip.

Mostly, the electrochemical cell consist of three electrodes; counter, reference, and working electrodes, to be fabricated using soft photolithography technology. This cell called; also, Microfluidic cell (MFC) and it was developed to have four electrodes instead of three to handle DNA detecting along with bacteria sensing. The core of MFC is interdigitated microelectrodes (IDMA) using Au as a good material for immobilization, and the collector (WCE) and generator (WEG) electrodes; as well, where the reference is made of Ag/AgCl and Pt for counter electrode.

In biomedical applications, dissolved oxygen measurement (DOM) is a key issue. In conventional electrochemical cell that consists of Au working electrode, Pt counter electrode and Ag/AgCl reference electrode; oxygen is reduced at the working electrode and released at the counter electrode. In such case where large distance between the working electrode and the counter electrode is strongly presented the response of the sensor is inadequate, thus, redox cycling does not achieve the target in biomedical applications. Credit for MEMS microfabrication techniques this issue is resolvable, where micro/nanotechnology is employing to fabricate tiny devices with very narrow gaps, which is a proper geometry for redox cycle to occur.

The narrowest gap between the working and counter electrodes the highest opportunity for redox cycle to increase, the biomedical particles that be inherent in the gap between microelectrodes can be readily detected; accordingly. MEMS technology allows measuring the amplification factor where the conventional techniques cannot. The amplification factor is the ratio of the redox cycling current flowing through the working electrodes to the current flowing through the reference electrode. In addition; to improve the mechanism of the redox cycle where interdigitated microelectrodes array are used in the electrochemical cell, a four microelectrodes are replaced instead of three electrodes. These microelectrodes are Working Electrode Collector (WEC), Working Electrode Generator (WEG), counter electrode, and reference electrode as shown in Fig. 6. Throughout the operation of the four-microelectrode configuration, where WEC and WEG are very close to each other, the redox cycling process can easily perform in a reversible operation; such as holding WEG at potential to drive the reduction, at the same time as holding WEC at potential to drive the oxidation [29].

Biosensors based amperometric sensors incorporating with interdigitated microelectrodes arrays (IDMA) are powerful technique for detecting biomolecules. The performance and high sensitivity of the biosensor rely on minimum gap between the fingers, which eventually lies on the fabrication process. In the other hand, the biosensor is limited by the capability of the biomolecules to be reversible redox cycling and the threshold space that require to perform the detection for the what wanted and undesirable reactions that cannot be avoided.

The patient usually suffers from either excess (hyperthyroidism) or deficiency (hypothyroidism) of thyroid hormone; T4/T3, therefore, testing sample of the patient blood to determine the level of thyrotropic hormone (TSH), which its standard reference range between 0.4 and 5.0 µIU/mL for adults. The curative target range TSH level for patients on treatment ranges between 0.3 to 3.0 µIU/L.

IV. Conclusion

The new configuration will help to detect the level of TSH and in the light of the result, a proper dose will recommend by the physician to the patient to bring him to the normal activity with no suffering from any symptom of this disease. The performance of the biosensor is achieved by adding the two working electrodes. MEMS technology allows fabricating tiny features to perform a reversible process of redox cycling. The IDMA are essential to be embedded in the MFC for detecting the biomolecular sample.

References

[1] Park, S. M Yoo, JS; Chang, BY; Ahn, ES, "Novel instrumentation in electrochemical impedance spectroscopy and a full description of an electrochemical system", Pure and Applied Chemistry 78 (5): 1069-1080, MAY, 2006.

[2] Chang, B.-Y.; Park, S. M.," Theory and applications of real-time electrochemical impedance spectroscopy measurements", 213th ECS meeting Abstracts 2008, 2, 1136-2008.

[3] Byoung-Yong Chang , Su-Moon Park, "Integrated Description of Electrode/Electrolyte Interfaces Based on Equivalent Circuits and Its Verification Using Impedance Measurements", pp 1052–1060, December 24, 2005.

[4] Jin-Young Park, Byoung-Yong Chang, Hakhyun Nam, Su-Moon Park, "Selective Electrochemical Sensing of Glycated Hemoglobin (HbA1c) on Thiophene-3-Boronic Acid Self-Assembled Monolayer Covered Gold Electrodes", pp 8035–8044, Oct 1, 2008.

[5] Jin-Young Park, Su-Moon Park, "DNA Hybridization Sensors Based on Electrochemical Impedance Spectroscopy as a Detection Tool", Sensors, 9, 9513-9532, 2009.

[6] Mehrab Mehrvar Mustafe Abdi, "Recent developments, characteristics, and potential applications of electrochemical biosensors", Analytical Sciences Vol. 20, No. 8 p.1113, 2004.

[7] Tomoko Kimura, Keiko Nakanishi, Terumichi Nakagawa, Akimasa Shibukawa, Katsumi Matsuzaki, "Simultaneous determination of unbound thyroid hormones in human

plasma using high performance frontal analysis with electrochemical detection", Journal of Pharmaceutical and Biomedical Analysis, Volume 38, Issue 2, 15 June 2005, Pages 204-209.

[8] L. Chung, R.C. Baxter, "Detection of growth hormone responsive proteins using SELDI–TOF mass spectrometry", Growth Hormone & IGF Research, Volume 19, Issue 4, August 2009, Pages 383-387.

[9] L. Chung, R.C. Baxter, "Detection of growth hormone responsive proteins using SELDI–TOF mass spectrometry", Growth Hormone & IGF Research, Volume 19, Issue 4, August 2009, Pages 383-387.

[10] Maria J. López de Alda, Damià Barceló, "Determination of steroid sex hormones and related synthetic compounds considered as endocrine disrupters in water by liquid chromatography–diode array detection–mass spectrometry", Journal of Chromatography A, Volume 892, Issues 1-2, 15 September 2000, Pages 391-406.

[11] Vitaliano Borromeo, Anna Berrini, Camillo Secchi, Gian Franco Brambilla, Alfredo Cantafora, "Matrix-assisted laser desorption mass spectrometry for the detection of recombinant bovine growth hormone in sustained-release form", Journal of Chromatography B: Biomedical Sciences and Applications, Volume 669, Issue 2, 21 July 1995, Pages 366-371.

[12] G. Csaba and É. Pállinger, "Thyrotropic hormone (TSH) regulation of triiodothyronine (T3) concentration in immune cells", Inflammation Research, Volume 58, Number 3, March, 2009.

[13] Kosuke Ino, Yusuke Kitagawa, Tsuyoshi Watanabe, Hitoshi Shiku, Masahiro Koide, Tomoaki Itayama, Tomoyuki Yasukawa, Tomokazu Matsue, "Detection of hormone active chemicals using genetically engineered yeast cells and microfluidic devices with interdigitated array electrodes", Electrophoresis 2009, 30, 3406–3412.

[14] Claudia Bich, Cédric Bovet, Natacha Rochel, Carole Peluso-Iltis, Andreas Panagiotidis, Alexis Nazabal, Dino Moras, Renato Zenobi, "Detection of Nucleic Acid–Nuclear Hormone Receptor Complexes with Mass Spectrometry", Journal of the American Society for Mass Spectrometry, In Press, Corrected Proof, Available online 28 December 2009.

[15] Yosef Yarden, Robert A. Weinberg, "Experimental approaches to hypothetical hormones: Detection of a candidate ligand of the neu protooncogene (growth factor/receptors/tyroine phosphorylation/ras oncogene)", Cell Biology, Proc. Nati. Acad. Sci. USA Vol. 86, pp. 3179-3183, May 1989.

[16] Georg Hennemann, Eric P. Krenning, "The kinetics of thyroid hormone transporters and their role in non-thyroidal illness and starvation", Best Practice & Research Clinical Endocrinology & Metabolism, Volume 21, Issue 2, June 2007, Pages 323-338.

[17] Josef Köhrle, "Local activation and inactivation of thyroid hormones: the deiodinase family", Molecular and Cellular Endocrinology, Volume 151, Issues 1-2, 25 May 1999, Pages 103-119.

[18] George Lovell, Patrick H. Corran, "Determination of L-thyroxine in reference serum preparations as the o-phthalaldehyde-N-acetylcysteine derivative by reversed-phase liquid chromatography with electrochemical detection" Journal of Chromatography B: Biomedical Sciences and Applications, Volume 525, 1990, Pages 287-296.

[19] Seyed Ahmad Mozaffari, Taihyun Chang and Su-Moon Park, "Diffusional Electrochemistry of Cytochrome c on Mixed Captopril/3-Mercapto-1-propanol Self-Assembled Monolayer Modified Gold Electrodes", pp 12434–12442, June 17, 2009.

[20] Omowunmi A. Sadik, Austin O. Aluoch, Ailing Zhou, "Status of biomolecular recognition using electrochemical techniques", Biosensors and Bioelectronics, Volume 24, Issue 9, Pages 2749-2765, 15 May 2009.

[21] Nair, P.; Alam, M., "Dimensionally frustrated diffusion towards fractal adsorbers", Physical Review Letters 2007, 99(25), 256101, 2007.

[22] Dorothee Grieshaber, Robert MacKenzie, Janos Voros, Erik Reimhult, "Electrochemical Biosensors-Sensor Principles and Architectures", Sensors, 8, 1400-1458, 2008.

[23] Eugenii Katz, Itamar Willner, "Probing Biomolecular Interactions at Conductive and Semiconductive Surfaces by Impedance Spectroscopy: Routes to Impedimetric Immunosensors, DNA-Sensors, and Enzyme Biosensors Electroanalysis", Volume 15, Issue 11, Date: July 2003, Pages: 913-947.

[24] Mehrab Mehrvar Mustafe Abdi, "Recent developments, characteristics, and potential applications of electrochemical biosensors", Analytical Sciences Vol. 20, No. 8 p.1113, 2004.

[25] Karel Stulík, Christian Amatore, Karel Holub, Vladimír Mareček, Włodzimierz Kutner, "Microelectrodes. Definitions, Characterization, and Applications", Pure Appl. Chem., Vol. 72, No. 8, pp. 1483–1492, 2000.

[26] Tashtoush, A. Landsberger, LM; Essalik, A.; Kahrizi, M. Paranjape, M. Currie, JF; Pandy, A. "Experimental Characterization of the Si/Al/Tetramethylammonium Hydroxide System", Journal of the Electrochemical Society 148: (7) p. C456-C460, 2001

[27] Jian Wu and Willy Sansen, "Micro Oxygen Sensor With Redox Cycling", Meeting Abstracts- Electrochemical Society -All Divisions-; 1; 1564 Electrochemical Society Meeting; 201st, Electrochemical Society, ISBN#1566773660, 2002.

[28] Edgar D. Goluch, Bernhard Wolfrum, Pradyumna S. Singh, Marcel A. G. Zevenbergen, Serge G. Lemay, "Redox cycling in nanofluidic channels using interdigitated electrodes", Anal Bioanal Chem (2009) 394:447–456, DOI 10.1007/s00216-008-2575-x.

[29] Adam E. Cohen, Roderick R. Kunz, "Large-area interdigitated array microelectrodes for electrochemical sensing", Sensors and Actuators B 62, 23–29, 2000.

Implementing hardware for new genetic algorithms

Fariborz ahmadi[1,a]
[1]Department of computer science, Islamic Azad University,Ghorveh branch,Ghorveh, Iran
sanandajstudent@gmail.com

Reza Tati[2,b]
[2]Department of computer science, Islamic Azad University, Mianeh branch, Mianeh, iran
tati_r@yahoo.com

Abstract— **Genetic algorithm is a soft computing method that works on set of solutions. These solutions are called chromosome and the best one is the absolute solution of the problem. The main problem of this algorithm is that after passing through some generations, it may be produced some chromosomes that had been produced in some generations ago that causes reducing the convergence speed.**

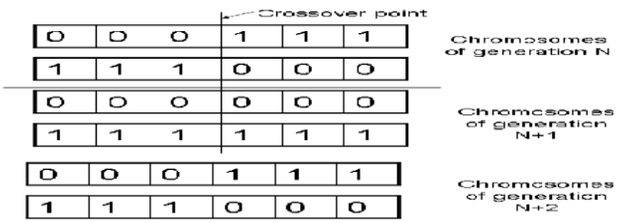
From another respective, most of the genetic algorithms are implemented in software and less works have been done on hardware implementation. Our work implements genetic algorithm in hardware that doesn't produce chromosome that have been produced in previous generations. In this work, most of genetic operators are implemented without producing iterative chromosomes and genetic diversity is preserved. Genetic diversity causes that not only don't this algorithm converge to local optimum but also reaching to global optimum. Without any doubts, proposed approach is so faster than software implementations. Evaluation results also show the proposed approach is faster than hardware ones.

Keywords: Genetic algorithm, Hardware engine, Chromosome, Ovum

I.INTRODUCTION

Genetic algorithm is a soft computing method that works on set of solutions. Each of these solutions are called chromosome and each population consists of a certain number of them. By applying some operators like selection, crossover, and mutation on the chromosomes of current population, the next generation is produced.

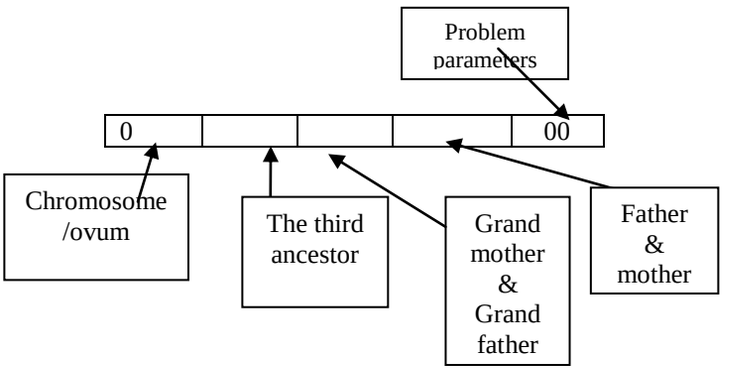
In the GA explained above, it is observed that by passing through a number of generations, some chromosomes may be produced that are the same as the chromosomes in the previous generations. It is clear that these chromosomes are not suitable ones because they were eliminated in the previous generations because of low fitness value. The main problem of these chromosomes is increasing calculations of each generation because GA operators in current generation are applied on chromosomes that were produced and deleted in the previous generations [1]. This process also causes decreasing of convergence speed toward problem solutions. Assume two individuals in generation N; produce two offspring by applying crossover and mutation. Now in generation N+1, if these two offspring recombine together, it may be produced chromosomes that are similar to ones in generation N, consider the figure 1[1].



Figure(1). Producing of repeated chromosomes [1]

In [1] individuals are divided into two groups namely, chromosomes and ovums. In order to prevent from producing of iterative individuals, the recombination of the same sex individuals is forbidden. Also, the parent of each individual together with grandparent and third-ancestor are put in the individual's structure. Figure2 shows the individuals structure presented in [1]. In this method, the following recombination is denied to avoid from producing of repeated individuals.

- parent(ovum) = = parent (chor)
- parent (ovum) = = grand parent (chor)
- parent (ovum) = = grand-grand parent (chor)
- ovum be in the previous generations of chor
- chor be in the previous generations of ovum
- parent (chor) = = parent (ovum)
- parent (chor) = = grand parent (ovum)
- parent (chor) = = grand-grand parent (ovum)
- chor=ovum



Figure(2). the structure of individuals

Therefore, selection algorithms related to ovum and chromosome are changed as following.

I. Select (Chromosome)

II. Move Chromosome to the intermediate generation to crossover

Algorithm(1). Pseudo code relating to chromosome selection

Select(ovum)

if parent (ovum) = = parent (chromosome) OR

parent (ovum) = = grand parent (chromosome) OR

parent (ovum) = = grand-grand parent (chromosome) OR

chromosome = = ovum OR

vice versa

then

Repeat from first line 'select another ovum'

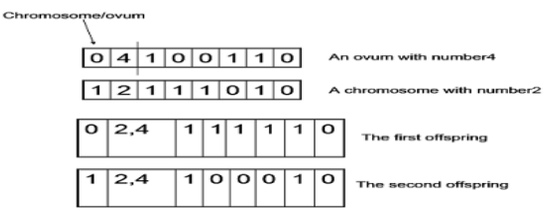
else

DELETE(ovum)

move ovum to the intermediate generation to be crossover with chor that selected in algorithm۱

Algorithm(2). Pseudo code relating to ovum selection

Also in [1], the recombination method put the information of recombined individuals into next stage of produced individuals. Figure 3 shows the process of crossover method.



Figure(3).recombination operation in proposed approach[1]

For another respective, almost all of the genetic algorithms are implemented in software. But in this work, we propose parallel architecture of hardware for above mentioned genetic algorithm and compare it with previous works. The evaluation results show the usefulness of the researchers enterprise work. Previous works [1,2,4,5] can be divided into three categories. 1- Those that implement fitness function on hardware and other parts of genetic algorithm like crossover, mutation, and selection in software. The disadvantage of this method is that hardware depends on a problem. 2- Those that implement fitness on software and other operators in hardware. In this method hardware is independent of a problem but the running time is increased [6, 7]. 3- Those that implement both fitness and operators in hardware [2,4,16]. In this work, we proposed an approach that improves this type of method for new approach to standard genetic algorithm [1].

This paper divided into 5 sections. In section 2, the proposed approach is described and in section 3, architectures of genetic operators are explained. Sections 4, 5 evaluate the results and draw some conclusion, respectively.

II. OVERALL VIEW OF PROPOSED ARCHITECTURE

The difference of genetic algorithm with other soft-computing methods is individuals' representation. Individuals in genetic algorithm are a stream of binary bits that are very attractive to hardware implementation. The selection, crossover, and mutation are generic operators and never depend on the problem types. These characteristic make them easy to implement. The main problem of implementing of genetic algorithm is fitness function. For this reason, to calculate it, hardware of neural networks is used [2, 9, 10]. Ideal estimation of each individual can be obtained using this network. Another reason to use neural network to compute fitness is its simplicity and stochastic representation of signals that reduces the hardware area. In figure4 the overall view of

proposed approach is showed. Genetic operators of this architecture are as follow.

- Selection operator: Rolette wheel & tournament
- Crossover operator: single point & two-point
- Mutation operator: bit complement
- Replacement operator: steady state & generalization

The main and novel contribution of this work consists of using all genetic operators in hardware implementation without producing repeated individuals in alternative generations. Hardware implementations of these operators are described in section 3. The main advantage of this architecture over previous works [16] is using fitness of next generation in replacement operator to avoid re-computing of fitness function in each generation. For this reason, fitness of each chromosome is kept in bank of registers and individuals' fitness of next generation is provided as input. It is necessary to remind that each register of register bank involve all part of individual structure mentioned above. When replacement operator is steady state and the fitness of input chromosomes are greater than ones that kept in the register bank, the input chromosomes are replaced. It should be noted that Fitness of each chromosome in the first population is considered as 0 [1].

I. OPERATORS ARCHITECTURE

The architectures of selection, crossover, and mutation operators and shared memory are described in this section.

A. shared memory for generational population

All of the chromosomes are kept in synchronized bank of registers. Each register keeps individuals as described in figure 2. These registers can be updated and written. The registers can be writte by the following steps.

1- Whenever the comparator output is 0 and replacement operator is generality.

2- Whenever the comparator output is 0 and replacement operator is steady state and fitness measurements of chromosome in next generation are greater than one's in current generation.

It is necessary bearing in mind that the output of comparator is 0 if the number of generation (*gene*) is greater than *counter* or fitness of all chromosomes is less than *fitness.* In such condition genetic algorithm continues to work and current generation is replaced by next generation.

B. Random number generator

This generator is used to produce crossover & mutation probability. In this work *linear feedback shift register* are used. More details about random number generator can be found in [11, 12].

C. Roulette selection components

In this subsection the roulette operator is elaborated. Inputs of this component are generational chromosomes($I_1, I_2, \ldots, I_n$), fitness function ($F_1, F_2, \ldots, F_n$), and cumulative sum of these fitness. This component selects two individuals and passes it to crossover unit. The hardware of this component is easy to implement .for more details see [1, 16].

D. Tournament selection component

One of the novel contributions of this work is tournament operator. This component takes several chromosomes together with their fitness and their information and output the fittest one that not satisfy any of the following conditions.

- parent (ovum) = = parent (chor)
- parent (ovum) = = grand parent (chor)
- parent (ovum) = = grand-grand parent (chor)
- ovum be in the previous generations of chor
- chor be in the previous generations of ovum
- parent (chor) = = parent (ovum)
- parent (chor) = = grand parent (ovum)
- parent (chor) = = grand-grand parent (ovum)
- chor=ovum

algorithm 3 shows the hardware algorithm of this unit.

1. Choose 3 individual in the population
2. Compare them according to their fitness's
3. Output the best one to individual1
4. repeat step 1 to 3 to select the second one
5. If two selected individuals doesn't satisfy any of the conditions, put them into final individual1 and final individual2

Algorithm 3. Algorithm of tournament selection

In this algorithm, three chromosomes are selected and raced to be one of the individual intermediate generations. This routine is also repeated to select the second chromosome. The hardware implementation of this component is shown in figure 5. Finally, if the selected individuals don't have any above mentioned properties, they are selected and put into final individual1 and final individual 2.

The synchronized bank of registers that keep the individuals are input of the multiplexers, *comp* unit compare the fitness of these chromosomes and also relationships between them. After passing from these steps, comp unit outputs the number of those individuals. This number is applied to selection signal of multiplexer in the middle of the figure 5. In other word, the appropriate chromosome is selected and sent to individual1. The same process is repeated for selecting ovum. After that, the properties of two individuals are compared to judge about their relationship. In this research iterative process is implemented using state machine (controller) [1, 16].

E. Crossover component

Another contribution of this work is implementing a unit for new crossover using both one point and also two-point recombination. To crossover, Firstly, the Random Number Generator produce a random number say *p*. if $p < \mu_{rate}$ then crossover operator is applied. In the case that crossover applied, the bits of the less significant half of the randomized number is used as the first crossover point and the most significant part as the second one. It should be reminded that in the single-point crossover, the bits of the most significant

part are considered as 0. Finally, the information of selected individuals are shifted to next stage of produced individuals like as figure 2. Figure 6 shows the hardware implementation of crossover. In this hardware if the value of *two-point* register be one, two-point crossover is applied; otherwise, one-point crossover is applied.

F. mutation components

This component has architecture like as crossover. To implement this component, a random number is produced using Random Number Generator. When this value is less than μ_{rate}, the mutation operator is applied. For more detail see [1, 16].

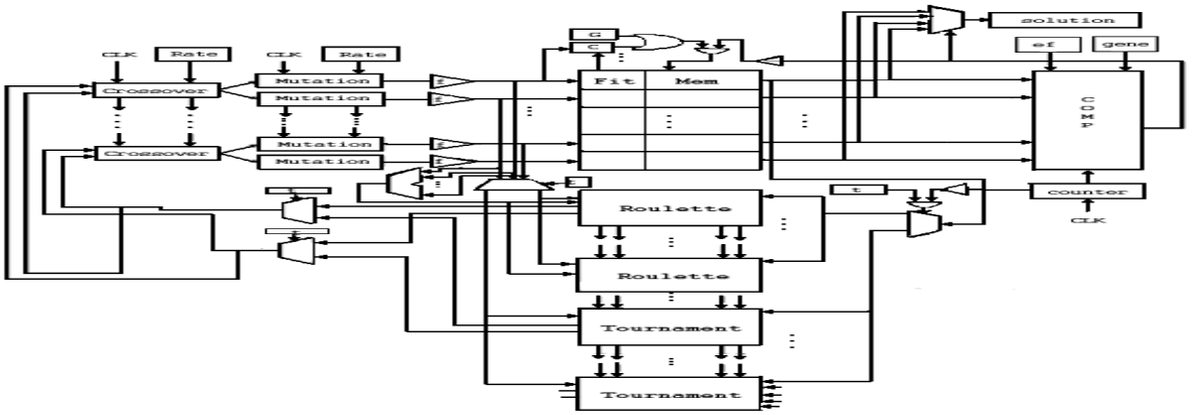


Figure 4. Overall view of architecture [2]

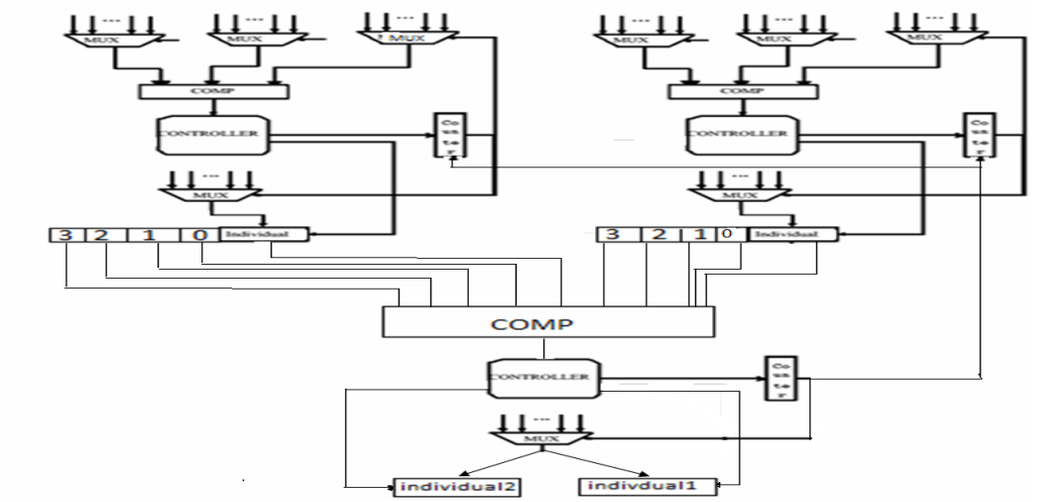


Figure 5. Architecture of tournament selection

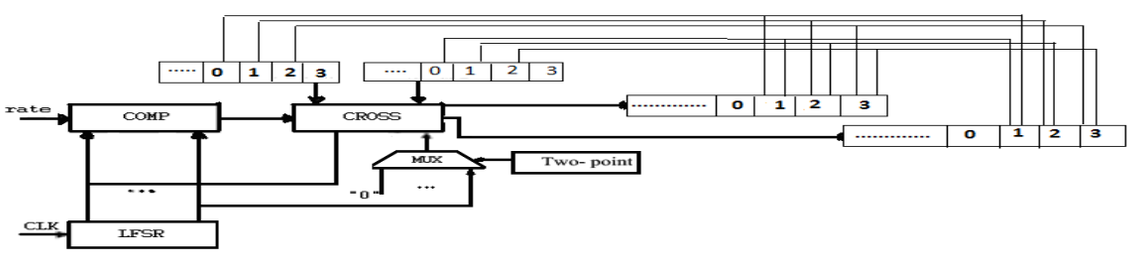


Figure 6. Hardware of crossover operator

G. *Fitness components*

For the reason that the fitness function will not be depending on a problem, the neural network is used to estimate fitness measure [10] in this work.

In [10] neural network has been implemented using stochastic signals and therefore reduces very significantly the hardware area required for the network. For the genetic hardware implementation, the number of input neurons is the same as the size of individuals in population. The output neuron is augmented with a shift register to store the final result. The training phase is supposed to be performed before the first use within the hardware genetic algorithm.

IV. EVALUATION RESULTS

Hardware for new genetic algorithm was simulated and then programmed into a Sparatan3 Xilinx FPGA [13]. To assess and evaluate of proposed architecture the function in [14] was used to maximize. This function was also used in previous works to evaluate their hardware implementation for genetic algorithm. This function is as follows.

$$F(x,y)=21.5+x\sin(4\Pi x)+y\sin(2\Pi y) \quad (1)$$

$$-3.9\leq x \leq 12.1$$

$$4.1 \leq y \leq 5.8$$

It is clear that this function is not easy to maximize and is a very attractive metric to evaluate and assess of proposed approach. The evaluation results are shown in table 1.

In this table the area is expressed in terms of CLBs and the time is in second, also previous works including the software and hardware implementation and our proposed approach have been compared.

The advantage of hardware approach to software approach is speed up. Therefore, in this work, besides software implementation, hardware implementations are studied and evaluated. It is clear that the area required by this work is more than [5, 15. 16]. From another respective, our proposed hardware is faster than [1, 15] and also approximately is faster than [16]. The main advantage of proposed approach over [15, 16] is applying almost all of the genetic operators that result in genetic diversity. In the [16] only roulette selection, two-point crossover and mutation have been implemented, but in our work and [1] not only have been these operators implemented but also tournament, one-point crossover, and steady state replacement have been designed and implemented. In our work, unlike [1, 5, 15, 16] producing of the iterative chromosome is avoided that causes this approach to increase in convergence speed in comparison to other approaches [1, 5, 15, 16]. In our work like [1] the operators can be changed in the next generation by changing a value of corresponding registers i.e. genetic diversity is respected. For example, in one generation, tournament selection, two-point crossover, and general replacement may be used and in next generation roulette selection, one-point crossover, and steady-state replacement be used. This diversity is implemented neither in [15] nor in [16].

Table1. evaluation results

Implementation	Time	Area	Area*Time	Absolute Answer
Software approach	42300	0	0	32.21
[15]	1012	850	860200	38.8211
[5]	212	1943	411916	38.8214
[2]	194	2016	391104	38.848912
Proposed Approach	181	2114	382634	38.8476

V. CONCLUSION

Genetic algorithm is an attractive method that used to solve NP-Hard problems. It is observed that by passing through some generation, it may be produced some chromosome that were produced in some generation ago. Therefore some methods are required to solve this problem. By putting ancestor information of each produced chromosome in its structure, this difficulty can be curable. From another respective, genetic algorithm is usually implemented in software and less works have been done to implement this algorithm in hardware. The main and novel contribution of this work is implementing almost all of the genetic operators that results in genetic diversity without producing of iterative individuals. In previous researches that work on hardware genetic algorithm, some of these operators are implemented and some of them are not implemented in hardware. To evaluate fitness measurements, neural network has been used and shown area required using this approach is significantly decreased in compare to previous works. In spite of previous works, proposed architecture preserve genetic diversity that causes to speed up in our architecture. From evaluation results can be drawn a conclusion that by changing genetic operators in specific conditions, convergence speed is significantly increased. For example, when some chromosome are produced in alternate generations, changing one-point crossover to two-point decreases production of these chromosome and improve genetic diversity and convergence speed. Evaluation results show advantage of our proposed architecture over previous works.

REFERENCE

[1]Fariborz Ahmadi, Amir Shikh Ahmadi, New approach to standard genetic algorithm,intenational journal of computer applications,vol 32-num10.pp 46-50

[2]Fariborz Amadi, Reza Tati, New hardware engine for genetic algorithm, In Proc 5th international conference on genetic and evolutionary computing,2012

[3] Liu, J., A general purpose hardware implementation of genetic algorithms, MSc.Thesis, University of North Carolina, 1993.

[4] Scott, S.D., Samal, A. and Seth, S., HGA: a hardware-based genetic algorithm,In Proc. ACM/SIGDA 3rd. International Symposium in Field-Programmable Gate Array, pp. 53–59, 1995.

[5] Turton, B.H. and Arslan, T., A parallel genetic VLSI architecture for combinatorial real-time applications – disc scheduling, In Proc. IEE/IEEE International Conference on genetic Algorithms in Engineering Systems, pp. 88–93, 1994.

[6] Bland, I.M. and Megson, G. M., Implementing a generic systolic array for genetic algorithms. In Proc. 1st. On-Line Workshop on Soft Computing, pp 268–273, 1996.

[7] Megson, G. M. Bland, I. M., Synthesis of a systolic array genetic algorithm. In Proc. 12th. International Parallel Processing Symposium, pp. 316–320, 1998.
[8] D.c Goldberg. Genic algorithm in search, optimization, and machine learning. Addison welsey, 1989.
[9] Gaines, B.R., Stochastic Computing Systems, Advances in Information Systems Science, no. 2, pp. 37–172, 1969.
[10] Nedjah, N. and Mourelle, L.M., Reconfigurable Hardware Architecture for Compact and Efficient Stochastic Neuron, Artificial Neural Nets Problem Solving
Methods, Lecture Notes in Computer Science, vol. 2687, pp. 17–24, 2003.
[11] Bade, S.L. and Hutchings, B.L., FPGA-Based Stochastic Neural Networks –Implementation, IEEE Workshop on FPGAs for Custom Computing Machines, Napa Ca, April 10–13, pp. 189–198, 1994.
[12] Brown, B.D. and Card, H.C., Stochastic Neural Computation I: Computational Elements, IEEE Transactions on Computers, vol. 50, no. 9, pp. 891–905, September 2001.
[13] Xilinx, http://www.xilinx.com/, 2004.
[14] Michalewics, Z., Genetic algorithms + data structures = evolution programs, Springer-Verlag, Berlin, Second Edition, 1994.
[15] Scott, S.D., Seth, S. and Samal, A., A hardware engine for genetic algorithms, Technical Report, UNL-CSE-97-001, University of Nebraska-Lincoln, July 1997.
[16] N.Nedjah, Pararllel evolutionary computations, springer 2006.

A Work-Optimal Parallel Connected-Component Labeling Algorithm for 2D-Image-Data using Pre-Contouring

Henning Wenke, Sascha Kolodzey, Oliver Vornberger
University of Osnabrueck, Germany, 49069 Osnabrueck
Email:hewenke@uos.de, skolodze@uos.de, oliver@uos.de

***Abstract*—Connected-component labeling (CCL) is a well-known problem with many applications, e.g. in image processing. In this paper, we describe a parallel algorithm to solve 2d-image-data CCL-problems resulting in linear overall work. It can be classified as an one-pass algorithm, since no temporary labels are required. Our algorithm initially extracts the connected components' independent contour-segments. Then, these are unified and labeled. Finally, the image is filled in order to label all non-contour pixels. Our approach is motivated by the observation that a line has, independent of its complexity, exactly two ends. Thus, two independently extracted lines (e.g. due to the parallel process) can be unified by only adjusting their ends, resulting in constant costs for this operation. Additionally, all contours are extracted as directed cyclic linked-lists, where the in and out degree of each node is always one. This topology is a simpler to deal with when compared to the original pixel grid.**

I. Introduction

One very common method in image processing is Connected Component Labeling (CCL). The CCL procedure assigns a unique label to each set of connected pixels in an image. In this paper, we will limit ourselves to 2d-rectilinear image data. There, a pixel's neighborhood can be defined in two ways: 4-connected or 8-connected [1]. Important applications of CCL algorithms are pattern recognition, computer vision and image processing. This includes, for example, character recognition [2], [3]. In addition, according to [4], the object's contours are also helpful for 2d-object recognition in many cases. Given an image consisting of n pixels, a CCL-algorithm is considered optimal if it executes in O(n) time. In general, considering the immediate vicinity of a pixel is not sufficient to determine global labels. Therefore, most approaches assign temporary labels also known as provisional labels. Hernandez-Belmonte [5] and He [6] identify three classes of CCL-algorithms: The first category is multi-pass algorithms, like [7]. Because the number of passes depends on the image's content, variants of this approach are typically not optimal. In contrast, two-pass algorithms make two distinct passes through the image. In a first pass, the image is scanned to determine equality of neighboring pixels, assign provisional labels and record the related equivalence information. Then, label equivalence information can be analyzed in order to determine the final labels. At last, a second pass assigns the final labels to the corresponding pixels. A theoretically optimal example of this class is the union-find algorithm [8]. The final category are one-pass algorithms, which go once through the image. Examples are algorithms based on contour-tracing, like [9] and its linear-time successor [4]. They scan a binary image from top to bottom and from left to right. Each time a contour is encountered in this process, a new label is generated. Then, the full contour is traced and the new label is applied to all the contour's points. Afterwards, pixels within a contour are labeled equal to the contour, using a scan-line algorithm. In case of unclassified regions completely surrounded by a connected-region, an inner-contour is generated and receives the same label as the corresponding outer contour. Thus, the scan-line algorithm can label pixels between the inner-contour and outer-contour using the outer-contour's label, while pixels within the inner-contour remain unlabeled. In the early 1990s, there was an interest in parallelizing CCL-algorithms to use them in image analysis applications. [10] considered CCL-algorithms to belong to the most time consuming algorithms used in pattern recognition. Recently, in line with the growing utilization of graphics processing units (GPUs) for general computation, again some attempts have been made to parallelize CCL-algorithms. Hawick [11] implemented a union-find CCL-algorithm for image data in Cuda. Their implementation was faster than an optimized CPU implementation for most of the tested instances. A drawback of their approach was the inability to guarantee that a single equivalence-tree is constructed for each connected component (CC) in the first pass, due to the parallelism. To solve this problem, the first pass has to be repeated several times, until all trees belonging to a single component are merged into a single one. The exact number of necessary passes is data-dependent and thus not known in advance, so after each iteration the labels' correctness has to be checked. At last, Stava et al. [12] have further enhanced this approach and optimized it for the Nvidia Fermi architecture. They noted a significant speedup when compared to Hawick.

Our contribution: We present a novel deterministic CCL-algorithm for a PRAM. As far as we know, there is no parallel contouring based CCL-algorithm in literature. Our algorithm presented in this paper preserves all properties of an optimal line-following CCL-algorithm like [4]. Plus, it is parallel and as such delivers a better asymptotic running-time. Furthermore, we have not found a CCL-algorithm in literature, which is both parallel and has an optimal asymptotic total complexity on a CRCW-PRAM. At last, we compare an early OpenCL implementation of our algorithm on a GPU to results of Stava [12] in practice.

II. Algorithm

A. Overview

We pursue a parallel one-pass approach to extract the CCs inner and outer contours in a first step. Therefore, contour-line-segments (CS) are extracted in parallel. These segments are subsequently unified, using a data-dependent or a data-independent conflict-free parallelism pattern. During the unification process, necessary information to distinct inner- from outer contours is aggregated and stored in the surviving elements. Then, surviving segments belonging to an outer contour receive a label. Afterwards, the lines' unification is reversed, using a scheme exactly reverse to their prior unification. In this process the surviving CSs' labels are passed to each splitted CS. So, all segments of one contour will receive this label. Finally, the image is filled in order to label all non-contour pixels, using a column-wise parallel and stack based scan-line approach. Our approach shares some properties, e.g. of the resulting lines, with the work of [4]. However, we do not classify our algorithm as line-following, since parallel extracted CS are never followed but merged in parallel to a complete contour.

B. Properties of the used contours

Our first sub-algorithm extracts all CSs of all CCs for all pixels in parallel. In order to enable subsequent sub-algorithms to unify these to complete contours, distinct inner from outer contours, label and finally fill them, several properties are required:

- A contour is always closed by definition.
- A contour (primarily) consists of the contained CC's outer pixels' outer edges.
- All properties of a CS can be computed based on its pixel's 8-way neighborhood.
- Each CS has one and only one unique successor and is predecessor to exactly one CS.
- The definition above results in distinct rotational directions for inner and outer contours.
- A pixel may contain up to 4 CS belonging to one or more contours. The final fill-algorithm decides within a pixel if a CC is entered and/or left. The in-out state-change must be recognized based on the entirety of a pixel's CSs' properties.

As a start, we describe one possible way to model the CS, which fits the specifications above. If the classification of one or more pixels of a pixel's 4-connected neighborhood differs from its own, one or more CSs for this pixel need to be extracted. The exact configuration of these extracted segments additionally depends on the pixel's 8-connected neighborhood. There are four potential *edge-parallel-CS-parts*, which are parallel to one of the pixel's left, right, upper or lower edges, respectively. One of these parts is generated if the pixel adjacent to the respective edge is classified differently. This is demonstrated left in figure 1. Light gray shaded squares indicate an identical classification or segmentation. Dark gray shaded pixels are not classified and therefore not part of any CC. Please note that this coloring scheme remains true for the rest of this paper. In addition to *edge-parallel-CS-parts* *link-CS-parts* are needed, if two cyclic consecutive pixels of a pixel's 4-connected neighborhood share the classification with

Fig. 1. Left: The 4 potential *edge-parallel-CS-parts* of a pixel p. c_l is set, if pixel l's classification differs from p's, for instance. Right:Link-CS-parts

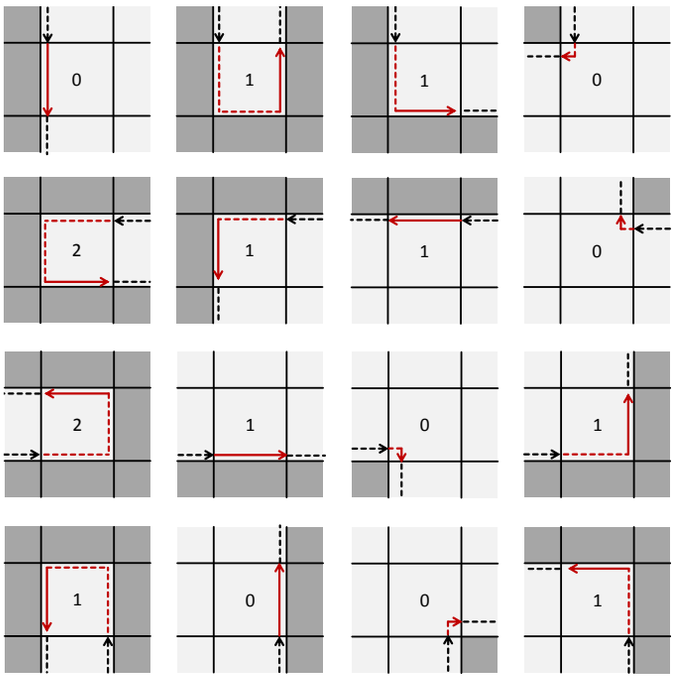


Fig. 2. 16 cases of a CS's course within a pixel. Numbers: ioCnt (vertical)

the central pixel under consideration and the pixel of the 8-neigborhood, which resides between the other two neighbors, has a different classification. As an illustration, see right part of figure 1. *Link-CS-parts* consist of a horizontal and a vertical half, which can be thought of as a extension of the neighboring pixel's *edge-parallel-CS-parts*, respectively. In order to unify individual CS, a unique successor has to be determined for every single CS. This has to be done in a parallel fashion, thus this operation is restricted to rely on local data. Thus, for each possible CS-part type a direction is defined as follows:

- Upper segments are directed left
- Lower segments are directed right
- Left segments are directed down
- Right segments are directed up

A unified CS within a pixel, consisting of edge-parallel-CS-parts and link-CS-parts, can be identified based on local data and without need of synchronization with other CS. Thus, the algorithm extracts unified CS within a pixel directly. Only these are considered in the remaining paper. All possible combinations of four succeeding and four previous directions result in 16 cases, which need to be distinguished, as shown in figure 2. Additionally, up to four of these CSs can appear within the same pixel. All combinations are possible where different CSs do not have a CS-part in common and do not cross each other. From this follows that at most one segment

can head to or come from the right, left and up and down pixel, respectively. Thus, it is possible to limit our further considerations to four types of CSs, based on the leaving-direction of their pixel. They are hence simply called *right*, *left*, *down* and *up*. Additional information, such as their course within their associated pixel can be stored as a property of a CS. However, one special case does not fit in this scheme. If a CC consists of a single Pixel, no CS leaves the pixel. This results in an undefined leaving direction. Since this situation can obviously be identified locally, no CS is extracted and the pixel can directly be labeled instead. Some possible combinations are shown in fig. 3. Notably, the right example shows the special

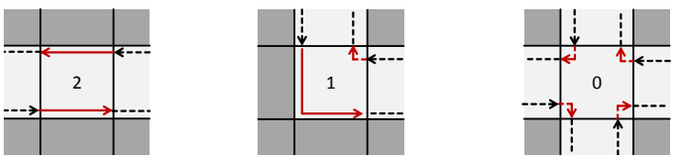


Fig. 3. Some combinations of CSs within a single pixel. Numbers: ioCnt (v)

case, where all four CSs exist. At last, inner and outer contours need to be dealt with in a different fashion. In case of one or multiple unclassified or differently classified areas enclosed completely by a CC, one or several inner contours occur (see figure 4). Since these contours are independent from each

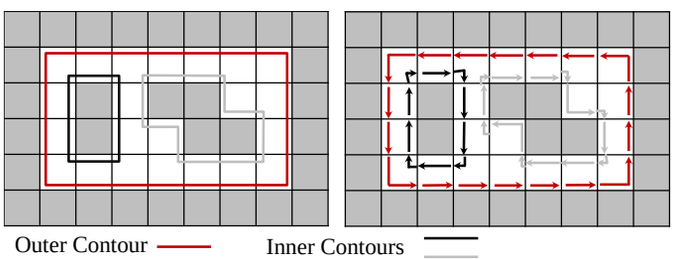


Fig. 4. Example of a CC containing enclosed unclassified areas. One outer contour and two inner contours are extracted.

other, they would receive different provisional labels. Thus, finding a distinct label for each associated pixel is difficult. So, we chose an approach based on the observation, that each CC has exactly one outer contour and zero, one or multiple inner contours. Obviously, once inner contours are classified as such and removed, the remaining outer contours can directly receive their final label. A line-following-approach, like [4], whose computation-scheme begins at the image's border and proceeds sequentially, can identify outer contours by simply finding them when firstly entering a new CC. Obviously, in case of a parallel approach, as desired here, this method of classification is not an option. Instead, a possible solution suitable for parallel execution may use the contour's rotational direction. The CSs' direction definition implies a counter-clockwise rotational direction in case of inner contours and a clockwise rotational direction in case of outer contours. As an illustration, see figure 4 (right).

C. Extraction of contour-segments (CS)

As outlined above, a CS is of type *left*, *right*, *up* or *down*, depending on the pixel, which contains the next CS in the defined contour-direction. For pixel p, a CS of one direction-type is created, if p shares the classification with the next pixel in the direction. Additionally, the next two pixels of the 8-neighborhood of pixel p starting after this neighbor-pixel and proceeding clockwise, need to be analyzed. Only if one of these has a different classification, the condition is completely met. Figure 5 shows all cases resulting in the generation of a CS of type *right*, for example. Other direction-types arise from rotationally-symmetric considerations. Some

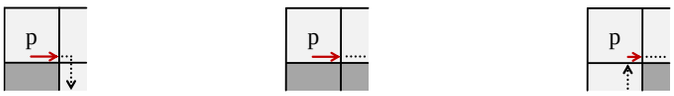


Fig. 5. All cases of a CS of direction-type *right* in pixel p being created

calculations require closed contours to operate. Following the above described procedure, CCs, which are in touch with an image's edge, would not receive a closed contour. Thus, these contours are closed by force. In order to do that, imaginary neighbors beyond the image's edges are considered. They are unclassified per definition. In the algorithm described in this paper, each CS has to be unified with its successor, which is contained in the pixel of the 4-connected neighborhood, matching its directional-type. This, for example, is the pixel residing left of its containing pixel if a CS's directional-type is *left* (see. figure 6). For this neighbor-pixel, the suitable of

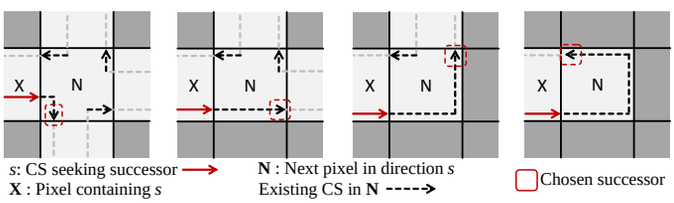


Fig. 6. CS of type *right* seeking a successor. Chosen is the first existing CS in a CCW-manner starting from down-direction.

the four possible CSs needs to be selected by the following way (see also figure 6): Find the first possible *exit* out of the pixel using a counter-clockwise test scheme. This correlates to the contour's rotational direction, as defined above. For this purpose, the first existing CS out of (*right*, *up*, *left*, *down*) is chosen, using a cyclic scheme and starting with the direction preceding its own. For example, in case of a CS of directional-type *right*, the right pixel's CSs are tested for existence in the order *down*, *right*, *up* and *left* (see figure 6). The detected successing CS is referenced by the field *suc*. The pseudocode responsible for each CS's extraction is contained in sub-algorithm 1.

D. Contour-Labeling: Data-Independent Approach

So far, individual CS have been extracted and each of those references its immediate successor. They are represented as directed cyclic linked-list, where each node has an in- and out-degree of one. Subsequently, closed contours have to be computed. Resulting in exactly one entity, here one single CS, representing the entire contour of a CC and receiving its label. Requested is a parallel computing-scheme for this task. As one alternative, we propose a data-independent approach where independent CS can be unified in parallel. As a start, let's take a look at the function unifyOp, which unifies two consecutive CSs. Here, each CS receives *head* and *tail*, that point to a

Algorithm 1: ExtractCS

```
Global Data: Pixel[] pa, CS[] ca;
pixelCnt n, image-width a, image-height b;
Precondition: p.class ∀ p of pa set;
for each Pixel p of pa in parallel do
    if p.class ≠ UNCLASSIFIED then
        for each Dir d ∈ { right, left, down, up } do
            if Condition for CS c in d fulfilled then
                c.status ← EXISTING;
                c.props ← calcProps(p,c);
                c.ioCnt ← calcIoCnt(p,c);
                c.head ← c; c.tail ← c;
                c.suc ← calcSuccessor();
            else
                c.status ← NOT_EXISTING;
            endif
            storeCS(p, d, c);
        end
    endif
end
```

Function unifyOp(CS c)

```
if c.status =EXISTING then
    CS c_n ← c.suc ;
    if c ≠ c_n.head then
        c.tail.props ← aggregateInfo(c.tail.props,
        c_n.props);
        c.tail.head ← c_n.head;
        c_n.head.tail ← c.tail;
    else if checkIfOuterContour(c_n.props) then
        c_n.label ← get1dAddress(c_n);
    endif
endif
```

start- and respectively an end-CS of a CS-strip. Initially, each CS's *head* and *tail* refer to itself. Afterwards, two subsequent CSs can be unified to a single segment representing both. This is done by letting the *head* reference of the contour-segment, which is referenced by the *tail* of the rear CS, point to the CS referenced by *head* of the fore CS it is being unified with. Analogous, the fore segment's *head* receives a reference to the rear CS's *tail*. In so doing, a CS may represent a CS-strip or even a complete contour, simply by letting *head* and/or *tail* point to other CSs. If a CS would be unified with itself, it is the last CS to be processed of the contour it belongs to and thereby represents the whole contour. In such a case, the label operation takes place instead of the unification operation. Therefore, inner and outer contours need to be distinguished. This can be done based on the direction of one of the contour's highest CS. Informally, it can be stated that, if the contour is directed left in one of its uppermost CS, it is an outer contour. Otherwise, it is an inner contour. This follows directly from the CS-direction definition as described in section II-B. The necessary information has been aggregated during the unification process of this sub-algorithm, so it can be decided locally for each surviving CS. These will, in case of a classification as an outer contour, receive the CC's final label.

As only the CS-strips' ends are of interest for the unification operation, we call this operation simply the unification of two CSs for the rest of this paper. From this follows also that two CS-strips' unification results, independently from their complexity, in constant costs. The parallel unification of multiple CSs is possible, but one restriction has to be fulfilled: The unification of a CS cs0 with its successor cs1 must not take place simultaneously to the unification of cs1 with its successor cs2. If so, the correct assignment of references to the resulting CS's head and tail cannot be guaranteed. In order to resolve potential synchronization problems, a data-independent scheme, processing as many CSs in parallel as are well-known in every step without any data-analysis is applied. In order to explain that, at first we define a tile as follows:

- A tile represents a rectangular pixel-region.
- All CSs within it, that can be unified, are unified.
- All CSs crossing one or more of its edges are not unified.
- Two tiles that share one edge can be unified and the result is one tile.
- Initially, after the extractCS sub-algorithm is executed, each pixel is a tile.

Then, the tile-merging sub-algorithm called unify is applied (see. algorithm 2). Additionally, figure 7 shows all iterations of this pattern applied to a 4x4 pixel-grid. The algorithm should be read together with the descriptions below. There, in every

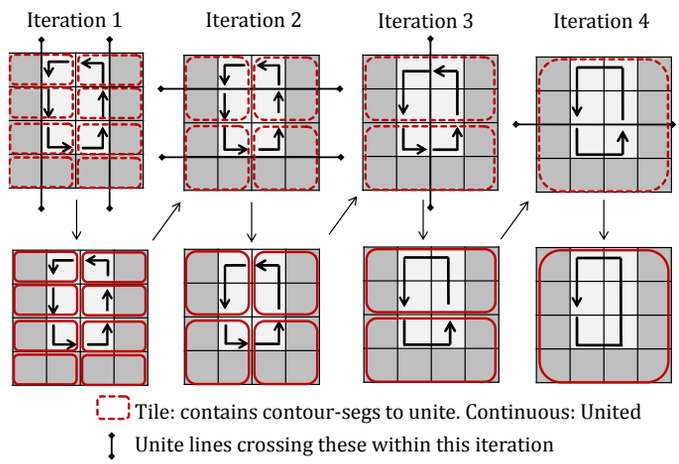


Fig. 7. Extraction of a contour with static parallelism-pattern for processing tiles of a 4×4 pixel-grid. Exactly one CS remains

Algorithm 2: unify

```
for all tile-iterations i ∈ { 1, ..., log2(a) + log2(b) } do
    for all Tiles t assoziated with i in parallel do
        for all Forth-Contours c of t in parallel do
            call unifyOp(c);
        end
        for all Back-Contours c of t ascending do
            call unifyOp(c);
        end
    end
end
```

iteration tiles are defined, which contain CSs, that can be processed independent from any CS contained in any other

tile. Thus, these tiles can be processed conflict-free in parallel. Tiles are defined in a way that in even iterations horizontal CS (directional-types *right/left*) and in uneven iterations vertical CS (directional-types *up/down*) within tiles associated with the iteration are to be processed. In the first iteration the pixel-grid is divided into tiles of size 2×1 pixel sized tiles, where CSs within every two adjoined pixels are united. Once the first iteration is accomplished, the CSs of the left pixels with direction-type *right* and the CSs of the right pixels with direction-type *left* of all tiles from the first iteration have been unified with their successors. As the pattern is alternating in the second iteration, a vertical unification is performed. Here every pair of tiles from the previous iteration lying on top of each other are unified. Obviously, all tiles of the second iteration have a size of 2×2 pixels. There are at most two CSs of directional-type *down* within the upper pixels and similarly two CSs of directional-type *up* within the lower pixels of each tile to be united with their respective successors, if these are existing. In the third iteration CSs of every pair of tiles of the previous iteration laying side by side are united. Hence, tiles associated with the third iteration are 4×2 pixel in size, and so on. Finally, we get one tile occupying all pixels. Since none of the CSs crossing one iteration's tiles borders have been united with their successors in iterations prior to this one, tiles of each iteration can be processed independently in parallel. Though the processing of tiles associated to one iteration can be done in parallel, the processing of CSs within one tile may lead to synchronization problems. For the remaining part of the paper, let direction-types *right* and *down* be called *forth*-directions and direction-types *up* and *left* be called *back*-directions. In case all CSs of direction-type *forth* of all tiles within an iteration are processed before all of the CSs of direction-type *back* are processed, no conflicts can occur during *forth*-CSs' processing. Consider the scenario of the unification of CS-strips with CSs of types *back*- and *forth* alternating along the separating edge between both tiles (see figure 8). Since all

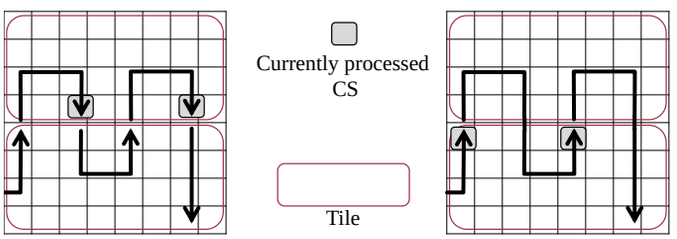


Fig. 8. Conflicts due to parallel unification of 2 tiles' CSs. Left: No conflicts possible in forth-direction. Right: Conflict in back-direction

CSs of type *back* are not unified until the unification of all CSs of type *forth* is finished, the simultaneous unification of three directly successing CSs is impossible during the *forth*-CSs' processing phase. As an illustration, see left part of figure 8. Thus, in each iteration, all CSs of direction-type *forth* can be unified in parallel if the CSs of direction-type *back* are not unified yet. However, the following unification of each tile's CSs of direction-type *back* cannot be done in parallel, as all CSs of direction-type *forth* associated with the current iteration already have been unified. Thus, possibly three CSs of type *back*, belonging to three directly subsequent CS-strips are to be unified in parallel within one tile. As an illustration, see figure 8 (right). The shown situation here results in a conflict. The postulated static pattern resolves these conflicts by sequential processing all CSs of type *back* per tile. The parallelism-scheme is described in pseudocode representing the unify sub-algorithm contained in the pseudocode-snippet of sub-algorithm 2.

Until now, exactly one CS per complete outer contour has received a label. These need to be passed subsequently to all other CSs of their corresponding contours. In order to achieve that, a parallelism-pattern exactly inverse to the one used in conjunction with the unify sub-algorithm is applied and in this process the surviving CS's label is passed on. The associated function passLabelOp is described in the pseudocode-snippet below. It is called instead of function unifyOp.

```
Function passLabelOp(CS c)
  if isLabeled(c.tail) then
    c.suc.label ← c.tail.label;
  endif
```

E. Contour-Labeling: Data-Dependent Approach

As an alternative to the data independent grid based contour-labeling approach, we want to introduce a data-dependent one, which enables a higher degree of parallelism. As a start, let's take a look at the pseudo-code of algorithm 3. It is very similar to the basic parallel list-ranking algorithm

```
Algorithm 3: labelContour_BasicPointerJumping
  for each CS c of ca in parallel do
    if c.status = EXISTING then
      c.minId ← get1dAddress(c);
      for step s ← 0, s < log2(n), s ← s + 1 do
        c.minId ← min(c.minId, c.suc.minId);
        c.props ← aggregateInfo(c.props,
        c.suc.props);
        c.suc ← c.suc.suc;
      end
      if checkIfOuterContour(c.props) then
        c.label ← c.minId;
      endif
    endif
  end
```

using a pointer-jumping technique. Input are directed cyclic linked-lists, where each node is a CS, as extracted by algorithm 1. At first, it aggregates data of all nodes of a linked list and applies it to each node. Afterwards, all nodes of each linked list share the minimal memory address of all nodes of the corresponding linked list in the field minId. Then, the information aggregated in the field props can be used to decide if a CS belongs to an outer contour. If so, it receives the aggregated minimal memory address as a label. Unfortunately, this algorithm does $O(n \cdot log_2(n))$ computations and is thus not work-optimal. Therefore, we apply a technique similar to Cole-Vishkin's[Quelle] optimal parallel list-ranking algorithm [13]. It can be described at a high level by the following three steps[13]:

1) Given a list L of size $O(n)$, create a list L' of length $O(n/log_2(n))$ by iterative deletion of at least half of each

lists nodes within each step. The selection of appropriate nodes can be done with a to be computed 2-ruling set in each iteration.
2) Solve the list-ranking problem for L' by application of the basic pointer jumping algorithm. This now takes O(n) operations.
3) Insert the nodes of L' back into L, using a pattern exactly reverse to their prior deletion from L. Thereby compute the ranks of all nodes of L

Cole-Vishkin's parallel algorithm is able to solve the list-ranking problem in $O(n/p)$ time using $p \leq n/log(n)$ processors, which is optimal. This scheme can be used for contour-labeling, too. All list-ranking specific computations need to be replaced by the operations given in algorithm 3. And in step 3, the found labels only need to be passed on. However, special care needs to be taken in case of small linked lists, which may be removed completely within step 1. If so, they will not reach step 2 and hence cannot be labeled. This is possible, since n represents the data-set's size and not the linked-list's length. To prevent that, whenever a node is to be removed, it is checked if it is its own successor. If that is the case, the node can be labeled directly as described in our data-independent approach.

As in Cole-Vishkin's original algorithm, all changed computations always compare two nodes, resulting in constant costs for these operations. Thus, all theoretical properties of Cole-Vishkin's algorithm remain true in case of our contour labeling algorithm.

F. Fill CCs within contours

Finally, pixels belonging to a certain CC and contained within a contour are to be labeled. Therefore, a filling-algorithm, originating from two opposing edges of the image, is applied. Here, this is done by processing half-columns starting from the upper and lower image's edges and proceeding sequentially to the center. These half-columns are independent from each other and can thus be dealt with in parallel. It is of importance to consider cases, where CCs are nested with one or multiple other CCs or unclassified regions. Regarding this, the algorithm described in this paper utilizes a stack-based fill-algorithm to manage labels of nested contours. The associated sub-algorithm 4 is described in the pseudocode-snippet below. Here, we assume a vertical processing direction, but a horizontal approach is also possible. Whenever a new pixel is entered, data from its associated four CSs needs to be gathered in case of their existence. This includes their label, which is guaranteed to be the same if more than one CS exists. Since the algorithm operates in a vertical direction, CCs can exclusively be entered or left by passing horizontal edge-parallel-CS-parts, which may be called io-CS-parts (io stands for in/out). Their total number needs to be determined for the current pixel, too. All possible courses of a single CS were already shown in figure 2. There, the numbers indicate the associated io-CS-part-counts (ioCnt for short). Obviously, it is 0, 1 or 2 in case of a single CS. Additionally, figure 3 (left) shows the only possible combination of CSs, where the ioCnt of more than one CS is nonzero. So, the total ioCnt of a pixel is also 0, 1 or 2. In case of a pixel containing exactly one io-CS-part, a CC is either entered or left. The proper operation can be identified by considering the label-stack. If the CS's label is currently not on top of the stack, the CC is entered and thus the label has to be put on top of the stack. Otherwise, the CC is left and the stack's topmost label is removed. In case of an ioCnt different from one, the stack remains unchanged, but in case of an existing CS-label it is applied to the pixel. If a pixel does not contain an io-CS-part but is classified, it receives the stack's top label.

Algorithm 4: fillContour

```
for each HalfColumn hc of pa in parallel do
    Stack labelStack;
    for each Pixel p of hc edge to center do
        p.ioCnt ← 0;
        for each CS c associated with p do
            if isLabeled(c) then
                p.ioCnt ← p.ioCnt + c.ioCnt;
                p.label ← c.label ;
            endif
        end
        if p.ioCnt = 1 then
            if labelStack.top() = p.label then
                labelStack.pop();
            else
                labelStack.push(p.label);
            endif
        else if p.class ≠ UNCLASSIFIED then
            p.label ← labelStack.top();
        endif
    end
end
```

G. Asymptotic analysis

Be n the number of pixels. We use the data-structures *pixels* and *CS*, consisting of a fixed number of fields each. And there are $4 \cdot n$ CS. Thus, our algorithm's memory consumption is linear with respect to the number of pixels. Asymptotic numbers of computations and running-times are data-independent and depend on the pixel-grid resolution only. These are:
Contour-segment extraction: All pixels can be processed independently in parallel. Thereby, up to four CS per pixel are extracted, resulting in constant costs per pixel. So, using $p \leq n$ processors delivers a running-time $O(n/p)$ on a PRAM.
Contour labeling: Here, we only consider our data-dependent approach, since it delivers better asymptotic properties due to its higher degree of parallelism. As described before, it preserves all properties of Cole-Vishkin's optimal list-ranking algorithm by which it was inspired: Given any number of $p \leq n/log(n)$ processors, all contours can be labeled in $O(n/p)$ time on a CRWC-PRAM.
Contour filling: All columns can be filled independently in parallel but each single column is processed sequential. If there are more rows than columns, it is reasonable to process in a row-wise scheme instead. Let a be image-width and b image-height. Then, any number of $p \leq max(a, b)$ processors can be utilized, resulting in $O(n/p)$ running-time on a PRAM.

Altogether, all sub-algorithms have a running-time guarantee of $O(n/p)$, and the processor-number is limited by the contour-filling algorithm to $p \leq max(a, b)$. Our algorithm

is work-optimal as it obviously does not more than $O(n)$ operations for any number of processors utilizable.

III. Implementation Details

We have implemented our algorithm utilizing OpenCL so it can be evaluated on a GPU or CPU. However, in this paper we limit our considerations to a GPU. The pseudo-code given before describes our implementation from a quite high level point of view. References are in most cases implemented as 1d memory-addresses. Others are only globally unique if the 2d-pixel-grid structure is taken into account, too. The stacks used by algorithm 4 are implemented as fixed-sized data-structures with a pointer to a current element. The 2-ruling sets necessary for Cole-Vishkin's processing scheme are implemented similar to their earlier paper [14], because a $log(log(log(log(n))))$-n ruling set already is a 2-ruling set for all earthly values of n. In order to implement our algorithm, dozens of distinct OpenCL-kernels are necessary. Especially Cole-Vishkin's algorithm requires numerous global synchronization points. Since OpenCL does not allow global synchronization within a kernel, whenever such a point is reached, the kernel is terminated and another one started thereafter. Additionally, GPUs typically offer some sort of user managed cache called local memory (or shared memory, CUDA term). Its use is important in many cases to receive good memory access patterns. For example, Stava [12] identify their shared memory use as one important reason delivering the speedup when their algorithm is compared to Hawick[11]. However, in case of linked-lists it is less straightforward to use local memory as it is in case of an algorithm working directly on a rectangular pixel-grid.

Contour labeling is done using a hybrid approach consisting of the data-independent and the data-dependent approach described in this paper. The data-independent one has no overhead to identify CS to be processed in parallel and it is possible to achieve convenient memory-access patterns at least to a certain degree. Both is not true in case of the our implementation of the data-dependent pattern. In exchange, it delivers a much higher degree of parallelism. Thus, our contour labeling algorithm always starts with the data-independent pattern and uses it as long as the delivered parallelism is sufficiently high to utilize the executing device. Due to the fact that it is a data-independent pattern, a proper number of iterations can be calculated in advance. Then, we switch to the data-dependent pattern. Of course, on the way back, the last iterations must use the static pattern again, since the unification needs to be exactly reversed. This hybrid approach runs about 50 percent faster than the faster approach alone. The overall memory-consumption of our current implementation is 77 bytes per pixel.

IV. Experimental Results

In order to get an idea of our algorithm's possible use, we compare it in regard to achievable throughput with Stava[12], who have published the fastest implementation of a CCL algorithm known to us. We choose data-sets that can be easily generated in different resolutions. So, repeatable results are guaranteed and they can be verified simply or compared by others. The measure covers the CCL-algorithm's entire CCL-processing time and excludes initialization time. We

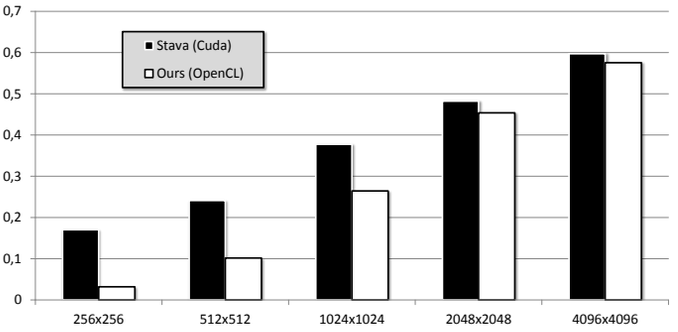


Fig. 9. Comparison of throughput [GPixel/sec] to Stava using spiral-data in different resolutions

TABLE I. Throughput [GPixel/sec] measured with different configurations of a 4096x4096 pixel checkerboard

Width of each square in pixels	1	2	8	32	128	512	2048
Number of CCs	16m	4m	262k	16k	1k	64	4
Our CS-count	0	16m	7,3m	2m	520k	131k	32k
Our throughput	0,92	0,31	0,49	0,56	0,69	0,83	0,88
Stava's throughput	1,09	0,91	0,81	0,65	0,58	0,56	0,51

have re-run Stava's original implementation with our data and measured the execution time of the method "FindRegions" in "ccl.cpp" only. We utilize a computer consisting of an Intel Core i7 2700k CPU, Windows 7 and a Nvidia GeForce GTX 480 GPU (Driver 337.88). Both our implementation and Stava's are executed on the GTX 480. However, in contrast to us, Stava's implementation uses the vendor specific API CUDA.

The first data-set contains a rectangular spiral with a width of one pixel and the same size as the data resolution. It is considered to be especially hard to label in literature. The achieved throughput with respect to the data-resolution is compared to Stava in figure 9. Obviously, both algorithms process much more data per unit time, if a higher data-resolution is applied. This scaling behavior is especially true in case of our algorithm. On the other hand, if a low resolution is measured, Stava's algorithm is much faster than ours. In case of a high resolution, our algorithm delivers comparable throughput.

In order to get an idea how the running-time measured is composed of the CS-extraction, Contour labeling and Contour filling, the running times of the corresponding kernels are measured using OpenCL-events and aggregated to the three categories. In case of the 4096 x 4096 Spiral, we get the following results for our algorithm:

- CS-extraction: 21 percent of running-time measured
- Contour labeling: 61 percent of running-time measured
- Contour filling: 18 percent of running-time measured

Now, let's take a look at data-dependence. In order to do that, we determine the throughput when processing a checkerboard-pattern, where the squares' size can be configured. Here, four different classifications are applied to separate all squares as distinct CCs. In doing so, the impact of different numbers and sizes of CCs can be evaluated. Table I shows the results compared to Stava. Here, Stava's algorithm delivers the highest throughput in case of the smallest CCs and it decreases as the

CCs' size increases. This behavior was already reported in their original paper. Contrasting these results, the throughput of our algorithm increases, if the CCs size increases. An exception to this behavior is the CC size one, where our algorithm does not extract any CS. So, our algorithm performs better, if the processed data consists of less contour-segments.

V. Conclusion

We have presented a parallel algorithm to solve 2d-image-data CCL problems in linear overall complexity. Similar to contour-tracing approaches, the extracted object-contours may also be helpful for certain applications like 2d-object recognition. Our current OpenCL-implementation performs comparable to the fastest previous approaches, when executed on a GPU, if a sufficiently high resolution is applied. Additionally, it scales superior with respect to the data-resolution in case of the tested data-set.

VI. Future Work

First and foremost, we will research alternatives to the fill contour sub-algorithm, since it limits the asymptotic running-time due to the small number of processors utilizable, when compared to our other sub-algorithms. Perhaps a per-column unification of linked-lists is a better idea. However, these would have to be re-arranged at first in order to solve problems regarding nested connected-components.

Besides, we will improve our implementation. One problem identified is the amount of OpenCL kernel calls (hundreds per processed image). And another one is the limited use of a GPU's local memory provided by OpenCL.

References

[1] A. Rosenfeld, "Connectivity in digital pictures," *J. ACM*, vol. 17, no. 1, pp. 146–160, Jan. 1970. [Online]. Available: http://doi.acm.org/10.1145/321556.321570

[2] J. H. Kim, K. K. Kim, and C. Y. Suen, "An hmm-mlp hybrid model for cursive script recognition," *Pattern Analysis and Applications*, vol. 3, pp. 314–324, 2000. [Online]. Available: http://dx.doi.org/10.1007/s100440070003

[3] J. S. Suri, S. Singh, and L. Reden, "Computer vision and pattern recognition techniques for 2-d and 3-d mr cerebral cortical segmentation: A state-of-the-art review," *JOURNAL OF PATTERN ANALYSIS AND APPLICATIONS*, p. 2002, 2001.

[4] F. Chang, C.-J. Chen, and C.-J. Lu, "A linear-time component-labeling algorithm using contour tracing technique," *Comput. Vis. Image Underst.*, vol. 93, no. 2, pp. 206–220, Feb. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2003.09.002

[5] U. H. Hernandez-Belmonte, V. Ayala-Ramirez, and R. E. Sanchez-Yanez, "A comparative review of two-pass connected component labeling algorithms," in *Proceedings of the 10th international conference on Artificial Intelligence: advances in Soft Computing - Volume Part II*, ser. MICAI'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 452–462.

[6] L. He, Y. Chao, K. Suzuki, and K. Wu, "Fast connected-component labeling," *Pattern Recogn.*, vol. 42, no. 9, pp. 1977–1987, Sep. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2008.10.013

[7] R. Haralick, "Some neighborhood operations," in *Real Time Parallel Computing: Image Analysis*, 1981, pp. 11–35.

[8] C. Fiorio and J. Gustedt, "Two linear time union-find strategies for image processing," *Theor. Comput. Sci.*, vol. 154, no. 2, pp. 165–181, Feb. 1996. [Online]. Available: http://dx.doi.org/10.1016/0304-3975(94)00262-2

[9] F. Chang and C.-J. Chen, "A component-labeling algorithm using contour tracing technique," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, ser. ICDAR '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 741–. [Online]. Available: http://dl.acm.org/citation.cfm?id=938980.939556

[10] H. M. Alnuweiri and V. K. Prasanna, "Parallel architectures and algorithms for image component labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 10, pp. 1014–1034, Oct. 1992. [Online]. Available: http://dx.doi.org/10.1109/34.159904

[11] K. A. Hawick, A. Leist, and D. P. Playne, "Parallel graph component labelling with gpus and cuda," *Parallel Comput.*, vol. 36, no. 12, pp. 655–678, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.parco.2010.07.002

[12] B. B. Ondrej Stava, "Connected component labeling in cuda," in *GPU Computing Gems, Emerald Edition*, W.-M. W. Hwu, Ed. Morgan Kaufmamnn, 2011, pp. 569–581.

[13] R. Cole and U. Vishkin, "Faster optimal parallel prefix sums and list ranking," *Inf. Comput.*, vol. 81, no. 3, pp. 334–352, Jun. 1989. [Online]. Available: http://dx.doi.org/10.1016/0890-5401(89)90036-9

[14] ——, "Deterministic coin tossing with applications to optimal parallel list ranking," *Inf. Control*, vol. 70, no. 1, pp. 32–53, Jul. 1986. [Online]. Available: http://dx.doi.org/10.1016/S0019-9958(86)80023-7

Geo-Morphology Modeling in SAR Imagery Using Random Fractal Geometry

Ali Ghafouri
Dept. of Surveying Engineering, Collage of Engineering,
University of Tehran,
Tehran, Iran,
ali.ghafouri@ut.ac.ir

Jalal Amini
Dept. of Surveying Engineering, Collage of Engineering,
University of Tehran,
Tehran, Iran,
jamini@ut.ac.ir

***Abstract*— Geological formations has different behaviors against weathering and erosion and it causes difference in geo-morphology. Geological mapping needs the ability of lithological discrimination on the basis of geo-morphology, and this capability is not fully accessible via optical remote sensing. Since radar spectral windows in electromagnetic spectrum is independent of solar energy and can penetrate clouds and particularly sensitive to surface parameters, they are considered to be useful for studies of the surface geological morphology. In order to discriminate the surface geometric pattern and differentiate top-geological formations surface, it is required to model the softness and roughness of surface according to the radar signal backscattering. Fractal geometry is much more capable to describe natural phenomena than conventional geometry. Fractal geometry has been used several times in literature in order to improve the radar backscattering models. This paper compares application of different autocorrelation functions for the most famous model in this manner, integral equation model (IEM) benefiting random fractal geometry. Trying to improve geological mapping of Dehloran geological formation (western boundary of Ilam in IRAN), the results display the level of effectiveness of the conventional autocorrelation function.**

Keywords—geological formations, SAR images, roughness modeling, backscattering coefficient

I. INTRODUCTION

Detection of top-geological structures cannot be possible via optical imagery especially in large regions; since study of geological morphology to some extents is not possible by passive remote sensing. Because of independence of microwave sensors to climate changes, and especially their sensitivity to surface parameters, SAR technology is suitable for geomorphology and earth surface studies.

In Dehloran geological formation, some geological structures containing lithologies like Marne, are more affected by alteration and weathering and consequently are physically smooth. In contrary, there are some other structures which are less affected by physical and chemical erosion, and have rough and rigid face, such as Anhydride lithology. In the process of mapping this region on geological maps, discrimination among the different top-geological structures cannot be possible via available optical imagery; since geological morphologies to some extents are not differentiable by passive remote sensing. Geological morphology modeling by SAR data needs to have topography and micro-topography model of the surface.

Geological morphology modeling by SAR data needs to have topography and micro-topography model of the surface. Roughness parameters are highly dependent to measurement scale which is the SAR signal wavelength in this study. Natural phenomena cannot be qualitatively modeled via conventional geometry; in contrast, Random Fractals Geometry is much more powerful in modeling natural shapes [1].

In this paper different autocorrelation functions for the most famous model in this manner, integral equation model (IEM) is applied and by using fractal autocorrelation function, instead of using the Gaussian and exponential functions [1], we try to improve geological mapping of morphology. In other words, this paper tries to improve precision of parameters estimation in Integral Equation Model (IEM) [2], and then by considering geomorphology, to increase quality and precision of geological maps. Verification of modeling processes are applied to ALOS SAR data of Dehloran geological structure to improve geological mapping precision.

II. INTEGRAL EQUATION MODEL (IEM) AND ROUGHNESS PARAMETERS

Standard theoretical models of backscattering, are: Geometric Optics Model (GOM) and Physical Optics Model (POM) and Small Perturbation Model (SPM). Geometric Optics Model, for very rough surfaces, Physical Optics Model, for medium roughness and Small Perturbation Model, for very smooth surfaces are used. Fung and Chen have developed Integral Equation Model (IEM) as a physically based electro-magnetic transfer model IEM via combination of the GOM, POM and SPM, and constructed a more applicable model which can tolerate a really wide range of roughness dimensions, theoretically, IEM is not restricted to any special situation [1]. As defined, IEM relates backscattering coefficients to roughness parameters of the surface, dielectric permittivity and magnetic permeability, and the local incidence angle. The co-polarized backscattering coefficient has been explained as [3]:

$$\sigma_{pp}^{0} = \frac{k^2}{4\pi} e^{-2k^2\sigma^2 cos^2\theta} \sum_{n=1}^{+\infty} \left| I_{pp}^{n} \right|^2 \frac{W^{(n)}(2ksin\theta, 0)}{n!} \qquad (1)$$

where

$$I_{pp}^{n} = (2k\ \sigma\ cos) f_{pp} \exp(-k^2\sigma^2 cos^2\theta) + (k\ \sigma\ cos)^n F_{pp} \quad (2)$$

and pp, polarization (hh or vv); k, wave number ($k = \frac{2\pi}{\lambda}$: λ is the wavelength), θ is the local incidence angle, σ, the surface rms-height, $W^{(n)}$, fourier transform of n^{th} power of the correlation function, and f_{hh}, f_{vv}, F_{hh} and F_{vv} are approximated by:

$$f_{hh} = \frac{-2R_h}{cos\theta}$$
$$f_{hh} = \frac{2R_v}{cos\theta}$$
$$F_{hh} = 2\frac{sin^2\theta}{cos\theta}\left[4R_h - \left(1 - \frac{1}{\varepsilon}\right)(1 + R_h)^2\right]$$
$$F_{vv} = 2\frac{sin^2\theta}{cos\theta}\left[\left(1 - \frac{\varepsilon\, cos^2\theta}{\varepsilon - sin^2\theta}\right)(1 - R_v)^2 - \left(1 - \frac{1}{\varepsilon}\right)(1 + R_v)^2\right]$$

The horizontally and vertically polarized Fresnel reflection coefficients, R_h and R_v, are described as:

$$R_h = \frac{cos\theta - \sqrt{\varepsilon - sin^2\theta}}{cos\theta + \sqrt{\varepsilon - sin^2\theta}}$$
$$R_v = \frac{\varepsilon\, cos\theta - \sqrt{\varepsilon - sin^2\theta}}{\varepsilon\, cos\theta + \sqrt{\varepsilon - sin^2\theta}}$$

ε is the dielectric constant. Considering $C(\rho)$, the surface autocorrelation function (ACF), and J_0 as zeroth order Bessel function, surface power spectrum, $W^{(n)}$ in IEM can be defined as:

$$W^{(n)}(K) = \int_{\rho=0}^{\rho=+\infty} C(\rho).\rho.J_0(K_\rho)d\rho \qquad (3)$$

Gaussian and exponential functions are two special cases of ACF, which will be described in the next section.
Calculation of the model parameters from SAR signal backscattering coefficient is not directly possible, because of the model complexity and some other strategies must be pursued. In this paper Look Up Table (LUT) method is employed; so for this purpose a table of different possible values of roughness parameters/dielectric constant and corresponding backscattering coefficients tabulated.
In sections *A*, *B* and *C*, respectively, the three main specifications for the radar Backscattering study, the rms-height, correlation length and autocorrelation function are defined.

A. Heights Root Mean Square (rms-height)

The root mean square of surface heights (rms-height) defines the variation in surface elevation above an arbitrary plane and is used to be calculated on the basis of a one-dimensional discrete surface profile measurement consisting N points with elevations z_i [4]:

$$s = \sqrt{\frac{1}{N}\left[\left(\sum_{i=1}^{N} z_i^2\right) - N\bar{z}^2\right]} \qquad (4)$$

where

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \qquad (5)$$

In (4), s represents standard deviation of the surface micro-topography discrete heights.

B. Correlation Length

The level of surface heights uniformity over a finite profile of the surface is usually described by Correlation Length [4]. In other words, the horizontal variations of the surface heights is called correlation length. Unlike the simplicity of this definition, measurements of the correlation length is complicated. The calculated values for this parameter via different ways are extremely variable and also greatly depends on the length of the sampling profile length [6]. As a typical methodology, Davidson et al. (2003) has proposed a linear interpolation on the correlation function of the heights:

$$l = \left(e^{-1} - C(\rho_1)\right)\frac{\rho_2 - \rho_1}{C(\rho_2) - C(\rho_1)} + \rho_1 \qquad (6)$$

where $C(\rho)$ is the autocorrelation function, ρ_1 and ρ_2 are two arbitrary points. On the basis of the parameter definition, in this equation, it is considered that $C(l) = e^{-1}$.

C. Autocorrelation Function

The normalized autocorrelation function, for $\rho = j\Delta x$, where Δx is the spatial resolution of the profile, is given by:

$$C(\rho) = \frac{\sum_{i=1}^{N-j} z_i z_{i+j}}{\sum_{i=1}^{N} z_i^2} \qquad (7)$$

In order to fully characterize the ACF of a surface, a discretization interval, used to sample the profile, should be at least as small as one tenth of the correlation length.
In backscattering models, often two types of ACFs, the exponential and the Gaussian autocorrelation functions are being used. The exponential ACF is given by:

$$C(\rho) = e^{-|\rho|/l} \qquad (8)$$

and the Gaussian function;

$$C(\rho) = e^{-\rho^2/l^2} \qquad (9)$$

where *l*, is the correlation length.

III. IEM Modeling Method using Fractal Geometry

The most famous improvement for backscattering modeling using fractal geometry is using fractal ACF instead of Gaussian or exponential one.
Eqs. (8) and (9), show the exponential and Gaussian ACFs, respectively. Likewise fractal correlation function is [1]:

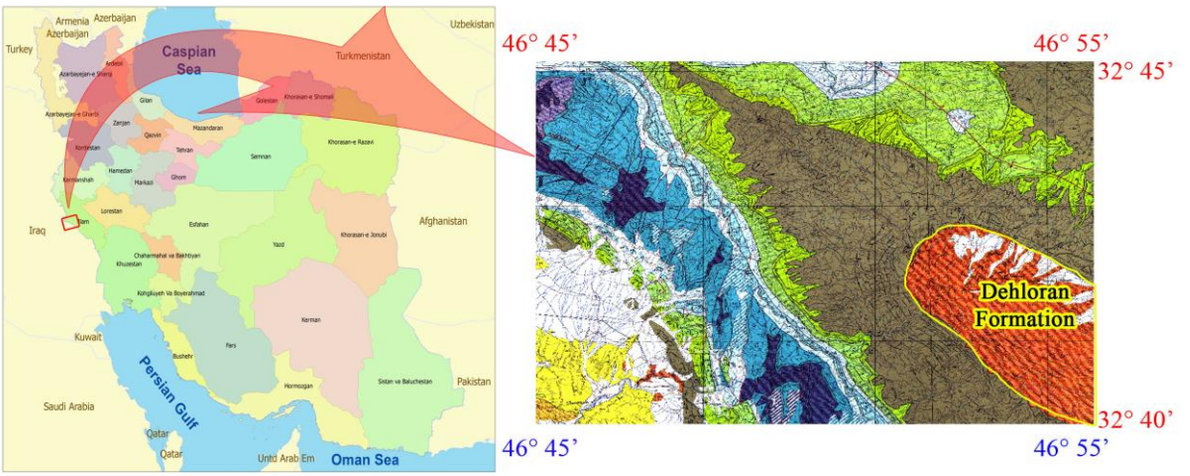


Fig. 1 Case Study region in western part of Dehloran geological formation

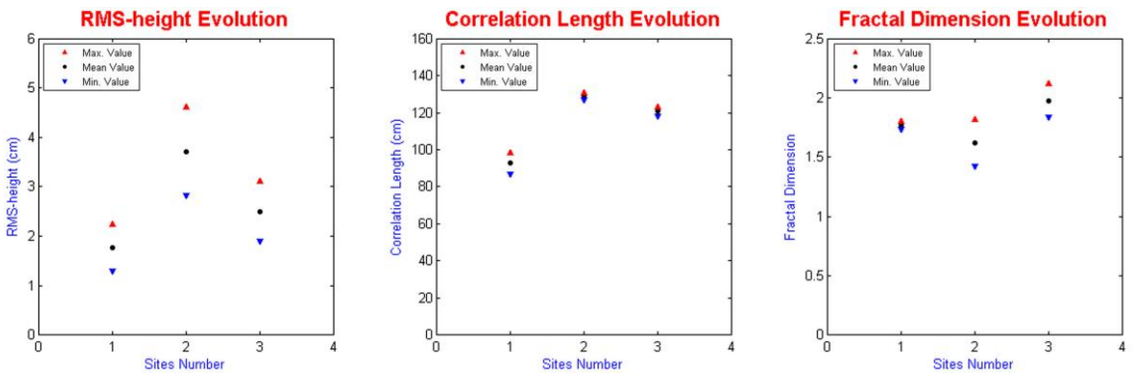


Fig. 2 Case Studies’ surface roughness parameters, rms-height, Correlation Length, Fractal Dimension

$$C(\rho) = e^{-\rho^{\tau}/_{l^{\tau}}} \qquad (10)$$

where, on the basis of experimental data, parameter τ has a linear relation with fractal dimension and the relation is computable via: $\tau = -1.67\ D + 3.67$, in which, D is the fractal dimension of the earth surface and can be calculated via various methods. In natural surfaces which have more complexity, fractal functions are more appropriate than the exponential and Gaussian correlation functions. Since the exponential and Gaussian functions are particular forms for particular situations of fractal function [1].

IV. Implementation, Results and Discussion

A. Case Study

As the application case study, western part of Dehloran geological formation is selected which is located in the following coordinates:
Longitude: *46°45′* to *46°55′*
Latitude: *32°40′* to *32°45′*

Fig. 1 illustrates geographic and geological position of the case study.Geomorphology of the region, which is depicted in the image of Fig. 1 is a geological section of Dehloran structure and obviously different members of Pabdeh, Asmari and Kalhor formations are figured out. Different decay properties of these geological units are the reason of different surface morphology. Regional lithologies contain of limestone, dolomite, marl and anhydride. Discrimination of the units for geological mapping and interpretation of optical images needs in situ hardness and softness measurements; and without considerations of the surface morphology is approximately impossible.

Fig. 2 depicts the variation of the three roughness parameters: rms-height, Correlation Length, Fractal Dimension. The rms-height has the most variations among other two parameters. Site 1 has more diversity range among the sites, In contrary, fractal dimension of the sites 2 and 3 are more diverse than the site 1.

Fig. 3 illustrates the three case studies position on the SAR image. Site1 is on a plain region, site2, on a mountainous region and site 3 is on foothill, which can be considered as a transition zone between mountain and plain regions.

B. Implementation and Results

As previously mentioned, the three study sites position are shown in Fig. 3. Surface roughness has been measured based on digital elevation model of the site for a total number of 20 pixels and the dielectric constant has been extracted from the presented tables of [5]. Fig. 4 illustrates backscattering

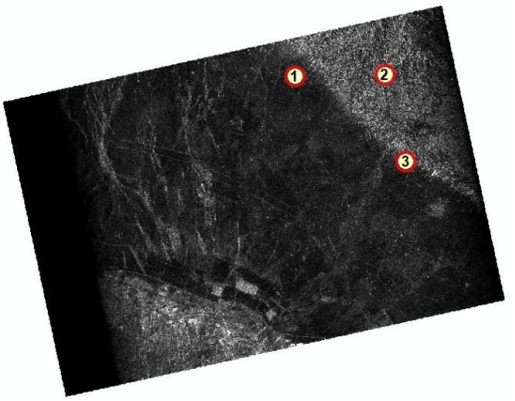


Fig. 3 Case Studies' surface roughness parameters, rms-height, Correlation Length, Fractal Dimension

coefficient calculated by IEM based on conventional geometry (Eq.1) in both *hh* and *vv* polarizations as a function of backscattering coefficient measured from SAR images. Distance of points far from the diagonal line shows the error of the simulation.

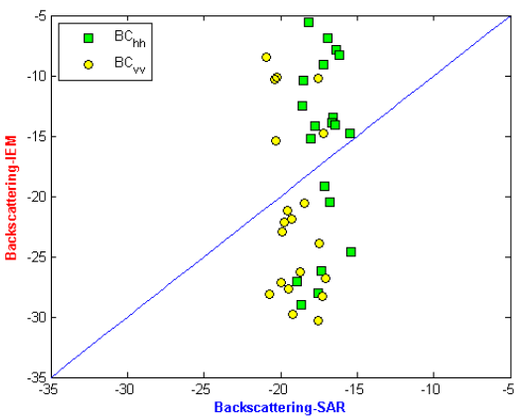


Fig. 4 Simulation accuracy using backscattering equation IEM in two hh and vv polarizations for 20 sample points of the study area, based on conventional geometry. Number and distance of points far from the diagonal line shows the error of the simulation.

As mentioned in the previous section, a method is already presented to utilize the benefits of fractal geometry in backscattering electromagnetic models; which is implemented on the case study data and the comparative results are presented in Figs. 5 and 6. As described for Fig. 4, these graphs show the accuracy level of the IEM simulation.

According to the section A, typically, surface spectrum in IEM equation must be calculated through the Fourier spectrum of the correlation function in the equations (8), (9) and (10). Also, equation (10) as a general function depending on the value of τ, can replace the functions (8) and (9). For the data of study area, the value of $\tau = 1.2$, represents the optimum results comparing to measured backscattering coefficients on SAR image, which is used in the graph of Fig. 6.

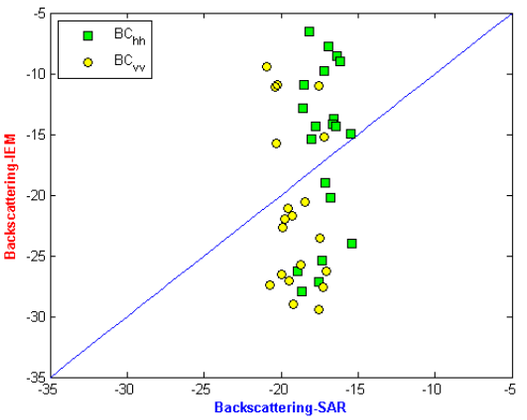


Fig. 5 Simulation accuracy using backscattering equation IEM in two hh and vv polarizations for 20 sample points of the study area, using fractal autocorrelation function ($\tau = 1.2$) instead of Gaussian and exponential functions.

Fig. 6, illustrates the simulation accuracy using IEM backscattering equation in two *hh* and *vv* polarizations for sample points of the study area, using the correlation length calculated via fractal dimension parameter.

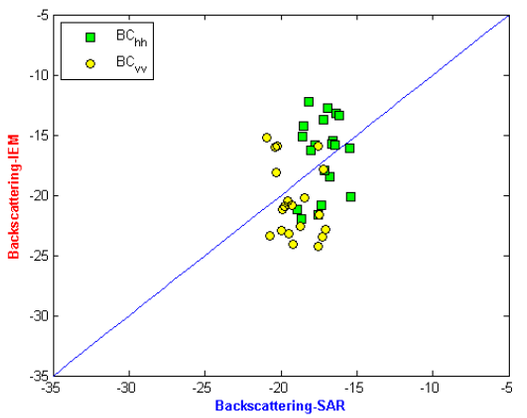


Fig. 6 Simulation accuracy of backscattering equation IEM in two hh and vv polarizations for sample points of the study area, using the correlation length calculated via fractal dimension parameter.

Table 1 presents the statistical analysis of the results acquired through these methods compared to each other.

TABLE I. METHODS RESULTS STANDARD DEVIATION FOR INTEGRAL EQUATION MODEL (IEM)

	Original IEM - Gaussian ACF	Original IEM - Exponential ACF	Original IEMm - fractal ACF ($\tau = 1.2$)
hh-polarization	10.023	7.518	6.891
vv-polarization	9.666	7.249	6.645

V. CONCLUSION

Due to irregular and fractal nature of the surface roughness, electromagnetic backscattering modeling of radar signals using fractal geometry calculates surface parameters closer to actual values. The model IEM is for three types of ACFs and for 20

sample points on three different sites is tested. The graphs and the deviation table, demonstrate obviously the effectiveness of fractal ACF. The studied fractal ACF is implemented with the available linear interpolation which relates fractal dimension and correlation length, more studies on this interpolation can be planned for future studies.

ACKNOWLEDGMENT

The authors are thankful to the University of Tehran for providing financial assistance to carry out this research.

REFERENCES

[1] Baghdadi, N., I. Gherboudj, M. Zribi, M. Sahebi, C. King, F. Bonn, "Semi-empirical calibration of the IEM backscattering model using radar images and moisture and roughness field measurements", International Journal of Remote Sensing, Vol. 25, Iss. 18, 2004, DOI: 10.1080/01431160310001654392

[2] Fung, A., Z. Li, and K. Chen "Backscattering from a randomly rough dielectric surface", IEEE Geoscience and Remote Sensing Letters vol.30, no.2, pp. 356,369, 1992, DOI: 10.1109/36.134085

[3] Fung, A. and K. Chen, "An update on the IEM surface backscattering model", IEEE Geoscience and Remote Sensing Letters vol.1, no.2, pp. 75,77, 2004, DOI: 10.1109/LGRS.2004.826564

[4] Gupta, V. K., and R. A. Jangid, "Microwave response of rough surfaces with auto-correlation functions, RMS heights and correlation lengths using active remote sensing", Indian Journal of Radio & Space Physics, Vol 40, pp 137-14, 2011.

[5] Martinez, A. and A. P. Byrnes, "Modeling Dielectric-constant values of Geologic Materials: An Aid to Ground-Penetrating Radar Data Collection and Interpretation", Current Research in Earth Sciences, Bulletin 247, part 1, 2001.

[6] Verhoest, N. E., H. Lievens, W. Wagner, J. Álvarez Mozos, M. S. Moran, and F. Mattia, "On the soil roughness parameterization problem in soil moisture retrieval of bare surfaces from synthetic aperture radar". Sensors 8.7, pp.4213,4248, 2008, DOI: 10.3390/s8074213

A Method for Road Area Detection in High Resolution SAR Images

Mehdi Saati, Jalal Amini
Department of Geomatics Engineering, College of Engineering, Tehran University, P.O. Box. 11365-4563, Tehran, Iran
msaati@ut.ac.ir, jamini@ut.ac.ir

***Abstract*— Automatic extraction of road from satellite images is one of the most important researches in the field of remote sensing. The method proposed in this study is based on a fuzzy method for detection of road areas from high resolution SAR images. In this method, the multiple features are extracted first based on the backscatter coefficient of each pixel and its neighbor pixels from the input image. The extracted features are combined with each other in the next step using a fuzzy algorithm and the desired road areas are selected separately in the last step considering the spatial and spectral criteria. The favorite results and root mean square error of 87% were obtained by applying this algorithm on high resolution SAR images obtained from the TerraSAR satellite.**

Keywords—component; Multiple features, road detection, fuzzy algorithm, high resolution, synthetic aperture radar.

I. INTRODUCTION

Manual extraction of features manually from the satellite images by expert operators takes cost and time. Therefore, automatic extraction of the features from the images is a basic research in the contexts of remote sensing for automatic generation of the spatial information and mappings. Roads and buildings, which are the most significant features are the most frequent ones in preparing maps of the urban, suburban and rural regions.

SAR images for the weather independence and the ability to operate during both day and night, has indeed advantages over optical images. Currently, automatic road extraction from VHR SAR data is a research topic in demand. However, the task complexity increases and is not eased by the enhanced resolution due to speckle noise. The fact that buildings along the road may mask them, or reduce their visibility, further complicates the road detection problem. Noteworthy here is that some satellites like TerraSAR-X, RadarSAT-2 and Cosmo/Skymed incorporate an appropriate resolution for extraction the roads even in the urban areas.

Many researchers have previously addressed this topic since the 1990s [1, 2].

A nearly automatic detection algorithm using Markov random field was proposed by Tupin et al. for linear features such as the main axes of road networks [3]. Jeon et al. [4] developed a technique for the detection of roads in a SAR image using a genetic algorithm.

Tupin et al. [5] presented a technique for the detection of roads in dense urban areas using SAR imagery based on multiscale framework. Dell'Acqua et al. [6] presented an algorithm for road extraction based on multiple road detectors and logical feature fusion in fine resolution SAR imageries. Bentabet et al. [7] updated road vectors by using SAR imagery without human–computer interaction, with comprehensive knowledge provided by a road database.

Wessel [8] also studied automatic road extraction from airborne SAR imagery supported by context information. Chen et al. [9] applied particle filtering in tracking consecutive road segments from SAR images. Lisini et al. [10] present a road extraction method comprising fusion of classification results and structural information in form of segmented lines. The approach was tested for airborne SAR data of resolution better than 1m.

Li et al.[11] presented a road extraction method from high-resolution dual-polarization SAR data over urban areas based on two road detector extraction and feature-level fusion. Stilla & Hedman [12] used a Bayesian network into an already existing road extraction approach for roads extraction from SAR imagery. Hedman et al. [13] proposed combination of two road extractors from VHR SAR scenes: one more successful in rural areas and one explicitly designed for urban areas. In order to get the best combination of both, a rapid mapping filter for discriminating rural and urban scenes is utilized.

To improve the performance of road extractors on SAR images, Zhou et al. [14] used the image of polarimetric SAR systems which measures a target's reflectivity using four polarizations. Liu et al. [15] Presented a road extraction method from SAR imagery based on an improved particle filtering and snake model.

From above paragraphs, we consider many works on road extraction have been done using the automatic or the semi-automatic methods in aerial or satellite SAR images. As basic step of the all mentioned methods, the local detection procedure plays a critical role and determines the overall performance of road detection procedures.

In the present work, we aim at improving the road area detection task in SAR images by exploiting the multiple spatial and spectral feature extraction defined in the range of a window in vicinity of a pixel and fuzzy-based feature fusion method in comparison of the methods using segmentation like Negri [16]. In fact, with meter or sub-meter spatial resolution, roads in SAR data may be more precisely modeled as dark elongated areas surrounded by bright edges, which are due to double-bounce reflections by surrounding buildings or uniform backscattering by the vegetation. The effect is more pronounced for roads oriented in range direction. As a result, there are many bright areas with high contrast adjacent to them. Therefore, the features in this paper are defined exactly based on the above mentioned issues which include aligned, dark and high contrast pixels with their adjacent regions.

At the rest of this paper, section II introduces the materials and methods used in the current research while the results are discussed in section III. Finally, section IV will give some conclusions and suggestions.

II. Materials and Methods

The method proposed in this paper is divided into two stages, as depicted in Fig. 1.

Stage 1 is road area detection. In this stage, the features adapted to road properties in high resolution SAR imageries are extracted and fused based on a fuzzy algorithm to detect road areas. Then spatial and spectral criteria are used for refinement of extracted roads.

Stage 2 is accuracy assessment. In this stage, some numerical parameters: road detection correctness coefficient (RCC), background detection correctness coefficient (BCC) and root mean square error (RMSE) are defined. And then the algorithm is tested for some speckled and despeckled images to evaluate the method.

In the following sub sections, we describe in details the method, emphasizing the adaptation of the algorithm to the suburban areas.

A. Road Area Detection

It has been tried in this research to achieve desirable results by using multiple detectors. Thus, straight linear features of the image is investigated considering the spatial relation between each pixel and its neighbors, then diagnosis of the road segments is improved in the next step by fuzzy combination of the results derived from each detector.

It was mentioned before that the road appears as the extended dark areas with relatively light edges in high resolution SAR images [14]. Therefore, one should search for pairs with parallel edges or homogeneous or extended dark areas. It is obvious that taking into account each one of the following conditions may generate undesirable results. Thus, a more precise answer would be combination of the above mentioned conditions and ideas.

1) Multiple Features Extraction

In order to extract pixels of a road, the first step will be to calculate some spatial features in a square window around the central pixel p(i,j). However, each of these features is a function of the window size R×R.

Fig. 2 depicts how to select the neighbors of pixel p(i,j) on the image.

The total radiance in a specific direction of θ in the selected window is given as [16]:

$$r(\mathrm{i},\mathrm{j},\mathrm{R},\theta) = \sum_{\mathrm{k}=-\mathrm{R}/2}^{\mathrm{k}=\mathrm{R}/2} \mathrm{p}([\mathrm{i}+\mathrm{k}\cos(\theta)],[\mathrm{j}+\mathrm{k}\sin(\theta)]) \qquad (1)$$

Direction of the least total radiance as first feature is defined as:

$$\theta_0(i,j,R) = \arg\min_\theta r(i,j,R,\theta)\,,\ \theta\epsilon[0^\circ \quad 180^\circ) \qquad (2)$$

The corresponding total radiance of θ_0 as second feature is also defined as:

$$r_0(\mathrm{i},\mathrm{j},\mathrm{R}) = r(\mathrm{i},\mathrm{j},\mathrm{R},\theta_0) \qquad (3)$$

θ_0 and r_0 demonstrate dark areas elongated around the pixel p(i,j) which is extended in the current window.

Assuming that these areas are appropriate candidates for the road areas, their contrast with the other areas (average total radiance of the other directions) must be high. Therefore, the 3rd feature is value of the contrast which can be calculated from Equation (4) as:

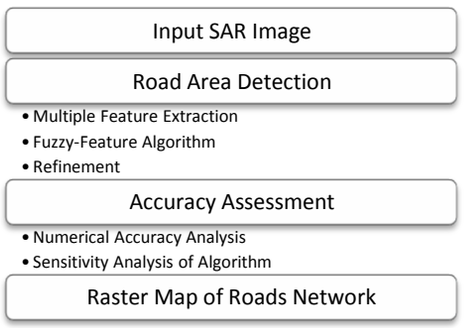


Fig. 1. Diagram of the proposed road area detection method

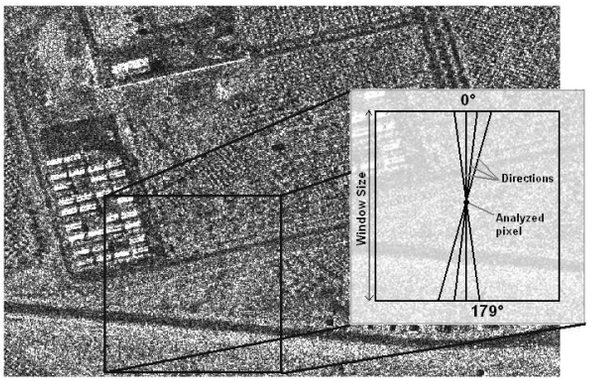


Fig. 2. Selection of the neighbors of pixels

$$c_0(i,j,R) = \left\| \frac{\sum_\theta r(i,j,R,\theta)}{n} - r_0(i,j,R) \right\| \quad (4)$$

Where, n represents number of the directions for which the value of total radiance has been calculated.

It is clear that the features θ_0, r_0 and c_0 display different information, all of them are a function of the window size R×R and each of them may reproduce incorrect results so that the incorrect results are expected to be deleted through combining them. Noteworthy here is that the wider and longer road areas are detected from dimensions of the larger windows, whereas the narrow and shorter road areas are characterized by smaller dimensions.

2) Fuzzy-Feature Algorithm

The road areas in SAR images are determined with a fuzzy algorithm called fuzzy-feature algorithm based of three features. Fuzzy logic provides a simple way to arrive at a definite conclusion based upon imprecise, uncertain, ambiguous, vague or missing input information. Fuzzy set theory, introduced by L. Zadeh in the 1960s, resembles human reasoning in its use of approximate information and uncertainty to generate decisions [17].

A classical fuzzy inference system consists of a rule base, membership functions, and an inference procedure that is showed in Fig. 3.

We define three linguistic variables: least total radiance (LTR), contrast (Co) and direction of least total radiance (DoLTR) as input and variable Road as output of the fuzzy system. Table I shows the terms of each linguistic variable.

IF-THEN rules are statements that make fuzzy logic useful. Generally a single fuzzy IF-THEN rule can be formulated according to:

IF *x* is ***A***; THEN *y* is ***B***.

TABLE I. LINGUISTIC VARIABLES AND LABELS FOR THE FUZZY-FEATURE ALGORITHM

	Linguistic variables	Fuzzy Sets	Type	Range	Parameters
Input	***Co***	Low	Trapmf*	[0 1]	[-0.1 0 0.2 0.3]
		Middle	Trimf**	[0 1]	[0.25 0.35 0.45]
		High	Trapmf	[0 1]	[0.4 0.5 1 1.1]
	LTR	Low	Trapmf	[0 1]	[-0.1 0 0.4 0.55]
		Middle	Trimf	[0 1]	[0.5 0.65 0.8]
		High	Trapmf	[0 1]	[0.7 0.85 1 1.1]
	DoLTR	Close to local average	Trapmf	[0 1]	[-0.1 0 0.05 0.0625]
		Moderate	Trimf	[0 1]	[0.058 0.1 0.2]
		Far to local average	Trapmf	[0 1]	[0.15 0.25 1 1.1]
Output	***Road***	True	Trimf	[0 1]	[-0.1 0 0.5]
		Probably	Trimf	[0 1]	[0.4 0.65 0.8]
		False	Trimf	[0 1]	[0.7 1 1.1]

a. *Trapezoidal membership function

b. **Triangular membership function

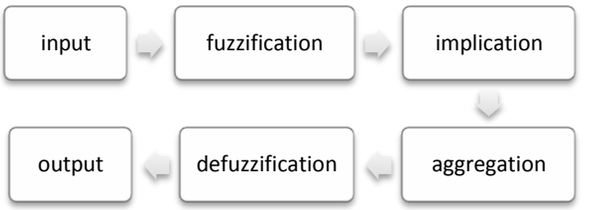


Fig. 3. Overall flow of a fuzzy inference system

A and ***B*** are linguistic labels defined by fuzzy sets on the range of all possible values of *x* and *y*, respectively. The first sentence is called *antecedent*, second one is called *consequent*.

For recognizing a road area in HR-SAR images, Some suggested fuzzy rules are:

If (Co is Low) Then (Road is False).

If {(Co is High) AND (LTR is Low) AND (DoLTR is Moderate)} Then (Road is True)

If {(Co is High) AND (LTR is Middle) AND (DoLTR is Moderate)} Then (Road is Probable)

These examples reveal an important aspect of fuzzy reasoning: the rule base should include observations of the important features.

Formulating the rules is more a question of the expertise of an operator than of a detailed technical modeling approach. Given the rules and input features, the degree of membership to each of the fuzzy sets has to be determined. Fuzzy processing of the input features requires a specification of the linguistic labels which represent fuzzy sets. The linguistic variables: Three linguistic input variables and one linguistic output variable and their fuzzy sets used in the investigations for each extracted feature are listed in Table I. Three fuzzy sets are assigned to each input variable which reflect an interactively carried out examination of all possible values of the features. In practice, this assignment is mostly a mixture of expert knowledge and examination of the desired input-output data. Also three fuzzy sets are chosen for the linguistic output with True, Probably and False. Widely applied membership functions are triangular and trapezoidal functions with maximum equal to 1 and minimum equal to 0. Sufficient overlap of neighboring membership functions is taken into account to provide smooth transition from one linguistic label to another.

The fuzzy AND or OR operators combine the membership values of the input features in each rule which results in one for the antecedent of that rule. Next step is the implication of the antecedent to the consequent. Implication is carried out for all rules and another step is to aggregate the output fuzzy sets over all rules. Inputs of aggregation are the truncated output functions returned by the implication process for each rule. The result of the aggregation process is one fuzzy set for each output variable. What remains in the final step is to defuzzify the fuzzy set and to produce a crisp output. The mostly applied defuzzification method is to calculate the centre of gravity which determines the centre of the area under the aggregated

output function (centroid). In our approach for fuzzy-feature algorithm as shown in Fig. 4 we follow the Max-Min approach proposed by Mamdani because it offers some advantages with regard to intuitive, widespread acceptance and well suited to human input [18]. As depicted in Fig. 4, for three feature values of a pixel: LTR=0.52, Co=0.3 and DoLTR=0.06 as input the model, the process of the fuzzy system is done to make a fuzzy resonance result. This Fig. shows the pixel is "Road" and corresponding defuzzified value is equal to 0.903. For a comprehensive study of fuzzy logic and fuzzy inference systems (FISs) please refer to [19].

3) Refinment

The last step of the road detection algorithm is selection of proper road areas on the image of features. Two criteria are utilized for this purpose, namely spatial and spectral.

Spatial Criterion: According to the ground spatial resolution of the sensors small detected areas cannot be part of the roads network, so they are negligible enough to be ignored. To this aims, areas containing 40 pixels or less can be deleted from the road areas.

Spectral Criterion: Areas detected with very high average radiance cannot be part of the roads network (considering the average total radiance of the image), so they are ignored. In other words, the areas of greater average radiance which are greater than a fraction of average total radiance are deleted.

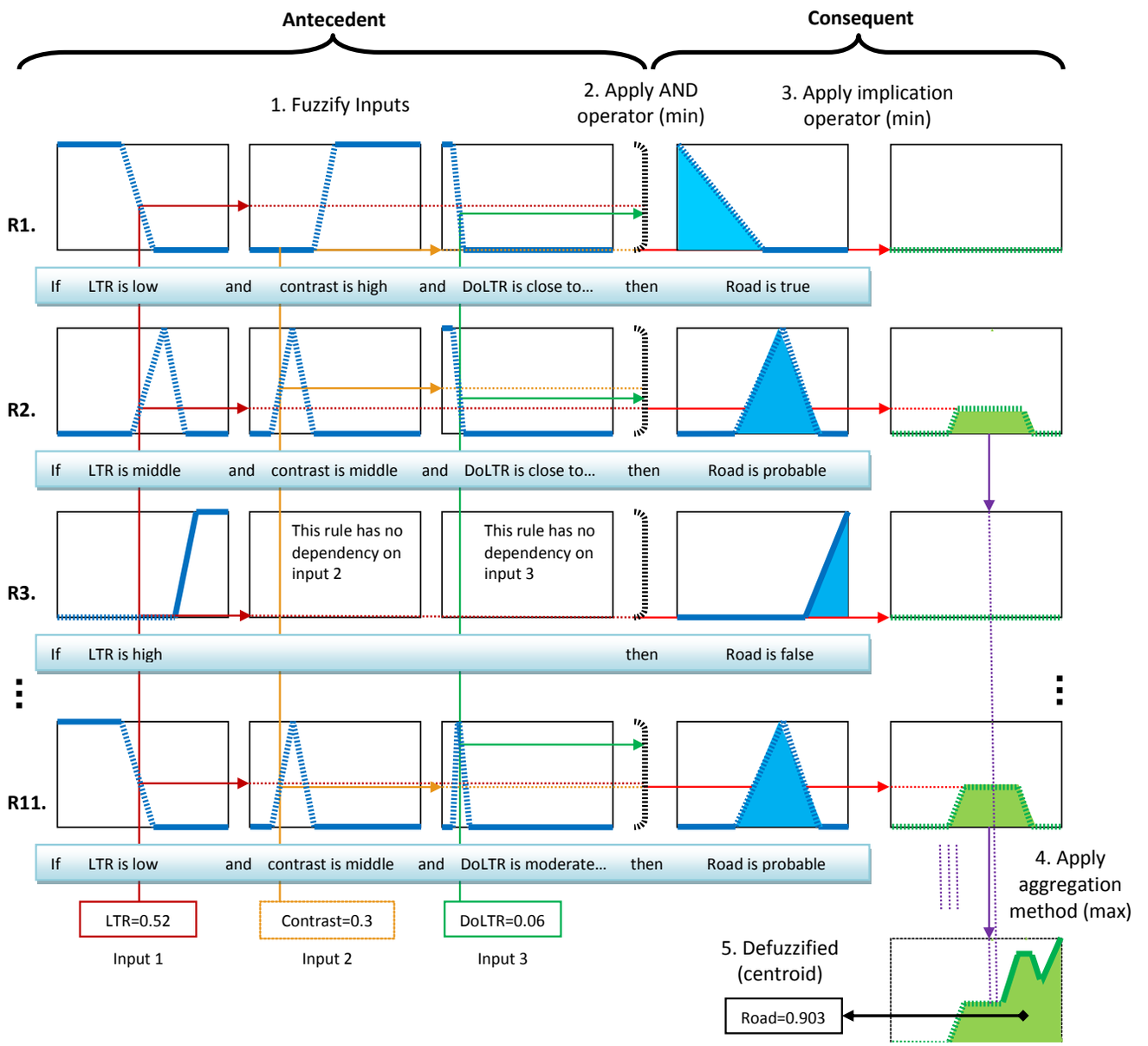


Fig. 4. Fuzzy inference system based on our approach

B. Sensitivity analysis and accuracy assessment

1) Accuracy Assessment

One of the basic requirements of the systems which perform a task automatically is analysis of accuracy. Therefore, various criteria have been employed to assess accuracy of the surveyed road. Wiedemann has provided a thorough discussion in this regard [20].

The corresponding pixels of the road area image are compared with the pixels of the reference road image in order to numerically analyze accuracy of the road detection obtained from logical combination of the images related to the radar imagery features. For this purpose, a binary image of the reference road network must be created by human operator in the first step.

Multiplying pixels of the final image in those of the binary image and calculating their average products will give a RCC. This RCC criterion can be deemed as a measure for average ability of the above mentioned algorithm in detection of the existing pixels of the road in this image. In other words, for showing nature of the pixels in terms of the road detection, this criterion demonstrates the acceptable percentage of performance applied.

Inversion of the values associated with the binary image obtained from reference road network and repeating the discussed operations for calculating the RCC, a similar criterion called BCC will be developed. This parameter acts as a criterion for showing how this algorithm performs to diagnose and distinguish the background road pixels.

2) Sensitivity Analysis of Algorithm

Inherent with all SAR imageries is speckle noise which is nothing else but variation in backscatter from inhomogeneous cells. Speckle noise reduces the image contrast which has a direct negative effect on texture based analysis of the imageries. Meanwhile, speckle noise also changes the spatial statistics of the underlying scene backscatter which in turn makes the classification of imageries a difficult task.

There are different filters for reduction of the noise effect on the SAR images. One of the most common filters for this purpose is Gamma-MAP [21]. The maximum a posteriori (MAP) filter is based on a multiplicative noise model with nonstationary mean and variance parameters. This filter assumes that the original digital number (DN) value lies between the DN of the pixel of interest and the average DN of the moving kernel.

Equation of this filter is the following cubic equation (5):

$$\hat{I}^3 - \bar{I}\hat{I}^2 + \sigma(\hat{I} - DN) = 0 \quad (5)$$

Where

$\hat{I}$: sought value ,

$\bar{I}$: local mean ,

$DN =$ input value ,

$\sigma =$ the original image variance

The filtering is controlled by both the variation coefficient and the geometrical ratio operators extended to the line detection.

III. Results and Discussion

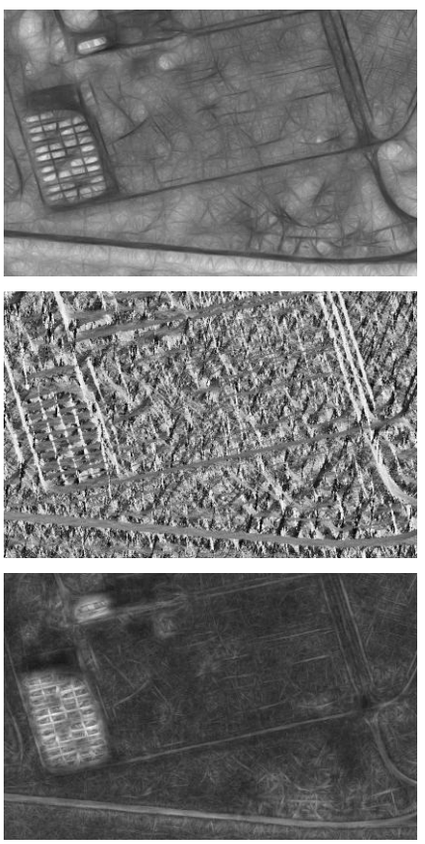
In the present work, the data set is a TerraSAR-X image dated 19th April 2011 with resolution of 1m×1m. This image covers various outspreads regions around Jam, Bushehr, Iran.

The multiple features were obtained in the first step according to (1) to (4). Images of each feature are illustrated in Fig. 5.

In the next step, the image of features are fuzzified first and then combined with each other by fuzzy logic. Finally, the defuzzification stage is used in the model to assign the expected (crisp) value for the output image.

The obtained image from combining the features is shown in Fig. 6.

In the following, spectral and spatial criteria were separately applied on the image in order to improve the performance and select proper road areas. Thereby, those areas with average radiance greater than 70% average total radiance of the images or smaller than 40 pixels were ignored from the road areas. The obtained results are depicted in Fig. 7.



(a) (b) (c)

Fig 5. Images of extracted features: (a), (b) and (c) belong to images related to features of r_0, θ_0 and c_0 respectively

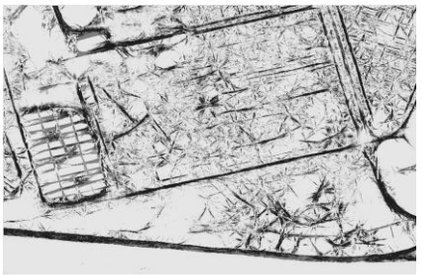


Fig 6. Image obtained from combination of features based on fuzzy method (dark pixels represent areas candidate for road, grey pixels show areas which might be road, and light pixels are representative of background or non-road areas)

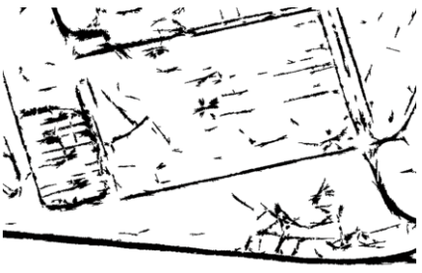


Fig 7. Image obtained from applying spectral and spatial criteria for selecting proper road areas (areas smaller than 40 pixels or those with average radiance greater than 70% average total radiance are deleted)

The results of applying the above mentioned criteria entitled TSX-Refinmented are listed in Table II.

A binary image is made from the reference network of road which is prepared by human operator to evaluate the results obtained from the road detection step. It has been demonstrated in Fig. 8a that all the road pixels are colored by green or red, while all the background pixels are colored by blue or white. By multiplying the corresponding pixels of reference image in the final binary image, the parameters of RCC, BCC and RMSE were calculated and the obtained numerical results are summarized in Table III.

Moreover, *ERDAS IMAGINE 9.1* software with gamma filter and windows of 3×3 and 5×5 sizes was used for studying sensitivity of the algorithm to speckle noise. All steps of the algorithm from beginning to analysis of accuracy were applied on the above mentioned image. The obtained results are listed in Table II. As can be seen from Table II below, no improvement was made in the RCC parameter by applying gamma filter and/or reduction of the speckle noise. Thereby, as was expected before, it can be concluded that the abovementioned algorithm has no sensitivity to the existing speckle noise in the SAR images and this insensitivity is due to nature of the extracted features which use information of the neighbor pixels. By taking into consideration the results obtained from numerical accuracy analysis it can be concluded that the sensitivity of this algorithm is mainly focused on detection of the road areas and that applying the spatial and spectral criteria in choosing the proper road areas has added to the aforementioned sensitivity. As usual, the roads which are parallel to the incident direction are much more visible than the other ones. [5].

TABLE II. INDEXES OF ACCURACY ANALYSIS FOR DIFFERENT INPUT IMAGES

Method	SAR image	RCC	BCC	RMSE
Proposed algorithm	TSX-Speckled	0.93	0.67	0.81
	TSX-Despeckled	0.93	0.72	0.83
	TSX-Refinmented	0.93	0.81	0.87
Negri's method	TSX-and	0.59	0.87	0.74

TABLE III. PARAMETERS FOR FEATURE BINARIZATION

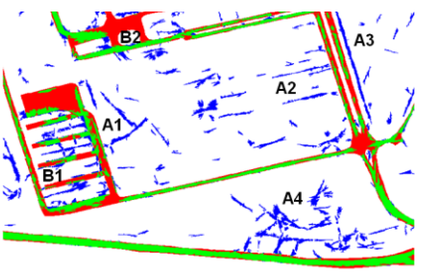
Features	Threshold	Window size
θ_0	$T_\theta=8^o$	22 pixels
r_0	$T_r=0.90$	17 pixels
c_0	$T_c=0.90$	17 pixels

As shown in Fig. 8a, the areas A1 to A4 which indicate regular rows of trees and probably traces of passing of agricultural machinery are detected as road areas. Furthermore, the algorithm is able to detect B1 to B2 areas as undetected ones due to one of these reasons:

- Areas without sufficiently high contrast in comparison with their surroundings,
- Areas completely covered by adjacent trees, or
- Areas located at margins of the image,

The feature combination method is compared with logical AND operator used by Negri[16] which was utilized considering proper threshold values indexed in Table III. The numerical results from this work is given in Table II with the name of "TSX-and". In total the result are strongly dependent to width and visibility of roads as well as the scene content.

(a)



(b)

Fig 8. (a) Classified image of detected areas (green: correctly detected areas, blue: incorrectly detected areas, red: undetected areas, white: non-road areas); and (b) input radar image

IV. Conclusion

Recovery of the road networks using remote sensing of high resolution optical and SAR images has become one of the frequent applications of these images. The areas candidate for road were obtained in this study by surveying multiple feature extraction based on the difference between road pixels as compared to their surroundings and then combining them together in a fuzzy method. The obtained results indicate success of the algorithm in detection of the road areas in comparison with the method of using logical AND operator, as well as insensitivity to the speckle noise. Thereby, an accuracy of 87% was met by a strict accuracy strategy. It is thus possible to contribute to regularization of the road areas by using mathematical morphology functions and applying them on the final image.

References

[1] F. M. Henderson and Z.-G. Xia, "SAR applications in human settlement detection, population estimation and urban land use pattern analysis: a status report," Geoscience and Remote Sensing, IEEE Transactions on, vol. 35, pp. 79-85, 1997.

[2] F. Caltagirone, et al., "SkyMed/COSMO mission overview," in Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS'98. 1998 IEEE International, 1998, pp. 683-685.

[3] F. Tupin, et al., "Detection of linear features in SAR images: application to road network extraction," Geoscience and Remote Sensing, IEEE Transactions on, vol. 36, pp. 434-453, 1998.

[4] B.-K. Jeon, et al., "Road detection in spaceborne SAR images using a genetic algorithm," Geoscience and Remote Sensing, IEEE Transactions on, vol. 40, pp. 22-29, 2002.

[5] F. Tupin, et al., "Road detection in dense urban areas using SAR imagery and the usefulness of multiple views," Geoscience and Remote Sensing, IEEE Transactions on, vol. 40, pp. 2405-2414, 2002.

[6] F. Dell'Acqua, et al., "Improvements to urban area characterization using multitemporal and multiangle SAR images," Geoscience and Remote Sensing, IEEE Transactions on, vol. 41, pp. 1996-2004, 2003.

[7] L. Bentabet, et al., "Road vectors update using SAR imagery: A snake-based method," Geoscience and Remote Sensing, IEEE Transactions on, vol. 41, pp. 1785-1803, 2003.

[8] B. Wessel, "Road network extraction from SAR imagery supported by context information," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science, vol. 35, pp. 360-366, 2004.

[9] Y. Chen, et al., "Particle filter based road detection in SAR image," in Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications, 2005. MAPE 2005. IEEE International Symposium on, 2005, pp. 301-305.

[10] G. Lisini, et al., "Feature fusion to improve road network extraction in high-resolution SAR images," Geoscience and Remote Sensing Letters, IEEE, vol. 3, pp. 217-221, 2006.

[11] S.-y. Li, et al., "Road extraction from high resolution dual-polarization SAR images over urban areas," in International Conference on Earth Observation Data Processing and Analysis, 2008, pp. 72850Q-72850Q-10.

[12] U. Stilla and K. Hedman, "Feature fusion based on bayesian network theory for automatic road extraction," in Radar Remote Sensing of Urban Areas, ed: Springer, 2010, pp. 69-86.

[13] K. Hedman, et al., "Road network extraction in VHR SAR images of urban and suburban areas by means of class-aided feature-level fusion," Geoscience and Remote Sensing, IEEE Transactions on, vol. 48, pp. 1294-1296, 2010.

[14] G. Zhou, et al., "Linear feature detection in polarimetric SAR images," Geoscience and Remote Sensing, IEEE Transactions on, vol. 49, pp. 1453-1463, 2011.

[15] J. Liu, et al., "Road extraction from SAR imagery based on an improved particle filtering and snake model," International Journal of Remote Sensing, vol. 34, pp. 8199-8214, 2013.

[16] M. Negri, et al., "Junction-aware extraction and regularization of urban road networks in high-resolution SAR images," Geoscience and Remote Sensing, IEEE Transactions on, vol. 44, pp. 2962-2971, 2006.

[17] L. A. Zadeh, "Fuzzy algorithms," Information and control, vol. 12, pp. 94-102, 1968.

[18] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," International journal of man-machine studies, vol. 7, pp. 1-13, 1975.

[19] H. J. Zimmermann, Fuzzy set theory-and its applications: Springer, 2001.

[20] C. Wiedemann, "External evaluation of road networks," International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences, vol. 34, pp. 93-98, 2003.

[21] M. Mansourpour, et al., "Effects and performance of speckle noise reduction filters on active radar and SAR images," in Proc. ISPRS, 2006, pp. 14-16.

Object Recognition Using Hough-transform Clustering of SURF Features

Viktor Seib, Michael Kusenbach, Susanne Thierfelder, Dietrich Paulus
Active Vision Group (AGAS), University of Koblenz-Landau
Universitätsstr. 1, 56070 Koblenz, Germany
{vseib, mkusenbach, thierfelder, paulus}@uni-koblenz.de
http://homer.uni-koblenz.de

***Abstract*—This paper proposes an object recognition approach intended for extracting, analyzing and clustering of features from RGB image views from given objects. Extracted features are matched with features in learned object models and clustered in Hough-space to find a consistent object pose. Hypotheses for valid poses are verified by computing a homography from detected features. Using that homography features are back projected onto the input image and the resulting area is checked for possible presence of other objects. This approach is applied by our team homer[at]UniKoblenz in the RoboCup[at]Home league. Besides the proposed framework, this work offers the computer vision community with online programs available as open source software.**

I. Introduction

The ability to memorize and redetect object instances is a key feature in modern computer vision applications and robotics. Such systems need to cope with many difficulties in practical scenarios like a home environment. Typically, objects appear in different poses and in cluttered environments. Further, occlusion makes it difficult to determine exact borders between objects and highly heterogeneous backgrounds are easily mistaken for object parts. Finally, different illuminations challenge the feature extraction process and add an additional uncertainty to the object recognition process.

In our proposed approach we use Speeded Up Robust Features (SURF) [1] to cope with these challenges. SURF is a point feature detector which also provides a descriptor for matching. Its main advantage is the fast computation while the features are distinctive enough to enable robust object recognition even under difficult circumstances such as partial occlusion and cluttered background.

The descriptors from the extracted interest points are matched with objects from a database of features using nearest-neighbor matching. The identification of clusters for a certain object was accomplished using Hough-transform clustering to obtain valid object pose hypotheses. In the final step a homography is built using the clustered features to verify consistent pose parameters and select the most probable object pose for each recognized object.

While [2] already presented our approach briefly, it focused on a comparison with a dense-statistical approach without detailed algorithm description. Therefore, the contributions of this paper are as follows. We describe our object recognition approach in detail with a thorough explanation of each step. We then evaluate our algorithm with respect to changes in important algorithm parameters. Finally, we provide an open source implementation of our approach as a Robot Operating System (ROS) package that can be used with any robot capable of interfacing ROS. The software is available at [3].

In Sec. II related approaches are briefly presented. These approaches either use similar techniques or are used by other teams competing in the RoboCup@Home league. Hereafter, Sec. III presents our approach in great detail, an evaluation follows in Sec. IV. Finally, Sec. V concludes this paper.

II. Related Work

Similar to our approach, Leibe and Schiele [4] use the Hough-space for creating hypotheses for the object pose. Later, this concept was extended to the *Implicit Shape Model* formulation in Leibe et al. [5]. Leibe et al. use the Harris corner detector [6] for key point extraction and image patches of 25×25 pixels as local descriptors. In contrast to SURF descriptors, image patches used in the Implicit Shape Model are very high dimensional. Thus, Leibe et al. use a special descriptor clustering technique to reduce the number of key points into so-called *codewords*. Together with relative vectors pointing to the object's center these codewords form the data basis for object recognition. During recognition, matching codewords cast their votes for the object center into a continuous Hough-space. Maxima in Hough-space are considered true object positions. While this approach provides good results in general, it is not invariant to object rotation. This restriction impedes its usage in the RoboCup@Home for service robots.

Team Golem [7] and their object recognition approach placed second in the RoboCup Technical Challenge of 2012. Golem uses the Multiple Object Pose Estimation and Detection (MOPED) framework proposed by Collet et al. [8] for object recognition. Similar to our approach, several images of each object are acquired during training. Scale Invariant Features Transform (SIFT) [9] features are extracted from each image and a 3D model of each object is reconstructed using structure from motion. In the recognition phase hypotheses for the object pose are obtained with SIFT features extracted from the scene and the previously created 3D model. In principle, a 3D object model provides advantages over a method solely working with 2D images, especially if object manipulation is desired. However, 3D model creation in the MOPED framework makes the training phase error-prone and time consuming. The camera has to be moved carefully around the object, while the background has to contain sufficient matching features in subsequent images. In contrast, our approach does not rely on

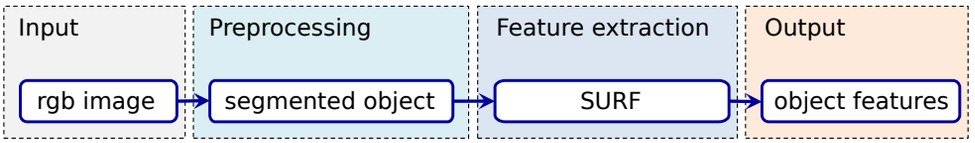


Fig. 1: Image processing pipeline for the training phase

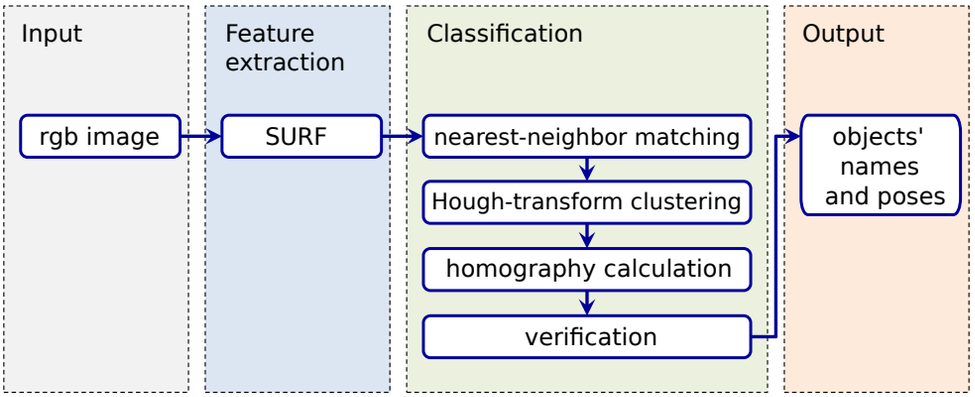


Fig. 2: Image processing pipeline for the recognition phase

background information for training and an object is usually trained in less than a minute.

The object recognition approach by team NimbRo@Home consist of a combined descriptor of SURF features and a color histogram [10]. As usual, descriptors are obtained from several object views in the training phase. During recognition, the extracted descriptors are compared to the database of learned object descriptors. If matching descriptors are found, object hypotheses are tracked with a multi-hypothesis tracker. hypotheses confirmed over several frames are considered valid.

III. Hough-Transform Clustering of SURF Features

Our object recognition approach is based on 2D camera images. In the training phase, SURF features are extracted and saved in a database. The recognition phase calculates SURF features on an input image and matches features with the database using nearest-neighbor matching. Wrong correspondences are removed and object pose transformations are clustered in a multi dimensional Hough-space accumulator. Finally, the matched features are verified by calculating a homography.

From each image we extract a number of SURF features f. A feature is a tuple f = (x, y, σ, θ, δ) containing the position (x, y), scale σ and orientation θ of the feature in the image, as well as a descriptor δ. Thus, the features are invariant towards scaling, position in the image and in-plane rotations. They are also robust towards changes in illumination and lesser off-plane rotations.

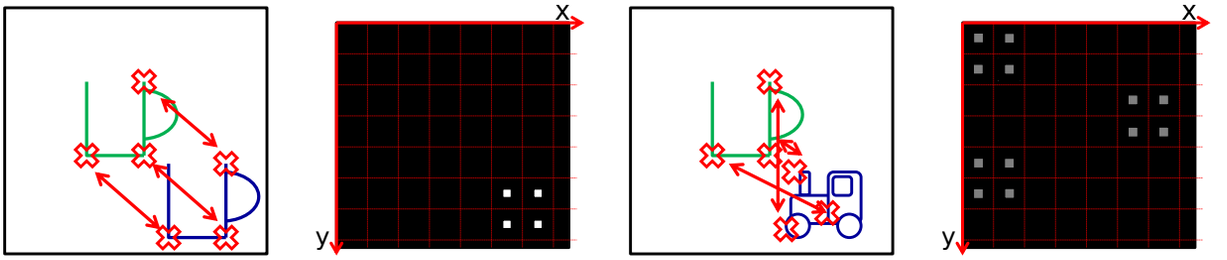
A. Training

The image processing pipeline for the training phase is shown in Fig. 1. In order to train the object recognition classifier an image of the background has to be captured. Subsequently, an image of the object is acquired. From this two images a difference image is computed to separate the desired object from the background. Depending on the light conditions, the object might cast a shadow on its surroundings. Naturally, this shadow would appear in the difference image and thus be considered as part of the object. Therefore, the borders of the extracted object are adjusted to reduce the area contributed to the object and thus remove shadows from the foreground. From the acquired object image SURF features are extracted and stored in an object database. Further, images with a different object view are acquired and added to the object model in the database. In their original publication, Bay et al. recommend 30° as an optimal rotation between subsequently acquired images of an object for SURF computation [1]. However, it is not necessary to acquire different rotations of the same object view, since SURF features and the presented algorithm are rotation invariant.

B. Recognition

The image processing pipeline for the recognition phase is shown in Fig. 2. During object recognition no information about the background is available. Thus, SURF features are computed on the whole input image. The obtained features are then matched against the features stored in the object database using nearest neighbor matching.

1) Nearest-Neighbor Matching: For each feature in the input image the best feature in the database is obtained by calculating a distance measure based on the euclidean distance of the feature descriptors. Since simple distance thresholds do not perform well in high dimensional space, Lowe introduced the distance ratio [9], which is used here. The distance ratio is the ratio of the euclidean distance to the best fitting and the second best fitting descriptor. If this quotient is low enough, the best fit is considered a matching feature. If the quotient is higher than a given threshold, the best and the second



(a) (b) (c) (d)

Fig. 3: The four dimensional accumulator is depicted as a red grid in (b) and (d), where the cell position along the two axes indicates the relative feature position. Inside each cell the horizontal and vertical translation encode the scale and rotation, respectively. Objects are recognized and hypotheses about the object pose are clustered in Hough-space using the accumulator. With correct objects, the feature position, rotation and scale match (a). Thus, the same accumulator bins in Hough-space are incremented for each feature, leading to a maximum, shown white (b). On the other hand, if wrong features are matched, their relative positions, as well as scale and rotation differ (c). Thus, their votes in the accumulator are scattered (gray) (d). These votes are considered as outliers.

best descriptor fit almost equally well. This leads to the assumption that they are very likely the best matches only by chance and not because one of them actually matches the query descriptor. The distance ratio also sorts out ambiguous matches, which may result from repeated textures on objects. For the fast nearest-neighbor and distance computation in the high dimensional descriptor space we use an approximate nearest neighbor approach [11].

Since SURF features are calculated on the whole input image, including a potentially different background than in the training phase, not all features are matched in this step. The result of feature matching is a set of matches between features extracted from training images and the scene image. This set may still contain outliers, i. e. matches between features which do not correspond to the same object point. These erroneous correspondences are discarded in the next step.

2) Hough-transform Clustering: Each feature match gives a hint of the object's pose in the scene image. To cluster this information from all feature matches, a four dimensional Hough-space over possible object positions (x, y, σ, θ) is created. Here, (x, y) is the position of the object's centroid in the image, σ it's scale, and θ it's rotation. The goal is to find a consistent object pose in order to eliminate false feature correspondences from the previous step. This four dimensional accumulator is visualized in Fig. 3 (b, d). The red grid represents translation in x- and y-directions. Inside each grid cell, the x-axis represents scale and the y-axis represents rotation. Thus, each pixel inside a grid cell corresponds to a bin.

Each feature correspondence is a hypothesis for an object pose and is added to the corresponding bin in the accumulator. As suggested in [9] and [12], to reduce discretization errors, each hypothesis is added to the two closest bins in each dimension, thus resulting in 16 accumulator entries per feature correspondence. Clusters of maxima in the Hough-space correspond to most probable object poses, whereas bins with erroneous object poses get only little votes (Fig. 3 (b,d)). Thus, outliers are removed and correct object poses are obtained. For the next steps only bins with at least five entries are considered, since we want to find a consistent pose applying homography calculation. This low threshold helps finding small as well as partially occluded objects in the scene.

We chose 10 bins for each dimension in the Hough-space, resulting in 10^4 bins describing different possible object positions in the image. Each feature correspondence votes into 16 bins (the one it falls into and the closest ones of each dimension to avoid discretization errors). More bins per dimension would allow for a finer quantization of feature poses. However, this would also lead to potential maxima being scattered among neighboring bins. Objects with sufficient feature correspondences would be recognized with less confidence or would not be recognized at all. Choosing less bins on the other hand would lead to feature correspondences voting for wrong object poses or even wrong objects.

To calculate the bin for the object position the centroid of the object in the scene $\boldsymbol{c}_s$ has to be estimated. In the following, values with index o describe properties of an object acquired during the training phase, whereas values with index s refer to key points found in a test scene during the recognition phase. The translation vector $\mathbf{v}'$ from the centroid of the learned object $\boldsymbol{c}_o$ to the position of the feature key point $\boldsymbol{p}_o$ on the object, normalized with the scale ratio of the scene key point σ_s and the object key point σ_o is calculated according to Eq. 1.

$$\mathbf{v}' = (\boldsymbol{c}_o - \boldsymbol{p}_o)\frac{\sigma_s}{\sigma_o} \tag{1}$$

The resulting vector $\mathbf{v}'$ has to be rotated to account for possible object rotation in the scene. This is done by applying Eq. 2

$$\mathbf{v} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \mathbf{v}' \tag{2}$$

where $\alpha = |\theta_o - \theta_s|$ is the minimal rotation angle between the feature rotation θ_o of the object and the feature rotation θ_s in the scene. Finally, the estimated centroid of the object in the scene $\boldsymbol{c}_s$ is obtained by adding the feature's position in the scene $\boldsymbol{p}_s$ to the calculated translation vector $\mathbf{v}$ (Eq. 3).

$$\boldsymbol{c}_s = (c_{xs}, c_{ys})^T = \mathbf{v} + \boldsymbol{p}_s \tag{3}$$

The bins for the x- and y-location in the accumulator, i_x and i_y, are calculated as shown in Eq. 4 and Eq. 5,

$$i_x = \left\lfloor \frac{c_{x,s}}{w} b_x \right\rfloor \tag{4}$$

$$i_y = \left\lfloor \frac{c_{y,s}}{h} b_y \right\rfloor \tag{5}$$

where w and h refer to the image width and height, and b_x and b_y is the number of bins in the x- and y-dimension of the Hough-accumulator, respectively. Apart from i_x and i_y also the corresponding bins with position indices $i_x + 1$ and $i_y + 1$ are incremented to reduce discretization errors.

The index for the rotation bin is calculated using the difference between the angles of the key point correspondences α and the total number of bins reserved for the rotation in the accumulator b_r. Originally, α is in $[-\pi, \pi]$, thus Eq. 6 normalizes the angle to be in $[0, 2\pi]$.

$$i'_r = \frac{(\alpha + \pi) b_r}{2\pi} \tag{6}$$

To allow for the rotations by $-\pi$ and π to be close together in the accumulator the final index for the rotation bin is calculated according to Eq. 7.

$$i_r = \lfloor i'_r \rfloor \mathrm{mod} b_r \tag{7}$$

Again, a second bin is used to reduce discretization errors. It's index is calculated according to Eq. 8:

$$i_r = \lfloor i'_r + 1 \rfloor \mathrm{mod} b_r. \tag{8}$$

The fourth dimension in the accumulator encodes the scale the point of interest was found at. To determine the accumulator bin for scale, first the ratio q between the scales of the key point in the scene σ_s and in the learned object σ_o is needed (Eq. 9):

$$q = \frac{\sigma_s}{\sigma_o}. \tag{9}$$

Further, the index is determined by the total number of bins used to represent scale b_s and the number of octaves n used for SURF extraction and is calculated according to Eq. 10:

$$b_s = \left\lfloor \frac{\log_2(q)}{2(n-1)} + 0.5 \right\rfloor b_s. \tag{10}$$

As before, discretization errors are reduced by using a second bin with the index $b_s + 1$. All scales that go beyond the range represented by the last bin are subsumed in the bin for the biggest scale of the accumulator.

As a result of the Hough-transform clustering all features with consistent poses are sorted into bins, while most outliers are removed because they do not form maxima in Hough-space (Fig. 3). So far, all features were processed independently of all other features without taking into account the geometry of the whole object. With the resulting possible object poses from the Hough-transform clustering, the goal in the next step is to find the best geometric match with all features in one accumulator bin.

3) Homography Calculation: In this step, bins representing maxima in Hough-space are inspected in order to find the bin that matches best the object pose. All bins containing five key point correspondences or more are considered as maxima. A perspective transformation is calculated between the features of a bin and the corresponding points in the database under the assumption that all features lie on a 2D plane. As most outliers were removed by discarding minima in Hough-space, a consistent transformation is obtained here. Random Sample Consensus (RANSAC) is used to identify the best homography for the set of correspondences. The homography with most point correspondences is considered to be the correct object pose. Using the obtained homography the recognized object can be projected into the scene (Fig. 8). Since homography calculation is computationally expensive the runtime of the object recognition algorithm would increase considerably if a homography was calculated for each bin. To speed up the algorithm all bins are sorted in descending order considering their number of features. A homography is calculated starting with the bin containing the highest number of features. The calculation terminates if the next bin contains less features than the number of found point correspondences in the calculation of the best homography so far. The result is a homography describing the relative position, orientation and scale of the best fitting training image for a test image, as well as the number of features supporting this hypothesis.

4) Verification of Results: The last step of our object recognition pipeline verifies the results. Using a threshold of a minimal matched feature number to verify the presence of an object in the scene is futile since large and heterogeneously structured objects contain more features than small and homogeneously structured objects. Instead, an object presence probability p is calculated as

$$p = \frac{f_m}{f_t} \tag{11}$$

where f_m is the number of matched features of that object and f_t is the total number of features that are present in the area of the object. The number of features in the object area is calculated by projecting the object into the scene using the homography and then counting all features in the bounding box of the projected object.

IV. Evaluation

This Section describes the different experiments performed to evaluate the presented object recognition approach. Experiments were performed to test the influence of the accumulator size, variable background and light conditions, as well as partial occlusion on the performance of the classification.

For the verification step we used a threshold of $p = 15\,\%$ (Eq. 11) and a minimum of 5 matched features per object. These two values have the most influence on the number of false positive recognitions. If they are not chosen restrictively

(a) (b) (c)

Fig. 4: Object view used for time measurement, green arrows indicate key points (a). Key point correspondences after nearest neighbor matching: some key points of the learned object view are matched to the background (b). Recognition result after eliminating erroneous correspondences (c).

TABLE I: Calculation time for each algorithm step, measured with the object from Fig. 4.

Algorithm Step	# Key Points	Time [ms]
Detection	500	697
NN-Matching	139	61
Hough-clustering	102	13
Homography	98	12

enough, the number of false positive recognitions increases. On the other hand, if they are chosen too restrictively, no objects would be recognized or a higher number of training views per object would be required to provide enough features for matching.

When not stated otherwise the accumulator size is 10 bins in each dimension. For the evaluation objects from the database presented in [13] were used. All images in this database have a resolution of 640×480 pixels.

A. *Performance*

The evaluation was performed on an off-the-shelf notebook with an Intel Core 2 processor with 2 GHz and 2048 MB RAM. We measured the processing time of the algorithm for one object view and a scene image with weak heterogeneous background and the learned object in the same pose. The object view used in this test is depicted in Fig. 4. The processing time needed for each step of our object recognition algorithm is presented in Tab. I.

Initially, 500 key points are detected in the scene image. The key point detection step takes the most time, since key points are not only extracted from the object, but also from the background. For instance, the nearest neighbor matching yields 139 key point correspondences between the scene image and an example object view (Fig. 4). However, some of these correspondences are erroneous as some of them are matched with the background. After the Hough-transform clustering only 102 correspondences remain. Finally, after calculating two homographies in 12 ms the best is found with 98 remaining correspondences to form the object recognition result (Fig. 4).

A total of 783 ms is needed for the calculation. This time increases if multiple object and object views are loaded into the object database, as the extracted features have to be compared

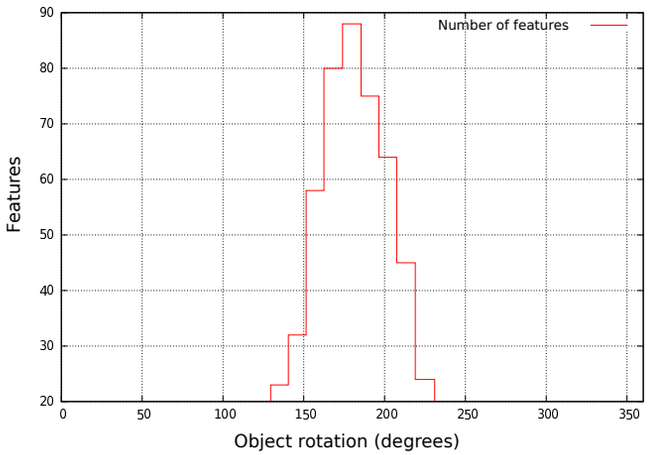


Fig. 5: Number of matched features of an object view depending on the object's rotation. The object view was acquired at a rotation of 180°.

TABLE II: Object recognition results on images with different backgrounds. The numbers in brackets indicate the number of false positive recognitions.

Object	hom. back.	weak het. back.	strong het. back.
bscup	100 %	90 % (1)	100 %
nizoral	100 %	100 % (2)	90 %
perrier	100 %	100 % (1)	100 % (2)
ricola	100 %	100 % (1)	100 %
truck	100 %	90 %	70 % (1)

with all data in the database. It is therefore crucial not to learn too many views of an object (see next Subsection). However, the step consuming the most time (key point extraction) has to be performed only once per scene image, regardless of the number of learned objects in the database.

B. *Learning Object Views*

We performed experiments to determine how many object views are necessary for reliable object recognition. In order to do this it has to be determined by what angle an object can be rotated without loosing too many key points. For the evaluation, a single object view was acquired. Without loss of

Fig. 6: Objects from [13] used for evaluation. From left to right: *bscup*, *nizoral*, *perrier*, *ricola*, *truck*. All objects are shown with homogeneous background.

generality the object view was defined as depicting a pose with a rotation of 180° about its vertical axis. Subsequently, images from the database showing the object at different rotations were used for testing. As shown in Fig. 5 the number of matched features decreases rapidly for rotations beneath 150° and above 220°. Thus, a rotation of 30° between subsequently acquired image views is a good trade-off between the number of found features and image views.

C. Accumulator Size

The size of each dimension of the accumulator is a crucial parameter for the performance of the algorithm. In our approach 10 bins per dimension proved to be a good trade-off between quantization errors (if too many bins are used) and insufficient accuracy (if too little bins are used). More than 10 bins lead to a shorter runtime as the features are distributed among a greater bin number, thus leading to less bins with a sufficiently large number of features for further processing. However, at the same time less features can be matched leading to unreliable recognition results.

D. Background Variation

The experimental results of our algorithm with different backgrounds are presented in Tab. II. A comparison of our algorithm with a statistical object recognition approach was given in [2]. The algorithm was trained with 5 different objects (Fig. 6) and 5 views per object from [13]. The classification was performed on the same 5 objects, but with 10 different views per object. The employed database contains images of the same objects with homogeneous, weak heterogeneous and strong heterogeneous backgrounds (Fig. 7). Different light conditions are present in the images with non-homogeneous backgrounds.

With increasing heterogeneity of the background, more erroneous correspondences are matched. If their number is very high, a false positive recognition occurs. A challenge in recognition is posed by the objects *perrier* and *truck* as they are small compared to the overall image size. With the low image resolution only few pixels remain for the object and thus only a little number of features is extracted. During the Technical Challenge of the RoboCup we used a higher image resolution. Please refer to Sec. IV-F for more details.

E. Partial Occlusion

Another experiment was performed to test the algorithm with partially occluded objects. Occlusion was simulated by partially replacing the object in the test data with the corresponding background. The results are presented in Fig. 8. The

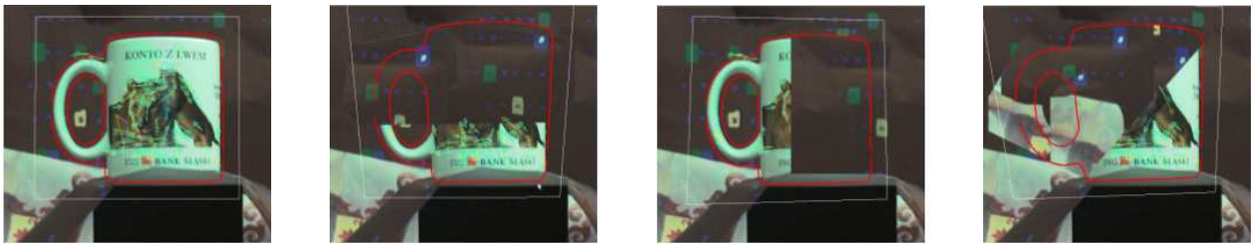
Fig. 7: Object *nizoral* from [13] with different backgrounds. From left to right: homogeneous, weak heterogeneous, and strong heterogeneous backgrounds.

unoccluded object is recognized with a total of 98 matched features and a confidence of 38 % (Eq. 11). With increasing occlusion the number of features decreases, but is still high enough for a correct recognition of the object. However, with increasing occlusion the accuracy of the computed homography (red lines in Fig. 8) and thus of the bounding box decreases.

F. RoboCup@Home: Technical Challenge

This object recognition approach was also applied during the Technical Challenge in the @Home league of the RoboCup world championship that took place in Mexico-City in 2012. 50 objects were placed on a table containing randomly selected 15 of 25 previously known objects. Our robot could correctly identify 12 of the 15 present known objects correctly, while at the same time having no false positive recognitions. This recognition result was achieved with a single scene view. With this result our robot places first in the Technical Challenge and won the Technical Challenge Award. The input image for object recognition as well as the recognition results are shown in Fig. 9.

We use a difference of 30° between object views to minimize training time and the number of images in the database. Objects were trained and recognized with an off-the-shelf digital camera (Canon PowerShot SX100 IS) and an image resolution of 8 megapixels (MP). Since the object recognition took a long processing time, further tests with the RoboCup@Home objects were performed after the Technical Challenge (Tab. III). The total recognition time depends on the resolution in the training phase as well as on the resolution of the scene image. However, the resolution in the training has a greater influence on the total recognition time. According to Tab. III it is sufficient to create an object database where features are extracted from 4 MP images, but use a resolution of 8 MP for recognition. This is not surprising since the object



(a) (b) (c) (d)

Fig. 8: Example images for detection of partially occluded objects. The unoccluded object is recognized with 98 matched features and 38 % confidence (a). The occluded object have less features, but are still recognized correctly: 49 with 33 % (b), 17 with 17 % (c), and 34 with 34 % (d).

(a)

(b)

Fig. 9: The input image for object recognition as acquired by our robot during the Technical Challenge of the RoboCup (a) and the output image depicting the recognition results (b). During training and recognition an image resolution of 8 MP was used.

to camera distance is usually smaller during training than during recognition. Thus, even with a lower image resolution a sufficient number of features can be extracted and saved in the database during training.

V. Conclusion

We presented our object recognition approach that we use in the RoboCup@Home league. Our approach is based on SURF features that are clustered in Hough-space. Hough-clustering allows us to discard most erroneous correspondences in the detection step. Subsequently, homographies are calcu-

TABLE III: Comparison of different image resolutions and their effect on recognition time and recognition quality.

Resolution Training [MP]	Resolution Scene [MP]	Recognition Time [s]	True Positives	False Positives
4	4	20	5	1
4	8	26	12	0
8	4	53	6	1
8	8	117	12	0

lated to take into account the relative positions of all features on the object. The homography that best matches the object geometry is back projected into the scene image to indicate the position of the recognized object.

Our recognition approach performs well on images with cluttered background and partially occluded objects. Objects at different scales and with arbitrary poses in the scene image are recognized reliably. For best results, it is recommended to use high resolution images in order to extract sufficient features for object representation.

This object recognition algorithm is available as an open source ROS package and can be downloaded from [3]. Apart from the recognition algorithm itself, the ROS package provides a convenient GUI to learn new object models and test the recognition. With this GUI new object models can be learned in less then a minute.

Our future work will concentrate on further evaluating and optimizing our approach. We plan to test several key point detectors and descriptors, as well as test different Hough-clustering techniques.

REFERENCES

[1] H. Bay, T. Tuytelaars, and L. J. V. Gool, "SURF: Speeded up robust features," *ECCV*, pp. 404–417, 2006.

[2] P. Decker, S. Thierfelder, D. Paulus, and M. Grzegorzek, "Dense Statistic Versus Sparse Feature Based Approach for 3D Object Recognition," *Pattern Recognition and Image Analysis*, vol. 21, no. 2, pp. 238–241, 2011.

[3] AGAS. (2014, Jul.) Ros package for object recognition based on hough-transform clustering of surf. [Online]. Available: http://wiki.ros.org/agas-ros-pkg

[4] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," in *BMVC*, 2003.

[5] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV' 04 Workshop on Statistical Learning in Computer Vision*, 2004, pp. 17–32.

[6] C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, Manchester, UK, 1988, pp. 147–151.

[7] L. A. Pineda, C. Rascon, G. Fuentes, V. Estrada, A. Rodriguez, I. Meza, H. Ortega, M. Reyes, M. Pena, J. Duran *et al.*, "The golem team, robocup@ home 2014."

[8] A. Collet, M. Martinez, and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *The International Journal of Robotics Research*, p. 0278364911401765, 2011.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] J. Stuckler and S. Behnke, "Integrating indoor mobility, object manipulation, and intuitive interaction for domestic service tasks," in *Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on*. IEEE, 2009, pp. 506–513.

[11] M. Muja, "Flann, fast library for approximate nearest neighbors," 2009, http://mloss.org/software/view/143/.

[12] W. E. L. Grimson and D. P. Huttenlocher, "On the sensitivity of the hough transform for object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-12, no. 3, pp. 255–274, 1990.

[13] M. Grzegorzek and H. Niemann, "Statistical object recognition including color modeling," in *2nd International Conference on Image Analysis and Recognition*. Toronto, Canada: Springer, 2005, pp. 481–489.

Plane Segmentation in Discretized Range Images

Viktor Kovacs[1], Gabor Tevesz[2]
Department of Automation and Applied Informatics
Budapest University of Technology
Budapest, Hungary
Email: {kovacsv[1], tevesz[2] }@aut.bme.hu

***Abstract*—In this paper we show a method for segmenting planes in discretized range images. Heavy quantization results in range images having a small number of different depth values. The range value encoded in each image pixel is not treated to be having random additive noise. For noise reduction purposes and feature extraction we use morphologic operations on each layer of the range image. Based on the features (layer skeletons) extracted from the images we show an algorithm for segmenting planes. We construct sets of linear skeleton segments and using a region growing technique we combine the linear segments into planar regions. We present segmentation results of simulated and captured real range images.**

I. Introduction

Advancements in 3D sensors and algorithms go hand in hand, supporting each other. In recent years 3D sensing products (time of flight, triangulation based sensors) have become more affordable thus finding their way to more applications. Based on the features of the sensor (horizontal, vertical, depth resolution, sampling rate, measurement errors, output data format) different post-processing techniques may be used to improve or filter data for specific applications.

Output data format may be ordered or unordered. The latter (point cloud) treats significance of each data point equal and considers only the coordinates of the points not their order. Range images are ordered structures, similar to conventional color images however pixels encode distance or depth coordinates. Color and range information may also be combined (RGB-D images). Layered depth images (LDI) describe not only the visible (from a given viewpoint) but the occluded surfaces as well. LDIs are mostly used in computer graphics to improve rendering performance by precalculating datasets.

Range images are based on regularly sampled data and implicitly contain information about adjacency thus surfaces, while point clouds need to be processed to reconstruct surfaces. On the other hand point cloud data implicitly comprise of world Cartesian coordinates (x,y,z), range image pixels must be projected back to (x,y,z) from image coordinates using the field of view (and may also be corrected for aberrations) and (u,v,z) data.

Segmentation is an important step in image understanding, incorporating the localization of parts of an image that fits a specific model or parts that fit different models. In order to map for example an indoors environment that consists of mostly planar surfaces, it is essential to detect planes in the sensor data. Detected planes may be used to interpolate missing data points, match viewpoints etc.

In this paper we deal with range images that comprise only a small number of different depth values due to quantization, compression, short stereo baseline or the sensor's principal of operation. We present an algorithm for extracting features from these range images such as planes. Our novel approach first extracts each layer of the input image to form binary images. One pixel wide skeleton lines are extracted from each layer and a skeleton segment graph is built up. We estimate the local surface normal vectors for the skeleton pixels. We further segment the skeletons to linear sections based on the estimated normal vectors. Linear segments are then connected by a region growing algorithm to localize planar surfaces. Other methods such as random sampling or Hough transformation based techniques are misled by quantization error.

The organization of the paper is as follows. In section II we present related work. Section III. deals with skeleton extraction, local surface normal estimation and plane extraction. In section IV we show results of processing simulated data and captured images. Section V contains conclusions.

II. Related work

3D sensing (time of flight cameras, structured light based reconstruction) is extensively used in 3D scene (urban or indoors) reconstruction, object detection, autonomous vehicle navigation or other robotics applications such as SLAM (Simultaneous Localization and Mapping). In case of artificial environments a large portion of surfaces can be described as planes thus plane detection is widely used as a first step in range image understanding.

Hough transformation is widely used for pattern recognition in image processing, especially for line or circle detection. Based on the n number of parameters, an n dimensional discretized accumulator space is formed. Detected features in the original image vote for all possible parameter sets that describe the feature. In case of lines $y = ax + b$ each edge or active point vote for all the a, b parameters that contain the (x, y) point. Finally the accumulator space is evaluated for local maximums, parameter sets that were voted by most features. It is clear that the number of parameters and the parameter space resolution restricts the applicability, having more than two or three dimensional accumulator spaces leads to huge memory requirements. In [1] Iocchi et al. present 3D Hough transformation used for plane fitting. Every point corresponds to a curve in the accumulator space. Still in case of carefully chosen discretization of parameters it is possible to achieve real-time performance. Borrmann et al. [2] proposed a new accumulator space design (accumulator ball) in order to improve performance by regularizing parameter resolution.

Weingarten et al. [3] used laser scanner to acquire unordered datasets of indoors environments. Data were decomposed to cubic regions and stored in a dynamic list structure. The side length of the cubes were about 30 cm, each containing between 1 and 51 vertices. Random Sample Consensus (RANSAC [4]) was used to fit planes in each cubic region. Model parameters were evaluated based on the least squares method. The advantage over the PCA method is calculation speed. For the pixels, supporting the plane model, a 2D quad bounding box is calculated. Neighboring quads are merged together based on matching orientation and translation. In [5] authors used this approach in a SLAM application. The plane segments were used as landmarks for a SLAM. In case of heavily quantized range images RANSAC methods are not easily applicable at first. As each layer is described perfectly by a plane perpendicular to the camera axis, sampling must be consciously designed. The region growing segmentation step in our method could be substituted by RANSAC, by taking samples from skeleton pixels.

A fast plane detection algorithm was presented by Poppinga et al. in [6]. This method is based on region growing and local plane fitting as an eigenvector problem, however in order to improve efficiency, instead of a naive implementation, an incremental calculation method was proposed. The algorithm also exploits the neighborhood information that is available in range images. The authors also present a polygonalization method that is applied in 2D image space and later transferred to a 3D model.

A coarse to fine method for plane segmentation is used in [7]. First surface elements (surfels) are extracted at multiple resolutions from coarse to fine. In each step surfels are associated to planar regions that were detected in the previous step. Planes are created by grouping new unassociated surfaces to planes by Hough-transform. Connected planes are refined by RANSAC to find the best fitting plane and reject outliers.

Hulik et al. [8] modified three known depth image segmentation methods and proposed two new algorithms. They compared all five methods for accuracy and time consumption. They showed that while accuracy measures gave similar results, visual comparison showed significant differences.

Guillaume et al. [9] presented a surface segmentation method based on curvature tensors. Curvature is a widely used local descriptor for surfaces due to its invariant nature. Authors segmented triangle meshes into constant curvature regions. Based on curvature not only planar (infinite curvature) but other special surface types such as spherical or cylindrical surfaces may be easily segmented. Again, the estimation of curvature in discretized depth maps is a challenge.

In a previous work [10] we proposed a method to smooth range images that suffer from a significant amount of quantization noise. After smoothing we used a Hough-space method, orientation histogram to cluster local normals. Segmenting in Hough-space gave us parallel plane segments. We segmented point distribution along the normal axis to identify parallel planes.

We presented a method to identify salient points and corners using our layer based method in [11]. We used higher level heuristics to identify points that meet certain set of criteria in order to recognize corners and classify their type.

III. Plane segmentation

In order to detect planes we first create iso-depth skeletons from the layers. Next we approximate the local surface normals, then segment the skeletons to linear sections. Finally we grow planar regions from the linear segments.

A. Skeleton extraction

Our method extracts each layer (given by the available range values) from the image and processes these layers individually. These layers are considered as binary images. In order to reduce noise we remove small connected regions and apply morphological dilation to the images. The connected components algorithm is used to identify small regions. The region size threshold and the amount of dilation must be selected based on the noise. After dilation skeletons are extracted [12]. The applied algorithm has excellent properties for our purposes such as few side branches, compared to other simple thinning algorithms.

Layer skeletons are one pixel wide lines resembling the features of each layer. These skeletons are similar to what we would see in case of extracting edges from an ideal but quantized range image, however the skeletons due to the dilation-erosion step, are more noise tolerant. Also due to the dilation as an unfortunate side effect, sharp corners become curved.

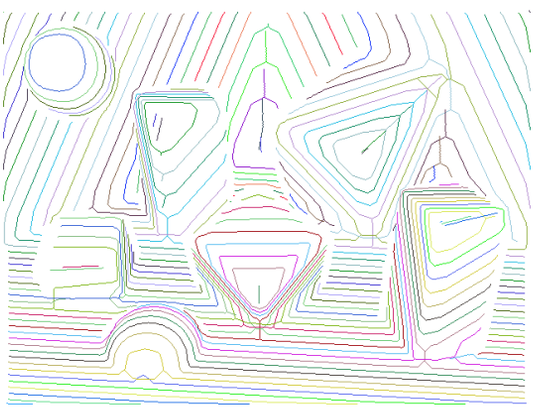


Fig. 1. Skeleton map generated from a quantized range image (44 layers)

Skeletons are broken into segments, which connect junctions or endpoints and do not contain other junctions. Short segments are likely to be unwanted side-branches. Segments containing pixels less than n_{MinSeg} are removed and segments are reevaluated, reconnected.

B. Local surface normal estimation

In case of ideal range images it is possible to sample a small region of pixels in image space, back-project them to world coordinates. The local surface normals may be estimated by PCA, the least significant direction points to the normal direction. Based on the error of the fitting it may be decided weather the sample is part of a plane.

Plane detection in range images having few depth values (quantized) is more challenging. In such case most of the estimated normals would be directed towards the camera $(0, 0, -1)$

direction as the layers are planar, having the same depth value. Similarly RANSAC based methods would find planar patches in the x-y plane to be best fitting. Noise model should not be considered random, quantization error is systematic. Pixels should not be handled equally: those that are on the edge or the center of quantization layers should be used for sampling. One option is to interpolate the pixel values within a layer [10] and estimate the local normals. Still, the noise might be high and the interpolation may be costly and might introduce other artifacts.

A 2D tangential orientation α is calculated for each pixel along all the segments by fitting a line to given number of $n_{FitLine}$ neighbors of the skeleton pixel. The tangent vector v is given by the eigenvector of the covariance matrix corresponding to the largest eigenvalue. This 2D orientation α equals the 3D tangential projected to the x-y plane. The b binormal direction is needed to estimate the local surface normal n.

$$C = cov(X,Y) = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \tag{1}$$

$$\lambda_{1,2} = \frac{\sigma_x^2 + \sigma_y^2 \pm \sqrt{(\sigma_x^2+\sigma_y^2)^2 - 4(\sigma_x^2\sigma_y^2 - \sigma_{xy}^2)}}{2} \tag{2}$$

$$\lambda^* = \max(|\lambda_1|, |\lambda_2|) \tag{3}$$

$$\mathbf{v}^* = \begin{bmatrix} -(\sigma_x^2 - \lambda^*)/\sigma_{xy} \\ 1 \end{bmatrix} \tag{4}$$

$$\mathbf{v} = \frac{\mathbf{v}^*}{|\mathbf{v}^*|} \tag{5}$$

$$\alpha = \arctan(v_2/v_1) \tag{6}$$

In order to determine b direction let us consider a P plane that is perpendicular to the previously defined tangent vector v. We look for the b binormal in P. We evaluate the signed distance d eq. (7) from P (given by the selected point p_0 and the tangent vector v) to all the p skeleton points along the segments in the next layer. We select a point where the signed distance changes sign and has the shortest Euclidean distance $d(p_0, p)$. This p^* point is considered adjacent to p_0. The binormal is given by $b = p^* - p_0$. The normal n is given by $n = b \times v$. Normals are flipped if needed to point to the front faces as the direction of v is undetermined. All vectors are normalized to unit vectors. To improve noise tolerance, a combination of several binormals may also be used. Not only the following but several adjacent layers' binormals may be used to estimate n. Note that in this case details such as edges, corners may be filtered.

$$d(p_0, p) = v \cdot (p - p_0) \tag{7}$$

where v is the tangent at point p_0.

The algorithm deals with pixels of layer skeletons, however to help visualization we assigned each pixel in the image space to the closest skeleton pixel in the same layer.

C. Plane extraction

In our model we assume that quantization levels are given by $x-y$ planes, perpendicular to the z axis. Considering P_1 $(\mathbf{n_1}, \mathbf{p_1})$ and P_2 $(\mathbf{n_2}, \mathbf{p_2})$ planes. P_1 is arbitrary, P_2 represents the quantization layer. Their intersection line is $\mathbf{l}$, given by:

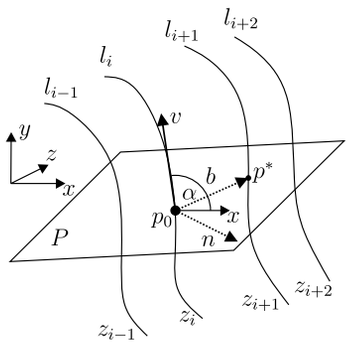


Fig. 2. l_i: layer skeletons, v tangent, b binormal, n normal direction

$$\mathbf{l} = \mathbf{p_0} + t\mathbf{v} = (x,y,z)^T \tag{8}$$

$$\mathbf{v} = \mathbf{n_1} \times \mathbf{n_2} \tag{9}$$

$$\mathbf{n_1} \cdot \mathbf{p_0} = d_1 = \mathbf{n_1} \cdot \mathbf{p_1} \tag{10}$$

$$\mathbf{n_2} \cdot \mathbf{p_0} = d_2 = \mathbf{n_2} \cdot \mathbf{p_2} \tag{11}$$

Assuming $\mathbf{n_2} = (0,0,z)^T$, P_1, P_2, l all contain $\mathbf{p_0}$, leads to:

$$y = ax + b = -\frac{n_x}{n_y} x \frac{d_1 - n_z z}{n_y} \tag{12}$$

Taking perspective projection into account in the image space (u, v) where $u = x/z$, $v = y/z$:

$$v = -\frac{n_x}{n_y} u \frac{d_1 - n_z z}{n_y} \tag{13}$$

Meaning that by changing z quantization edge position, it does not change the orientation of the line in image space, that is representing P_1 plane. Thus all $P(\mathbf{n}, \mathbf{p})$ planes are represented by n_x/n_y slope lines.

As a next step in the proposed algorithm we extract linear sections of the skeleton segments. To do this we first define $\beta(p)$ breakage angle for each skeleton pixel by fitting a line to n_β number of pixels to both the left and right neighbors of the pixel. The intersection angle is stored for each skeleton point. Next we apply a region growing algorithm for each segment: select the skeleton pixel where (14) the sum of all i adjacent pixels' breakage angle is minimal. In case this sum is larger than a β_{min} threshold, the segment is omitted. A linear segment is formed by adding p_0 and its adjacent skeleton point until the angle between the normal of the segment and the normal of the examined point is less than a given γ_{max} threshold. Each time a point is added, a given weight is associated to modify the normal direction and center point for the linear segments. The weight is given by the 2D Euclidean distance of the point and the center of the segment. $w(d_{2D}) = (\sqrt{(2\pi)}\sigma)^{-1} \exp(-d_{2D}/2\sigma^2)$ where σ is a predefined constant.

$$p_0 = \underset{p \in S}{\operatorname{argmin}} \sum_{q \in N_i(q)} \beta(q) \tag{14}$$

The next step is to grow parallel linear segments into planes. First the longest unprocessed linear segment l_0 is

LinearSegment 2 Linear segment extraction algorithm

input: S
output: P
while $S \neq \emptyset$ **do**
  $lookl, lookr \leftarrow$ true
  $n \leftarrow 1$
  $p_0 \leftarrow$GetFirstPoint(S)
  $L_i \leftarrow p_0$
  while not $finished$ **do**
    $l \leftarrow$GetNthLeftNeighbor(S,p_0,n)
    $r \leftarrow$GetNthRightNeighbor(S,p_0,n)
    if $lookl$ and Angle($l.n$, $L.n$) $< \gamma_{max}$ **then**
      $L_i \leftarrow l$, Weight($L_i.p, l$)
    else
      $lookl \leftarrow$ false
    end if
    if $lookr$ and Angle($r.n$, $L.n$) $< \gamma_{max}$ **then**
      $L_i \leftarrow r$, Weight($L_i.p, r$)
    else
      $lookr \leftarrow$ false
    end if
    Update $L_i.n$, $L_i.p$
    $finished \leftarrow$ (not ($lookl$ or $lookr$))or($\{l, r\} = \emptyset$)
    $n \leftarrow n + 1$
  end while
  if Count($L_i > n_{min}$) **then**
    $L \leftarrow \{L, L_i\}$
  end if
end while

PlaneSegment 4 Plane segmentation algorithm

input: L
output: P
loop
  $l_0 \leftarrow \underset{l \in L}{\text{argmax}}$length($l$)
  if length(l_0)¡n_{min} **then**
    Exit.
  end if
  $P_i \leftarrow l_0$
  $j \leftarrow 0$
  $c \leftarrow 0$
  loop
    $L_s = \{l | Angle(P_i, l) < \gamma_{max} \wedge |d(P_i, l)| \leq d_{max}\}$
    $l \in L_{l_0+j} \cup L_{l_0-j}$
    if $\#L_s > 0$ **then**
      $P_i \leftarrow L_s$, Weight(P_i, L_s)
      $c \leftarrow 0$
    else
      $c \leftarrow c + 1$
      if $c \geq c_{max}$ **then**
        End loop
      end if
    end if
  end loop
  if $\#P_i > p_{min}$ **then**
    $P \leftarrow P_i$
  end if
end loop

selected. Next the linear segments on the neighboring layers $L_{l_0 \pm j}$ are searched for similar normal directions and distances. Linear segments that fulfill the requirements are added to the plane (P_i). Each time a linear segment is added a weight is also given to modify the plane parameters. $w = 1/\sqrt{2\pi}\sigma \exp(-\text{length}_{3D}(l_i)^2/2\sigma^2)$, $\sigma = \text{length}_{3D}(l_0)$. Next the adjacent layers are evaluated and so on. The process stops when no more fitting segments are found in the last c_{max} layers. A new (P_{i+1}) plane segmentation starts.

$$\text{length}_{3D}(l) = \max_{p,q \in l} d(p, q) \tag{15}$$

IV. Results

In this section we investigate the accuracy of the algorithm for plane segmentation. For evaluation purposes first we use simulation images at different discretization levels with ground truth data. The algorithm was also tested on real captured data.

A. Simulation

Simulated scenes (Fig. 3.) were constructed with ground truth data to test the algorithm at different quantization levels. The scenes contained several distorted cubic and spherical objects. In order to simulate quantization a non-linear, tangential function was used. Although the image did not suffer from any noise, we used the same parameters (dilation) as for real images. This resulted in significant smoothing of skeletons around edges (Fig. 1). Figure 4. shows estimated local surface normals for skeleton pixels. Due to the quantization of the

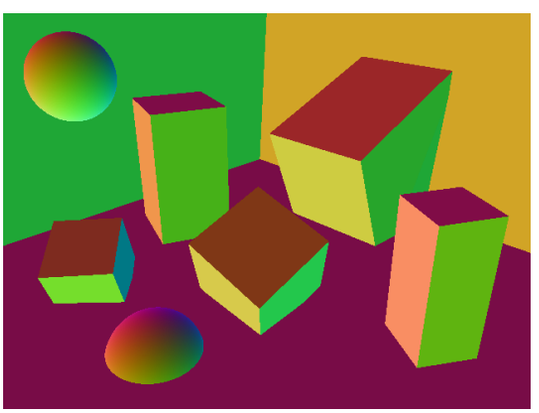


Fig. 3. Simulated scene, colors represent normal direction (RGB:XYZ)

layer skeletons' pixel positions, especially near 0° and 90° orientation, local surface normals tend to fluctuate. By increasing the number of pixels used for α estimation, this effect may be handled. Normal estimation on non-planar surfaces (spheres) showed good results. The resulting segmentation of the algorithm can be seen in Fig.5. Each color represents a segmented planar surface. Results show uncertainty near crease edges, in some cases these regions are oversegmented. By choosing an appropriate γ_{max} value it is possible to restrict segmentation from edges. The normals estimated for these individual planes are shown in Fig. 6 .

By increasing the depth resolution layers become less continuous causing linear skeleton segments to break into several pieces and distributed on several adjacent layers. This affects

the performance of the algorithm significantly as increased number of adjacent layers are needed to be searched for skeleton segments. In case of higher depth resolution range images, other conventional methods should be used.

In case of decreasing the number of layers the algorithm still gives acceptable results. The number of planar segments identified depends on the size of the regions and their orientation respect to the camera. A plane near perpendicular to the camera would be described by only a few layer segments thus increasing localization uncertainty.

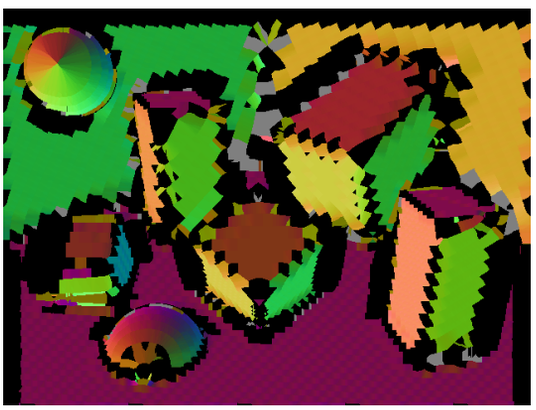


Fig. 4. Estimated local surface normals

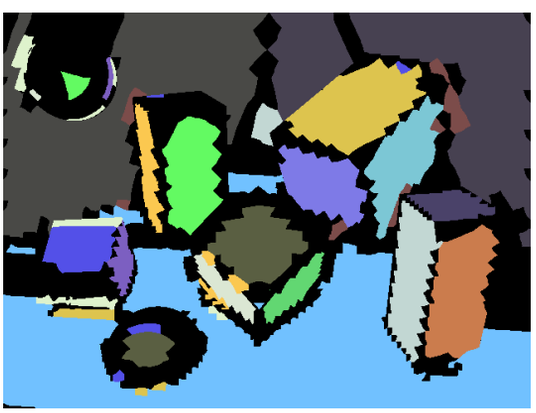


Fig. 5. Plane segmentation results

Fig. 6. Reconstructed, segmented normal direction image, 44 layers, steep regions were omitted

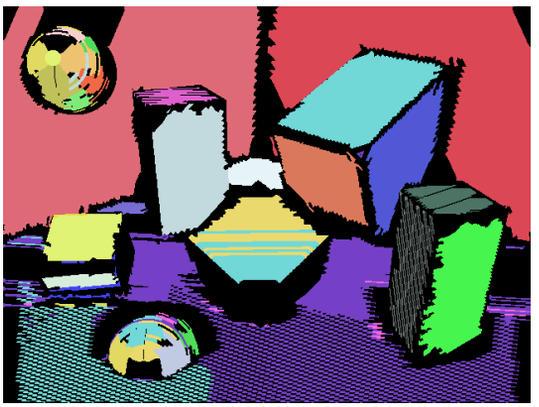


Fig. 7. Oversegmentation at 192 layers

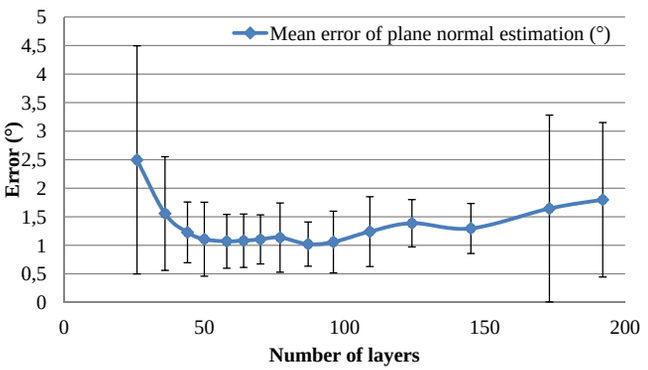


Fig. 8. Normal approximation error at different depth resolution

Figure 8. shows the mean error of the normal vector estimation for planes in the sample image. The measurements are based on different locations that are meant to be part of planar surfaces. In case the selected point was not identified as part of a plane, the value was omitted. As the number of layers increase the mean error drops, until about 80-100 layers. Having even more layers slowly increases the error as layers become more fragmented, less continuous (Fig. 7.). The optimal number of layers depend on the image contents and resolution. In case of the mean plane errors we see lower estimation error, as low as 1°.

Figure 9. shows the distribution of normal error for individual pixels that are part of a segmented plane, and for different quantization levels. We see that for less range layers the error falls faster, giving more accurate estimations. For more layers we see less drop in the error distribution due to the problem of quantization in image space: layer skeletons are very close thus estimation is more uncertain. The most significant portion of the estimation error lies in the 0-5° region.

B. Real data

We also evaluated the algorithm by using real data acquired by the Microsoft Kinect sensor. Optical distortions were not corrected. We examined the algorithm by further reducing the number of depth layers. Due to noise, oversegmentation is more likely compared to simulated images. By selecting algorithm parameters for a specific image acquisition system and application, segmentation errors may be greatly reduced.

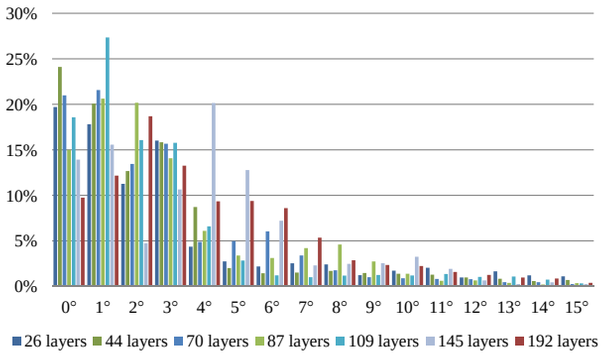


Fig. 9. Normal approximation error distribution on planes

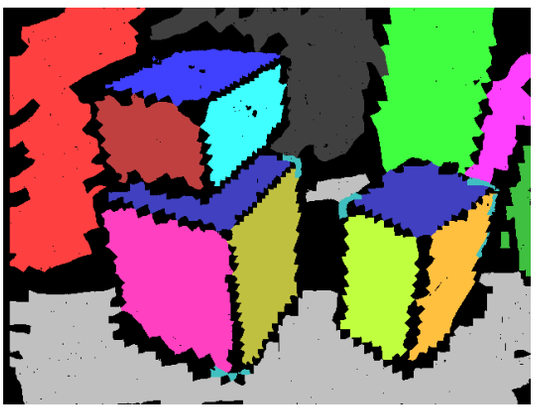


Fig. 10. Segmentation of a captured range image, consisting of only 41 layers

Figure 10. shows a segmentation result of a real image consisting of only 41 depth levels. All planar regions were identified, however small regions were omitted due to uncertainty. The top face of the two boxes in the image were nearly in an identical plane thus these were segmented as one. The algorithm may be easily modified to separate these regions. However the farthest part of the floor would also be segmented in a different plane.

V. Conclusion

In this paper we presented an algorithm capable of identifying planar surfaces in heavily quantized range images. Such data is not suitable for traditional algorithms as quantization error treated as random noise leads to undesired results. Our proposed algorithm utilizes a low amount of different depth values, creates skeleton structures from layers and estimates local surface normals. Finally skeletons are grown into planar regions.

We evaluated the performance of the algorithm by using simulated images as ground truth. We examined the accuracy at several quantization levels and concluded the algorithm works well between 40-100 layers - depending on the contents of the image. High and low slope planes may not be identified properly. High slope regions should be handled in the traditional way as these parts of the image are not considered heavily quantized.

Real captured images were also used in the evaluation process, where the algorithm showed excellent segmentation results even in images suffering from significant noise (uncertain transient between layers).

Acknowledgment

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013) organized by VIKING Zrt. Balatonfured. This work was partially supported by the Hungarian Government, managed by the National Development Agency, and financed by the Research and Technology Innovation Fund (grant no.: KMR-12-1-2012-0441).

References

[1] L. Iocchi, K. Konolige, and M. Bajracharya, "Visually realistic mapping of a planar environment with stereo," in *Experimental Robotics VII*, ser. ISER '00. London, UK, UK: Springer-Verlag, 2001, pp. 521–532.

[2] D. Borrmann, J. Elseberg, K. Lingemann, and A. Nchter, "The 3d hough transform for plane detection in point clouds: A review and a new accumulator design," *3D Research*, vol. 2, no. 2, 2011.

[3] J. Weingarten, G. Gruener, and R. Siegwart, "A Fast and Robust 3D Feature Extraction Algorithm for Structured Environment Reconstruction," in *None*, 2003.

[4] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[5] J. Weingarten and R. Siegwart, "3d slam using planar segments," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, Oct 2006, pp. 3062–3067.

[6] J. Poppinga, N. Vaskevicius, A. Birk, and K. Pathak, "Fast plane detection and polygonalization in noisy 3d range images," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, Sept 2008, pp. 3378–3383.

[7] B. Oehler, J. Stueckler, J. Welle, D. Schulz, and S. Behnke, "Efficient multi-resolution plane segmentation of 3d point clouds," in *Intelligent Robotics and Applications*, ser. Lecture Notes in Computer Science, S. Jeschke, H. Liu, and D. Schilberg, Eds. Springer Berlin Heidelberg, 2011, vol. 7102, pp. 145–156.

[8] R. Hulik, V. Beran, M. Spanel, P. Krsek, and P. Smrz, "Fast and accurate plane segmentation in depth maps for indoor scenes," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Oct 2012, pp. 1665–1670.

[9] L. Guillaume, D. Florent, and B. Atilla, "Curvature tensor based triangle mesh segmentation with boundary rectification," in *Computer Graphics International, 2004. Proceedings*, June 2004, pp. 10–25.

[10] V. Kovacs, "Landmark detection in low depth resolution range and intensity images," in *Automation and Applied Computer Science Workshop (AACS 2012)*, June 2012, pp. 169–181.

[11] V. Kovacs and G. Tevesz, "Corner detection and classification of simple objects in low-depth resolution range images," *Periodica Polytechnica, EECS*, vol. 57, no. 1, pp. 9–17, aug. 2013.

[12] G. Németh and K. Palágyi, "2d parallel thinning algorithms based on isthmus-preservation," in *ISPA 2011: 7th International Symposium on Image and Signal Processing and Analysis: Dubrovnik, Croatia, 4 - 6 September 2011. IEEE*, 2011, pp. 585–590.

GHSS iterative method for image restoration

Mehdi Bastani
Department of Applied Mathematics
Azarbaijan Shahid Madani University
Tabriz, Iran
Emails: bastani.mehdi@yahoo.com
bastani@azaruniv.ac.ir

Nasser Aghazadeh
Research Group of Processing and Communication
Azarbaijan Shahid Madani University
Tabriz, Iran
Email: aghazadeh@azaruniv.ac.ir

***Abstract*—In this study, we introduce the special Hermitian and skew-Hermitian splitting (SHSS) iterative method. The generalized version of HSS (GHSS) iterative method is also introduced to use in image restoration problem. Moreover, we present a new splitting for coefficient matrix in image restoration problem to apply in the GHSS method. Convergence of the proposed method is also investigated. Finally, three numerical examples are given to illustrate the efficiency and accuracy of the GHSS iterative method.**

***Keywords*—image restoration; Hermitian and skew-Hermitian splitting iterative method; generalized Hermitian and skew-Hermitian splitting iterative method; boundary conditions**

I. Introduction

In image restoration process, a priori knowledge of the degradation phenomenon is used to reconstruct or recover the degraded image. Image restoration is a major problem in many fields of applied sciences such as medical and astronomical imaging, engineering, optical systems and many other areas [4], [7]. Several methods have been applied to investigate the solution of this problem such as wavelet transform method [4], total variation (TV) method [3] and least square method [6]. In this paper, we consider the matrix-vector form of the image restoration problem which can be written as [4]:

$$g = Af + \eta, \tag{1}$$

where A is a blurring matrix of size $n^2 \times n^2$ and f, η, and g are n^2-dimensional vectors representing the original image, additive noise, and degraded image, respectively. Note that we use the point spread function (PSF) to construct the blurring matrix A. Since the observed image g is of finite dimension and Eq. (1) is obtained from a convolution process, we consider some boundary conditions (BCs) on the f outside the observed image domain. The structures of the matrix A is depend on these BCs.

In this work, we consider zero, periodic and reflexive BCs to restore images [5]. In zero BCs, the outside boundary of the exact image is assumed to be black. In this case, the blurring matrix A is a block Toeplitz with Toeplitz blocks (BTTB) matrix. Periodically extension of the data outside the domain of consideration leads to periodic BCs. For this case, the matrix A has block circulant with circulant blocks (BCCB) structure. In zero and periodic BCs, Fast Fourier Transforms (FFTs) are effectively applied to matrix-vector multiplications. In reflexive BCs, the original scene inside the boundary immediately reflects itself outside of the boundary. In this case, the matrix A is block Toeplitz-plus-Hankel with Toeplitz-plus-Hankel blocks (BTHTHB). Moreover, for symmetric PSFs, the two-dimensional fast cosine transform (DCT) can be used to diagonalize the matrix A.

It is noted that the linear system (1) is ill-conditioned. The Tikhonov method is one of the most known methods for solving ill-conditioned linear systems [5]. In the proposed method, the linear system (1) is substituted by solving the following problem:

$$\min_f \|Af - g\|_2^2 + \mu^2 \|Lf\|_2^2, \tag{2}$$

where L is a regularization matrix and $\mu > 0$ is called the regularization parameter. Note that the balance between the minimization of $\|Af - g\|_2^2$ and the regularization term $\|Lf\|_2^2$ is controlled by regularization parameter. In this study, we consider Eq. (2) when $0 < \mu < 1$ and $L = I$. Hansen et al. have shown that the Tikhonov minimization problem is mathematically equivalent to solving the following equation [5]:

$$(A^T A + \mu^2 I) f = A^T g. \tag{3}$$

To find the solution of the system (3), Lv et al. presented the following equivalent system [8]:

$$\underbrace{\begin{bmatrix} I & A \\ -A^T & \mu^2 I \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} e \\ f \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} g \\ 0 \end{bmatrix}}_{b}, \tag{4}$$

where K is $2n^2 \times 2n^2$ non-Hermitian positive definite matrix, I is the $n^2 \times n^2$ identity matrix and the additive noise is represented by auxiliary variable e, i.e $e = g - Af$.

The Hermitian and skew-Hermitian splitting (HSS) iterative method has been presented to solve non-Hermitian positive definite linear systems [1]. A generalization of the HSS (GHSS) iterative method has been presented by Benzi [2] to solve a class of non-Hermitian linear systems [2]. In the GHSS method, the Hermitian part of the coefficient matrix of linear system is splitted to positive definite and positive semi-definite matrices. Lv et al. used the idea of the HSS iterative method and presented a special case of the HSS (SHSS) method to solve the image restoration problem (4).

In this paper, we present a new splitting to the Hermitian part of coefficient matrix in (4) to use in the GHSS method. Convergence of the proposed method for image restoration problem (4) is also investigated.

This paper is organized as follows. In Section 2, we give a description of the SHSS and the GHSS methods to solve linear systems and image restoration problem. We present a new splitting of iteration matrix for the application in GHSS method. In Section 3, three examples are given to show the effectiveness and accuracy of the proposed method. Finally, some concluding remarks are presented in Section 4.

II. Description of the method

Suppose that A be a non-Hermitian matrix. To implement the HSS method, we first split the matrix A as

$$A = H + S, \tag{5}$$

where

$$H = \frac{1}{2}(A + A^T), \qquad S = \frac{1}{2}(A - A^T). \tag{6}$$

Then, the two splittings of A are presented as:

$$A = (H + \alpha I) - (\alpha I - S), \quad A = (S + \alpha I) - (\alpha I - H),$$

where α is a positive parameter. By alternating between the two splittings, the HSS iterative method for solving the proposed system is given for $k = 0, 1, \ldots$ as

$$\begin{cases} (H + \alpha I)x_{k+\frac{1}{2}} = (\alpha I - S)x_k + b, \\ (S + \alpha I)x_{k+1} = (\alpha I - H)x_{k+\frac{1}{2}} + b, \end{cases} \tag{7}$$

where x_0 is a given initial guess. Lv et al. [8] presented SHSS method by substituting $\alpha := 1$ in the second equation of Eq. (7). In other word, the SHSS method can be written as a following two-step iterative method for $k = 0, 1, \ldots$:

$$\begin{cases} (H + \alpha I)x_{k+\frac{1}{2}} = (\alpha I - S)x_k + b, \\ (S + I)x_{k+1} = (I - H)x_{k+\frac{1}{2}} + b. \end{cases} \tag{8}$$

Note that Hermitian and skew-Hermitian parts for the image restoration problem presented in Eq. (4) are given as follows:

$$\begin{aligned} K &= \begin{bmatrix} I & A \\ -A^T & \mu^2 I \end{bmatrix} \\ &= \begin{bmatrix} I & O \\ O & \mu^2 I \end{bmatrix} + \begin{bmatrix} O & A \\ -A^T & O \end{bmatrix} = H + S. \end{aligned} \tag{9}$$

where O is the $n^2 \times n^2$ zero matrix.

Remark 1. ([8]) If σ_1 and σ_n be the largest and smallest singular values of the matrix A, respectively, then for $\alpha > \frac{\sigma_1^2 - \mu^2 - 2\mu^2\sigma_1^2}{2 - \mu^2 + \sigma_1^2}$ the SHSS iteration method is convergent for any initial vector x_0. Moreover, the optimal value of α in the proposed method for image restoration problem is given by:

$$\alpha^* = \frac{\sigma_1^2 + \sigma_n^2 + 2\sigma_1^2\sigma_n^2}{2 + \sigma_1^2 + \sigma_n^2}. \tag{10}$$

Note that for the optimal value α^*, we have the most convergence speed of the SHHS method. It has been shown that the SHSS method is more effective than the HSS method for image restoration [8].

Benzi presented the GHSS method to solve the linear systems with non-dominant Hermitian part and skew-Hermitian part of the coefficient matrix [2]. In the GHSS method the Hermitian part of H is decomposed as

$$H = G + P = \epsilon L + P, \tag{11}$$

where L is a Hermitian positive definite, P is a Hermitian positive semidefinite and $\epsilon > 0$ is a small constant. The GHSS iteration is given as follows for $k = 0, 1, \ldots$:

$$\begin{cases} (G + \alpha I)x_{k+\frac{1}{2}} = (\alpha I - P - S)x_k + b, \\ (S + P + \alpha I)x_{k+1} = (\alpha I - G)x_{k+\frac{1}{2}} + b, \end{cases} \tag{12}$$

where x_0 is an initial guess. Now, we introduce the following lemma for the convergence of the proposed GHSS iterative method to solve linear systems.

Lemma 1. ([2]) Suppose that the matrix A is splitted as $A = H + S = (G + P) + S$, where G and P are Hermitian positive semidefinite and S is skew-Hermitian. If either G or K is positive definite, alternating iteration (12) converges unconditionally to the unique solution of Ax = b.

Now, to implement the GHSS method in image restoration problem, a new splitting for the Hermitian part of K is presented. This splitting is given as follows:

$$\begin{aligned} K = &\begin{bmatrix} \frac{1}{2}I & O \\ O & \beta\mu^2 I \end{bmatrix} + \begin{bmatrix} \frac{1}{2}I & O \\ O & (1-\beta)\mu^2 I \end{bmatrix} \\ &+ \begin{bmatrix} O & A \\ -A^T & O \end{bmatrix} = G + P + S, \end{aligned} \tag{13}$$

where β is a positive constant. By using the splitting (13), we can implement the GHSS iteration (12) to solve image restoration problem (4). Note that a simple matrix-vector multiplication can be used to solve the first equation in (12):

$$\begin{aligned} x_{k+\frac{1}{2}} = &\begin{bmatrix} \frac{1}{0.5+\alpha}I & O \\ O & \frac{1}{\alpha+\beta\mu^2}I \end{bmatrix} \\ &\times \left(\begin{bmatrix} (\alpha - 0.5)I & -A \\ A^T & (\alpha + (\beta - 1)\mu^2)I \end{bmatrix} x_k + b \right). \end{aligned} \tag{14}$$

Since the Krylov subspace methods are effective on problems containing matrix-vector multiplications. We use the GMRES method [9] to solve the second equation in GHSS iteration (12). After some simple manipulations of proposed algorithm in [8], we can implement GHSS method as the Algorithm 1.

Remark 2. In the third step of Algorithm 1, we use $e_{k+1}^{(0)} = e_{k+1/2}$ and $f_{k+1}^{(0)} = f_{k+1/2}$ as the initial guesses for GMRES method. Moreover, consider the residual of the proposed method as:

$$q_j = \begin{bmatrix} (\alpha - 0.5)e_{k+\frac{1}{2}} + g - (\alpha + 0.5)e_{k+1}^{(j)} - Af_{k+1}^{(j)} \\ (\alpha - \beta\mu^2)f_{k+\frac{1}{2}} + A^T e_{k+1}^{(j)} - (\alpha + (1-\beta)\mu^2)f_{k+1}^{(j)} \end{bmatrix}.$$

The GMRES method acts until $\|q_j\|_2 / \|q_0\|_2 < \zeta$, where ζ is a very small positive value.

In the next theorem, we show that proposed GHSS method with our splitting is convergent for image restoration problem.

Theorem 1. *Assume that $0 < \beta \leq 1$ and the matrices G, P and S are defined as Eq. (13). Then the iteration (12) unconditionally converges to the unique solution of $Kx = b$.*

Algorithm 1: The GHSS iterative method

1. Choose the initial guess of original image $f_0 = g$, initial value of noise $e_0 = g - Af_0$, maximum number of outer iteration M and very small positive τ ;
2. $r_0 := b - Kx_0$;
3. **for** $k = 0, 1, 2, \ldots,$ *until* $\frac{\|r_k\|_2}{\|r_0\|_2} > \tau$ *or* $k < M$ **do**

$$e_{k+\frac{1}{2}} := \frac{1}{0.5+\alpha}((\alpha+0.5)e_k - Af_k + g),$$

$$f_{k+\frac{1}{2}} := \frac{1}{\alpha+\beta\mu^2}(A^T e_k + (\alpha + (\beta-1)\mu^2)f_k),$$

$$\text{Solve} \begin{cases} (\alpha+0.5)e_{k+1} + Af_{k+1} = (\alpha-0.5)e_{k+\frac{1}{2}} + g, \\ -A^T e_{k+1} + (\alpha + (1-\beta)\mu^2)f_{k+1} \\ \qquad = (\alpha - \beta\mu^2)f_{k+\frac{1}{2}}. \end{cases}$$

$$r_{k+1} := b - Kx_{k+1},$$

end

Proof: For $0 < \mu, \beta \leq 1$, it can be easily seen that G is Hermitian positive definite and P is positive semidefinite matrix. Moreover, S is skew-Hermitian part of splitting. Hence from lemma 1, the proposed method unconditionally converges to the unique solution of $Kx = b$. ■

III. Numerical Examples

In this section, we consider three examples to show the efficiency and accuracy of the proposed method. All examples are implemented in Matlab 8.1 software.

The peak signal-to-noise ratio (PSNR) is defined as follows to compare the original image with the restored one:

$$\text{PSNR} = 10\log_{10}\frac{4 \times 255^2 \times n^4}{\|f - f_{\text{true}}\|_2^2}$$

where f_{true} and f are true and restored images, respectively. Furthermore, the relative error is given by $\|f_{\text{true}} - f\|_2 / \|f_{\text{true}}\|_2$. In all examples, the optimal values of α and β are approximately estimated by some tests on an image in the GHSS method. The optimal values of α in the SHSS method, is obtained by the Eq (10). The restarted GMRES method of MATLAB with $restart = 15$ and $\zeta = 10^{-6}$ has been used to solve the proposed system in Algorithm 1. Furthermore, the stopping tolerance of the proposed algorithm is $\tau = 10^{-6}$ and the maximum number of outer iteration is $M = 15$.

Example 1. In this example, the 256×256 cameraman grayscale image is used to investigate the proposed methods. The symmetric truncated Gaussian PSF is used to blur the true image. Moreover, 1% Gaussian white noise is added to blurred image to give degraded image. Finally, the observed image domain is shown by white lines. The proposed PSF function is considered as:

$$h_{ij} = \begin{cases} ce^{-0.1(i^2+j^2)}, & \text{if } |i-j| \leq 8, \\ 0, & \text{otherwise}, \end{cases}$$

where c is a normalization parameter. The true image and degraded image are shown in Fig. 1. The PSNR value of degraded image in this example is 26.06.

TABLE I. Values of (α, β) in Example 1

Method/BC	Zero	Periodic	Reflexive
SHSS	$(0.3333, -)$	(0.3333,-)	$(0.3255, -)$
GHSS	$(0.9, 0.95)$	$(2.1, 0.98)$	$(0.06, 0.1)$

TABLE II. PSNR values for various methods in Example 1

Method/BC	Zero	Periodic	Reflexive
SHSS	21.50	24.29	28.06
GHSS	22.55	26.22	28.51

TABLE III. Relative error of the SHSS and GHSS methods for Example 1

Method/BC	Zero	Periodic	Reflexive
SHSS	0.3296	0.2389	0.1547
GHSS	0.2916	0.1911	0.1470

Fig. 1. True (left) and degraded image (right) in Example 1

Fig. 2. Restored image with GHSS method for zero (left) periodic (middle) and reflexive (right) BCs in Example 1

The chosen values of (α, β) for SHSS and GHSS methods in this example are given in Table I. The PSNR of restored images with proposed methods are available in Table II. Moreover, the relative error of the SHHS and the GHSS methods is given in Table III. The restored images by using the GHSS method for zero, periodic and reflexive BCs are shown in Fig. 2. As the results show, the GHSS method is more accurate and effective than SHSS method.

Example 2. In this example, we consider the 128×128 simulated MRI of a human brain which is available in the MATLAB Image Processing Toolbox. To blur the image, the introduced out-of-focus PSF function in [5] is used with $dim = 9$ and $R = 4$. The degraded image is obtained by adding 2% Gaussian noise to blurred image.

The true and degraded images are shown in Fig. 3. In this example, the PSNR value of the degraded image is 32.39. The chosen values of (α, β) are given in Table IV. The PSNR values of restored images are presented in Table V. Moreover, the relative error of the SHHS and the GHSS methods is given in Table VI. As the numerical results show, the GHSS method is more accurate than SHSS method to restore images. For more investigation, the restored images with GHSS method for zero, periodic and reflexive BCs are shown in the Fig. 4.

TABLE IV. VALUES OF (α, β) IN EXAMPLE 2

Method/BC	Zero	Periodic	Reflexive
SHSS	$(0.3377, -)$	$(0.3333, -)$	$(0.3255, -)$
GHSS	$(0.11, 0.18)$	$(0.13, 0.25)$	$(0.21, 0.3)$

TABLE V. PSNR VALUES FOR VARIOUS METHODS IN EXAMPLE 2

Method/BC	Zero	Periodic	Reflexive
SHSS	34.99	35.14	35.13
GHSS	35.76	35.92	35.68

TABLE VI. RELATIVE ERROR OF THE SHSS AND GHSS METHODS FOR EXAMPLE 2

Method/BC	Zero	Periodic	Reflexive
SHSS	0.2340	0.2302	0.2304
GHSS	0.2147	0.2103	0.2162

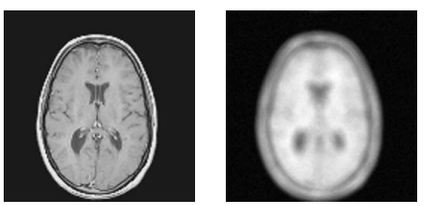


Fig. 3. True (left) and degraded image (right) in Example 2

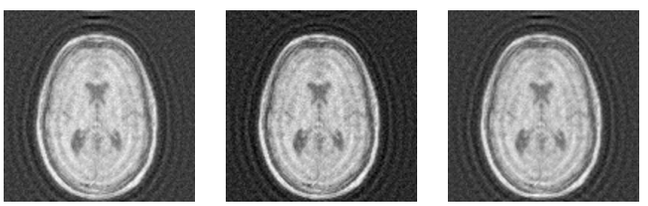


Fig. 4. Restored image with GHSS method for zero (left) periodic (middle) and reflexive (right) BCs in Example 2

Example 3. In this example, we consider an astronomical image from a ground-based telescope. The proposed image is blurred by using Keck telescope PSF. Degraded image is obtained by adding 6% Gaussian noise to the blurred image. The true image and degraded image are given in Fig. 5. Now, we consider the SHHS and GHSS methods and compare them for periodic BCs. The number of outer iteration in both methods is supposed as $M = 25$. In Fig. 6, the restored images with SHSS and GHSS methods are shown for $\alpha = 0.3383$ and $(\alpha, \beta) = (0.1, 0.3)$, respectively.

The residual error in k-th outer iteration of Algorithm 1 is computed by $e_k = g - Af_k$. In Fig. 7, $\|e_k\|_2/\|e_0\|_2$ versus the outer production number k is plotted to consider the convergence speed of proposed methods for periodic BCs. As we can see, the convergence speed of the GHSS method is more faster than the SHSS method.

IV. CONCLUSION

In this paper, image restoration problem was reduced to solve a non-Hermitian positive definite linear system. A new splitting is presented for the coefficient matrix to use in GHSS method. The numerical results were also given. As the numerical results show, our method is accurate and effective in image restoration problem.

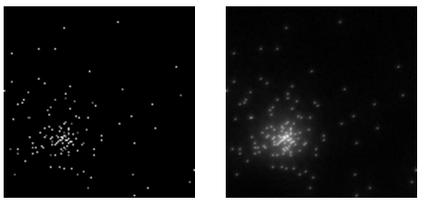


Fig. 5. True image (left) and degraded image (right) in Example 3

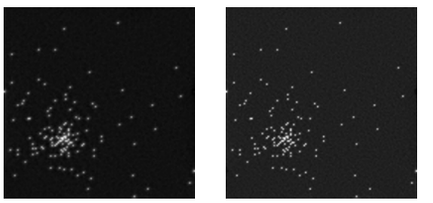


Fig. 6. Restored image with SHHS (left) and GHSS (right) methods in Example 3

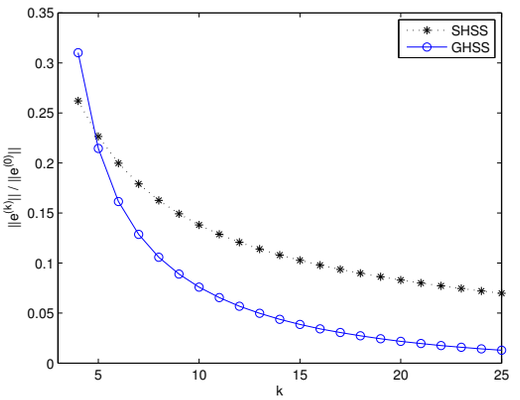


Fig. 7. Comparison between residual errors for periodic BC in Example 3

REFERENCES

[1] Z.Z. Bai, G.H. Golub, M.K. Ng, "Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems," SIAM J. Matrix Anal. Appl. vol. 24, pp. 603–626, 2003.

[2] M. Benzi, "A generalization of the Hermitian and skew-Hermitian splitting iteration," SIAM J. Matrix. Anal. Appl. vol. 31, pp. 360–374, 2009.

[3] T. Chan, A. Marquina, P. Mulet. "High-order total variation-based image restoration," SIAM J. Sci. Comput., vol. 22, pp. 503–516, 2000.

[4] R.C. Gonzalez, R.E. Woods, Digital Image Processing, 2nd edition, Prentice Hall, New Jersey, 2002.

[5] P.C. Hansen, J.G. Nagy, D.P. OLeary, Deblurring Images: Matrices Spectra and Filtering, SIAM, Philadelphia, 2006.

[6] C.W. Helstrom "Image restoration by the method of least squares," J. Opt. Soc. Am. vol. 57, pp. 297–303, 1967.

[7] A.K. Jain, Fundamentals of digital image processing, Prentice Hall, Englewood Cliffs, NJ, 1989.

[8] X.G. Lv, T.Z. Huang, Z.B. Xu, X.L. Zhao, "A special Hermitian and skew-Hermitian splitting method for image restoration," Appl. Math. Model. vol. 37, pp. 1069–1082, 2013.

[9] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd edition, SIAM, Philadelphia, 2002.

Energy Efficient OFDM Transmission Scheme based on Continuous Phase Modulation

Mohammad Irfan†, Soo Young Shin*
Wireless and Emerging Network system Laboratory
Kumoh National Institute of Technology
Gumi, South Korea
Email: isapzai@gmail.com†,
wdragon@kumoh.ac.kr*

***Abstract*—In this paper we propose an energy efficient OFDM transmission scheme. Two CPM (continuous phase modulation) schemes are introduced to OFDM, keeping in view the constant envelope and continuous phase changing nature of CPM modulation. Energy efficiency of the proposed scheme is investigated by simulating peak to average power ratio (PAPR), and is compare with the conventional OFDM scheme. The impact of PAPR on BER is also investigated. Bit error rate (BER) of the proposed scheme is simulated in different channels and is compared with conv-OFDM (QPSK-OFDM) scheme.**

I. Introduction

Orthogonal frequency division multiplexing (OFDM) is the most popular high data rate and spectral efficient multicarrier communication system. In OFDM the subcarriers are orthogonal to each other, so they will not interfere with each other and more number of subcarriers can be packed to the same spectrum to accomodate multiple users [3],[4] . it's high data rate, spectral efficiency and robustness against multipath fading make it one of the most promising candidate for portable computing devices and smartphones. These devices demands high data rate but low power consumption as they are usually battery operated.

Despite many advantages of OFDM, high peak to average power ratio (PAPR) of an OFDM signal is its major limiting factor in practice. In practical OFDM systems a high power amplifier (HPA) is used to obtain sufficient transmit power [1]. In a typical OFDM system the transmit power is only 8% of the total consumed power, 41% of power is wasted in HPA and the rest of the circuit consume about 51% power [5],[1]. There are two reasons of 41% of power loss in HPA, low efficiency of HPA's and high PAPR of OFDM signal which further reduces the efficiency of HPA. HPA has limited linear range, such non-linear effects of HPA's seriously degrades OFDM signal of high PAPR [6]. To solve this problem the linear range of HPA should be enlarged, which means the input back off (IBO) of HPA should be greater than the PAPR of the signal (unless some pre-distortion technique is used) to avoid such signal distortions [7],[1]. But increasing IBO means increasing power consumption of HPA and may decrease its efficiency [5]. Improving efficiency of HPA by reducing PAPR of OFDM signal will not only result in power saving, but the probability of bit error will also reduce as the signal will be less distorted.

The $PAPR_c$ for a continuous time OFDM transmit signal $x(t)$ is given by (1), where N represents number of subcarriers, symbol duration is T, NT represents duration of the signal.

$$PAPR_c = 10\log_{10}\left(\frac{\max\limits_{0\le t\le NT}|x(t)|^2}{\frac{1}{NT}\int_0^{NT}|x(t)|^2dt}\right) \quad (1)$$

Various schemes have been developed to reduce PAPR of OFDM signal. These schemes are summarize in [2]. Method of coding is discussed in [8] and tone reservation in [9]. Predistortion scheme in [10] and clipping in [11]. All these methods can be divide into two categories distortion and distortionless. clipping [11], clips the peak power $\max\limits_{0\le t\le NT}|x(t)|^2$ to some pre-defined value and thus distort the signal which results in high BER at receiver and reduces throughput. The distortion-less methods like tone reservation [9] and selective mapping [12] allows perfect recovery of the original signal at the cost of transmitting some side information which reduces throughput, increase bandwidth requirement and increases signal overhead. [13] shows an approach which transforms the OFDM signal into constant envelope by phase modulating the signal after its inverse discrete Fourier transform (IDFT). Our approach is different from [13]. We replace the conventional modulation schemes used in OFDM such as QPSK, 16-QAM and 64-QAM by Gaussian minimum shift keying (GMSK) and minimum shift keying (MSK) (known as scheme 1 and scheme 2 in this work) Fig. 1.

In our proposed scheme the input data is first modulated by a CPM modulator, the modulated symbols are then transformed from serial to parallel, an IDFT is then applied to the parallel modulated symbols. After IDFT cyclic prefixes are inserted the data is then again changed from serial to parallel followed by amplification by a HPA and is transmitted, at the receiver side a viterbi detector is used to detect the symbols after parallel to serial conversion.

The rest of the paper is organized as follows. Section II presents fundamental of CPM, two CPM modulation schemes that we have select for this work and some previous work is discussed. Section III presents our system model, CPM based OFDM. In section III we design CPM based transmitter, PAPR analysis of CPM based OFDM, receiver structure for CPM based OFDM and the effect PAPR on BER. In section IV we present our numerical results and conclusion is drawn in section V.

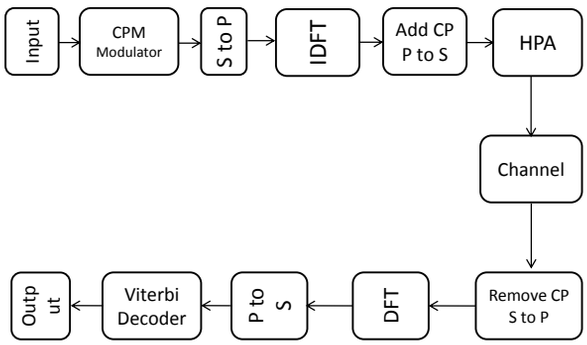


Fig. 1: CPM-OFDM, Block Diagram

II. Fundamentals of CPM

As compared to other modulation schemes like QPSK, which is one of the modulation scheme used in LTE [14]. In such modulation schemes the carrier phase is abruptly reset to zero at the start of every symbol. In QPSK when the message signal changes by one bit a phase shift of 90 degrees occurs. This discontinuity of QPSK modulated signal lead to high power consumption and some percentage of power occurs outside of the intended band, which leads to poor spectral efficiency [7].

While CPM has constant envelope and the phase of the carrier signal is modulated in a continuous manner, which yields high spectral efficiency [7]. Due to its constant envelop the PAPR of such waveform is unity [7]. A CPM waveform is given by (2), [15].

$$S(t;\beta) \triangleq exp\{j\phi(t;\beta)\} \tag{2}$$

ϕ is the phase of the signal and is given by the following equation.

$$\phi(t;\beta) \triangleq 2\pi \sum_{i} \beta_i h_i q(t - iT) \tag{3}$$

$\beta \triangleq \beta_i$ represents the discrete time symbol sequence of $M-ary$ symbols form i_{th} user each symbol carries $m = log_2 M$ bits, T is the symbol duration, h_i is the modulation index. Smaller the modulation index the more narrower bandwidth the signal occupies. M is the value of M-ary signaling. Throughout this work is $M = 4$. Two CPM schemes are selected for this work.

- Scheme 1: $4-ary$ signaling is used, modulation index $h_i = 0.3$, Raised cosine frequency pulse with Pulse length $L = 4$. and minimum square Euclidean distance, $d^2_{min} = 1.48$.
- Scheme 2: $4-ary$ signaling with modulation index $h_i = 0.6$, Gaussian frequency Pulse with $BT = 0.25$, Pulse length, $L = 3$, and minimum squared Euclidean distance, $d^2_{min} = 4.6$

The above two scheme differs in PAPR and BER performance. Scheme 1 has smaller modulation index than scheme 2 so it will occupy a narrow bandwidth as compare to scheme 2, its power spectral density will drop more abruptly which will result in better PAPR performance. But on the other hand scheme 1 has small value of d^2_{min}, so Scheme 2 will have better BER performance compare to Scheme 1. Increasing d^2_{min}, lowers the probability of bit error [15].

III. CPM Based OFDM

A. Previous work

This work is loosely based on the observations presented in [7][16]. In [7] the authors have proposed CPM for single carrier frequency division multiple access (SC-FDMA), the uplink of LTE, with interleaved subcarrier mapping. In [16] the authors have phase modulate an OFDM signal by a CPM modulator after assigning sub-carriers. The purpose of this work is to extend CPM to OFDM. This work is different from the previous as we are using two different schemes of CPM as baseband modulation schemes , and do performance evaluation of PAPR, the effect of PAPR on probability of bit error in the absence of amplifier input back-off, and evaluate BER in different channels and compare them with conventional OFDM scheme.

B. Transmitter

Consider an OFDM system as shown in Fig 1. N number of symbols and sub-carriers. Each symbol is mapped to one sub-carrier. In our proposed scheme, the input data is first modulated by a CPM modulator. The modulated samples are then converted from serial to parallel by a serial to parallel converter as shown in Fig 1.

$$S_i = [s_{i,0}, s_{i,1}, ...s_{i,N-1}]^T$$

where i is the index of a user.

$$S_{i,l} = exp\{j\phi(i;l)\} \tag{4}$$

This is discrete time equivalent of (2). The continuous time transmitted OFDM signal $x(t)$ is expressed as

$$x(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi k \Delta f t}, 0 \leqslant t \leqslant T, \tag{5}$$

where T is the duration of the OFDM symbol, $X_K = [X_0, X_1, ..., X_{N-1}]$ is the input data block, $\Delta f = \frac{1}{T}$ is the frequency spacing between two adjacent subcarriers. The same way a discrete time CPM-OFDM signal $x(n)$ with Nyquist rate is obtained by N-point IDFT operation of $S_{(i,l)}$.

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} S_{(i,l)} e^{j2\frac{\pi kn}{N}}, n = 0, 1, ...N-1 \tag{6}$$

The equation above shows the waveform of a CPM based OFDM. After insertion of cyclic prefix and digital to analog conversion the signal is amplified by a high power amplifier (HPA) and is transmitted.

C. PAPR Analysis

PAPR stands for Peak to Average Power Ratio. It is basically used as performance metric for envelop variations in a signal. As OFDM is a multi carrier system, A set of orthogonal frequencies are used. Mathematically these frequencies will not interfere with each other but in practice they interfere with each other. In this paper N number of sub carriers are used. Assuming amplitude of each sub-carrier is 1, as the constellation points are taken from MSK and GMSK modulators. These sub carriers can constructively add up resulting in high peak powers. So the range of PAPR is 1 to N.
In case of non unity amplitude A like the case of QAM, the max PAPR is given by equation below [17].

$$PAPR_{max} = A_{max}^2 N \tag{7}$$

Another reason of high PAPR of an OFDM signal is the spectral leakage of DFT. the sampling period

$$\Delta t = \frac{1}{f_s} \tag{8}$$

The nth array element in frequency domain is known as bucket or bin. The frequency of bucket n can be related to the frequency of input signal $S_{i,l}$ by equation below [18].

$$f_n = \frac{n}{N\Delta t} \tag{9}$$

where n is the number of bin, N is the number of samples and f_n is the frequency of bin n. Re arranging (8) and (9) we get

$$n = \frac{Nf_n}{f_s} \tag{10}$$

N is the total number of sub-carriers which is equal to 16 for this work, f_s sampling frequency and f_n is the frequency of nth bin. for 900Hz signal the maximum amplitude occurs at $n = Nf_n/f_s = 16 * 900/3000 = 4.8$. As each bin represents a frequency not a range of frequencies and bin number is an integer, then the question is how can DFT represent energy at bin number 4.8, actually it shows some leakage of energy between bin 4 and bin 5. In DFT for some frequencies there is some energy leakage between bins, while in other frequencies there is no energy leakage. for example for a 1500 Hz waveform, $n = Nf_n/f_s = 16*1500/3000 = 8$, there is no spectral leakage between bin 8 and bin 7 or bin 8 and bin 9. In OFDM multi-carriers are generated by IDFT and due to the spectral leakage problem of DFT and interference of the multi carriers OFDM exibits high PAPR.
As the input $S_{i,l}$ is random so PAPR of $x(n)$ given by equation below is also a random variable.

$$PAPR = 10\log_{10}\frac{\max\limits_{0\leq n\leq \tau\cdot N-1}|x(n)|^2}{E\left[|x(n)|^2\right]} \tag{11}$$

Therefore cumulative distribution function (CDF) is used to describe its properties.

$$\begin{aligned} CDF &= Prob(PAPR \leqslant \gamma) \\ Prob(PAPR \leqslant \gamma) &\approx 1-(1-e^{\gamma})^N \end{aligned} \tag{12}$$

where γ is a constant. $PAPR$ of a discrete time signal $x(n)$ is an accurate approximation of the continuous time $PAPR_c$, if the up-sampling factor $\tau \geq 4$, [7]. For this work we are using up-sampling factor equal 4.

D. Receiver

The received signal y is first sampled to generate the discrete time signal. After removing the cyclic prefix the received signal r can be expressed as below

$$y = \sum_{k=0}^{N-1} h \times x(n) + n_0 \tag{13}$$

Where h is the channel matrix, $x(n)$ is discrete version of transmitted signal and n_o is the additive white Gaussian noise with zero mean. A discrete time Fourier transform of the received signal y is taken

$$y(n) = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} ye^{-j2\pi kn/N} \tag{14}$$

and after parallel to serial conversion Viterbi decoder is used to decode the signal.

E. Effect of PAPR on BER Performance

In this section we determine the relation of PAPR and BER performance of an OFDM signal in the absence of IBO from HPA. A HPA exibits two kinds of distortions, AM-AM and AM-PM distortion.

When the HPA is operating in the linear region, the output voltage is only a scaled version of the input voltage. However when the input signal has high peaks in the compression region, the output of the HPA is amplitude distorted. The resultant distortions are known as AM-AM distortions. Similarly AM-PM is used for phase non linearity. In such kind of distortion the phase of the signal is distorted. But here we will only consider the effect of AM-AM distortion.

The efficiency η of HPA depends on the input voltage or input power. An amplifier operates more efficiently if the input is within the range of operating point. CPM has constant envelop, but due to the PAPR problem of multi-carrier systems there is a large envelope fluctuation. So the efficiency η depends on PAPR and class of the amplifier [17].

$$\eta = G \cdot e^{-g\cdot PAPR_{dB}} \tag{15}$$

Where G is the gain of an amplifier. As the probability of bit error depends on SNR of the signal.

$$SNR = \frac{\eta\cdot[x(n)]^2\cdot T}{{n_o}^2} \tag{16}$$

where n_o^2 represents noise power and $[x(n)]^2$ is the transmitted signal power SNR depends on the efficiency η of HPA. And the efficiency of HPA depends on $PAPR$ of the signal. So the probability of bit error depends on PAPR of the signal [17].

$$SNR = \frac{G}{\zeta}\cdot\frac{[x(n)]^2\cdot T}{{n_o}^2} \tag{17}$$

Where ζ is equal to.

$$\zeta = \begin{matrix} 1 & PAPR \leq SatLevel \\ PAPR^{(10g/\ln 10)} & PAPR > SatLevel \end{matrix} \tag{18}$$

(17) shows that low PAPR results in better HPA efficiency and high SNR which results in low energy consumption by HPA and low error rates.

IV. Numerical Results

A. PAPR Numerical Results

In this section we first numerically analyze PAPR and BER properties of our proposed two schemes of CPM-OFDM and then compare them to conventional OFDM scheme. Relation of PAPR to BER in the absence of IBO is also numerically analyze in this section. All the simulations performed in this section consists of (N=16) N = number of allocated sub carrier to one user. number of users = 1;

Fig 2. shows numerically calculated CDF of PAPR of our two proposed schemes and conventional OFDM. Due to low modulation index of scheme 1 its power spectral density will drop abruptly as compare to scheme 2 of high modulation index and conventional OFDM. Therefore scheme 1 has low PAPR as compare to scheme 2 and conventional OFDM. The previous discussion in section II, PAPR analysis shows that CPM-OFDM will have constant envelop but as Fig 2 shows CPM-OFDM still has envelope variations. The reasons are already discussed in the section II. Spectral leakage and sub-carriers interference which results in high peaks. PAPR statistics of these three schemes are provided in Table I. Proposed scheme 1 has almost 4 dB performance gain over conv-OFDM and almost 0.5 dB performance gain over scheme 2 in terms of their max PAPR.

In order to operate the HPA in linear region so that it will not distort the OFDM signal, the input back off must be higher than the PAPR of the signal. Input back-off is the amount of power drawn from direct current (dc) source of HPA in-order to keep the HPA in linear region. Table II shows required amount of minimum input back-off calculated from CDF plot of Fig 2. Table 2 shows that scheme 1 will draw little amount of power from the dc source which happens to be the most power efficient scheme of all the three schemes.

B. BER vs PAPR

Numerical results of PAPR and BER relation are presented in this section. Analytically it is already developed in the previous section (18). Fig 3 shows AM-AM distortions of a HPA as discussed in the previous section. Three different HPA with input saturation level of 2, 4 and 6 dB are considered here. As shown in Fig 3. the output of amplifier is saturated and distorted once the input cross it saturation limit. for such kind of HPA with no IBO, BER is simulated and is shown in Fig 4. The fig shows that BER is at low when the input signal PAPR is with in HPA input saturation level. As the input signal PAPR crosses the input saturation of HPA the probability of bit error increases abruptly. The probability of bit error increases exponentially with increasing PAPR.

TABLE I: PAPR statistics

Scheme	Max[dB]	Min[dB]	Mean[dB]
Scheme 1	6.23	1.95	4.82
Scheme 2	6.68	2.24	5.36
Conv-OFDM	10.16	3.32	6.14

TABLE II: Required IBO from CDF plot

Scheme	IB-90%[dB]	IB-99%[dB]
Scheme 1	5.5	6
Scheme 2	6.3	6.5
Conv-OFDM	7.6	9.4

C. Bit Error Rate Performance

Fig. 5, Fig. 6, and Fig.7 shows BER performance of Scheme 1, Scheme 2 and conv-OFDM in three different kinds of channels AWGN, Ped-A and Veh-A channel. Some parameters of the said channels are listed in Table III. The effect of HPA distortions to OFDM signal are also added in the form of IB-90% from Table II. The actual BER performances are plotted by solid lines for different channels. The BER performance with HPA distortions are plotted by dotted lines as a function of $E_b/N_o + IB_{90\%}$. As shown without taking the effect of HPA non-linearity, scheme 2 has better BER performance as compare to scheme 1 and Conv-OFDM. The reason is the large minimum square Euclidean distance $d^2_{min} = 4.6$, of scheme 2 as compare to $d^2_{min} = 1.48$ of scheme 1. At BER of 10^{-4} in AWGN channel scheme 2 has almost 1.5 dB gain over conv-OFDM and almost 2 dB gain over scheme 1. The same way scheme 2 out performs scheme 1 and conv-OFDM in Ped-A and Veh-A channel.

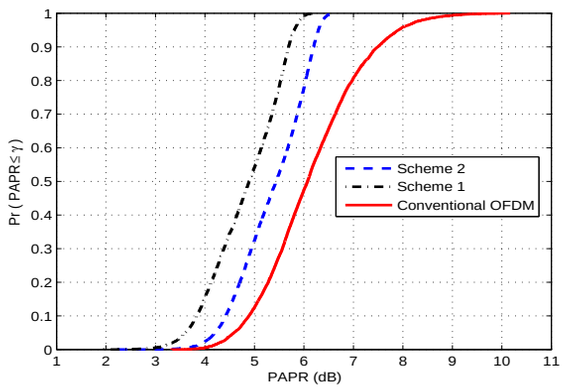


Fig. 2: PAPR,CDF plots of scheme 1, scheme 2, and conv-OFDM

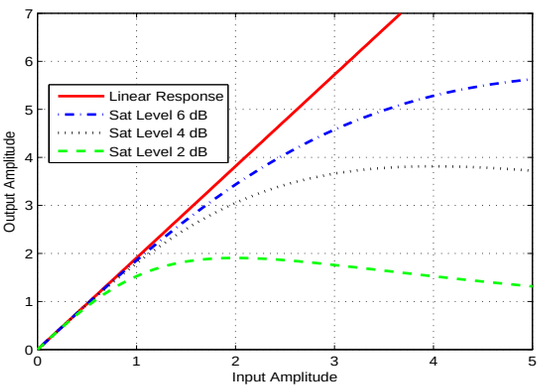


Fig. 3: AM / AM, Response of HPA for different input saturation levels

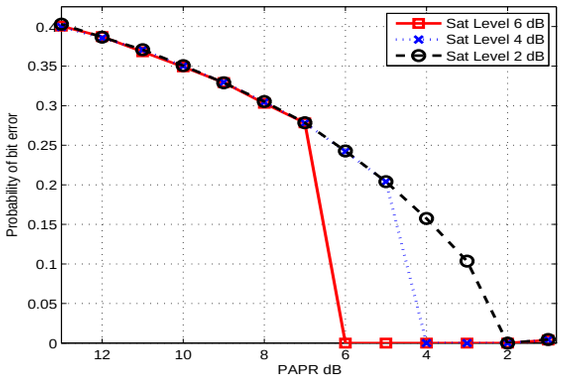


Fig. 4: Effect of PAPR on BER, under constant SNR of 10dB

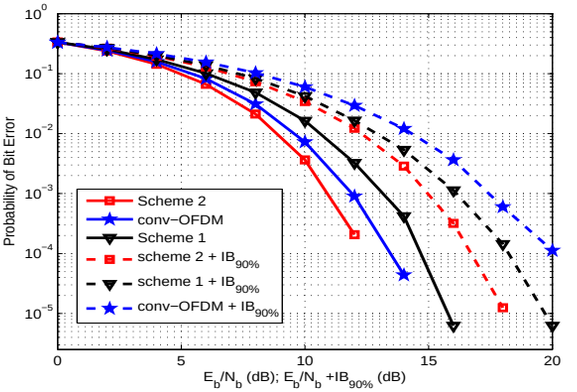


Fig. 6: BER plots of Scheme 1, Scheme 2 and conv-OFDM in Ped-A channel

When the HPA non-linearity is taken into account in the form of $E_b/N_o + IB_{90\%}$ we notice that in all channels scheme 1 and scheme 2 has almost the same BER performance although the previous discussion shows that scheme 2 has better BER performance. Scheme 2 has better BER performance than scheme 1 but it has high PAPR as well so in presence of HPA non-linearity both the scheme has almost the same BER performance. The conv-OFDM shows the worse performance of all due to its high PAPR as compare to scheme 1 and scheme 2.

V. CONCLUSION

In this work, we have compared PAPR properties of our proposed two schemes with conv-OFDM, and have shown that how the proposed schemes will have better PAPR performance as compare to conventional scheme. High PAPR results in worse BER performance in the absence of HPA Input Back-off, and high power consumption and low efficiency in the presence of input Back-off . BER performance is also evaluated in the presence and absence of HPA non-linearity. In the presence of HPA non-linearity which is more realistic approach our proposed scheme outperforms the conv-OFDM (QPSK-OFDM).

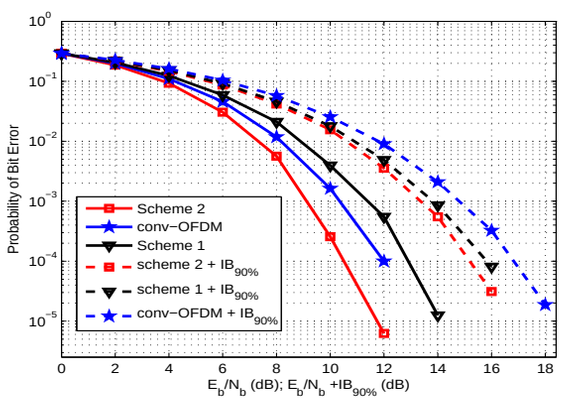


Fig. 5: BER plots of Scheme 1, Scheme 2 and conv-OFDM in AWGN channel

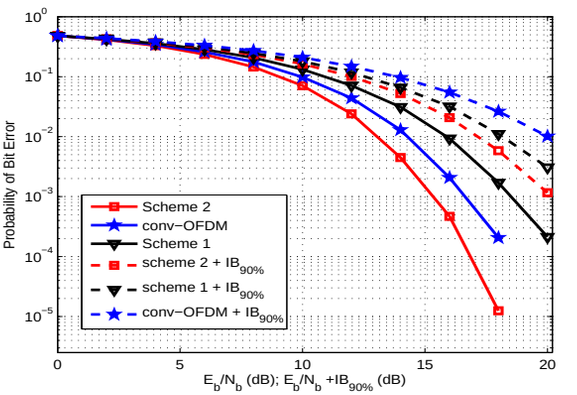


Fig. 7: BER plots of Scheme 1, Scheme 2 and conv-OFDM in Veh-A channel

ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (2012R1A1A1009442).

TABLE III: Parameters of Ped-A and Veh-A channel

Ped-A		Veh-A	
Tap delay (ns)	Relative power (dB)	Tap delay (ns)	Relative Power (dB)
0	0	0	0
30	-1.0	310	-1
70	-2.0	710	-9.0
110	-9.7	1090	-10.0
190	-19.2	1730	-15.0
410	-22.8	2510	-20.0

REFERENCES

[1] Jiang, Tao, Cai Li, and Chunxing Ni. "Effect of PAPR reduction on spectrum and energy efficiencies in OFDM systems with Class-A HPA over AWGN channel." Broadcasting, IEEE Transactions on 59.3 (2013): 513-519.

[2] Han, Seung Hee, and Jae Hong Lee. "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission." Wireless Communications, IEEE 12.2 (2005): 56-65.

[3] Peled, Abraham, and Antonio Ruiz. "Frequency domain data transmission using reduced computational complexity algorithms." Acous-

tics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80.. Vol. 5. IEEE, 1980.

[4] Saltzberg, B. "Performance of an efficient parallel data transmission system." Communication Technology, IEEE Transactions on 15.6 (1967): 805-811.

[5] Jiang, Tao, et al. "Multicast broadcast services support in OFDMA-Based WiMAX systems [Advances in Mobile Multimedia]." Communications Magazine, IEEE 45.8 (2007): 78-86.

[6] Krongold, Brian S., and Douglas L. Jones. "An active-set approach for OFDM PAR reduction via tone reservation." Signal Processing, IEEE Transactions on 52.2 (2004): 495-509.

[7] Wylie-Green, Marilynn P., Erik Perrins, and Tommy Svensson. "Introduction to CPM-SC-FDMA: a novel multiple-access power-efficient transmission scheme." Communications, IEEE Transactions on 59.7 (2011): 1904-1915.

[8] Davis, James A., and Jonathan Jedwab. "Peak-to-mean power control in OFDM, Golay complementary sequences, and Reed-Muller codes." Information Theory, IEEE Transactions on 45.7 (1999): 2397-2417.

[9] Krongold, Brian S., and Douglas L. Jones. "An active-set approach for OFDM PAR reduction via tone reservation." Signal Processing, IEEE Transactions on 52.2 (2004): 495-509.

[10] D'Andrea, Aldo N., Vincenzo Lottici, and Ruggero Reggiannini. "Nonlinear predistortion of OFDM signals over frequency-selective fading channels." Communications, IEEE Transactions on 49.5 (2001): 837-843.

[11] Armstrong, Jean. "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering." Electronics letters 38.5 (2002): 246-247.

[12] Lee, Yung-Lyul, et al. "Peak-to-average power ratio in MIMO-OFDM systems using selective mapping." Communications Letters, IEEE 7.12 (2003): 575-577.

[13] Tan, Jun, and Gordon L. Stuber. "Frequency-domain equalization for continuous phase modulation." Wireless Communications, IEEE Transactions on 4.5 (2005): 2479-2490.

[14] Yu, Chen, et al. "Research on the modulation and coding scheme in LTE TDD wireless network." Industrial Mechatronics and Automation, 2009. ICIMA 2009. International Conference on. IEEE, 2009.

[15] Anderson, John B., Tor Aulin, and Carl-Erik Sundberg. Digital phase modulation. Springer, 1986.

[16] Thompson, Steve C., et al. "Constant envelope OFDM." Communications, IEEE Transactions on 56.8 (2008): 1300-1312.

[17] Wulich, Dov. "Definition of efficient PAPR in OFDM." Communications Letters, IEEE 9.9 (2005): 832-834.

[18] Lyon, Douglas A. "The discrete fourier transform, part 4: spectral leakage." Journal of object technology 8.7 (2009).

Adaptive Double-Threshold Based Energy and Matched Filter Detector in Cognitive Radio Networks

Ashish Rauniyar*, Soo Young Shin†
Wireless and Emerging Networking System (WENS) Lab.
School of Electronic Engineering
Kumoh National Institute of Technology, Gumi-si, South-Korea 730-701
(*ashish.rauniyar,†wdragon)@kumoh.ac.kr

***Abstract*—In this paper we propose a new cooperative spectrum sensing method using adaptive double threshold energy and matched filter detector in cognitive radio networks. The energy detector (ED) performance is highly degraded under noise uncertainty condition. Also, ED cannot well differentiate between the signal and noise, if the detected observational value of the primary user (PU) lies in the confused region i.e., between signal and noise. We propose a scheme based on adaptive double threshold that uses matched filter (MF) detector for the reliable detection in the confused region and energy efficient ED for the clear region where the detector can easily differentiate between signal and noise and makes its own local decision. The fusion center collects the local decisions and observational values of the secondary users and then makes the final decision to ascertain whether the primary user is present or not. Simulation results shows that our proposed method has higher detection performance as compared to other spectrum sensing methods.**

I. Introduction

In a survey conducted by the Federal Communications Commission (FCC) on spectrum utilization has indicated that the actual licensed spectrum is largely under-utilized in vast geographical dimensions [1]. Cognitive Radio (CR) provides opportunistic access to unused licensed bands [2][3]. CR allows secondary users (SU) to utilize the free portions of licensed spectrum while ensuring no interference to primary users (PU) transmissions. In the recent years, cooperative spectrum sensing scheme(CSS) has become a popular technique to solve the efficiency of spectrum usage and provide high level of protection to the PU from SU. In CR, sensing accuracy is important for avoiding interference to the primary users in CR technology. Reliable spectrum sensing is not always guaranteed due to multipath fading, shadowing and hidden terminal problem. Cooperative spectrum sensing has thus been introduced for quick and reliable detection [4][5][6][7]. The CSS has two successive stages, sensing and reporting. In sensing stage, spectrum sensing is done by several local SU. Then in next stage, PU sensing decisions or measurements are sent to fusion center (FC) to combine them and make a better overall decision.

Among several spectrum sensing techniques, energy detector (ED) is the most popular method employed for spectrum sensing. Measuring only the received signal power and comparing it with a pre-fixed threshold, the ED is a non-coherent detection device with low implementation complexity and is more power efficient. But ED performance is highly degraded under noise uncertainty condition [8]. Also, ED cannot well differentiate between the signal and noise, if the detected observational values lies in the confused region i.e., between signal and noise.

In [9], a censoring method using double threshold based on ED was proposed. If the detected observational energy values (O_i) by the SU lies in the confused region, they will not report to the fusion center. This method can reduce the sensing time and cause sensing failure problem. Paper [10] also proposed a method using double threshold based on ED to increase the detection performance as compared to the conventional ED. In this method, first SU will make the local decision by comparing their O_i of the clear region with the pre-defined threshold of ED. If the O_i lies in the confused region then the SU will forward it to the fusion center (FC). The FC will make overall decision by considering the local decision of SU of clear region and comparing the O_i of confused region with another threshold value of ED.

To overcome the noise uncertainty problem of ED and to increase the detection performance, we propose a new CSS technique using adaptive double threshold based on ED and matched filter (MF) detector. Our proposed scheme has higher detection performance as compared to other conventional methods and the method described in [10]. We take the advantages of energy efficient ED to make the local decision in the clear region and reliable MF to take the decision in the confused region. Simulation results shows that our proposed scheme has higher detection performance, lower miss-detection probability and it can perform well in low SNR as compared to other methods.

The rest of the paper is organized as follows: Section II presents the system description. Section III describes our proposed model. Simulation results are shown in Section IV. Finally conclusion is drawn in Section V.

II. System Description

The main aim of CR is to correctly identify the presence of PU and allows the SUs to utilize the unused spectrum if it is not used by licensed PUs. Under binary hypothesis testing, we consider the occurrence of two input events in observing

signal x_i in some observation interval denoted by

$$\begin{aligned} H_0 &: x_i = n_i \\ H_1 &: x_i = s_i + n_i, \end{aligned} \tag{1}$$

where $i = 1, 2, 3, ...N$ is number of samples. H_0 represents the hypothesis that the observation vector consists of noise. H_1 represents the hypothesis that the observation vector consists of noise and signal. The noise component n_i is assumed to be Additive White Gaussian random variable which is independent and identically-distributed (i.i.d) with zero mean normal distribution with variance $\sigma^2 \sim \mathcal{N}(0, \sigma^2)$, and s_i is the signal.

A. Energy Detector

The ED is non-coherent detector and consumes less amount power. ED detects the presence of signals by simply squaring its energy and comparing that energy around the carrier frequency with certain threshold [11]. The ED is not so accurate as the detected signal can be affected by noise level. The performance of ED is highly degraded under noise uncertainty condition.

The ED consists of a quadrature receiver with y_I and y_Q representing samples from In-phase and Quadrature branch respectively. The samples after passing the squaring device, output of the integrator is denoted by

$$y_I = y_Q = (\frac{1}{N_0}) \int_0^T r^2(t)dt, \tag{2}$$

where $r(t)$ is input signal, N_0 is noise spectral density.

Within observed sensing period, test statistic ED can be approximated as $Y_{ED} = y_I + y_Q$. At the observation time t, decision variable Y_{ED} will be compared to a detection threshold of ED denoted by λ^{ED}. Threshold value is set to meet the target probability of false alarm p_f according to the noise power. The probability of detection p_d can be also identified. The expression for p_f and p_d can be given as [12]

$$p_{fa}^{ED} = 1 - F_\chi \left(\frac{\lambda^{ED}}{\sigma^2}, 2n \right), \tag{3}$$

where F_χ is cumulative distribution function (CDF) of standard chi-square random variable with k degree of freedom.

$$p_d^{ED} = \mathcal{Q} \left(\sqrt{n(SNR)}, \sqrt{\frac{\lambda^{ED}}{\sigma^2}} \right), \tag{4}$$

where $\mathcal{Q}$ is generalized Marcum-Q function.

B. Matched Filter

MF is a reliable detector but consumes high amount of power. MF works using receivers bank of L matched filters, which runs together to correlate the incoming signals [13]. At each sampling instant t, de-correlators process signal $x(t)$, the output on interval $(0, T)$ that contains two sample output from a module is given by

$$Y_{MF} = y_{I_i}^2 + y_{Q_i}^2, \quad i = 1, 2, ..., L \tag{5}$$

The Y_{MF} forms L de-correlators output in which we find the decision variable V from the maximum of Y_{MF} over M offset

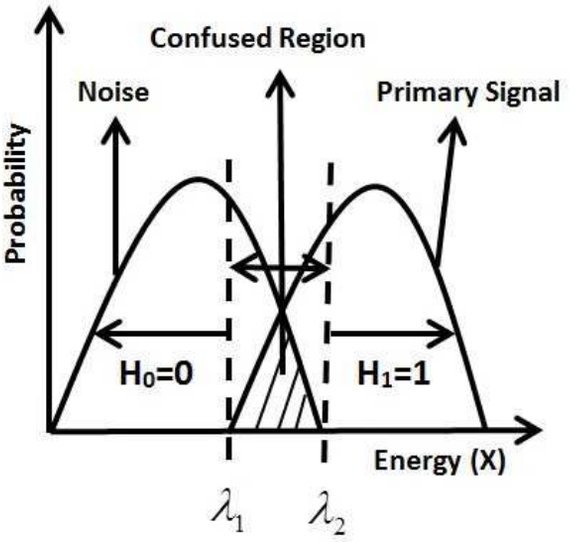


Fig. 1: Energy distribution of primary user signal and noise

bits. Variable V is compared to threshold λ^{MF} to decide the presence or absence of signal.

$$V = \max\{Y_{MF}^m\}, \quad m = 1...M \tag{6}$$

The acquisition process of MF will give probability of false alarm and probability of detection that can be calculated as [12]

$$p_{fa}^{MF} = 1 - F_\chi \left(\frac{\lambda^{MF}}{\sigma}, 2 \right), \tag{7}$$

$$p_d^{MF} = \mathcal{Q} \left(\sqrt{2n(SNR)}, \sqrt{\frac{\lambda^{MF}}{\sigma^2}} \right), \tag{8}$$

where λ_{MF} is the threshold setting for MF, the non-centrality parameter $s^2 = 2n(SNR)$ is the output of the filters in I and Q branches at the correct offset. The correlation process of MF has a central chi-square distribution with 2 degree of freedom with a variance $(\sigma = \sqrt{n})$.

III. Adaptive Double-Threshold Based Energy and Matched Filter Detector

In conventional ED, each SU makes their own local decision whether PU is present (H_1) or PU is absent (H_O) by comparing the O_i with single predefined threshold. Fig. 1 shows the energy distribution of the primary user signal and noise. [10] proposed a method using double threshold based on ED. In this method, cooperating SU will make the local decision by comparing their O_i of the clear region with the predefined threshold of ED i.e., λ_1 and λ_2. As shown in Fig. 1, if the O_i crosses the λ_2 then the local decision taken will be H_1 i.e., PU is present. Similarly, if the O_i is less than λ_1 then the local decision taken will be H_0 i.e., PU is absent. But if the O_i lies in the confused region i.e., between λ_1 and λ_2 then the SU will forward it to the FC along with the local decisions. In FC, the O_i of the confused region will be compared with another threshold value of ED λ and the overall decision will be taken by FC considering all the decisions.

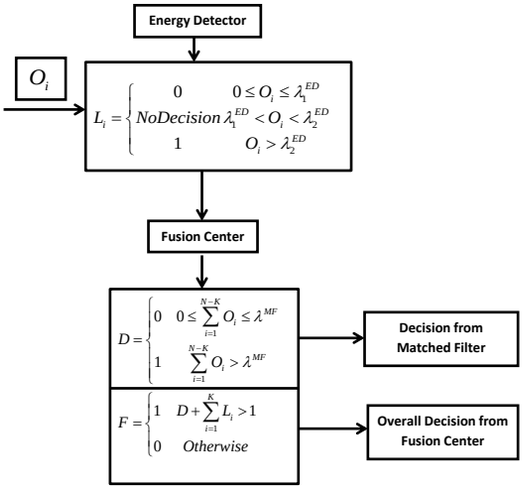


Fig. 2: Working model of proposed adaptive double-threshold based on energy and matched filter detector

The main idea of our proposed scheme is that we take the advantage of energy efficient ED to make decision in the clear region and reliable MF to make decision in the confused region. Our proposed scheme is based on adaptive double threshold of both ED and MF detector.

Fig. 2 shows the working model of our proposed method. Each SU of the CSS are equipped with ED and performs spectrum sensing individually. The observational value O_i of SU is checked with the threshold values λ_1^{ED} and λ_2^{ED} of ED and the decision will be taken accordingly. If O_i satisfies $\lambda_1^{ED} \le O_i < \lambda_2^{ED}$ then no decision will be taken and further the SU will forward its observational value O_i to the fusion center. The local decision L_i is given by

$$L_i = \begin{cases} 0 & 0 \le O_i \le \lambda_1^{ED} \\ \text{No Decision} & \lambda_1^{ED} \le O_i \le \lambda_2^{ED} \\ 1 & O_i \ge \lambda_2^{ED} \end{cases} \tag{9}$$

Without the loss of generality, we assume that fusion center FC receives K local decisions out of N SUs. Then $N-K$ observational values O_i will be reported to the FC to make the decision. The fusion center will now apply more reliable MF on those $N-K$ observational values of the signal for the decision process as $N-K$ secondary users could not distinguish between the presence or absence of the primary users. The threshold of MF λ^{MF} at the FC is chosen according to the appropriate false alarm probability of MF as given by Eq. 7. The decision D at the FC using the MF detector on $N-K$ O_i is as follows

$$D = \begin{cases} 0 & 0 \le \sum_{i=1}^{N-K} O_i \le \lambda^{MF} \\ 1 & \sum_{i=1}^{N-K} O_i > \lambda^{MF} \end{cases} \tag{10}$$

The FC has the local decision L_i of K SUs using ED and decision D of $N-K$ SUs using MF. Let us denote the total decision at FC by Z, i.e., Z= D + $\sum_{i=1}^{K} L_i$. The FC makes a final decision using a hard decision OR rule for deciding the presence or absence of PU. As per the hard decision OR rule, if total decision Z is greater or equal to 1 then signal is detected (H_1) and if Z is smaller than 1 then signal is not detected (H_0). The mathematical expression of hypothesis at the FC can be written as

$$FC = \begin{cases} \text{Z} < 1, & H_0 \\ \text{Z} \ge 1, & H_1 \end{cases} \tag{11}$$

1) Cooperative Detection and False Alarm Probabilities of Proposed Method: First each secondary user decides either '0' or '1' or "No Decision" on the basis of comparison of O_i with pre-defined threshold value of energy detector. Decision goes in favor of '0' if PU is absent. Similarly decision goes in favor of '1' if the PU is present. let us denote the probability of deciding '1', probability of "No Decision" and probability of deciding '0' under hypothesis H_1 is represented by p_{d1}^{ED}, $\triangle_{1,i}^{ED}$ and p_m^{ED} respectively. Similarly, Probability of deciding '1', probability of "No Decision" and probability of deciding '0' under hypothesis H_0 is denoted by p_{fa0}^{ED}, $\triangle_{0,i}^{ED}$ and p_{d0}^{ED} respectively. The expressions for different probabilities are given below considering the AWGN channel [9].

$$p_{d1}^{ED} = P\{O_i > \lambda_2^{ED} | H_1\} = \mathcal{Q}\left(\sqrt{n(SNR)}, \sqrt{\lambda_2^{ED}}\right), \tag{12}$$

$$p_{d0}^{ED} = P\{O_i < \lambda_1^{ED} | H_0\} = F_\chi\left(\lambda_1^{ED}, 2\right), \tag{13}$$

$$\triangle_{1,i}^{ED} = P\{\lambda_1^{ED} < O_i < \lambda_2^{ED} | H_1\} \tag{14}$$

$$\triangle_{0,i}^{ED} = P\{\lambda_1^{ED} < O_i < \lambda_2^{ED} | H_0\} \tag{15}$$

$$p_m^{ED} = P\{O_i \le \lambda_1^{ED} | H_1\} = 1 - \triangle_{1,i}^{ED} - p_{d1}^{ED}, \tag{16}$$

$$p_{fa0}^{ED} = P\{O_i > \lambda_2^{ED} | H_0\} = 1 - F_\chi\left(\lambda_2^{ED}, 2\right), \tag{17}$$

The cooperative probability of detection Q_d of the FC using OR rule as indicated in Eq. 11 can be expressed as

$$\begin{aligned} Q_d &= P\{Z \ge 1 | H_1\} \\ Q_d &= P\{(\sum_{i=1}^{K} L_i + \sum_{i=1}^{N-K} D) \ge 1 | H_1\} \\ Q_d &= 1 - \sum_{K=0}^{N-1} \binom{N}{K} \prod_{i=1}^{K} p_m^{ED}. \\ &\quad \prod_{i=K+1}^{N} \triangle_{1,i}^{ED} [1 - \mathcal{Q}_{(N-K)u}\left(\sqrt{2n(SNR)}, \sqrt{\lambda^{MF}}\right)] \\ &\quad + \prod_{i=1}^{N} p_m^{ED} \end{aligned} \tag{18}$$

where u is the time bandwidth product. The cooperative probability of miss-detection Q_m of the FC is given by

$$Q_m = 1 - Q_d \tag{19}$$

The cooperative probability of false alarm Q_f of the FC using OR rule as indicated in Eq. 11 can be expressed as

$$\begin{aligned} Q_f &= P\{Z \geq 1|H_0\} \\ Q_f &= P\{(\sum_{i=1}^{K} L_i + \sum_{i=1}^{N-K} D) \geq 1|H_0\} \\ Q_f &= -\sum_{K=0}^{N-1} \binom{N}{K} \prod_{i=1}^{K} (1 - \triangle_{0,i}^{ED} - p_{fa0}^{ED}). \\ &\quad \prod_{i=K+1}^{N} \triangle_{0,i}^{ED} [1 - F_{\chi(N-K)u}\left(\lambda^{MF}, 2\right)] \end{aligned} \tag{20}$$

Algorithm 1 Adaptive Double-Threshold Based Energy and Matched Filter Detector Method

1: Given $\{x_1, x_2, ...x_N\}$
2: Given $\{SU_1, SU_2, ...SU_N\}$
3: Define the value for $\triangle_{0,i}^{ED}$ and $\triangle_{1,i}^{ED}$
4: Find threshold value λ_1^{ED} and λ_2^{ED} for Energy Detector for a given probability of false alarm p_{fa}^{ED} and using step 3
/* O_i is Observational Signal Value */
5: $If\ O_i \geq \lambda_2^{ED}$
L_i=H_1;
$elseif\ O_i \leq \lambda_1^{ED}$
L_i=H_0;
$else$
L_i=$NoDecision$;
$endif$
/* Forward the Local Decision L_i of clear region sensed by K SUs and $N-K$ "No Decision" observational values O_i of confused region using Energy Detector to the fusion center to make overall decision by employing Matched Filter */
6: Set the threshold value for Matched Filter λ^{MF} by fixing the probability of false alarm p_{fa}^{ED} to a pre-defined threshold and setting the boundary value for λ_1^{ED} and λ_2^{ED} from step 4
7: $If\ \sum_{i=1}^{N-K} O_i \leq \lambda^{MF}$
$D = 0$; /* PU is not present */
$else$
$D = 1$; /* PU is present */
$endif$
8: Z= D + $\sum_{i=1}^{K} L_i$; /* Total Decision at fusion center */
9: $If\ Z < 1$
$FC = H_0$; /* PU is not present */
$else$
$FC = H_1$; /* PU is present */
$endif$

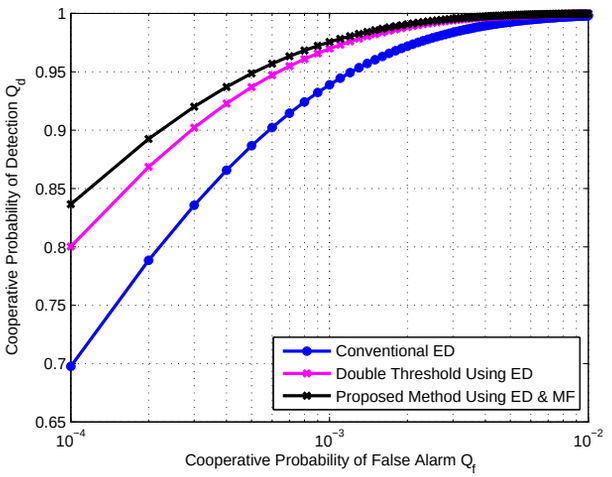


Fig. 3: ROC of our proposed scheme using adaptive double threshold energy and matched filter detector

IV. Simulation Results

Our simulation was conducted in MATLAB to investigate the performance of our proposed scheme. AWGN is imposed on the original signal x_i either for H_0 or H_1 condition. We assume that there is error free control channel available between the secondary users and the fusion center at the base station for sending local decisions and observational values O_i of the confused region.

The receiver operating characteristics (ROC) curves of our proposed scheme as compared to other schemes is shown in Fig. 3. The ROC curve is obtained with $SNR = 10dB$, Number of cooperative SUs=10, $\triangle_{0,i}^{ED}$=$\triangle_{1,i}^{ED}$=0.1, time bandwidth product u =5. Clearly our proposed scheme has the higher detection performance compared to other double threshold method using ED only and conventional ED. Our scheme takes the advantage of reliable MF detector in the confused region to take the decision.

Fig. 4 shows cooperative miss-detection probability curve of our proposed scheme as compared to other schemes. With the use of MF, our scheme is able to differentiate the signal and noise in the confused region and it can take decision accordingly. As expected our proposed scheme miss-detection probability is lower as compared to the previous schemes explained in the literature.

Fig. 5 shows the cooperative probability of detection curves against different SNR values. Fig. 5 is plotted using the probability of false alarm of energy detector is set at 0.01 i.e., $p_{fa}^{ED} = 0.01$, SNR values ranges from -10 dB to 10 dB, number of cooperative SUs= 10, $\triangle_{0,i}^{ED}$=$\triangle_{1,i}^{ED}$=0.01, time bandwidth product u =5. It is clear from Fig. 5 that our proposed scheme outperforms the other schemes at different SNR ranges. Even at -10 dB SNR value, our scheme is clearly able to detect the signal as compared to other schemes. The scheme using double threshold energy detector and conventional energy detector suffer greatly at low SNR region is due to the fact that the energy detector is highly susceptible to the noise uncertainty at the low SNR. The

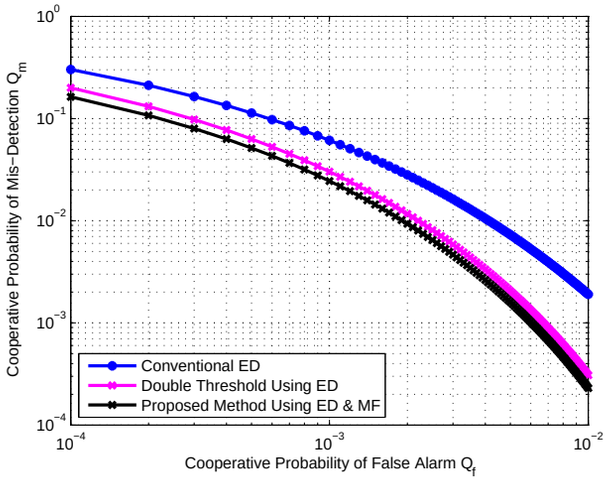


Fig. 4: Comparision of cooperative miss detection probability of our proposed scheme with other methods

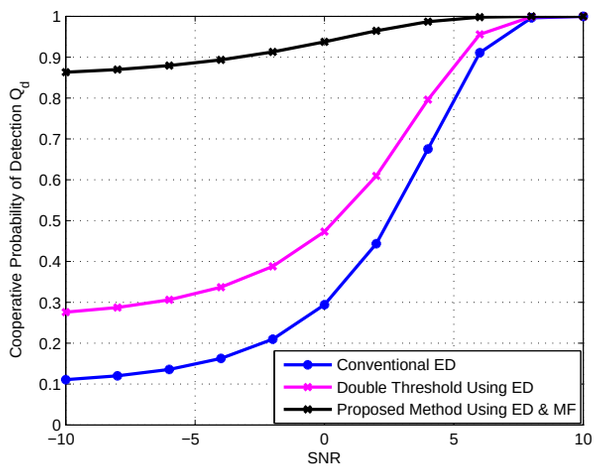


Fig. 5: Comparision of cooperative probability of detection of our proposed scheme with other methods at different SNR

decision in the confused region is clearly indicated by the MF in our scheme. Hence, our proposed scheme performance is superior to all other scheme.

V. Conclusion

In this paper, we have proposed a new adaptive double-threshold based energy and matched filter detector for cognitive radio networks. The proposed method gives significantly better detection performance compared to other methods. Also, the cooperative probability of miss-detection of our proposed scheme is lower than other scheme. At lower SNR region, energy detector cannot differentiate between signal and noise and is susceptible to noise uncertainty. Our proposed scheme takes the advantage of energy efficient energy detector to take decision in the clear region and reliable matched filter detector to take decision in the confused region. Hence our proposed scheme performance is better compared to other schemes.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (2012R1A1A1009442).

References

[1] Federal Communications Commission et al. Spectrum policy task force, rep. et docket no. 02-135, 2002.

[2] Joseph Mitola and Gerald Q Maguire Jr. Cognitive radio: making software radios more personal. *Personal Communications, IEEE*, 6(4):13–18, 1999.

[3] Simon Haykin. Cognitive radio: brain-empowered wireless communications. *Selected Areas in Communications, IEEE Journal on*, 23(2):201–220, 2005.

[4] Ghurumuruhan Ganesan and Ye Li. Cooperative spectrum sensing in cognitive radio, part ii: Multiuser networks. *Wireless Communications, IEEE Transactions on*, 6(6):2214–2222, 2007.

[5] Rongfei Fan and Hai Jiang. Optimal multi-channel cooperative sensing in cognitive radio networks. *Wireless Communications, IEEE Transactions on*, 9(3):1128–1138, 2010.

[6] Wei Zhang, Ranjan K Mallik, and K Letaief. Optimization of cooperative spectrum sensing with energy detection in cognitive radio networks. *Wireless Communications, IEEE Transactions on*, 8(12):5761–5766, 2009.

[7] Shridhar Mubaraq Mishra, Anant Sahai, and Robert W Brodersen. Cooperative sensing among cognitive radios. In *Communications, 2006. ICC'06. IEEE International Conference on*, volume 4, pages 1658–1663. IEEE, 2006.

[8] Ke Y Park. Performance evaluation of energy detectors. *Aerospace and Electronic Systems, IEEE Transactions on*, (2):237–241, 1978.

[9] Chunhua Sun, Wei Zhang, and Khaled Letaief. Cooperative spectrum sensing for cognitive radios under bandwidth constraints. In *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*, pages 1–5. IEEE, 2007.

[10] Jiang Zhu, Zhengguang Xu, Furong Wang, Benxiong Huang, and Bo Zhang. Double threshold energy detection of cooperative spectrum sensing in cognitive radio. In *Cognitive Radio Oriented Wireless Networks and Communications, 2008. CrownCom 2008. 3rd International Conference on*, pages 1–5, May 2008.

[11] Janne J Lehtomäki, Johanna Vartiainen, Risto Vuohtoniemi, and Harri Saarnisaari. Adaptive fcme-based threshold setting for energy detectors. In *Proceedings of the 4th International Conference on Cognitive Radio and Advanced Spectrum Management*, page 33. ACM, 2011.

[12] Dimas Triwicaksono and Soo Young-Shin. Energy detector and matched filter as cascaded clear channel assessment in wireless network. In *Information and Communications Technologies (IETICT 2013), IET International Conference on*, pages 551–556. IET, 2013.

[13] GB Giannakis and MK Tsatsanis. Signal detection and classification using matched filtering and higher order statistics. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(7):1284–1296, 1990.

Superframe Scheduling with Beacon Enable Mode in Wireless Industrial Networks

Oka Danil Saputra*, Soo Young Shin†
Wireless and Emerging Networking System (WENS) Lab.
School of Electronic Engineering
Kumoh National Institute of Technology, Gumi-si, South Korea
(*okadanil,†wdragon)@kumoh.ac.kr

***Abstract*—In this paper, we develop a method to check the schedulability superframe with beacon-enabled mode in ISA100.11a Wireless Industrial Networks environment. Maximum length of time slot is added in the superframe to reduce the overhead, and deadline monotonic scheduling approach is proposed to analyze the system scenario. The simulation results indicate that our proposed method required less number of beacon for message scheduling. Therefore, more data can be sent in the network with our proposed method.**

***Keywords*—Superframe, Scheduling, Beacon, ISA100.11a.**

I. Introduction

In recent years, wireless technology has grown rapidly that offer low cost, flexibility, reliability and ease of installation. The process automation at industrial applied wireless system to saving place and reduce cost for cable. The increase enforcement of wireless has made many kind of organizations to develop a wireless standard for industrial, such as ISA100.11a and WirelessHART. The embedded technology in ISA100.11a has more interoperability with other technology compared to WirelessHART [1]. The latest version of IP, IPv6, is placed in network layer to connect with internet network likewise support by ISA100.11a. Moreover, the upper layer made the ISA100.11a protocol easier to hybrid with wired standard such as PROFIBUS, Modbus, and fieldbus network. Paper [2] is conducted comparison between ZigBee Pro and ISA100.11a. The framework is set inside aerospace. The motivation is to test protocol in high uncertainty condition. The outcome is proven ISA100.11a more stable under interference as compared with ZigBee Pro.

More about the history of ISA100.11a, the protocol established in April 2009 through ISA100 committee in International Society of Automation (ISA). The purpose of the establishment of ISA100.11a is to meet the challenges that exist within the industry that require a protocol that was not only safe but also robust in communication to be applied in the field. ISA100.11a adopt physical layer from IEEE 802.15.4 using 2.4 GHz unlicensed frequency band with 16 channel inside. Direct-Sequence Spread Spectrum (DSSS) is used in the modulation scheme. In media access control (MAC) Layer, ISA100.11a used carrier sense multiple access with collision avoidance (CSMA/CA) to detect the interference from other frequency and equipped time division multiple access (TDMA). 250 kbps was the maximum data rate and channel hopping is defined by standard as the features. ISA100.11a support 3 channel hopping (slotted hopping, slow hopping, the last hybrid hopping is a combination of slotted and slow hopping). The slotted hopping use different frequency to send the data transmission where this method usually for the accommodation of periodic data, slow frequency hopping requires the same frequency for data transmission so that the method is suitable for data applications that are sporadic [3],[4].

Paper [5] distinguished two kind of strategy operational in IEEE 802.15.4 MAC layer called beacon enable mode and non beacon model. In beacon enable mode, coordinator produced beacon periodically to synchronize node connected with it and to defined the superframe formation. Non beacon enable mode is determined no superframe structure. The advantage of this mode, node instantly sent data without synchronization needed, more scalability and self-organization. Nevertheless, non beacon mode cannot ensure to transmitted data frames. ISA100.11a define the MAC layer is supported by beacon enable mode, as this mode accommodate to consign data frame. So, the schedulability of superframe become an interesting research topic to be investigated.

Message scheduling based on modify superframe has been proposed in [6]. This approach decided that periodic real-time message sent by dedicated time slot when shared time slot are transmitted non real-time message and aperiodic real-time message. The result showed that using message scheduling technique require more beacon in system and increased overhead. Paper [4] investigated in more detail, the parameter which effect the ISA100.11a performance. Superframe period was one of the performance parameter with the rise of period, it will increase throughput of the network but at the same time increase the overhead.

To solve the same problem, our paper proposed the scheme have concern to check schedulability of superframe. The main contributions: 1) this paper discuss clearness superframe scheduling inspired by deadline monotonic scheduling and 2) Our scheme develop based on real scenario in field, not require any special hardware and easy to implement.

The rest of this paper is organized as follows. In section II, scheduling technique with consist rate monotonic and deadline monotonic scheduling are discussed. System model provided in section III to depict traffic scenario. Section IV and section V present about scheduling analysis and simulation result, respectively. Finally, the conclusions are given in section VI.

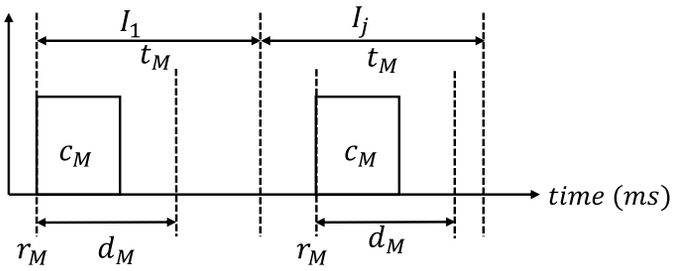


Fig. 1. Message characteristics.

II. Scheduling Technique

We concisely converse about rate monotonic and deadline monotonic scheduling. In past, both scheduling technique is developed to check scheduling process of processor. Now, our paper developed new application of deadline monotonic is implemented to check superframe scheduling in ISA100.11a wireless industrial networks. The limitation of rate monotonic is expressed with clear way. Deadline monotonic is forwarded to break the constraints.

The scheduling technique one of solution to address issues exchange data in industrial without overlap from other node. Rate monotonic scheduling scheme is proposed by paper [7] to analysis real-time computing in plant. The term of predictability is introduced by paper for describe ability to determine for a gave set of tasks whether the system meet all of the timing requirement of those tasks. There are two type of task, dynamic priority and static priority. A dynamic priority permitted task to switch priorities at any time, in contrast with static priority allocated fixed priority for all task. Another classification was on-line and off-line scheduling. On-line scheduling is arranged to make scheduling outcome during the run-time of the system. Finally, this algorithm hard for non-periodic processes. The offline scheduling is feasible for the all processes in fact periodically for example task monitoring in factory [8]. For those reason, our paper focus only in off-line scheduling rules and all of task assume have static priority.

In our paper, we are consider beacon enable mode, therefore rate monotonic unsuitable for superframe scheduling because policy "period equal deadline". The regulations "period less than or equal to deadline" is used correspond with deadline monotonic (more about the argument used of deadline monotonic is explained in section 3). In the end, deadline monotonic scheduling is prevailed for analysis superframe scheduling in ISA100.11a networks. Generally, our paper is reclaimed "message" to represent "task". Fig. 1 illustrate the characterization of the message with follows the concept according paper [9].

$$c_M \leq d_M \leq t_M \tag{1}$$

where c_M is computation time, r_M is the start of release time, d_M is deadline time, and t_M is the period of message m in superframe. The number of iteration $I = (I_1, I_2, ..., I_j)$, j indicated end of iteration while system reset or stop. In case of to check scheduling 2 process, schedulability of deadline monotonic is defined by theorem II.1.

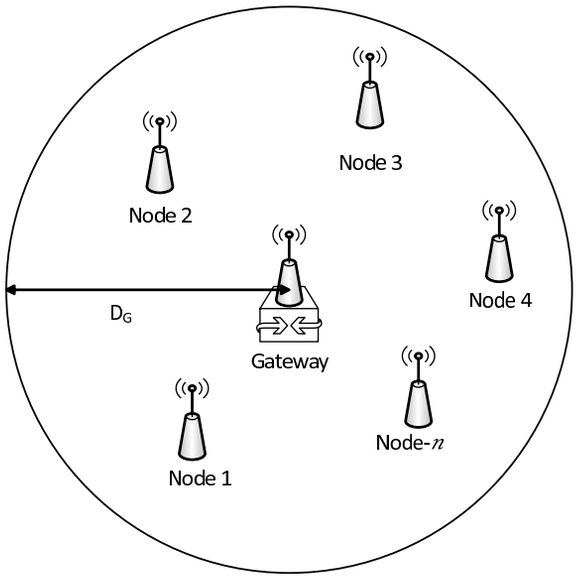


Fig. 2. Network model consists of one gateway and node.

$$\forall i : 1 \leq i \leq 2 : \frac{c_M}{d_M} + \frac{p_M}{d_M} \leq 1 \tag{2}$$

Theorem II.1. *The schedulability test given in equation (2) is sufficient for two process.*

Proof: The detail proof can be find in paper [10]. ■

III. System Model

The network model is discussed in this section. The model is verified to close as well as real wireless industrial networks environment. Indeed, this section also describe the traffic scenario and exposed the superframe form.

Our proposed scheme is expanded tackle problem to check superframe scheduling mention section 2 for many process. There are fixed node $N = (N_1, N_2, ..., N_\eta)$ with η denote the maximum number of node located randomly in network. The maximum area that can be reached by gateway is D_G. An example of a network model is described in Fig. 2. The setup industrial in our paper is build under star topology. Less complexity is effected star topology become favorite compared with mesh and tree topology. The service of network is created to serve the system monitoring in plant where interconnecting among node to finish process very essential. For illustration, suppose node 1 transmitted data to node 3 but range of node 1 is not enough to reach node 3. In that case, the function of gateway is needed to manage the data. The gateways is associated as the bridge to distribution data from node 1 to node 3.

The traffic scenario is started by gateway as coordinator sent beacon periodically to node N in network. We assume known all period message M_N for each node. Node frequently generate message with every message $M_N = (M_{N_1}, M_{N_2}, ..., M_{N_j})$ go along after beacon. Message from node has parameters $M_N = (r_{M_n}, c_{M_n}, d_{M_n}, t_{M_n})$ as present section 2. Furthermore, beacon and all the message determined shape of the superframe. The first iteration of superframe is

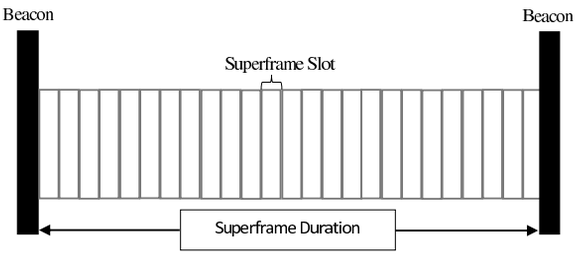


Fig. 3. The conguration of superframe.

started after first beacon generate by gateway and the end of superframe occur if the network reset or stop by admin authority. The superframe is constructed as express equation (3).

$$S = B + \sum_{N=1}^{N=\eta} M_N \tag{3}$$

From equation (3), S is superframe form, B is beacon. Consequently, the configuration of superframe is composed through time slot as shown in Fig. 3. Each box time slot is dedicated only for one node to occupy, which shared data from different node is forbidden. The length of each time slot is varying from 10-12 ms [3]. To simplify the calculation, the length time slot 10 ms is selected in our paper. Latter, the duration of one superframe is determined by the interval time between two beacons. In paper [6] scheme is defined 7 time slot in one superframe, where the superframe devise into dedicated time slot, shared time slot and beacon. This scheme has drawback more overhead because the relation beacon proportionally with number iteration of superframe. Based on knowledge of field, most data monitoring in industrial is generated periodically and ISA100.11a standard assign maximum number of time slot in one superframe as 25 timeslot [3]. Through those reason, in our paper, we implemented with maximum timeslot to reduce the overhead.

IV. Scheduling Analysis

In this section, discussion of superframe scheduling are presented. The basic of deadline monotonic has been describe in the section 2. The goal this section is to check superframe scheduling and guaranteed all message can be transmitted without interruption from each other.

We consider a set of message M_N according model in Fig. 2. The scheduling analysis of M_N is reached according theorem IV.1 based on deadline monotonic rules. "Message with lowest deadline is executed first until computation time done" is ruled of schedulability. Since we define our system as the off-line scheduling, so all the parameter $M_N = (r_{M_n}, c_{M_n}, d_{M_n}, t_{M_n})$ already known. We assume that, scheduling analysis is calculated before started the system. Thus, the failure cause of unschedulable never come out. The aim of pick up off-line scheduling to settle during run-time no error happen in case of conflict data involving node inside system.

Theorem IV.1. *Superframe is schedulable if and only if sufficient* $\forall M_N : c_{M_n} \leq d_{M_n}$

TABLE I. Simulation Parameter Using Deadline Monotonic Scheduling

	r_{M_n} (ms)	c_{M_n} (ms)	d_{M_n} (ms)	t_{M_n} (ms)
Beacon	0	10	10	250
Node 1	10	20	20	150
Node 2	20	20	80	80
Node 3	30	30	100	100
Node 4	40	10	50	50

Proof: Suppose that the condition is insufficient thereby $c_{M_n} \geq d_{M_n}$, then message M_N cannot finish computation time before deadline d_{M_n}. Therefore node misses to deliver data and unschedulable. The condition schedulable is passed while message have enough time to finish the execution c_{M_n}. ■

The schedulability analysis is passsed by using algorithm 1.

Algorithm 1 Scheduling Analysis

1: Initialize maximum number of node N.
2: Initialize $r_{M_n}, c_{M_n}, d_{M_n}, t_{M_n}$.
3: Gateway generated beacon B.
4: Node transmitted message M_N.
5: Message M_n the lowest deadline d_{M_n} is assigned to superframe first and followed until all message finish.
6: $Superframe$ equation (3) $\longleftarrow$ $granted$ schedulalable.
7: if $\sum_{N=1}^{N=\eta} \forall M_N$: sufficient Theorem IV.1.
else
$\forall M_N$: unschedulalable.
end if

V. Simulation Result

In other to evaluate the performance of proposed scheme in ISA100.11a Wireless Industrial Networks, matlab program [11] is developed for simulation. The objective of investigate to perform the schedulability of superframe ISA100.11a with static priority among fixed node and one gateway in network. The parameters is used for simulation list in Table I.

Superframe scheduling with beacon constrain is considered in simulation. The ornament show in Fig. 4 and Fig. 5 to check whether superframe scheduling or unscheduling under deadline monotonic scheduling method. The result are presented consists of 6 layers, layer 4 to layer 1 represent for node message $M_n = (M_{N_1}, M_{N_2}, M_{N_3}, M_{N_4})$, while beacon and superframe in layer 5 and layer 6, respectively. We assume beacon B have the lowest deadline, thus beacon B is transmitted always in first and never interruptions by any other message M_n.

Fig. 4 present that reproduce of the simulation result paper [6]. That paper is defined the length of each time slot 10 ms, while the period of t_{M_n} 70 ms. From Fig. 4 is presented for the most of message complete the computation time c_{M_n} before the deadline time d_{M_n}, except for the message M_{N_3} in layer 2. In the third to forth of beacon iteration, message M_{N_3} still leaves 10 ms of computation time. The reason of due M_{N_3} does not have any time slot to assign execution time and too small length interval of is gave also another creator. According to the theorem IV.1 superframe is unschedulable.

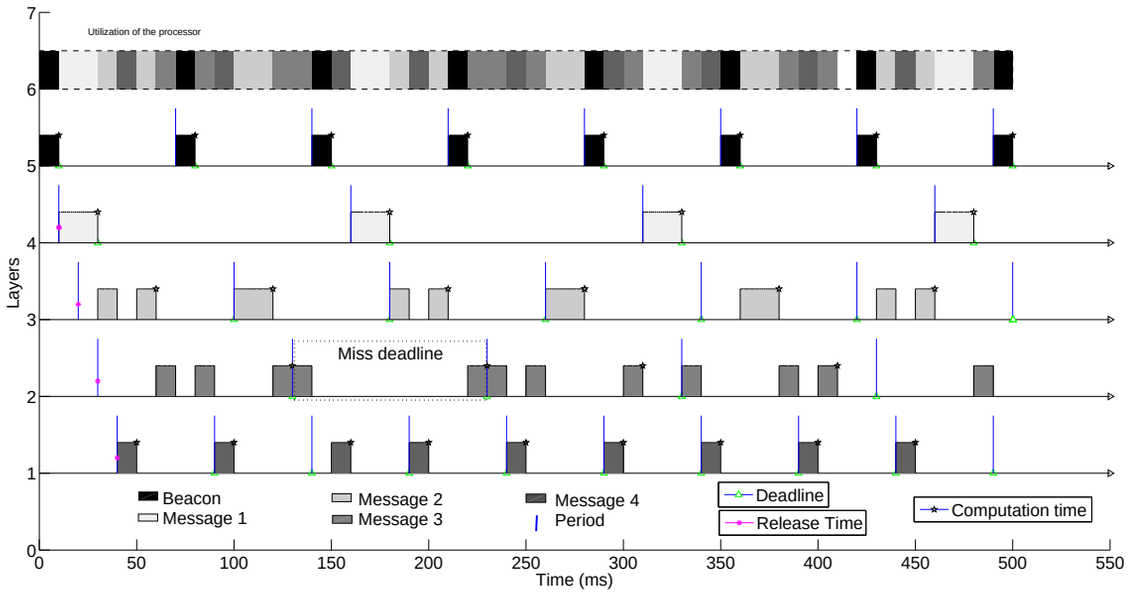


Fig. 4. Message scheduling [6] is applied by deadline monotonic scheme.

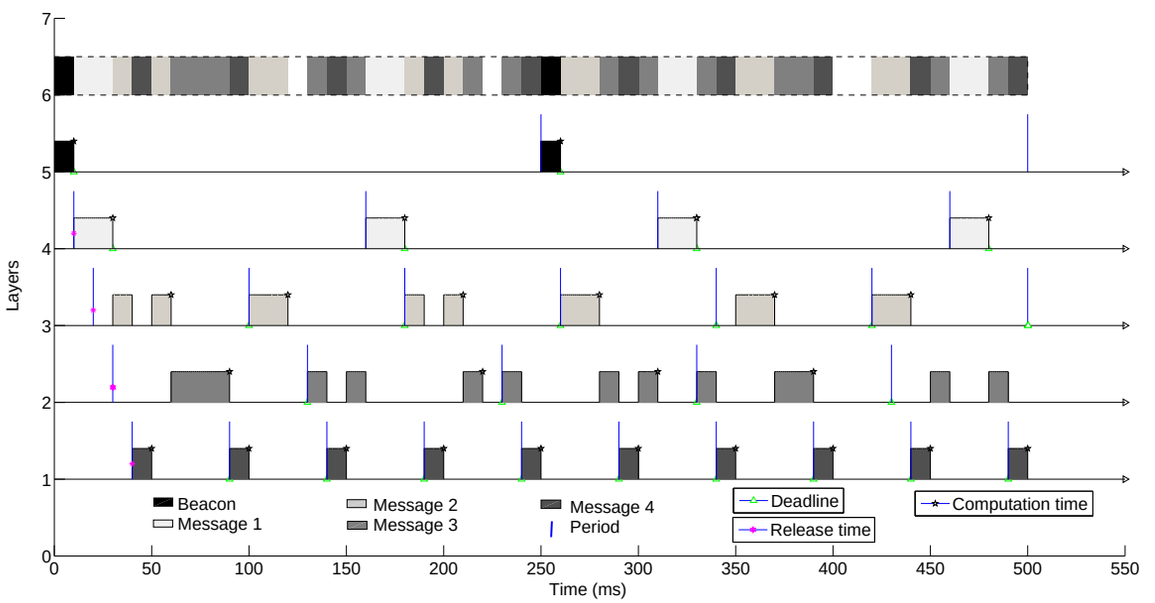


Fig. 5. The proposed scheme, superframe scheduling, is applied by deadline monotonic scheme.

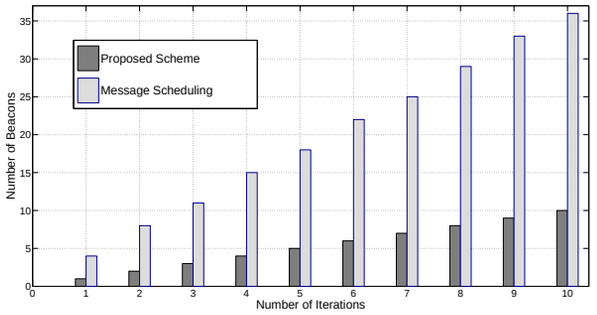


Fig. 6. Comparison beacon required between message scheduling scheme with proposed method.

The setup parameter in Fig.5 is followed by length of superframe 250 ms [3]. Thus, we saw in Fig. 5 the deliver of message M_{N_2} is deferred since message M_{N_4} have the deadline d_{M_4} less 30 ms as compared with deadline d_{M_2}. The similar behavior between message M_{N_3} and message M_{N_4}. In contrast to other message, message M_{N_1} always completed without irritation. The reason of this phenomenon, because computation time c_{M_1} eternally fulfilled and the deadline d_{M_1} second lowest after deadline d_B. Finally $\forall M_n$ is schedulable. Furthermore, the conform to theorem IV.1. Fig. 6 show compared simulation result between our scheme and message scheduling technique [6]. That our technique is needed less amount of beacons as we extend the length of superframe up to 250 ms, so more number of message be transferred in ISA100.11a Wireless Industrial Networks environment. Our methods guarantee that the exchange of data across the network successfully without interference or overlap among data in one time slot of the superframe.

VI. Conclusion

In this paper, a new application of deadline monotonic scheduling is proposed to check and test superframe scheduling and to reduce the overhead without degrading the network performance in ISA100.11a Wireless Industrial Networks environment. The performance of the proposed method is compared with the other scheme, which is message scheduling. In addition, beacon constraints are also considered in this paper. We also demonstrated the schedulability test by using the deadline monotonic policy. The simulation results showed that our proposed method required less number of beacons, compared to message scheduling. We added maximum length of time slot in superframe to reduce the overhead. Hence, the proposed method could assign more data to be sent in the network. For future work, we will examine multi-tree and multi-channel by apply Nash equilibrium approach from game theory.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2012R1A1A1009442).

References

[1] H. Hayashi, T. Hasegawa, and K. Demachi. Wireless technology for process automation. In *ICCAS-SICE, 2009*, pages 4591–4594, Aug 2009.

[2] R.S. Wagner and R.J. Barton. Performance comparison of wireless sensor network standard protocols in an aerospace environment: Isa100.11a and zigbee pro. In *Aerospace Conference, 2012 IEEE*, pages 1–14, March 2012.

[3] Wireless systems for industrial automation: Process control and related applications. *ISA100.11a Working Group*, pages 1–817, 2009.

[4] F.P. Rezha and Soo Young Shin. Performance evaluation of isa100.11a industrial wireless network. In *Information and Communications Technologies (IETICT 2013), IET International Conference on*, pages 587–592, April 2013.

[5] A. Koubaa P. Jurcik. The ieee 802.15.4 opnet simulation model: Reference guide v2.0. pages 1–13, May 2007.

[6] F. Dewanta, F.P. Rezha, and Dong-Sung Kim. Message scheduling approach on dedicated time slot of isa100.11a. In *ICT Convergence (ICTC), 2012 International Conference on*, pages 466–471, Oct 2012.

[7] M.H. Klein, J.P. Lehoczky, and R. Rajkumar. Rate-monotonic analysis for real-time industrial computing. *Computer*, 27(1):24–33, Jan 1994.

[8] R. Davis K. Tindell and A. J. Wellings N.C. Audsley, A. Burns. Real-time system scheduling. In *Predictably Dependable Computing Systems*, pages 41–52, 1995.

[9] M. F. Richardson N.C. Audsley, A. Burns and A. J. Wellings. Hard real-time scheduling: The deadline-monotonic approach. In *in Proc. IEEE Workshop on Real-Time Operating Systems and Software*, pages 133–137, 1991.

[10] N.C. Audsley. Deadline monotonic scheduling. In *Department of Computer Science, Univ. of York*, pages 1–38, 1990.

[11] C. Vincent. Task scheduler beta. Jan 2013. Available at http://www.mathworks.com/,.

Development of Electro-Hydraulic Servo Drive Train System for DORIS Robot

Khaled Sailan, Klaus D.Kuhnert , Saeed Sadege
Siegen University.
Electrical engineering department
Real time system institute.
Siegen, Germany.
khaled.sailan@student.uni-siegen.de,kuhnert@fb12.uni-siegen.de,saeed.sadege@student.uni-siegen.de

***Abstract*— This paper presents an electro-hydraulic servo drive train system to improve better control tasks for speed and steering on Unmanned Ground Vehicle (UGV). The developed electro-hydraulic servo drive train system enables independent control of the speed at each wheel. The motivation for developing this drive train is to overcome the problems from using hydrostatic transmission that used in the previous project. The electro-hydraulic servo drive train system is designed, modelled and simulated and demonstrate a good result that meets our requirements vehicle speed and torque.**

***Keywords*—servo system, hydraulic, Unmanned Ground Vehicle**

I. INTRODUCTION

The University of Siegen in Germany working on the research and development of an autonomous DORIS Robot (Dual media Outdoor Robot Intelligent System) project as in figure 1. As it is still a work-in-progress, a problem has been identified in the designing and implement of speed and steering control system. The previous drive train system was Hydrostatic transmission system. There is two basic methods used for controlling speed of a hydraulic motor. First, a variable-displacement pump controls flow to the hydraulic motor. This configuration is commonly known as a hydrostatic transmission this system designed and implemented in the previous project [1]. Second, a proportional or servo valve powered by a constant-pressure source, such as a fixed displacement pump, drives the hydraulic motor and this will be the research point. Electro-hydraulic servo systems (EHSS) are essential components in a wide range of modern machinery, due to their high power-to-weight ratio, as well as their fast, accurate response, high stiffness, fast response, self cooling, good positioning and capabilities, etc. Some commonly encountered industrial applications of EHSS include industrial robots, aerospace flight-control actuators, automobile active suspensions, as well as a variety of automated manufacturing systems. The principal elements of EHSS are a pump, an accumulator, a relief valve, a servo valve, and a hydraulic actuator. The accumulator and the relief valve, respectively, add and remove fluid in the pressure line to maintain the supply pressure of the system. The servo valve controls the motion and the pressure of the hydraulic actuator, based on an electrical input signal. The hydraulic actuator drives the load, transmitting the desired displacement, velocity, and/or pressure to the load. The dynamics of EHSS are highly nonlinear and make control design for precise output tracking very challenging [2].

Fig. 1. DORIS robot

II. SYSTEM MODELLING

Figure 2 shows a schematic of the EHSS that is considered. The pump feeds the system with oil stored in the tank. The relief valve and the accumulator are intended to keep the supply pressure constant. The electrical control input acts on the electro-hydraulic proportional valve to move its spool. The spool motion controls the oil flow from the pump through the hydraulic motor which controls also the hydraulic pump speed for each side. Depending on the desired control objectives, the vehicle wheel is driven appropriately by the bidirectional hydraulic motor. Mechanical directional valve used to make short circuit to enable wheel free running during pushing the vehicle manually. 6/2 directional control valve used switch between the right side motor and the jetski motor which is used during driving on the water and controlling the jetski achieved through the dual acting cylinder of rudder which controlled by 4/3 directional valve.

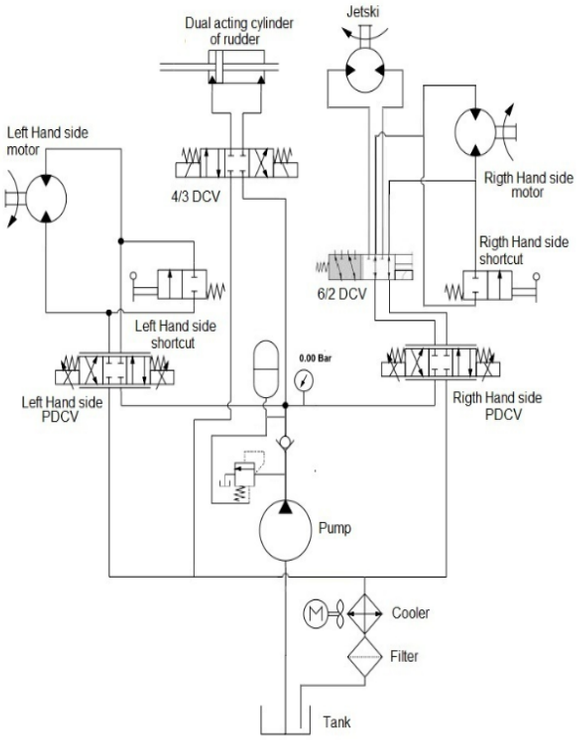


Figure 2 schematic of EHSS

III. DRIVE TRAIN DESIGN

To select drive wheel motors for our Unmanned Ground Vehicles, a number of factors must be taken into account to determine the maximum required torque.
The vehicle design criteria are
Vehicle weight (m) = 1000 Kg
Weight on each drive wheel = 500 Kg
Radius of wheel/tire (r) =30 cm
Desired top speed (V.max) = 30 km/h
Maximum incline angle (θ) = 45 degree
Worst working surface = concrete (good)
To choose motors capable of producing enough torque to propel our vehicle, it is necessary to determine the total tractive force (Ft) requirement for the vehicle[4]:

$$F_t = F_r + F_c + F_a \quad (1)$$

Where:
F_t = Total tractive effort [N]
F_r = Force necessary to overcome rolling resistance [N]
F_c = Force required to climb a slope [N]
F_a = Aerodynamic drag force [N]
The components of this equation will be determined in the following steps.

Step One: Determine Rolling Resistance
Rolling Resistance is the force necessary to propel a vehicle over a particular surface. The worst possible surface type to be encountered by the vehicle should be factored into the equation.

$$F_r = MC_{rr}g \quad (2)$$

Where:
F_r = rolling resistance [N]
M = vehicle weight [Kg]
C_{rr} = surface friction [3]
g = gravity acceleration [m/sec2]
F_r = 1000 × 0.012 (good Concret) ×9.81= 117.72 [N]

Step Two: Determine Grade Resistance
Grade Resistance F_c is the amount of force necessary to move a vehicle up a slope or "grade".
This calculation must be made using the maximum angle or grade the vehicle will be expected to climb in normal operation.

$$F_c = MG\sin\theta \quad (3)$$

Where:
F_c = grade force resistance [N]
M = vehicle weight [Kg]
θ = maximum incline angle [degrees]
g = gravity acceleration [m/sec2]
F_c = 1000×9.81 × sin (45°) = 6936.71 [N]

Step Three: Determine aerodynamic drag force
A force that acts parallel and in the same direction as the airflow on the front vehicle area.

$$F_a = 1/2\rho v^2 C_d A \quad (4)$$

Where:
F_a= Aerodynamic drag force [N]
ρ = Air density= 1.20 [kg·m−3]
v = Velocity
C_d= Aerodynamic drag coefficient = 0.82
A= Active area
F_a=0.5×1.20×1.5×0.82×69.44= 51.24 [N]

Step four: Determine Total Tractive force
The Total Tractive force F_t is the sum of the forces calculated in steps 1, 2, 3and 4.
Ft = 117.72 [N] + 6936.71 [N] + 51.24 [N] = 7105.67 [N]
Suppose that the weight divided between the two sides so

$$Ft1= \frac{7105.67}{2}=3552.84 \text{ [N]}$$

Step five: Determine Wheel Torque
To verify the vehicle will perform as designed regarding to the tractive force and acceleration, it is necessary to calculate the required wheel torque T_w based on the tractive force.

$$T_w = F_t R_w \quad (5)$$

Where:
T_w = wheel torque [N.m]
F_t = total tractive force [N]
R_w = radius of the wheel/tire [m]
T_w =3552.84×0.3=1065.85 N.m

Step six: Determine hydraulic motor Torque

The motor torque is the torque applied by the hydraulic motor to drive the vehicle[5].

$$T_w = T_m w_r \quad (6)$$

w_r =Ratio between vehicle wheel and motor shaft

Due to the dimensions from the mechanical components in the power train the gear ratio between the hydraulic motor and wheels are 2. So we can calculate the torque of the motor by:

$$Tm = \frac{1065.85}{2} = 532.92 \text{ Nm}$$

To put the effect of the mechanical efficiency in the calculation, the mean efficiency is supposed to be 85% so:

$$T \text{ motor} = \frac{T}{0.85} \text{ so T motor} = 626.97 \text{ Nm}$$

$$N = v/R_w 2\pi \quad (7)$$

N = 27.888/2 π =4.42 1/sec =265 rpm.

$$P = 2\pi T_m N \quad (8)$$

P = 626.97 nm × 4.42 s-1 × 2π= 17403.62 w

The hydraulic motor is available on our old project ,this motor is Danfoss OMR 160 [6].

Now all of the calculation will be done again to check how fast the car in a 45 degree slope .

$$F_t = F_r + F_c + F_a$$

Now we can calculate the maximum slope that the car can go up. So from the data sheet of the motor:

$$Tm = 380 \text{ Nm}$$

and

$$Tw = 380x2 = 760 \text{ Nm}$$

the force for one side is

$$Ft = \frac{760}{\text{wheel diameter}=0.3} = 2533.33 \text{ N}$$

The total force is

$$Ft = 2533.33 \times 2 = 5066.67[N]$$

With regard to our mean efficiency which is 85%

$$Ft = 5066.67[N] \times 0.85 = 4306.67[N]$$

$$F_r = 1000 \times 0.012 \times 9.81 = 117.72 \text{ [N]}$$

$$F_c = 1000 \times 9.81 \times \sin(x°) = 9810 \sin(x°) \text{ [N]}$$

$$Ft = 9810 \sin(x°)[N] + 117.72[N]$$

$$4306.67[N] = 9810 \sin(x°)[N] + 117.72[N]$$

So

Sin(x°) = 0.4270 and the working slope will be:

$$x° = 25° .$$

Due to the characteristic diagram of the hydraulic motor we can calculate the velocity for this slope:

From characteristic diagram of Danfoss OMR 160 Motor

The power is 12.5 kW and Working torque is 380 Nm

$$P = 2\pi T_m N$$

$$N = 314 \text{ rpm}$$

Now the velocity of the car can be calculated as :

$$N \text{ wheel} = \frac{314}{\text{Gear ratio}=2} = 157 \text{ rpm}$$

the linear velocity of the car will be 17.7 km/h which is acceptable for this slope.

IV. DRIVE TRAIN SYSTEM COMPONENTS

1) Hydraulic motor: from OMR160 [6] and according to the calculated angular velocity and the torque the Volume flow rate =60 l/min and pressure drop =60 bar as it shown in figure 3.Tow hydraulic Motors used one for right side and one for left side.

Figure 3 OMR 160 Hydraulic motor

2) Hydraulic pump: Because the typical gasoline engine power (from engine specifications) is 25.5 KW, the output power from the hydraulic pump is the power from gasoline engine multiply by the efficiency and suppose that typical mechanical efficiency is 85% the output power is 25.5 × 0.85 =21.6 KW = 21600Nm / s.

To calculate the Typical Torque the following formula can be used:

$$P = T \times W \quad (9)$$

P= Power of the engine

W= Nominal rotational velocity

T= Typical Torque

T = (21600 Nm/ s) / (2 x π x 67 1 / s) = 51.48 Nm .

To supply Two Hydraulic Motor we have to choose Hydraulic pump with 120 L/min , 51.48Nm . figure 4 shows the selected Hydraulic pump[7].

Figure4 Hydraulic pump Hydromot HM3

3) To control the velocity of the motors the best control system is to control the flow rate in the hydraulic system. The flow rate can be controlled by Proportional Directional Control Valve (PDCV) the selected Proportional valve shows in figure 5 [8].

Figure 5 Sauer Danfoss PDCV

V. DRIVE TRAIN SIMULATION AND RESULTS

This hydraulic system has been tested by Matlab/Simulink and Automation Studio Software. figure 6 shows on drive train side simulated using Automation Studio and the completed system simulation shown in figure 7 using Matlab/Simulink. The check out the results for the simulated system we put several scope in some certain place to check the Velocity, flow rate and position of the components.

The results from these scops have be evaluated and the related number due to the time have been released. figure 8 illustrates the Hydraulic Motor angular speed according to the valve spool position that shows in figure 9 and if we apply full spool position we got 16.5 rad/sec which is equal to 157 rpm which nearly the same calculated value .if the spool pushed to the other direction the Hydraulic Motor rotate in opposite direction .Figure 10 shows Hydraulic Pump flow rate needs to drive the Two Motors while figure 11 shows the Hydraulic Motor flow rate for forward and backward direction. Figure 12 shows the Hydraulic pump pressure and the overshoot cancelled by using pressure relieve valve and the accumulator keep the pressure constant.

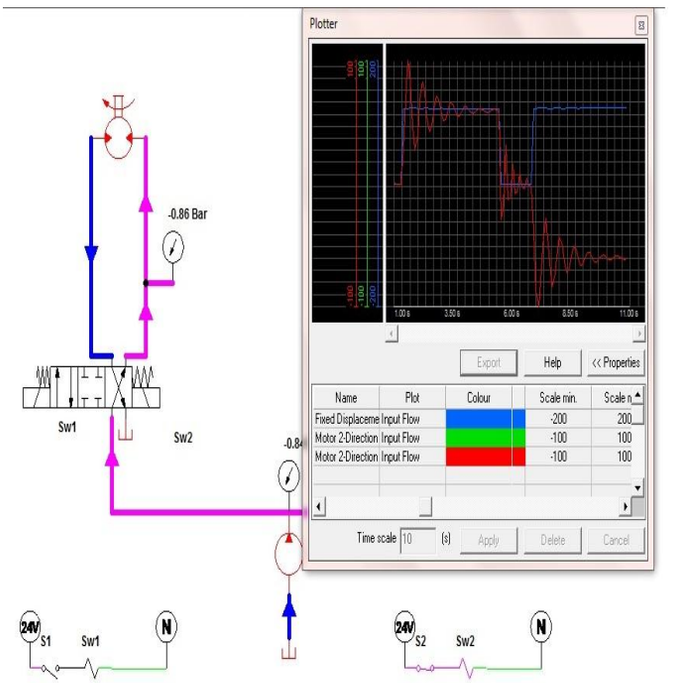


Figure 6 Simulated hydraulic system in AS5

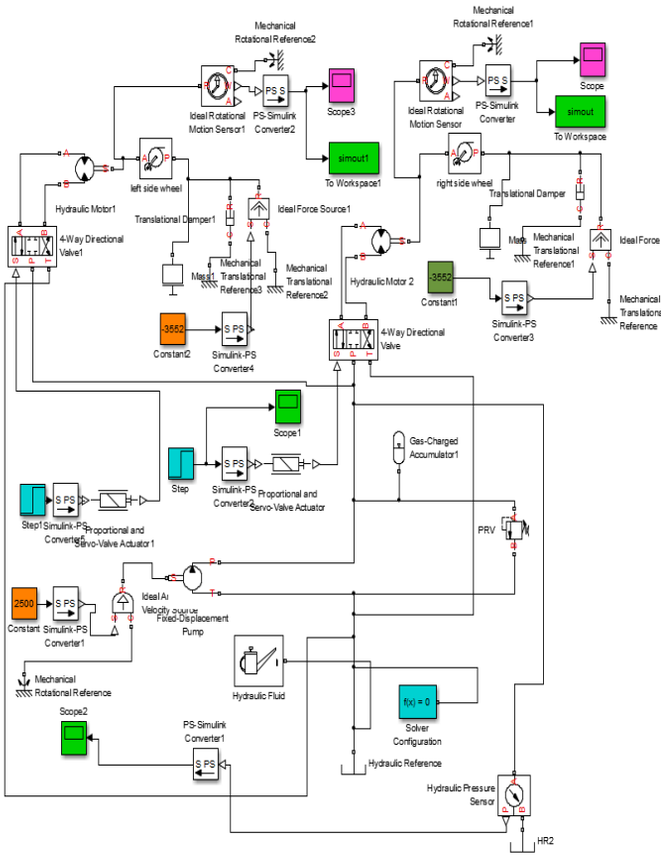


Figure 7 Simulated hydraulic system in Matlab

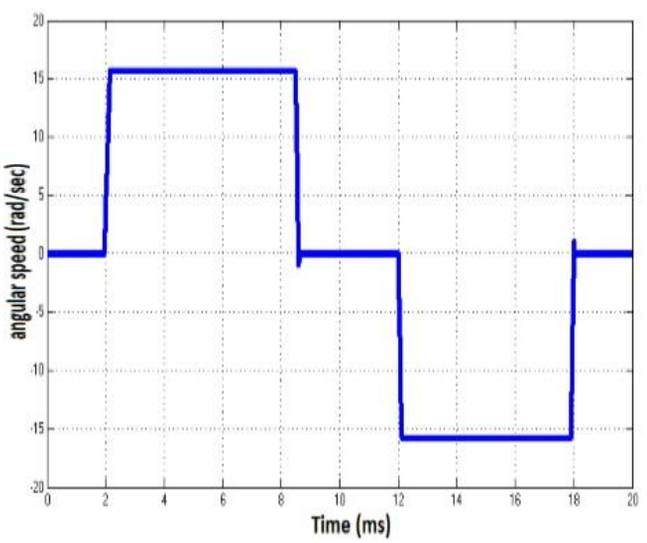


Figure 8 the angular speed of the Hydraulic Motor

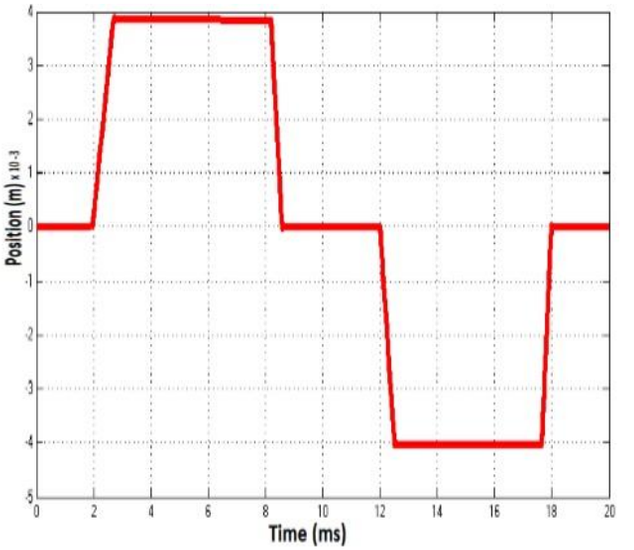


Figure 9 the PDCV spool position

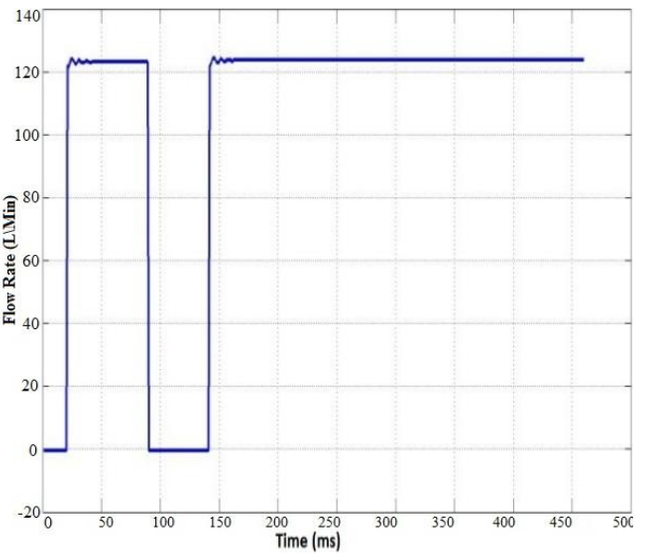


Figure 10 flow rate from hydraulic pump

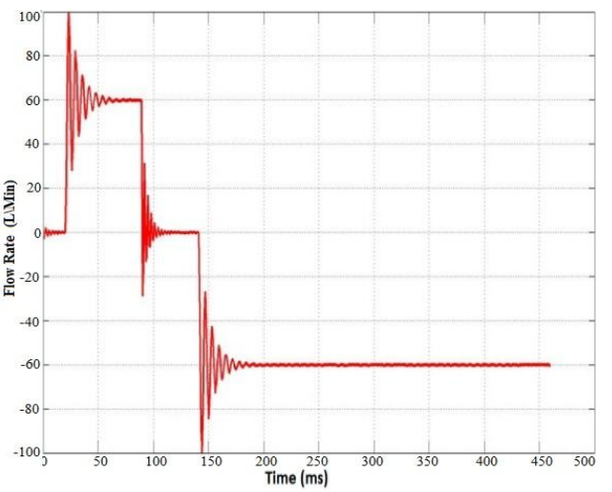


Figure 11 flow rate to the Hydraulic Motor

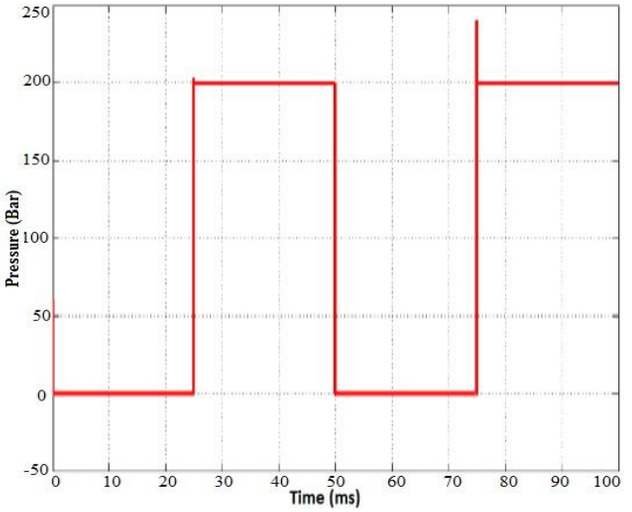


Figure 12 the Hydraulic pump pressure

VI. CONCLUSIONS

In this paper we have designed and developed a drive train system for our Robot (DORIS) to overcome the problems in the Hydrostatic transmission drive train system .All the calculations to choose the proper Pump ,Valves and Motors that meets the required specifications achieved . The system simulated using Matlab/Simulink and Automation Studio and the results represent that the calculations values meets our requirements.

REFERENCES

[1] Khaled sailan , Klaus-Dieter Kuhnert *"Design and Impliment of Powertrain Control System for the All Terrian Vehicle*"International Journal of Control, Energy and Electrical Engineering (CEEE) ,vol.1, pp. 50-56, Copyright – IPCO-2014.

[2] Honorine Angue-Mintsa , Ravinder Venugopal, Jean-Pierre Kenne and Christian Belleaub *"Adaptive position control of an Elecotro-Hydraulic servo system with load disturbance rejection and friction compensatio*n" Journal of Dynamic Systems, Measurement, and Control , , Vol. 133 / 064506-1 NOVEMBER 2011

[3] http://en.wikipedia.org/wiki/Rolling_resistance

[4] James D.Van de Van , Michael W. Olson and Perry Y. Li *"Development of a Hydro-Mechanical Hydraulic Hybrid Drive Train with Independent Wheel Torque Control for an Urban Passenger Vehicle*", Engineering Research Center for Compact and Efficient Fluid Power (CCEFP), EEC-0540834.

[5] Guo Fang , Li Xiwen, Xiong Hongxia *"Compound Gear Box Torque Analysis of Vertical Mixer*" Proceedings of 2012 International Conference on Mechanical Engineering and Material Science (MEMS 2012).

[6] Sauer Danfoss OMR Hydraulic Motor DKMH.PK.110.B4.02 520L0262. http://sauer-danfoss.cohimar.com/OMR.pdf

[7] http://www.hydromot.lu/pdf/deutsch/HM2_de.pdf

[8] http://powersolutions.danfoss.com/stellent/groups/publications/documents/product_literature/520l0344.pdf.

sailan khaled was born in Sana´a, Yemen, in 1978.He received the B.S. from Sana´a University in 2001 and M.S. degree in Mechatronics from Siegen University of in Germany in 2010. Since 20012, he is a PhD student in Siegen university institute of Real Time Systems

His research interests are control systems, embedded systems, Robotics and Autonomous Systems.

Prof. Klaus-Dieter Kuhnert Received his Dipl.-Ing. in Computer Science from the Technical University of Aachen (Germany), in 1981, and a Ph.D. Degree in Computer Science from the UniBw München (Germany) in 1988. After working as leading scientist with MAZDA-Motor Yokohama he is now full professor and head of the Institute of Real-Time Learning-Systems at the University of Siegen (Germany). He received several international awards: Nakamura Price for best paper of IROS 1989, Most influential paper of the decade award from IAPR/WMVA 1998, first price ELROB 2007 for autonomous driving in urban and non-urban terrain, innovation award of ELROB 2010. He has published over 90 refereed papers and chaired numerous conferences and sessions. referee for IEEE journals and program committees he also served for several research foundations. He is European Editor of the International Journal of Intelligent Systems Technologies and Applications (IJISTA). Beside others he is member of the graduate school MOSES, member of the center of sensor-systems NRW and founding member of the IEEE TC robotics learning. His research interests include: autonomous mobile robotics, driver assistance systems, 3D Environment Sensing, 3D modeling and visual object recognitio

Vibration Control of an Elastic Structure using Piezoelectric Sensor and Actuator with Cantilevered Beam as a Case Study

Mohammad Jafari
Graduate Student, Mechanical Engineering and Manufacturing Department
Universiti Putra Malaysia (UPM)
Serdang, Selangor DarulEhsan, Malaysia
al.m.jafari@gmail.com

Harijono Djojodihardjo
Professor and Corresponding Author
Aerospace Engineering Department
Universiti Putra Malaysia (UPM)
Serdang, Selangor DarulEhsan, Malaysia
harijono@djojodihardjo.com

***Abstract*—As a baseline for treating flexible beam attached to central-body space structure, the generic problem of a cantilevered Euler-Bernoulli beam with piezoelectric sensor and actuator attached as appropriate along the beam and its control is solved in great detail. For comparative study, three generic configurations of the combined beam and piezoelectric elements are solved by numerical methods and compared with available analytical and experimental results. The equation of motion of the beam is obtained by using Hamilton's principle, and the baseline problem is solved using finite element method. An in-house computational routine is developed for various applications.**

Keywords*—*Euler-Bernoulli Beam Theory, Finite Element Method, Hamiltonian Mechanics, Piezoelectric Material, Active Vibration Control, Structural Dynamics

I. INTRODUCTION

Vibration control of light-weight structures is of great interest of many studies and investigations[1-3]. The high cost of sending heavy masses and large volumes into space has prompted the wide utilization of light-weight structures in space applications, such as antennas, robot's arms, solar panels. A model of such set-up is exemplified in Fig. 1[2]. These kinds of structures are largely flexible, which results in lightly damped vibration, instability and fatigue. To suppress the adverse effect of vibration, sophisticated controller is needed.

Fig. 1. Solar Panel on a typical satellite

Active control approaches are widely reported in the literatures for the vibration control of structures. The active control approach makes use of actuators and sensors to find out some essential variables of the structure and suppress its vibration through minimizing the settling time and the maximum amplitude of the undesirable oscillation. This method requires a specific level of understanding about the dynamic behavior of continuous structures via mathematical modelling[4, 5]. Selecting adequate sensor and actuator is an important issue in active vibration control[6, 7]. The conventional form of sensor and actuator, such as electro-hydraulic or electro-magnetic actuator, are not applicable to implement on the light-weight space structures. Thus, in recent years, a new form of sensor and actuator has been studied using smart materials, such as shape memory alloys and piezoelectric materials. The definition of smart material may be expressed as a material which adapts itself in response to environmental changes. Among smart materials, piezoelectric materials are widely studied in literatures, since they have many advantageous such as adequate accuracy in sensing and actuating, applicable in the wide frequency range of operations, applicable in distributed or discrete manner and available in different size, shape and arrangement.

Space structures can be simplified mostly in the form of beam and plate. In this investigation, only beam theory is considered. From the fundamental beam theory Euler and Bernoulli developed one of the most practical and straightforward theories; however, as beam theory progresses, more sophisticated and accurate theories are developed like Rayleigh and Timoshenko beam theories. Euler-Bernoulli beam theory is applicable to thin and long span, for which plane sections can be assumed to remain plane and perpendicular to the beam axis, and shear stress and rotational inertia of the cross section can be neglected. Solar panel and antenna are very flexible and slender, so that Euler-Bernoulli beam theory can be considered. The equation of motion of the beam may be obtained using Newtonian mechanics, or analytical mechanic approaches such as Hamilton's method and Lagrange method[8, 9]. Hamiltonian mechanics is an elegant and convenient approach, since scalar equation of motion of the beam and boundary conditions are obtained simultaneously. The partial differential equation of motion of the beam can be solved by analytical methods such as separation of variables, or numerical methods such as finite element method. Since these structures are flexible, there is a need for control in order not to disturb the functionality of the space structure as a whole; for example solar panels should be

able to do some maneuver to point toward the sun, where vibration can occur in the panel. In order to facilitate maneuvering and attitude vibration control, this study is focused on how the flexibility can be controlled for well-behaved space structural dynamics. There are several ways to control the vibration[10]. Then the effort is aimed for devising a simple and effective controller to manipulate the vibration of a flexible structure. One of the adequate and simple controllers is Proportional-Integral-Derivative (PID) controller, which is classified as classic and linear controller[11]. PID controller minimizes the steady state error of the system. Linear Quadratic Regulator (LQR) controller is another convenient method. LQR is expressed as optimal and modern controller, which is based on minimizing the cost function of a dynamic system[11]. To develop a successful operation, most controllers have been developed for a finite number of natural modes where the controllability and observability conditions are met.

In order to design a system for controlling the vibration of structures, a good knowledge of several particular areas such as dynamics, control theory and dynamics of sensing and actuating transducers are required; however, many assumptions are considered in the literatures to simplify the problem. In this study, the need for the vibration control of beam structure is explored. The dynamic behavior of the beam under transverse vibration is studied using Hamilton's principle. Both analytical and finite element method are utilized to solve the equation of motion. First and second natural modes of the beam are considered for controlling the vibration of the beam, since these modes have more significant effect than other higher modes in the dynamic analysis of the beam. Also, the mode reduction can simplify the problem in control. PID and LQR control are considered to control vibration through PZT(Lead ZirconateTitanate – $PbZr_xTi_{1-x}O_3$) actuator.

In following sections, formulation of problem is represented in section two. The general mathematical model of the Euler-Bernoulli beam and finite element method to solve the equation of motion is discussed in section 3. In section four, modal order reduction and state-space form of the system is described. The control strategies utilized are defined in section five. Finally, the results from three case studies are discussed in section six.

II. Formulation of Generic Problems

Following a series of previous investigation on the analysis of impact resilient structure[12-15], and vibration analysis of an elastic clamped cantilever beam[16], the main aim of this investigation is to design a straightforward and convenient controller for suppressing the transverse vibration in a cantilever aluminum flexible beam through using sensing and actuating transducers. The Euler-Bernoulli beam theory is utilized to model the flexible beam with piezoelectric patches. Three different piezoelectric material configurations on the aluminum beam are considered for comparative study. Finite element method is utilized to achieve the natural frequencies and natural modes. Case study one is validated by analytical solution. Previous experimental result is used for validation of case study two.

To design the controller, two first major natural modes of the beam vibration are considered, since other natural modes has insignificant effect[4, 7, 17].The dynamic equation of the beam is transferred to state space form in order to design controllers. Two controllers are designed for each case study: PID controller and LQR controller with observer. These controllers are easy to perform and effective to suppress the vibration of the beam.

III. Modelling of The beam

The general equation of motion of the beam patched with piezoelectric material can be described as follows, which can be derived by using Hamilton's principle(a detailed elaboration is carried out in a companion paper [18]).

$$\bar{\rho}(x)\frac{\partial^2 w}{\partial t^2}+\frac{\partial^2}{\partial x^2}\left(\bar{I}(x)\frac{\partial^2 w}{\partial x^2}\right)=\frac{\partial^2 M(x,t)}{\partial x^2} \tag{1}$$

where $\bar{\rho}$ and $\bar{I}$ define as

$$\begin{aligned}\bar{\rho}(x)&=\rho_{bm}A_{bm}+\rho_{sn}A_{sn}K_{sn}(x)+\rho_{ac}A_{ac}K_{ac}(x)\\ \bar{I}(x)&=E_{bm}I_{bm}+E_{sn}I_{sn}K_{sn}(x)+E_{ac}I_{ac}K_{ac}(x)\end{aligned} \tag{2}$$

where subscripts *bm*, *sn* and *ac* represent the original beam, the sensor layer and the actuator layer, respectively. ρ,*A*, *E* and *I* are density, area of the cross section, elasticity modulus and the moment of the inertia, respectively. *K*(x) represents the sensor or actuator location on the beam through Heaviside function.*M*(*x*,*t*) is actuator moment on the beam, which defines as

$$\begin{aligned}M(x,t)&=bE_{ac}\left(\frac{h_{bm}}{2}\right)d_{31}V(x,t)\\ M(x,t)&=C_{ac}V(x,t)\end{aligned} \tag{3}$$

b and h is the width and thickness of the layer. d31 is piezoelectric strain constants and Cac is a constant that expresses the moment produced per unit control voltage. The voltage generated by sensor can be expressed as

$$\begin{aligned}V_{sn}&=-E_{sn}g_{31}\frac{h_{sn}}{(x_{sn,2}-x_{sn,1})}\left(\frac{h_{bm}+h_{sn}}{2}\right)\left(\frac{\partial w}{\partial x}\Big|_{x_{sn,1}}^{x_{sn,2}}\right)\\ V_{sn}&=C_{sn}.\left(\frac{\partial w}{\partial x}\Big|_{x_{sn,1}}^{x_{sn,2}}\right)\end{aligned} \tag{4}$$

where x_{sn}represents the location of beginning and end points of the sensor and g_{31} is the piezoelectric voltage constant. C_{sn}can be defined as the sensor constant.

To obtain the natural frequencies and modes of the system through finite element method, the external moment assumed to be zero.The finite element method due to Galerkin method is utilized to solve (1) for natural frequencies and natural modes[19]. To implement Galerkin method, a test function $\phi(x)$ is multiplied by (1) and integrated with respect to *x* over the domain.

$$\int_0^l \phi(x)\left\{\bar{I}\frac{d^4W(x)}{dx^4}-\bar{\rho}\omega^2 W(x)\right\}dx=0 \tag{4}$$

After integrating by parts and applying the boundary conditions, (5) is obtained.

$$\int_0^l \frac{d^2\phi(x)}{dx^2}\left(\bar{I}\frac{d^2W(x)}{dx^2}\right)dx - \int_0^l \phi(x)\left(\bar{\rho}\omega^2 W(x)\right)dx = 0 \tag{5}$$

which can be rewritten in following form.

$$\sum_{j=1}^{n}\int_{(j-1)h}^{jh}\bar{I}\left[\frac{d^2\phi(x)}{dx^2}\right]\left[\frac{d^2W(x)}{dx^2}\right]dx - \omega^2\sum_{j=1}^{n}\int_{(j-1)h}^{jh}\bar{\rho}\phi(x)W(x)dx = 0 \tag{6}$$

where*j* represents the number of element. Each integral equation can be used for each element to determine the stiffness matrices and mass matrices. The transverse deflection, *W*(*x*) is assumed as

$$W(x) = \mathbf{L}^T\bar{\mathbf{w}}_j,\quad \frac{d}{dx}W(x) = \left(\frac{1}{l_{elm}}\right)\mathbf{L}'^T\bar{\mathbf{w}}_j,\quad \frac{d^2}{dx^2}W(x) = \left(\frac{1}{l_{elm}}\right)^2\mathbf{L}''^T\bar{\mathbf{w}}_j \tag{7}$$

where**L** and **w** are the shape function and nodal vector, respectively, and can be expressed as

$$\mathbf{L} = \begin{bmatrix} 3\xi^2-2\xi^3 & l_{elm}\left(\xi^2-\xi^3\right) & 1-3\xi^2+2\xi^3 & l_{elm}\left(-\xi+2\xi^2-\xi^3\right)\end{bmatrix}^T \tag{8}$$

$$\bar{\mathbf{w}}_j = \begin{bmatrix} W_{j-1} & \theta_{j-1} & W_j & \theta_j\end{bmatrix}^T \tag{9}$$

$$(j-1)l_{elm} \le x \le jl_{elm},\quad \xi = j - \frac{x}{l_{elm}},\quad 0 \le \xi \le 1 \tag{10}$$

By setting the test function equal to *W*(*x*) for each element, and using the assumption (10), equation (6) can be written as

$$\sum_{j=1}^{n}\bar{\mathbf{w}}^T k_j\bar{\mathbf{w}} - \omega^2\sum_{j=1}^{n}\bar{\mathbf{w}}^T m_j\bar{\mathbf{w}} = 0,\quad k_j = \frac{\bar{I}}{h^3}\int_0^1\mathbf{L}''\mathbf{L}''^T d\xi,\quad m_j = h\bar{\rho}\int_0^1\mathbf{L}\mathbf{L}^T d\xi \tag{11}$$

The k_j represents the stiffness matrix and m_j describes the mass matrix for one beam element. By determining the integrals of (11), these matrices can be obtained as

$$k_j = \frac{\bar{I}}{l_{elm}^3}\begin{bmatrix} 12 & 6l_{elm} & -12 & 6l_{elm} \\ 6l_{elm} & 4l_{elm}^2 & -6l_{elm} & 2l_{elm}^2 \\ -12 & -6l_{elm} & 12 & -6l_{elm} \\ 6l_{elm} & 2l_{elm}^2 & -6l_{elm} & 4l_{elm}^2\end{bmatrix} \tag{12}$$

$$m_j = \frac{l_{elm}\bar{\rho}}{420}\begin{bmatrix} 156 & 22l_{elm} & 54 & -13l_{elm} \\ 22l_{elm} & 4l_{elm}^2 & 13l_{elm} & -3l_{elm}^2 \\ 54 & 13l_{elm} & 156 & -22l_{elm} \\ -13l_{elm} & -3l_{elm}^2 & -22l_{elm} & 4l_{elm}^2\end{bmatrix} \tag{13}$$

By considering the number of nodes, the stiffness and the mass matrices for each element can be assembled together and synthesized into the global stiffness and mass matrices. Thus, the eigen-value problem of (11) can be represented as (14) [19].

$$K\bar{\mathbf{w}} = \omega^2 M\bar{\mathbf{w}} \tag{14}$$

where K and M are global stiffness and mass matrices for an arbitrary beam. These matrices are valid for the part of the beam with symmetric piezoelectric patch. However, these matrices are applicable for the beam without piezoelectric patch by considering $\bar{I}$ and $\bar{\rho}$ equal to original beam E × I and ρ × A. Finite element formulation is utilized to write in-house MATLAB program. The actuator distributed moment on the beam element can be obtain through the virtual work. The virtual work done by moment is expressed as follows

$$\delta W = \int_{(j-1)l_{elm}}^{jl_{elm}} M_{ac}(x,t)\bullet\delta\left(\frac{\partial^2 w}{\partial x^2}\right)dx \tag{15}$$

By substituting (7) for actuator moment into (15), equation (16) is obtained

$$\delta W = -bE_{ac}\left(\frac{h_{bm}+h_{ac}}{2}\right)d_{31}V(t)\bullet\delta\left(\frac{\partial w}{\partial x}\right)\Bigg|_{(j-1)l_{elm}}^{jl_{elm}} \tag{16}$$

Substituting (7) into (6) and changing the integral band in order to local coordinates, (10), equation (16) can be rewritten as

$$\delta W = -bE_{ac}\left(\frac{h_{bm}+h_{ac}}{2}\right)d_{31}V(t)\bullet\delta\left(\bar{\mathbf{w}}^T\right)\mathbf{L}'\Big|_0^1,\quad \delta W = -\delta\left(\bar{\mathbf{w}}^T\right)bE_{ac}\left(\frac{h_{bm}+h_{ac}}{2}\right)d_{31}V(t)\begin{bmatrix}0 & 1 & 0 & -1\end{bmatrix}^T \tag{17}$$

where first term after equality shows the transverse variation and, thus, the actuation force expresses as

$$\{P_{ac}\} = bE_{ac}\left(\frac{h_{bm}+h_{ac}}{2}\right)d_{31}V_{ac}(t)\begin{bmatrix}0 & 1 & 0 & -1\end{bmatrix}^T,\quad \{P_{ac}\} = \{f_{ac}\}V_{ac}(t) \tag{18}$$

f_{ac} is the force vector of piezo-actuator, which maps the control voltage to the structure.

IV. Response of the System

For a Multi-degree of freedom (MDOF) system, the time response of the system can be expressed as matrix form[11, 20].

$$[m]\{\ddot{\eta}(t)\} + [c]\{\dot{\eta}(t)\} + [k]\{\eta(t)\} = \{P_{ac}\} + \{P_{ex}\} \tag{19}$$

Where m is the mass matrix, k is the stiffness matrix, $P_{ac}$is the actuation force and P_{ex} is the external force. c is damping matrix, which can be expressed as proportional damping, which is typically mentioned as Rayleigh damping. For damped structure, the closed forms of solution are not generally feasible. However, the idealized solution of damped can be assumed by utilizing classical damping. Classical damping is usually divided into category, Rayleigh damping and Caughey damping. Rayleigh damping method is widely used in beamlike structures, where shows acceptable model of structure damping. In this study, Rayleigh damping method is utilized which represents a linear combination of mass and stiffness matrices[21]

$$[c] = \alpha[m] + \beta[k] \tag{20}$$

whereα and β are known proportional Rayleigh coefficients respect to mass and stiffness matrices, respectively. α and β are defined as

$$\alpha = 2\xi_i \left(\frac{\omega_i \times \omega_j}{\omega_i + \omega_j} \right), \ \beta = 2\xi_i \left(\frac{1}{\omega_i + \omega_j} \right) \tag{21}$$

ξ is the modal viscous damping coefficient, which correspond to un-damped natural frequency, ω_i. The damping coefficient is the dynamic property of material, which cannot be determined theoretically. In this study, a uniform damping coefficient of 0.5% is assumed taking into account the results obtained from to previous experimental investigations and dynamic properties of metals [4, 17, 21, 22].

A. Modal Order Reduction

To facilitate the solution of the dynamical system (19), which involves very large matrices, resort is made to order reduction method. The concept is to estimate the high dimensional state space by using an appropriate low dimensional subspace to obtain a smaller system with approximately similar properties. To design the linear controller, first and second natural modes of the beam vibration are considered, since the other natural modes exhibit insignificant effect in comparison to the first two modes. In this regard, modal order reduction technique is utilized to reduce the large number of order of the system, which is obtained by finite element solution. Thus, first the coordinate of the system is reduced by considering the first two modes.

$$\begin{aligned} \{\eta(t)\} &= [T]_{n\times 2}\{g(t)\}_{2\times 1} \\ \{g(t)\} &= \begin{Bmatrix} g_1 \\ g_2 \end{Bmatrix} \end{aligned} \tag{22}$$

where$T$ is the reduction matrix of eigen-vectors based on first two modes, and g is the reduced coordinates. By substituting (22) into (19) and multiplying by $[T]^T$, the reduced order of the transfer function of the system can be obtained.

$$\begin{aligned} &[T]^T[m][T]\{\ddot{g}(t)\} + [T]^T[c][T]\{\dot{g}(t)\} \\ &+ [T]^T[k][T]\{g(t)\} = [T]^T\{f_{ac}\} + [T]^T\{f_{ex}\} \end{aligned} \tag{23}$$

where mass damping and stiffnessmatrices and force vectors can be defined as follows

$$\begin{aligned} [\hat{m}]_{2\times 2} &= [T]^T_{2\times n}[m]_{n\times n}[T]_{n\times 2}, \\ [\hat{c}]_{2\times 2} &= [T]^T_{2\times n}[c]_{n\times n}[T]_{n\times 2}, \\ [\hat{k}]_{2\times 2} &= [T]^T_{2\times n}[k]_{n\times n}[T]_{n\times 2}, \\ \{\hat{f}_{ac}\}_{2\times 1} &= [T]^T_{2\times n}\{f_{ac}\}_{n\times 1}, \\ \{\hat{f}_{ex}\}_{2\times 1} &= [T]^T_{2\times n}\{f_{ex}\}_{n\times 1}, \end{aligned} \tag{24}$$

Thus, equation (23) can be rewritten as

$$[\hat{m}]\{\ddot{g}(t)\} + [\hat{c}]\{\dot{g}(t)\} + [\hat{k}]\{g(t)\} = \{\hat{f}_{ac}\} + \{\hat{f}_{ex}\} \tag{25}$$

B. State-Space Representation

Equation (25) is transformed to a state space vector dynamic equation for designing the state feedback control system. To express in state space, it is solved for $\ddot{g}$.

$$\{\dot{g}(t)\} = \{\dot{g}(t)\} \tag{26}$$

$$\begin{aligned} \{\ddot{g}(t)\} &= -[\hat{m}]^{-1}[\hat{c}]\{\dot{g}(t)\} - [\hat{m}]^{-1}[\hat{k}]\{g(t)\} \\ &+ [\hat{m}]^{-1}\{\hat{f}_{ac}\} + [\hat{m}]^{-1}\{\hat{f}_{ex}\} \end{aligned} \tag{27}$$

Then, X vector is introduced in order to reduce the order of (27) as

$$\begin{aligned} \{g(t)\} &= \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \\ \{\dot{g}(t)\} &= \begin{bmatrix} \dot{g}_1 \\ \dot{g}_2 \end{bmatrix} = \begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \end{bmatrix} = \begin{bmatrix} X_3 \\ X_4 \end{bmatrix}, \\ \{\ddot{g}(t)\} &= \begin{bmatrix} \ddot{g}_1 \\ \ddot{g}_2 \end{bmatrix} = \begin{bmatrix} \ddot{X}_1 \\ \ddot{X}_2 \end{bmatrix} = \begin{bmatrix} \dot{X}_3 \\ \dot{X}_4 \end{bmatrix} \end{aligned} \tag{28}$$

Thus, equations(26) and (27) can be represented, respectively, as

$$\begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \end{bmatrix} = \begin{bmatrix} X_3 \\ X_4 \end{bmatrix} \tag{29}$$

$$\begin{aligned} \begin{bmatrix} \dot{X}_3 \\ \dot{X}_4 \end{bmatrix} &= -[\hat{m}]^{-1}[\hat{c}]\begin{bmatrix} X_3 \\ X_4 \end{bmatrix} - [\hat{m}]^{-1}[\hat{k}]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \\ &+ [\hat{m}]^{-1}\{\hat{f}_{ac}\} + [\hat{m}]^{-1}\{\hat{f}_{ex}\} \end{aligned} \tag{30}$$

Equations (29) and (30) can be demonstrated in state space form as

$$\begin{aligned} \begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \dot{X}_3 \\ \dot{X}_4 \end{bmatrix} &= \begin{bmatrix} [0]_{2\times 2} & I_{2\times 2} \\ -[\hat{m}]^{-1}[\hat{k}] & -[\hat{m}]^{-1}[\hat{c}] \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \\ &+ \begin{bmatrix} [0]_{2\times 1} \\ [\hat{m}]^{-1}\{\hat{f}_{ac}\} \end{bmatrix} + \begin{bmatrix} [0]_{2\times 1} \\ [\hat{m}]^{-1}\{f_{ex}\} \end{bmatrix} \end{aligned} \tag{31}$$

$$\{\dot{X}\} = [A]\{X\} + [B] + [B^*] \tag{32}$$

where A, B and B^* are termed as state matrix, input matrix correspond to actuator and input matrix correspond to external force, respectively. The output of the system is expressed as Y. In this study, the output of the system is sensor voltage, where it was obtained in previous part. Thus, the output can represent as

$$Y = [C]\{X\} \tag{33}$$

where C is output matrix, and shown as

$$[C] = C_{sn}\begin{bmatrix}\theta_1(x_{s2}) - \theta_1(x_{s1}) & \theta_2(x_{s2}) - \theta_2(x_{s1}) & 0 & 0\end{bmatrix} \tag{34}$$

where θ is derivative of displacement and x_s is the location of piezoelectric sensor on the beam. The benefits of state space approach are in the formulation of the appropriate control to obtain the desired output.

V. Control Strategies

In this investigation, two linear control methods are applied to suppress the vibration of the beam. The state space model consisting of the first two natural modes of the system is utilized to design the controller. First, PID control method is considered, which is a well-known classical control method. Then, LQR modern control is utilized to design an optimal controller in order to compare with PID control.

A. PID Control

In this study, Proportional-Integral-Derivative (PID) control is considered for controlling the flexible beam structure. PID is a well-known control tool, due to its robustness andsimplicity. The proportional feedback constant, P, controls the natural frequency of the system, and therefore control the amplitude of vibration. The integral constant, I, set the necessary adjustment for the damping, or energy dissipation of the system. The combination of proportional and integral control action gives the controller a way to minimize steady state error, while having the ability to minimize the effects of disturbances to the system. Proportional and integral constants manipulate past control error and cannot prognosticate the future control error. Thus the derivative constant, D, is proportional to the change in the error. In other words, it manipulates the speed or response of the controller. The convenient selections of the PID constants are a key aspect in the success of executing the PID controller. The transfer function of the PID controller is given by[11]:

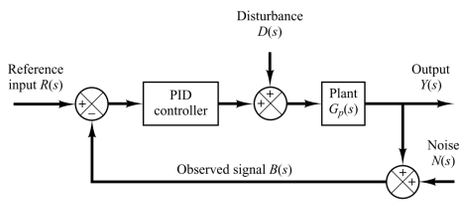


Fig. 2. Closed-loop system with PID control block diagram

$$u(t) = K_p e(t) + K_i \int e(t)dt + K_d \frac{de(t)}{dt} \tag{35}$$

where K_p is the proportional gain, K_i is the integral gain and K_d is the derivative gain. Fig. 2shows the block diagram of PID control system.

B. LQR Control with Observer

Linear Quadratic Regulator (LQR) is an optimal control method, which provides a symmetric way to determine the state feedback control matrix11. For controlling the beam vibration, all variables for LQR control are not available. An observer is necessary to design an estimator for the unavailable feedback values, since it can estimate the unavailable variables. In order to control the system with observer, it should be controllable and observable. The definition of system controllability, system observability, observer and LQR are elaborated subsequently in the following, respectively.

A system is called controllable, if a system, with regard to unconstrained input control vector, can be transferred from any initial value $\mathbf{X}(t_0)$ at time t_0 to the any specified value in a specific time $t_0 \le t \le t_1$. The controllability matrix is expressed as[11]

$$\left[\mathbf{B} \mid \mathbf{AB} \mid \ldots \mid \mathbf{A}^{n-2}\mathbf{B} \mid \mathbf{A}^{n-1}\mathbf{B}\right]_{n\times n} \tag{36}$$

The system is controllable, if and only if the controllability matrix (36) is a full rank matrix (rank of n); in other words, each vectors of the matrix (36) should be linearly independent. In this regard, MATLAB program can be utilized to determine the controllability matrix.

One system is called observable, if every state vector $\mathbf{X}(t_0)$ can be obtained from the observation output in a specific time, $t_0 \le t \le t_1$. In the control theory[11], the observability matrix of a system is shown as

$$\left[\mathbf{C} \mid \mathbf{CA} \mid \ldots \mid \mathbf{CA}^{n-2} \mid \mathbf{CA}^{n-1}\right]^T \tag{37}$$

One system is observable, if and only if the observability matrix has n linearly independent vectors, or it is full rank matrix. Same as controllability, MATLAB program is applicable to determine observability matrix of a system and check its rank.

1) Observer

Following the requirement of observer, the observer is designed based on pole placement method. By measuring the output of the system and control variables, it determines the estimated variables. The notation of $\hat{\mathbf{X}}$ and $\hat{Y}$ is used to assign the state vector of observer and the estimated output of observer, respectively. The observer gain, K_{ob}, can be obtained by pole placement method[11]. Thus, the equation of full state observer is expressed as

$$\dot{\hat{\mathbf{X}}} = [A]\hat{\mathbf{X}} + [B]f + K_{ob}(Y - \hat{Y}) \tag{38}$$

where A and B are state matrix and actuation matrix, which are same as the state space (32). f and Y are input and output of the system, respectively, where expressed as

$$Y = \mathbf{C}\mathbf{X}$$
$$\hat{Y} = \mathbf{C}\hat{\mathbf{X}} \quad (39)$$
$$f = V_{sn}$$

By substituting (39) into (38), it can be rewritten as

$$\dot{\hat{\mathbf{X}}} = [A - K_{ob}C]\hat{\mathbf{X}} + [B]f + K_{ob}Y$$
$$\dot{\hat{\mathbf{X}}} = [A_{ob}]\hat{\mathbf{X}} + [B \quad K_{ob}]\begin{Bmatrix} f \\ Y \end{Bmatrix} \quad (40)$$
$$\dot{\hat{\mathbf{X}}} = [A_{ob}]\hat{\mathbf{X}} + [B_{ob}]\begin{Bmatrix} f \\ Y \end{Bmatrix}$$

Equation (40) is illustrated in the block diagram form in Fig. 3.

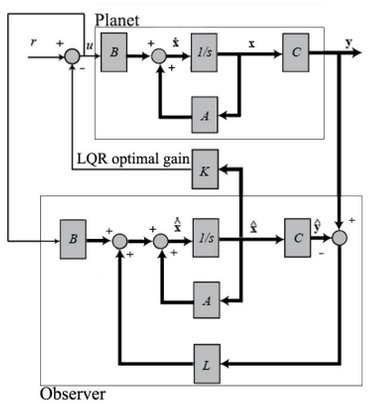


Fig. 3. Block diagram of closed-loop system with observer

2) LQR Optimal Gain

LQR control provides an approach to calculate the state feedback gain of the control system[11]. The system equation is given as

$$\dot{\mathbf{X}} = [A]\mathbf{X} + [B]\mathbf{f} \quad (41)$$

The optimal actuator input (control vector) can be determined as

$$\mathbf{f}(t) = -K_{lqr}\mathbf{X}(t) \quad (42)$$

The state feedback gain K_{lqr} is optimized to minimize the following objective function.

$$J = \int_0^\infty \left(\mathbf{X}^T\mathbf{Q}\mathbf{X} + \mathbf{f}^T\mathbf{R}\mathbf{f}\right)dt \quad (43)$$

where**Q**and **R** are the constant weighting matrices, which are real symmetric and positive-definite matrices. **Q**and **R** can be estimated by experiments; however, assigning **Q** large with regard to **R** represents that the response attenuation has more weight than the control effort and conversely. In LQR control, the minimized control gain is expressed as

$$K_{lqr} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P} \quad (44)$$

where **P**is the unique and positive solution of the well-known *Riccati* equation [11].

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{Q} = 0 \quad (45)$$

For multi degree of freedom system, the solution of (45) for P is difficult to achieve; however, it can be solved numerically. MATLAB software has a built-in command to solve *Riccati* (45) and therefore obtain$K_{lqr}$and *P*. LQR controller with observer is performed in MATLAB/SIMULINK.

VI. Results and Discussion

An Aluminum beam with three different configurations is considered here.The properties of Aluminum beam can be obtained from established references and experimental work[17]; these are tabulated in Table I.The natural frequencies and natural modes for each case study are obtained through the procedure discussed in previous section. The PID controller and LQR controller are utilized to manipulate the vibration of the beam, where controllers are designed using MATLAB/SIMULINK. The results are obtained by assuming 1 *N*impulse for the duration of 0.001 second at the tip of the beam. If the actuator voltage exceeds the maximum operating voltage of the piezoelectric material, the latter may lose its polarization and piezoelectricity property. In this regard, the input control voltage is limited to ± 90 volts, which is much less than maximum operating voltage of PZT and PVDF. Since the classical method for determining PID coefficients is not applicable here, these coefficients are obtained from available experimental, and tabulated in Table II for each case study. In the present work, the LQR controller with observer is designed and simulated to control the vibration of the system. The controllability and observability of systems were checked in first step, where state space forms of the assumed beams were all controllable and observable. The observer control method is based on pole-placement, thus, new poles should be chosen with respect to the poles of the system[11]. The most conventional approach is to obtain new poles based on experience.

TABLE I. Properties of Aluminium and Piezoelectric Materials[17, 22]

	Aluminum	***PZT5-H***	***PVDF***
Modulus of Elasticity (GPa)	71	61	2
Density (Kg/m^3)	2710	7500	1780
Width (m)	0.014	0.014	0.014
Thickness (mm)	0.66	0.75	0.11
Length(m)	0.319	-	-
Strain Constant d_{31}	-	-171×10^{-12}	23×10^{-12}
Voltage Constant g_{31}	-	0.0114	0.216

TABLE II. PID Coefficients that obtained by experiment for each case study

PID Coefficients	***Case Study I***	***Case Study II***	***Case Study III***
Proportional K_p	700	25	20
Integral K_I	675	22.5	18
Derivative K_D	900	30	24

To determine the LQR optimal gain, first, *Q* and *R* should be specified. *Q* and *R* are obtained to find the best settling time by experiment. In this study, *R* is only one number and *Q* is a

square diagonal matrix in which the entries of the main diagonal are all one and multiplied by a coefficient, *q*. *R* and *q* for each case study are presented in Table III.The results of each case study are given and discussed in the following paragraphs.

TABLE III. LQR PARAMETERS THAT OBTAINED BY EXPERIMENT FOR EACH CASE STUDY

LQR Parameters	***Case Study I***	***Case Study II***	***Case Study III***
q	50×10^{7}	10^{6}	9×10^{5}
R	10^{-4}	10^{-3}	10^{-3}

The first configuration is considered as a beam completely bonded with PVDF on the top and the bottom. The sensor is located at the tip of the beam, which is 15 *mm* long. Other PDVF patches are utilized as actuator. The length of PVDF on the upper surface is same as the length of the beam and the lower one is 304 *mm*. Schematic of case study one is illustrated in Fig. 4. The free vibration analysis of this case study is done by both analytical and numerical approaches. Properties of PVDF material is listed in Table I.

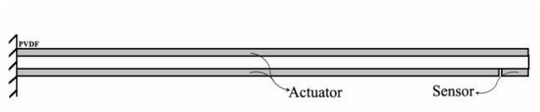
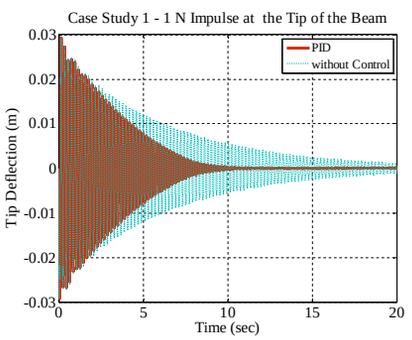


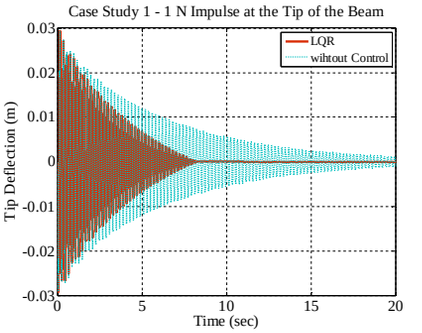
Fig. 4. Case Study one: beam with complete boned piezoelectric patch

TABLE IV. NATURAL FREQUENCIES OF CASE STUDY ONE

Natural Frequencies (Hz)	***Analytical[18]***	***FEM***	***Error%***
1st Mode	4.9501	4.9502	2.02×10^{-3}
2nd Mode	31.0197	31.0222	8.05×10^{-3}
3rd Mode	86.8618	86.8636	2.07×10^{-3}



(a)



(b)

Fig. 5. Time response of the beam, case study one (a) PID control (b) LQR control

Equation (14) is utilized to obtain the eigen-value and, consequently, the natural frequencies, via FEM and compared with analytical result [18]as shown in Table IV. The results of the FEM solution show very small error in comparison to the analytical solution.

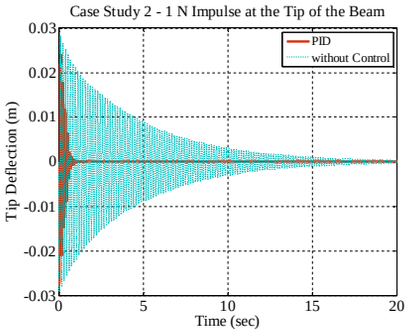
Controlled and uncontrolled responses of the case study one is shown in Fig. 5. Time response of PID and LQR controller are independently compared with uncontrolled system. In this case, LQR shows a little better settling time in comparison to PID. Table VII expresses settling time of these controllers.

In case study two, two PZT are bonded at the base of the beam, which are 38 *mm* long. PZT sensor is located right after the piezo-actuator on the top of the beam. The length of the sensor is 15 *mm*, as shown in Fig. 6. The natural frequencies of this case study are obtained by finite element methods and validated by experimental result in a previous study[17], and expressed in Table V.

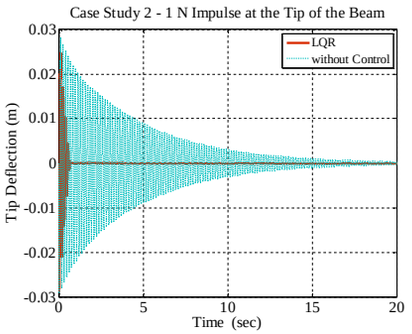
Fig. 6. Schematic of case study two

TABLE V. NATURAL FREQUENCIES OF CASE STUDY TWO COMPARED WITH AVAILABLE EXPERIMENTAL STUDY [17]

Natural Frequencies(Hz)	***Experimental Study[17]***	***FEM***	***Error%***
1st Mode	6.3	6.44	2.22
2nd Mode	38	39.45	3.81
3rd Mode	99	107.65	8.73



(a)



(b)

Fig. 7. Time response of the beam, case study two (a) PID control (b) LQR control

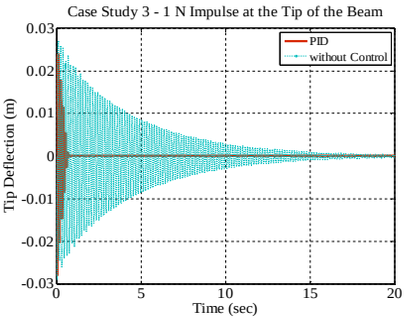
The result of PID and LQR controllers are given in Fig. 7, which exhibits significant performance. Both of them can manipulate the vibration of the system immediately; however, the results of LQR control are a little bit better than PID.

Two PZT actuators and one PVDF sensor are considered in case study three. The length of each actuator is 38 *cm* and both of them are laid on the upper surface of the beam. The sensor is located between two piezo-actuators with 2 *mm* distance from each one. The length of piezo-sensor is 15 *mm*. Fig. 8 shows the location of actuators and sensor on the beam. In this case, natural frequencies and natural modes are determined by finite element method, which are shown in Table VI.

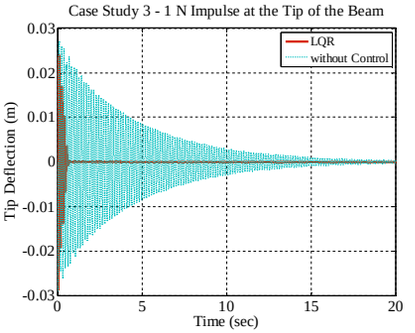
TABLE VI. NATURAL FREQUENCIES OF CASE STUDY THREE DETERMINED BY FINITE ELEMENT METHOD

Natural Frequencies	***FEM***
1st Mode	6.87
2nd Mode	37.17
3rd Mode	98.60

Fig. 8. Schematic of case study three



(a)



(b)

Fig. 9. Time response of the beam, case study three (a) PID control (b) LQR control

The controlled and uncontrolled systems of case study 3 are illustrated in Fig. 9. The performance of LQR is also shown to be better than that of PID. Three case studies are considered to evaluate which one has better performance in controlling the system. Settling time and root-mean-square (RMS) value of all of these case studies are exhibited in Table VII. The settling time is defined as the time taken for the response of the system to reach and stay in a range around the desired value, which is usually considered to be 2 to 5 percent of the final value. The RMS value is the square root of the arithmetic mean of the squared magnitudes of the waveform, which is a measure of the wave amplitudes.

As shown in the figures, it is obvious that the beam with PZT actuator has much better response in comparison to the beam bonded with PVDF. PZT can surpass the vibration better than PVDF, since it has higher stiffness and strain constant than PVDF. Stiffness has an effect on passive vibration of the beam, where in uncontrolled beam, the beam with PZT configuration has less settling time than beam bonded with PVDF, as exhibited in Table VII. Higher strain constant of actuator results in higher moment on the beam. Thus, the system can be controlled quickly with PZT actuator. The settling time and RMS of case study two and three are very close; however, case study three can surpass the vibration slightly better than case study two. By utilizing the PVDF sensor in the case study three, the weight of the system reduces in comparison to case study 2. Both controllers provide significant vibration suppression. For all of these case studies, LQR shows a slightly better settling time performance than PID. Case study three with LQR controller shows the best result in this study. Settling time and RMS value of it is less than the others and the weight of the beam is acceptable.

TABLE VII. Comparison of case studies and controllers

Response of the system		*Case study 1*	*Case study 2*	*Case Study 3*
Without Control	Settling Time (s)	24.29	17.70	17.43
	RMS × 10^{-5}	483	400	390
PID Control	Settling Time (s)	9.34	0.88	0.68
	RMS × 10^{-5}	414	138	135
LQR Control	Settling Time (s)	7.87	0.66	0.64
	RMS × 10^{-5}	385	134	129
Weight (g) of the Beam		9.7	10.2	9.8

VII. CONCLUSION

In this investigation, the Vibration Analysis of a Cantilevered Beam with Piezoelectric Actuator as a Controllable Elastic Structure has been elaborated by solving generic beam problem and three configurations of beam-piezoelectric elements composite beams. The active control of the vibration of an Aluminum flexible beam is studied through the bonding of the piezo-electric elements onto a beam as a smart material. Two widely utilized piezoelectric materials, PZT and PVDF, are considered for controlling the system. The dynamic of the beam bonded with piezoelectric material patches is investigated employing Euler-Bernoulli beam theory. The equation of motion of the beam is obtained through the use of Hamilton's principle. Three different configurations for the beam are selected for comparison and to gain insight into the issue. Free vibration of the first case study is determined by both analytical and finite element method. The second case study is solved by finite element and compared to an experimental work. The third one is determined by finite element only. To design the controller, state space form of each beam is formed. Two straightforward and convenient control method are investigated, PID and LQR with observer. The results thus obtained exhibit the effectiveness of the various controllers chosen and give a beneficial insight on the utilization of the stability and control of flexible beam structure.

Acknowledgments

The authors would like to thank Universiti Putra Malaysia (UPM) for granting Research University Grant Scheme (RUGS) No.9378200, and the ministry of higher education ERGS: 5527088 ; FRGS:5524250 under which the present research is carried out.

References

[1] M. Azadi, S. Fazelzadeh, M. Eghtesad, and E. Azadi, "Vibration suppression and adaptive-robust control of a smart flexible satellite with three axes maneuvering," Acta Astronautica, vol. 69, pp. 307-322, 2011.

[2] P. Gasbarri, M. Sabatini, N. Leonangeli, and G. B. Palmerini, "Flexibility issues in discrete on–off actuated spacecraft: Numerical and experimental tests," Acta Astronautica, vol. 101, pp. 81-97, 2014.

[3] M. Sabatini, P. Gasbarri, R. Monti, and G. B. Palmerini, "Vibration control of a flexible space manipulator during on orbit operations," Acta astronautica, vol. 73, pp. 109-121, 2012.

[4] S. Narayanan and V. Balamurugan, "Finite element modelling of piezolaminated smart structures for active vibration control with distributed sensors and actuators," Journal of sound and vibration, vol. 262, pp. 529-562, 2003.

[5] S. Xu and T. Koko, "Finite element analysis and design of actively controlled piezoelectric smart structures," Finite elements in analysis and design, vol. 40, pp. 241-262, 2004.

[6] C. C. Fuller, S. Elliott, and P. A. Nelson, Active control of vibration: Academic Press, 1996.

[7] N. Jalili, "Piezoelectric-based vibration control," From Macro to Micro/Nano Scale Systems Springer, 2010.

[8] H. Baruh, Analytical dynamics: WCB/McGraw-Hill Boston, 1999.

[9] L. Meirovitch, Principles and techniques of vibrations vol. 1: Prentice Hall New Jersey, 1997.

[10] R. Alkhatib and M. Golnaraghi, "Active structural vibration control: a review," Shock and Vibration Digest, vol. 35, p. 367, 2003.

[11] K. Ogata and Y. Yang, "Modern control engineering," 1970.

[12] H. Djojodihardjo, "Computational simulation for analysis and synthesis of impact resilient structure," Acta Astronautica, vol. 91, pp. 283-301, 2013.

[13] H. Djojodihardjo and A.Shokrani, "Generic Study And Finite Element Analysis Of Impact Loading On Elastic Panel Structure, Paper IAC-10.C2.6.2," presented at the 61th International Astronautical Congress, Prague, The Czech Republic, 2010.

[14] H. Djojodihardjo, P.M.Ng, and L.K.Soo, "Analysis And Simulation Of Impact Loading On Elastic Beam Structure with Case Studies, Paper IAC-09-C2.6.01," presented at the 60th International Astronautical Congress, Daejeon, Republic Of Korea, 2009.

[15] H. Djojodihardjo and I. Safari, "BEM-FEM Coupling For Acoustic Effects On Aeroelastic Stability Of Structures," CMES: Computer Modeling in Engineering & Sciences, vol. 91, pp. 205-234, 2013.

[16] M. Jafari, H. Djojodihardjo, and K. Arifin Ahmad, "Vibration Analysis of a Cantilevered Beam with Spring Loading at the Tip as a Generic Elastic Structure," presented at the Aerotech, Kuala Lumpur, Malaysia, 2014.

[17] S.-Y. Hong, "Active vibration control of adaptive flexible structures using piezoelectric smart sensors and actuators," 1992.

[18] H. Djojodihardjo and M. Jafari, "Vibration analysis of a cantilevered beam with piezoelectric actuator as a controllable elastic structure," presented at the accepted in 65 th International Astronautical Congress, Toronto, Canada., 2014.

[19] L. Meirovitch, Computational methods in structural dynamics: Springer, 1980.

[20] S. S. Rao, Vibration of continuous systems: John Wiley & Sons, 2007.

[21] A. K. Chopra, Dynamics of structures vol. 3: Prentice Hall New Jersey, 1995.

[22] A. Yousefi-Koma, "Active vibration control of smart structures using piezoelements," Carleton University, 1997.

A Parallel Robotic Mechanism Replacing a Machine Bed for Micro-Machining

Zareena Kausar Muhammad Asad Irshad Shaheriyar Shahid
Department of Mechatronics Engineering, Air University, Islamabad, Pakistan
corresponding author: zareena.kausar@mail.au.edu.pk

***Abstract*— Micromachining is meant for small accurate and precise parts production for many industrial applications. These parts vary from very simple to complicated shapes. As applications grow in complexity and shrink in size, machines need to be designed to meet the desired precision and accuracy. In this paper it is proposed to give a six degree of freedom to the bed of a machine instead to tool box which carry a heavy load. A parallel mechanism consists of six legs of variable lengths is proposed for the machine bed. The objective is to enhance accuracy and precision in micromachining in comparison to conventional machines. The paper presents a design of the machine bed, the kinematics, dynamics and a sim-mechanics model of the bed. Simulation results verify the working of a 6-DOF machine bed for micromachining.**

Keywords*— Parallel Mechanism; Robotics; Micromachining; Machine bed; Manufacturing.*

I. INTRODUCTION

Micromachining is a process of fabrication of microstructures using geometrically determined cutting edges techniques like, drilling, milling, de-burring and slotting. There are few limitations about geometry of work piece and also 3-D structures that are manufactured. The components produced on micro machine tools are of a size in a range of centimeter or millimeter [1]. Different machine tools [2, 3] are developed for micromachining. One of such a machine tool is developed at the Fraunhofer institute, named as 'MiniMill', for production sciences in Aachen [4].This is a high precision machine tool featuring one square meter of floor space. This machine tool focuses on high precision and thermal stability for achieving high accuracy [5]. Another device for micro production is developed at the National Autonomous University of Mexico [6] for the machining of micro structures of different work pieces with an accuracy of 50 micrometer. The fact that the micro-machines operate on very small parts using very small tools means that they are vulnerable to external and internal vibrations. In the presence of vibrations the cuts made and the parts created will be dimensionally inaccurate and relatively imprecise. To solve the problem, this research proposes to restrict the tool motion in a single dimension and instead replace the existing beds with a 6-DOF manipulator.

There are several types of manipulators serial and parallel. We propose a parallel manipulator, since it offers a high stiffness, high dynamics and prevents the accumulation of position errors. A parallel manipulator has received a lot of attention recently in the industry and the robotic community due to its high accuracy, high speed and high load capacity in comparison to conventional serial manipulators. Several serial links control a platform simultaneously to make a parallel manipulator. The most known configuration of parallel manipulator is hexapod used for multiple applications.

The proposed parallel mechanism for mechanical micromachining has same concept as a Stewart platform does [7]. The tool is fixed and all type of motions is being created by bed. With other conventional robots, adjustments to these stations often require shutting down the entire line to reset tool for a different part to be done. With the flexibility of a Parallel Mechanism tool might never need to be moved again. Imagine a process where, when a new part is accessible into the station, motion of platform can align each part individually.

The mechanism of the movable bed is presented in next section. Kinematics of the mechanism is described in the form of inverse kinematics along with the position and velocity analysis. A dynamics model of the mechanism is developed in section-IV and the Sim-mechanic model follows it. The model is used to simulate the results which are presented as results.

II. MECHANISM DESIGN AND ANALYSIS

Mechanical design is presented which may be used to determine the performance measures such as, positioning accuracy, repeatability and freedom from vibration. The Stewart platform mechanism is a parallel kinematic structure that can be used as a basis for controlled motion with 6 degrees of freedom, such as manufacturing process and precise manipulative tasks. Stewart platform are resourceful and effective solutions to complex motion applications that require high load capacity and accuracy in up to six independent axes [8].

The proposed Stewart platform like manipulator for this research consists of a rigid moving plate, connected to a fixed base through six independent legs [9]. These legs are identical kinematics chains, coupling the moveable upper platform and the fixed lower platform. The position and orientation of the end-effector is controlled by the lengths of six linear actuators which may be driven by servo motors that connect it to the base. At the base end, each actuator is connected by a two-degree-of-freedom universal joint. At the end-effector, each actuator is attached with a three-degree-of-freedom ball-and-socket joint. Thus, length of the legs is variable and they can be controlled separately to control the motion of the moving

platform. It exhibits characteristics of closed-loop mechanisms.

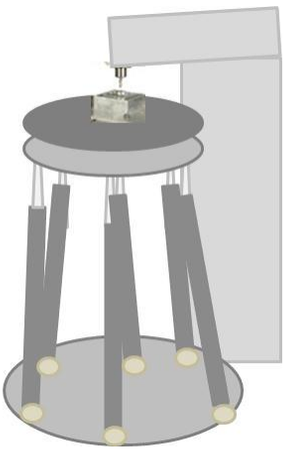


Figure1: Schematic model of the system

A. Workspace

The workspace of the manipulator is defined in two ways [9]: *reachable workspace* and *dexterous workspace*. The reachable workspace is the collection of all points $\{x\ y\ z\}^T$ that can be reached by the manipulator in any orientation. The dexterous workspace is the collection of all points that can be reached by manipulator in all orientations consequently; the dexterous workspace is a subset of reachable workspace. For the parallel manipulator the dexterous workspace is null, since this cannot reach all orientations at any position in the reachable workspace. Here we define the dexterous workspace as collection of all points that can be reached by the manipulator in all orientations. Consequently, the dexterous work space is a subset of reachable workspace. The workspace of the parallel manipulator is described based on assumptions that there is no actuator limitation, leg interference and singularities.

III. KINEMATICS ANALYSIS

To analyze kinematics of the mechanism, the manipulator motion is studied without regard to the forces that cause it. Within this analysis the position and the velocityof the manipulator are studied. This study of kinematics of manipulators also refers to all the geometrical and time-based properties of motion [5]. Here in this article, we considered inverse kinematics of the parallel manipulator. The inverse kinematics analysis of parallel manipulators gives the actuator displacement from the given position and orientation of a movable platform. The solution is unique, and can be simply determined. The forward kinematics analysis determines the position and orientation of moveable platform for the given actuator displacement which is not recommended for this study. The reason was that the solution becomes complicated because the problem involves system of higher order non-linear equations, as in [10].

A. Inverse kinematics

In inverse kinematics analysis given the desired position and orientation, we computed the set of joint angles. Inverse kinematics derived in this article is based on [1, 11]. The coordinate frame *A*(x, y, z) is attached to the fixed base and the coordinate frame *B* (x´, y´, z´) is attached to a moving platform. Furthermore, a local coordinate frame $C(x_i, y_i, z_i)$ is attached to each limb such that its origin is at point A_i , the z_i axis points from A_i to B_i, the y_i axis is parallel to the cross product of two unit vectors defined along the z_i and z axis and the x_i axis is defined by right-hand rule. For the convenience, the origin of the frame B is located at mass center P of the moving platform. The location of moving platform is described by a position vector, p and rotation matrix $^{A}\mathbf{R_B}$ as shown in Fig.2.

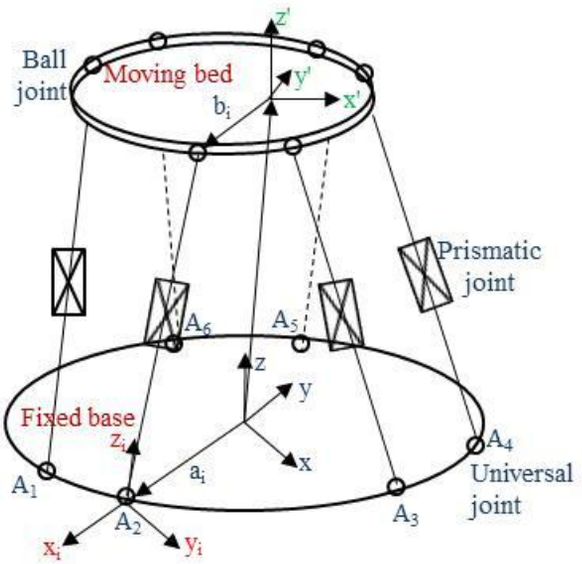


Figure 2: Schematic diagram of the parallel manipulator bed

The rotation matrix is defined by roll, pitch and yaw angles. These are a rotation of ϕ_x about the fixed x-axis, a rotation of ϕ_y about the fixed y-axis and a rotation of ϕ_z about the fixed z-axis. The rotation matrix is (1).

$$^{A}\boldsymbol{R}_{\boldsymbol{B}}=\begin{pmatrix} C\phi_z C\phi_y & C\phi_z S\phi_y S\phi_x - S\phi_z C\phi_x & C\phi_z S\phi_y C\phi_x + S\phi_z S\phi_x \\ S\phi_z S\phi_y & S\phi_z S\phi_y S\phi_x + C\phi_z C\phi_x & S\phi_z S\phi_y C\phi_x - C\phi_z S\phi_x \\ -S\phi_y & C\phi_y S\phi_x & C\phi_y C\phi_x \end{pmatrix} \quad (1)$$

Where $S\phi_i = \sin(\phi_i)$, $C\phi_i = \cos(\phi_i)$, $i = x, y, z$

B. Position analysis

From Fig. 2, a vector loop equation can be written for each limb as

$$a_i + d_i s_i = p + b_i \quad (2)$$

Where $a_i = [\ a_{ix}, a_{iy}, a_{iz}\ ]$ is a position vector of a ball joint A_i with respect to the fixed frame *A*

b_i =[b_{ix}, b_{iy}, b_{iz}] is a position vector of ball joint B_i with respect to fixed frame A
s_i = unit vector from A_i to B_i

Eq. (2) may be written for di as

$$d_i = |p + b_i - a_i| \qquad \text{Where i = 1, 2…6}$$

Solving equation 2 for s_i,

$$s_i = (p + b_i - a_i)/d_i \tag{3}$$

Since each limb is connected to the fixed base by a universal joint, its orientation with respect to the fixed base may be conveniently described by two Euler angles [11]. As shown in Fig. 3, the local coordinate frame of the *i* th limb can be thought of as a rotation of ϕ_i about the z_i axis resulting in a (u, v, w) followed by another rotation of θ_i about the rotated v axis. Hence the rotation matrix for the *i* th written as

$$^{A}\boldsymbol{R_i} = \begin{pmatrix} C\theta_i C\theta_i & -S\theta_i & C\theta_i S\theta_i \\ S\theta_i C\theta_i & C\theta_i & S\theta_i S\theta_i \\ -S\theta_i & 0 & C\theta_i \end{pmatrix} \tag{4}$$

Where i=1, 2,…6

Equating the third column of $^{A}\boldsymbol{R_i}$ to S_i gives

$$S_i = \begin{pmatrix} C\,\theta_i S\theta_i \\ S\theta_i S\theta_i \\ C\theta_i \end{pmatrix} \tag{5}$$

Where i=1, 2,… 6

Solving equation 5 gives

$$\theta_i = a\tan2\left(\sqrt{s_{ix}^2 + s_{iy}^2}\,, s_{iz}\right) \tag{6}$$

$$\phi_i = a\tan 2(s_{iy}/s_{\theta i}, s_{ix}/s_{\theta i})$$

Where s_{ix}, s_{iy}, and s_{iz} are the x, y and z components of s_i.
Eq. (5) and (6) determine the direction and Euler angles of the *i* th limb in term of the moving platform location.

C. Velocity analysis

Here we define the linear and angular velocities of all moving links from the independent Cartesian velocities of the platform v_{px}, v_{py}, v_{pz}, ω_x, ω_y , ω_z. The latter three scalar quantities are the components of the angular velocity of the moving platform ω_p. Let $p = [p_x \;\; p_y \;\; p_z]^T$ is a position vector of moving platform $\dot{p} = V = [v_{px} \;\; v_{py} \;\; v_{pz}]^T$ is a velocity vector of moving platform $\omega = [\omega_x \;\; \omega_y \;\; \omega_z]^T = [\phi_x' \;\; \phi_y' \;\; \phi_z']^T$ is an angular velocity vector of moving platform Now $\mathbf{b_i}$ is written in term of angular velocity vector of the ith leg ω_i as in (7).

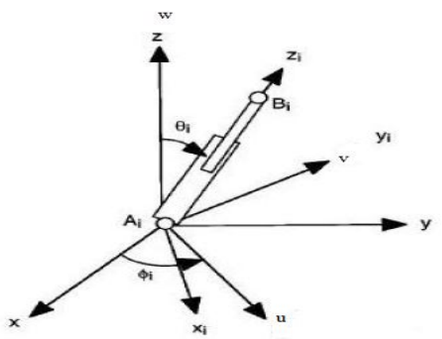


Figure 3: Euler angles of a leg

$$b_i' = d_{ir'} + \omega_i \times d_i \tag{7}$$

Where *i* = 1, 2, 3, 4, 5, 6

and

$$d_{ir}' = \begin{bmatrix} di'\cos\phi_i \sin\theta_i \\ di' \sin\phi_i \sin\phi_i \\ di'\cos\theta_i \end{bmatrix}, \quad d_i = \begin{bmatrix} di\cos\phi_i \sin\theta_i \\ di\sin\phi_i \sin\theta_i \\ di\cos\theta_i \end{bmatrix},$$

$$\omega = \begin{bmatrix} -\theta i' \sin\theta_i \\ \theta i' \cos\theta_i \\ \theta i' \end{bmatrix}$$

Eq. (7) is expressed in matrix form as (8).

$$C_{bi}\lambda_{bi} = b_i' \tag{8}$$

$$C_{bi} = \begin{bmatrix} \cos\phi_i \sin\theta_i & -di\sin\phi_i \sin\phi_i & di\cos\phi_i \cos\theta_i \\ \sin\phi_i \sin\phi_i & di\cos\phi_i \sin\theta_i & di\sin\phi_i \cos\theta_i \\ \cos\theta_i & 0 & -di\sin\theta_i \end{bmatrix},$$

$$\lambda_{bi} = \begin{bmatrix} di' \\ \phi_i' \\ \theta_i, \end{bmatrix}$$

Eq. 8 is expression of angular velocity and can be solved for λ_{bi} which leads to the determination of $di', \phi i' \text{ and } \theta i'$. Once these quantities are known, the computation of velocities of *i*th leg is straightforward.

$$\lambda_{bi} = (C_{bi})^{-1} b_i' \tag{9}$$

Where i = 1, 2,… 6

IV. Dynamic Analysis

In order to develop the dynamic equations of the Stewart platform manipulator, the entire system is divided in two parts: one is moving platform and the other is legs. The kinetic and potential energies for the both of these parts are figured out and the dynamic equations are derived using these energies [12].

A. Kinetic and potential energies of moving platform

The kinetic energy of moving platform has translation and rotational motion energies about three orthogonal axis (X, Y, Z). The translation energy occurring because of the translation motion of the center of mass of moving platform and is defined by (12).

$$K_{(trans)}=0.5m[\phi_x'^2\phi_y'^2\phi_z'^2] \quad (10)$$

Where m is mass of the moving platform.

For rotational motion of the moving platform around its center of mass, the rotational kinetic energy is as given in (11).

$$K_{(rot)}=0.5[b_i{}^T I_{mf} b_i'] \quad (11)$$

In (11) I_{mf} and bi´ are the rotational inertia mass and angular velocity of the moving platform, respectively. Angular velocity bi ´ is defined in eq. (8) and I_{mf} is as following:

$$I_{mf}=\begin{bmatrix} Ix & 0 & 0 \\ 0 & Iy & 0 \\ 0 & 0 & Iz \end{bmatrix}$$

As a result, the total kinetic energy of moving platform is given by

$$K=K_{(trans)}+K_{(rot)}=\frac{1}{2}m[\phi_x'^2\phi_y'^2\phi_z'^2]+\frac{1}{2}[b_i{}^T I_{mf} b_i'] \quad (12)$$

The potential energy of moving platform is given in (13).

$$P=[0\ 0\ m_{up}g\ 0\ 0\ 0]\begin{bmatrix} Px \\ Py \\ pz \\ \phi x \\ \phi y \\ \phi z \end{bmatrix} \quad (13)$$

Where g is a gravitational force.

B. Kinetic and potential energies of legs

Each leg consists of two parts: the moving part connected to moving platform through ball joint and the fixed part connected to fixed base through universal joint. As shown in Fig. 4, the center of mass is G_i for each part of leg (where i=1,2,…6) and G_{1i} denotes the center of mass of fixed part. $m_1$and l_1 are the mass and the length of the fixed part and δ is the distance between B_i and G_i. For the moving part of leg, G_{2i} express its center of mass. M_2 and l_2 are the mass and length of the moving part.

The length of leg is assumed to be constant. The total kinetic energy of a leg, L_i , is calculated as (14).

$$K_{Li}=K_{Li(trans)}+K_{Li(rot)} =\frac{1}{2}(m_1+m_2)[V_{Tj}^T\ h_i\ V_{Tj} - L_i'K_iL_i'] \quad (14)$$

Where $K_i=hi-(m_2/m_1-m_2)^2$, $hi=(I/Li+m_2/m_1+m_2)^2$ and $I=1/m_1+m_2\ (\&m_1l_1-1/2\ m_2l_2)$

Ui is the unit vector along the axis of the leg *(Li).* By using this vector calculation of leg velocity is as $Li'=V_{Tj}Ui$

Total potential energy of leg is

$$P_{legs}=(m_1+m_2)\,g\sum_{i=0}^{n}[\,I(\frac{1}{L_{2i}}+\frac{1}{L2_{i-1}})+\frac{2m_2}{m_1+m_2}](P_z+Z_{Tj}) \quad (15)$$

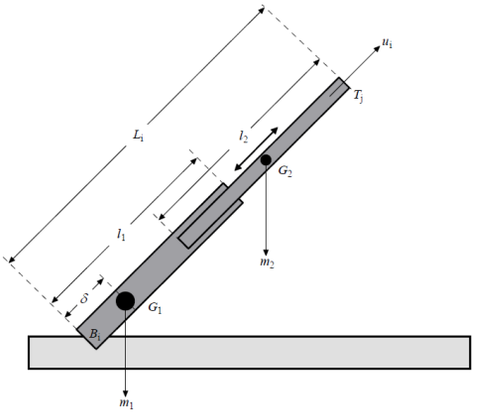


Figure 4: A leg of the parallel manipulator bed

C. Dynamic Equations of Motion of Parallel Manipulator Bed

The generalized dynamic model of the Stewart platform [13] is given as (16).

$$T=J_p^T F_p+\sum_{i=1}^{m}(\frac{\partial P_i}{\partial P_a})T(Hi) \quad (16)$$

With:

F_p is the total forces and momentum on the platform

J_p is the (6 x n) kinematic Jacobian matrix of the robot

The Jacobian matrix is computed following [10], which writes the platform velocity (translational and angular) as function of active joint velocities $V_p=J_p\dot{p_a}$.

H_i is the dynamic model of *i*th leg, it is a function of ($\boldsymbol{p}$ $\dot{\boldsymbol{p}}$ $\ddot{\boldsymbol{p}}$)which can be obtained in terms of the platform location, velocity and acceleration, using the inverse kinematic models of the legs. Any one of methods presented in [14-16] may be used to calculate these elements.

For a general case, where the platform has 6 degrees of freedom Fp is calculated using Newton-Euler equation [14, 15].

$$F_P=Is\begin{bmatrix} \dot{V}-g \\ \dot{\omega} \end{bmatrix}+\begin{bmatrix} \omega\ x\ (\omega\ x\ Msp) \\ \omega\ x\ (Ip\ x\ \omega) \end{bmatrix} \quad (17)$$

Where

Is is (6×6) spatial inertia matrix of the platform.

Ip(3×3) inertia matrix of the platform around the origin of the platform

M*sp* is (3×1) vector of first moments of the platform around the origin of the platform.

V the velocity vector of moving platform

V' the acceleration of moving platform

ω the angular velocity vector of moving platform

ω' angular acceleration of moving platform

V. Modeling, Control and Simmulation

In this section a model of the system described in a software, controller used for simulations and simulation results are presented.

A. The Model

The model is presented in Simmechanics, a tool of Matlab®. The model shown in Fig.5 is divided into subsystems: Plant subsystem; Controller subsystem. The plant subsystem consists of the parallel manipulator bed along with necessary actuators and sensors while a controller subsystem controls the motion of the parallel manipulator bed through a predefined motion profile with actuation signals. The controller keeps the actual motion close to the reference motion via sensor-actuator feedback.

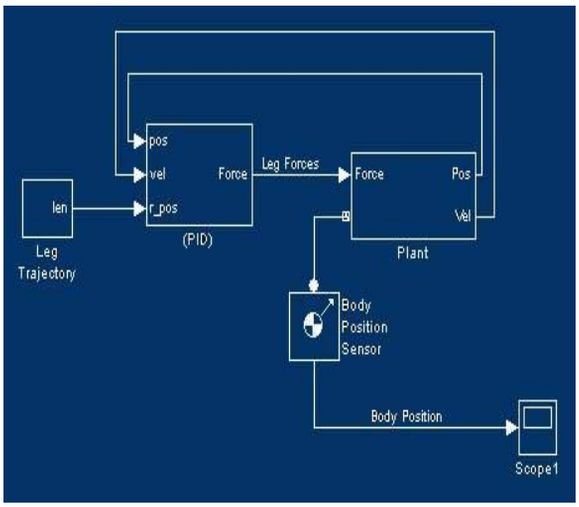


Figure 5: Block Diagram of Simechanic Model of the bed

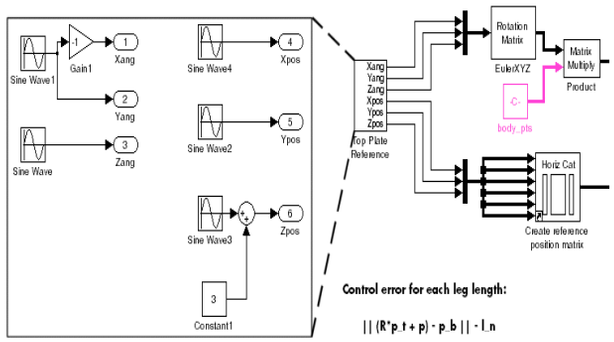
A reference trajectory is given to the controller through leg trajectory block. The reference trajectory provided for use in the simulations for this study is a sinusoidal function of time. A sinusoidal trajectory is selected, shown in Fig.6, to define the rotational as well as translational degrees of freedom. Any kind of trajectory can be designed and implemented for this sub-subsystem.

Figure 6: Subsystem of Desired Trajectory Block

B. The Controller

A PID controller is used to track the reference trajectory. PID stands for proportional, derivative and integral controller. The selection of PID for this research was due to its simplicity in implementation. The simplest implementation of a trajectory control is to apply forces to the plant proportional to the motion error. The error is measured through Joint Sensors for which separate joint sensor blocks are used in the model.

A PID control law is a linear combination of a variable detected by a sensor, its time integral, and its first derivative. The PID controller uses the leg position errors E_r and their



integrals and velocities for the proposed parallel manipulator bed. The control law for an r^{th} leg has the form:

$$F_{act,r} = K_p E_r + K_i \int_0^t E_r dt + K_d (dE_r/dt)$$

The controller applies the actuating force $F_{act,r}$ along the leg :

- If E_r is positive, the leg is too short, and $F_{act,r}$ is positive (expandive).
- If E_r is negative, the leg is too long, and $F_{act,r}$ is negative (compressive).
- If E_r is zero, the leg has exactly the desired length, and $F_{act,r}$ is zero.

The real, nonnegative K_p, K_i, and K_d are, respectively, the proportional, integral, and derivative gains that modulate the feedback sensor signals in the control law:

- The first term is proportional to the instantaneous leg position error or deviation from reference.
- The second term is proportional to the integral of the leg position error.
- The third term is proportional to the derivative of the leg position error.

The result is $F_{act,r}$, the actuator force applied by the controller to the legs. The proportional, integral, and derivative terms tend to make the leg's top attachment points $p_{t,r}$ follow the reference trajectories by suppressing the motion error.

C. Simulation Results

For simulations the leg lengths were assumed to be same initially. The reference trajectory is taken as sinusoidal wave and corresponding block is added to the model. This is the reference trajectory each point on the moving bed is desired to follow. The dynamic simulations were performed whereas a couple of clips are shown in Fig. 7a and 7b.

The difference in figures depicts the stretch or compression in leg lengths which the robotic mechanism performed to achieve the desired points on the moving bed. The blue traces on the top of the bed shows the traces of path the bed followed. These results confirms that the proposed mechanism for the bed is capable of following the complex trajectory expected as the contours on the work piece to be machined.

The error in the positions is plotted as shown in Fig.8. These are given in mm. The maximum error reaches to 5 micrometer which verifies the suitability of the design for micro machining operations. In order to increase the accuracy

and precision of the system the control algorithm may be improved.

VI. Conclusion

In this research we presented design, the kinematics and dynamic analysis of a parallel robotic mechanism for a micro machining bed. We used MATLAB Simmechanics for the dynamic analysis of parallel system. The model is simulated in order to verify the objective of the proposed mechanism. Simulations show promising results. The computed modeling error depicts the high accuracy of the developed model. It is concluded that the verified model of the proposed mechanism may be used for bed control and design purposes for micromachining. In future the mechanism will be developed yo verify the results in real time.

References

[1] Schubert A, Neugebauer R, Schulz B (2007) System concept and innovative component design for ultraprecision for assembly processes. Towards Synth Micro-Nano Syst Part 2

[2] Wulfsberg JP, Redlich T, Kohrs P (2010) Square foot manufacturing: a new production concept for micro manufacturing. Prod Eng 4(1):75–83

[3] Y Tanaka M (2001) Development of desktop machining microfactory. Riken Rev Nr. 34, S. 46–49 Journal Code J0877A.I,Internet:http://sciencelinks.jp/jast/article/200115/000020011501 A0520776.php. (14.1.2013

[4] Klar R, Brecher C, Wenzel C (2008) Development of a dynamic high precision compact milling machine. In: Proceedings of euspen international conference, Zurich/CH

[5] Wulfsberg JP, Grimske S, Kohrs P, Kong N (2010) Kleine Werkzeugmaschinen fu¨r kleine Werkstu¨cke - Zielstellungen und Vorgehensweise des DFG-Schwerpunktprogramms 1476. Wt Werkstattechnik online, Jahrgang 100, vol 11/12, pp 886–891

[6] Kussul E, Baidyk T, Ruiz-Huerta L, Caballero-Ruiz A, Velasco G, Kasatkina L (2002) Development of micromachine tool prototypes for microfactories. J Micromech Microeng 12:795–812

[7] Yao Wang. ” Symbolic Kinematics and Dynamics Analysis and Control of a General Stewart Parallel Manipulator” Department of Mechanical and Aerospace Engineering. State University of New York at Buffalo Buffalo, New York 14260, September 2008.

[8] http://en.wikipedia.org/wiki/Stewart_platform

[9] http://prsco.com/uofcalgary.html,http://prsco.com/uofalberta.html

[10] Domagoj Jakobović, Leo Budin . “Forward Kinematics of a Stewart Platform Mechanism” Faculty of Electrical Engineering and Computing Faculty of Electrical Engineering and Computing.Unska 3, 10000 Zagreb Unska 3, 10000 Zagreb Croatia.

[11] J.J. Craig, Introduction to Robotics: Mechanics and Control, Addison Wesley Publishing Co Reading, MA, 1989.

[12] Yang, D.C. and Lee, T. W., "Feasibility Study of a Platform Type of Robotic Manipulators from a Kinematic Viewpoint," Tmns. *ASME Journal of Mechanisms, Transmissions, and Automation in Design*, Vol. 106. pp. 191-198, June 1984.

[13] L.-w. Tsai, "Solving the Inverse Dynamics of a Stewart-Gough Manipulator by the Principle of Virtual Work," Journal of Mechanical Design, vol. 122, 2000.

[14] Serdar Kucuk "Serial and Parallel Robot Manipulators - Kinematics, Dynamics, Control and Optimization", Chapter 2 Dynamic Modeling and Simulation of Stewart Platform By Zafer Bingul and Oguzhan Karahan ,ISBN 978-953-51-0437-7, Published: March 30, 2012.

[15] Wisama KHALIL, Ouarda IBRAHIM “General Solution for the Dynamic Modeling of Parallel Robots” Journal of Intelligent and Robotic Systems, Vol. 49, pp. 19-37, 2007.

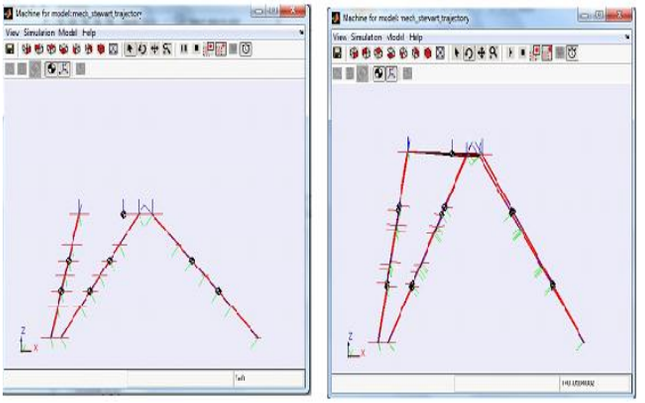


Figure 7a: Motion of upper platform according to reference trajectory

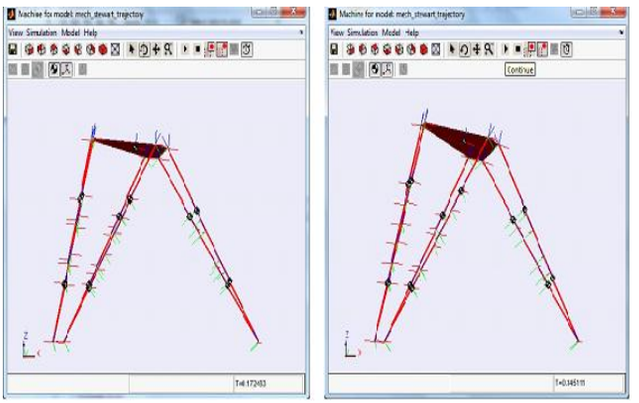


Figure 7b: Motion of upper platform according to reference trajectory

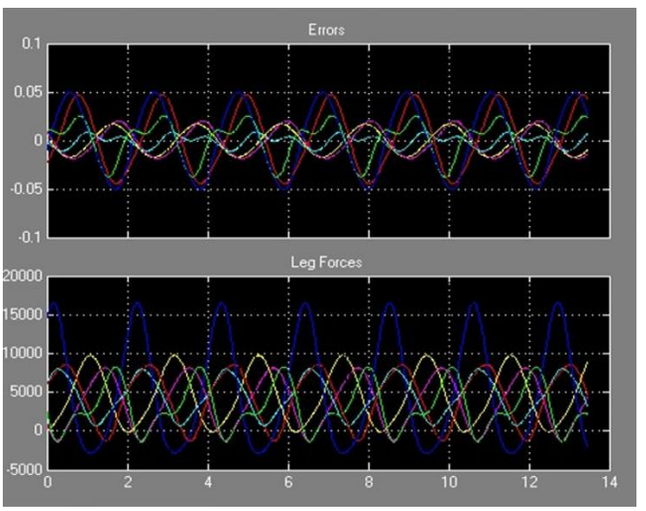


Fig.8 Showing errors and position of body with leg forces

[16] Merlet J.-P.: Parallel robots, Kluwer Academic Publ., Dordrecht, The Netherland (2000).

Improving Population Diversity in Parallelization of a Real-Coded Genetic Algorithm Using MapReduce

Takuto Enomoto
Shibaura Institute of Technology,
Department of Information Science and Engineering
3-7-5 Toyosu, Koto Ward, Tokyo 135-8548, Japan
l10020@shibaura-it.ac.jp

Masaomi Kimura
Shibaura Institute of Technology
Department of Information Science and Engineering
3-7-5 Toyosu, Koto Ward, Tokyo 135-8548, Japan
masaomi@sic.shibaura-it.ac.jp

***Abstract*—Genetic algorithms (GAs) whose solution space has many dimensions require a long execution time. In a previous study, although a GA was parallelized using MapReduce to reduce execution time, individuals were divided into subpopulations assigned to islands. Therefore, the diversity (search efficiency) of each population in subpopulations was less than the original GA, which has a single population. In this paper, we propose a method to improve individual diversity in parallelization of a GA using MapReduce. We have focused on migration, which is a method to improve population diversity by exchanging individuals among subpopulations. However, the MapReduce model does not allow such synchronization during parallel processing. Therefore, we realize migration during Shuffle tasks. We evaluated the proposed method by comparing it with a GA that has a single population and the GA proposed in the previous study. The results show improvement in solution accuracy.**

Keywords – Genetic Algorithm, Real-coded Genetic Algorithm, migration, MapReduce

I. INTRODUCTION

Combinatorial optimization problems are difficult to solve because of their significant cost. A genetic algorithm (GA) is a typical and powerful way to solve such problems [1][2][3]. A GA expresses solution candidates, i.e., *individuals*, as binary arrays and searches the solution space by their crossover, mutation, and selection.

This metaheuristic method is an effective way to search for a good solution that cannot be found analytically. However, if it is applied to a problem with a solution space that has a huge number of dimensions, the individuals must have the corresponding number of chromosomes.

Numerous studies have improved GAs to search a huge solution space. Such improvements include the Real-coded GA (RCGA) and the parallel GA [4]. RCGA is a GA whose solution candidates are expressed as arrays with real-value elements as chromosomes. The parallel GA reduces the total execution time by employing parallelization techniques.

There are several approaches to parallelize GAs. Typical methods to parallelize GAs are as follows [4]:

- Parallelizing individual evaluation
- Dividing the population into subpopulations and parallelizing (island model).

The former parallelizes a part of a GA, and the latter parallelizes the entire GA. The latter has a great advantage over the former because it can shorten the execution time and prevent convergence to a localized solution.

There are several techniques that can parallelize GAs. We introduce the following three techniques for parallelization.

- MapReduce [5]
- Message Passing Interface (MPI) [6]
- Bulk Synchronous Parallel (BSP) [7]

MapReduce is a parallelization model for large-size data that is effective for sequential processing such as batch processing. This model is scalable and has fault resistance. BSP is another computational model to perform a parallel algorithm. BSP realizes distributed processing and inter-task communication by iterating a processing unit called a SuperStep. MPI is a standard library for parallel programing that realizes an exchange of messages in each task. MPI can perform synchronous communication and/or asynchronous communication for each computer using functions for transmission and reception.

BSP and MPI can perform inter-task communication, which reoccurs (i.e., redoes) if obstacles are experienced in a given task. A GA with a solution space that has many dimensions requires a significant execution time. Assuming that obstacles constantly occur, a long execution time is disadvantageous because redoes can occur many times. Therefore, BSP and MPI are not suitable for our target. For this reason, we have focused on MapReduce, which has strong fault tolerance.

MapReduce is effective for sequential processing, such as batch processing, and excels at scalability and fault resistance. Its implementation, Hadoop [8][9], is an open source platform that has been studied in the context of the parallel GA by Keco et al.[10]. However, in their study, individuals were divided into subpopulations that were assigned to islands (i.e., computers used for parallelization). The number of individuals in each island is less than the original population; therefore, the population diversity is also less than the original GA. This degrades search performance.

Migration is one solution for this problem. Migration improves the population diversity by exchanging individuals among subpopulations. However, a method for migrating

individuals that is suitable for RCGA employing MapReduce has not been clarified.

Hurun et al. proposed a similar approach [11]. In their study, individuals were created or read from HDFS in the Map step, and GA operation was executed in the Reduce step. However, their method has three problems. First, since the keys were assigned in Reduce phase, it could not find the best individual effectively. Second, initial individuals were unnecessarily read from HDFS, since created individuals were saved to HDFS before GA was applied to them. Third, reduction of transferred data size during Shuffle step was not taken into account as we will discuss in Section II, A and Section II, B.

We propose a methodology that iteratively performs crossover, mutation, and selection of individuals in Map steps and migrates them in Shuffle steps.

II. METHODS

MapReduce is a programming model used to divide and process large-scale data. MapReduce consists of three steps: Map, Shuffle, and Reduce (Fig. 1).

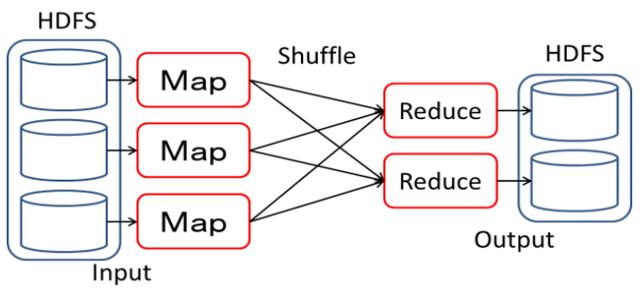


Fig. 1. MapReduce

All Map tasks receive inputs that are split from a data store, *e.g.,* HDFS for Hadoop, and execute parallel processing. After parallel processing is completed, the results are output to Shuffle tasks. Just prior to output, a key is assigned to each of the results to associate it to a Reduce task. Shuffle tasks receive results and their keys from the Map tasks and then unite results with the same keys. The Shuffle tasks then output the united results to Reduce tasks. The Reduce tasks aggregate them and output the result to the data store.

A. *Migration by Shuffle tasks*

In conventional studies of migration methods, parallelized tasks must be synchronized to exchange individuals. If we realize a migration method in the MapReduce model, we must exchange individuals during the Map and Reduce steps. However, the MapReduce model does not allow such synchronization during parallel processing because Map/Reduce tasks must work independently. Thus, we propose to realize migration in Shuffle tasks. When Shuffle tasks unite results and keys from Map tasks, they communicate with each other. Therefore, we are able to realize migration during Shuffle tasks.

Fig. 2 illustrates the method employed to realize migration in the MapReduce model. The GA is applied to a subpopulation in each island during the Map step. After convergence, the Map tasks shuffle all resultant individuals and output them to Reduce tasks.

To realize migration during the Shuffle step, we set an identification number of an island as a key to each individual in the subpopulations. Each individual is transferred to the island corresponding to the key during Shuffle tasks. Twenty percent of the individuals in each subpopulation are assigned the original keys to keep individuals that are unique to the island. If the subpopulations have many individuals, the effect of parallelization will be limited. Subpopulations with few individuals do not have good search efficiency. Therefore, we do not assign the identification key randomly because this can disproportionately assign individuals to islands. We assign keys so that the individuals originally in the same island are uniformly distributed to the remaining islands. In addition, we define a special key to indicate the individual with the highest adaptive value. Individuals that have an identical key are aggregated during the Shuffle step. The Reduce tasks receive aggregated individuals and output them to the data store. If individuals have the special key, the best one is recorded to the data store in the Reduce step. To judge if the best solution candidates converge to an exact solution, the best individual recorded in the data store is compared to the one in the previous iteration of MapReduce. If they coincide, the population is regarded as having converged, and then the MapReduce iteration is terminated. Otherwise, the MapReduce iteration is repeated.

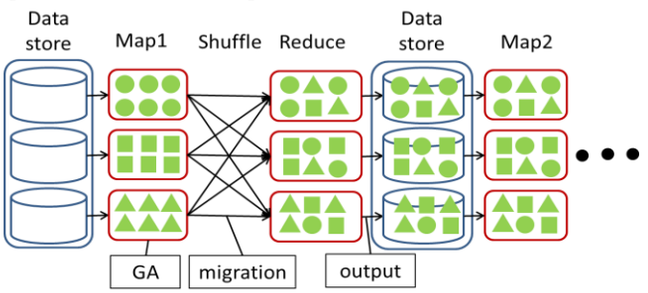


Fig. 2. Migration by iterative MapReduce

B. *Reduction of similar individuals*

To realize migration, we must determine the migration parameters as follows [4].

- Migration rate
- Migration interval

The migration rate is the ratio of the number of immigrated individuals per migration, and the migration interval is the number of generations between migrations.

If the migration rate is small and the migration interval is long, the effect of migration does not improve performance. If the migration rate is large and the migration interval is short, diversity of the population is lost [12].

Although we propose MapReduce iteration to realize migration, we must consider MapReduce overhead. MapReduce has processing overhead at start and end times, overhead related to I/O to the data store, and communication overhead in Shuffle tasks. In this study, we operate a mass of individuals; therefore, overhead is expected to be large. In this

subsection, we propose a method to reduce the amount of outputs from the Map tasks before executing the Shuffle tasks. Fig. 3 illustrates the proposed method. To reduce the number of MapReduce iterations and the outputs from Map tasks, the GA should be executed iteratively until convergence in Map tasks. Migration after GA-convergence increases the migration interval and leads to a reduction in the migration frequency. This imposes a condition whereby the migration rate cannot be small. After the GA converges, each population has approximately the same individuals, and it is obvious that not all of these individuals must be exchanged. We propose to eliminate half of the individuals in each population after completing Map tasks. If we eliminate individuals randomly, it is likely that we will eliminate good individuals. This will reduce search efficiency. Reduction of the individuals that have low adaptive values during GA process leads to convergence to a local solution. However our point is to reduce individuals after GA-convergence. After convergence, since we should utilize optimized individuals, we eliminate individuals that have low adaptive values to reduce individual data to be transferred.

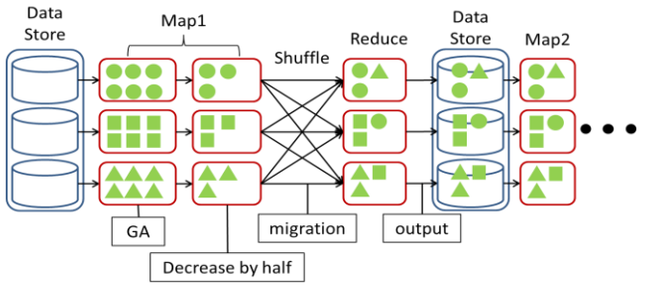


Fig. 3. Reduction of similar individuals

C. *Recovery of the number of individuals*

It is clear that the search efficiency decreases if the number of individuals is reduced by half every time MapReduce steps are iterated. To maintain search efficiency, it is necessary to add individuals to recover their original number.

The timing of this recovery is important from the viewpoint of system performance. The GA phase is included in the Map tasks; therefore, recovery must be executed prior to or during Map tasks. Recovery of individuals in Reduce tasks causes an increase of data output to the data store, which results in large I/O cost. Therefore, we propose recovery during Map tasks, as is shown by the Map2 step in Fig. 4.

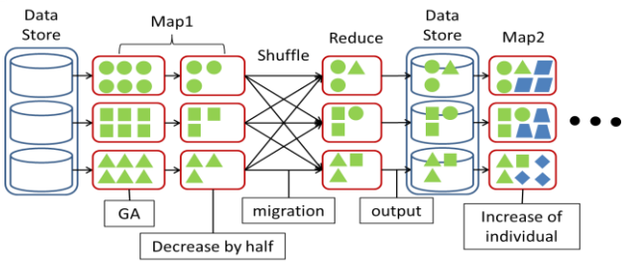


Fig. 4. Increasing individuals in Map tasks

If the recovered individuals are generated randomly, they do not necessarily have a good adaptive value. Consequently, we randomly duplicate individuals from a subpopulation in each island and then mutate them.

III. IMPLEMENTATION

We implemented the proposed method using Hadoop. Table I. shows the specifications of a master node and four slave nodes, which comprise the Hadoop cluster.

Table I. Master node and slave node specifications

Structures	Contents
CPU	Intel(R) Core(TM)2 Duo CPU @ 2.80GHz /Intel(R) Pentium(R) Dual CPU @ 1.80GHz
Memory	2GB
OS	CentOS6.2
Hadoop	Hadoop0.20

Fig. 5 shows a schematic of all MapReduce tasks for Hadoop. In our setting, the maximum number of Map and Reduce tasks was seven. The outline of the tasks is as follows:

1. Each Map task receives a subpopulation to which it applies the GA from the HDFS.
2. Each Map task performs the GA until convergence. They then eliminate half of the individuals in each population. The special key “0best” is assigned to the individual with the highest adaptive value. For all other individuals, an identification number associated with the island is assigned as a key. Then, the pair of identification number and individual is combined as a key–value pair and is output to partitioners. If Steps 1 to 6 are iterated more than once, the number of individuals increases by mutation at the beginning of this step.
3. Each partitioner receives the identification number and individual given by the corresponding Map task. The partitioners assign individuals to the Reduce task by referring to the hash of identification number of the island.
4. Shuffle tasks sort individuals by their keys. The individuals with the same key are stored in the same array.
5. Each Reduce task receives the array from the Shuffle tasks. To conserve the most highly adaptive individuals, we select the individual whose adaptive value is the highest from those having special keys. The chosen individual is output to the HDFS. All other individuals in the array are output to the HDFS.
6. To assess the convergence of the solution, the individual with the special key stored in Step 5 is compared to the one stored in the latest iteration. If the difference between their adaptive values is less than a threshold, then the solution is considered to have reached convergence.
7. If convergence is achieved, then this process returns the solution and terminates. Otherwise, it repeats Steps 1 to 6.

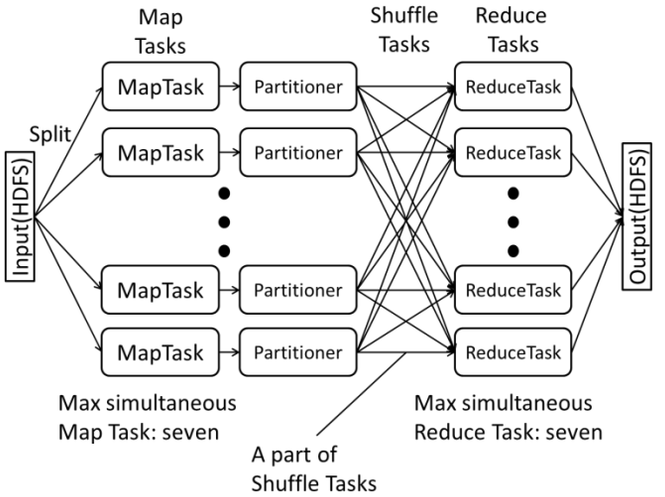


Fig. 5. Implementation of MapReduce on Hadoop

IV. EXPERIMENT

To evaluate the proposed method, we applied it to optimization problems of benchmarking functions.

The purpose of our study is to propose a new parallel GA that provides a more precise solution than previous work. Due to additional tasks, such as the migration of individuals, the proposed method may require a longer execution time than previous methods. However, we expect that the proposed method will effectively provide a more precise solution and that the execution time required to converge the solution is at least less than the original RCGA, which is not parallelized.

Based on this motivation, we have compared the accuracy of the obtained solutions and the execution time required to reach convergence using the original RCGA (GA1), a method from a previous study (GA2)[10], and the proposed method (GA3).

To investigate the effect of population reduction on the execution time, we compared the execution time for the version with reduction and a non-reduction version of the proposed method. We compared the execution time from part of a Shuffle task.

The first condition is commonly applied to all methods, and the second condition is applied to GA2 and GA3.

Our benchmarking functions were the Rastrigin function (n = 50) and the Rosenbrock function (n = 20), which are commonly used functions for evaluating optimization algorithms. Fig. 6 and Fig. 7 show graphs of a two-dimensional version of these functions. Equations (1) and (2) are their detailed expressions.

We chose the Rastrigin function for evaluation because this function is multimodal. A multimodal function has many quasi-optimum solutions; thus, it is difficult to seek an exact solution because finding a quasi-optimal solution can terminate the search. For this reason, the Rastrigin function is suitable for evaluating how the proposed method improves GA performance and solution accuracy.

The Rosenbrock function is a monophasic function that is shaped like a long valley. It is well-known that a GA is weak when optimizing a function whose value does not change in the neighborhood of its optimum solution. Thus, the Rosenbrock function is suitable for evaluating the accuracy in such a disadvantageous situation.

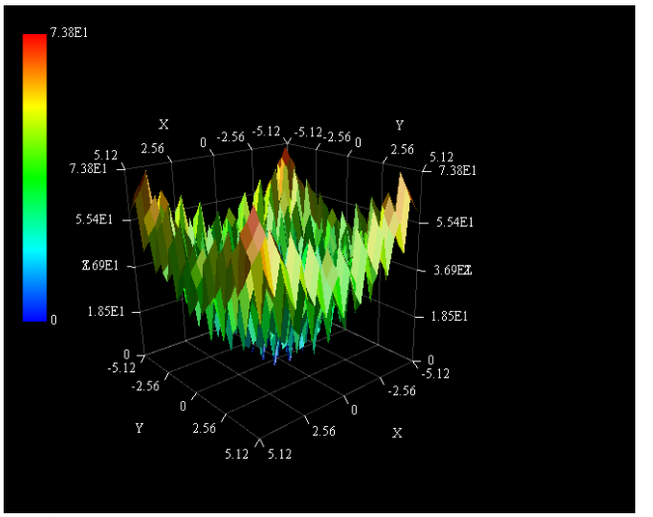


Fig. 6. Rastrigin function

$$F_{Rastrigin}(x) = 10n \sum_{i=1}^{n} \left(x_i^2 - 10\cos(2\pi x_i)\right) \quad (-5.12 \le x_i < 5.12) \quad Min\left(F_{Rastrigin}(x)\right) = F(0, 0, \ldots, 0) = 0 \tag{1}$$

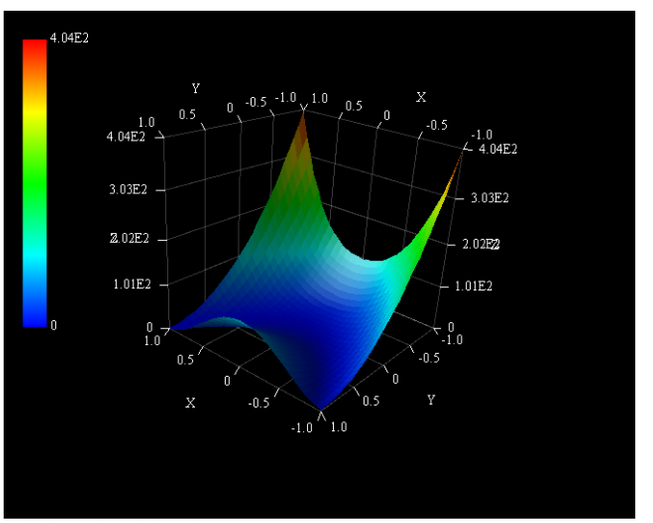


Fig. 7. Rosenbrock function

$$F_{Rosenbrock}(x) = \sum_{i=1}^{n-1} \left(100\left(x_{i+1} - x_i^2\right)^2 + (1 - x_i)^2\right) \quad (-2.048 \le x_i < 2.048) \quad Min\left(F_{Rosenbrock}(x)\right) = F(1, 1, \ldots, 1) = 0 \tag{2}$$

The parameters of the GA in this experiment are as follows.

- Total number of individuals: 50,000
- Total number of subpopulations: 250
- Crossover method: BLX-α [13]
- Selection method: Tournament selection
- Tournament size: 4
- Mutation method: Uniform mutation [14]
- Crossover probability: 0.98

- Mutation probability: 0.02
- Migration rate: 0.8

The convergence threshold used in this experiment was 10^{-5}.

A. *Accuracy of solutions and execution time*

Table II. and Table III. show the results for the execution time, accuracy, and migration count. The values shown are the averages obtained over 50 trials.

For the Rastrigin function, the results of GA1 were the worst for both accuracy of solutions and execution time. It should be noted that a single population tends to increase the same individuals compared to a case of many subpopulations. This inhibits search in a sufficiently large region and gives a local optimal solution rather than an exact solution. As the errors of the solution are not zero, this definitely occurred in the GA1 case.

The execution time of GA2 was the shortest of the three methods. This method divided the whole population into 250 subpopulations and parallelized the processes. Although this method is better than GA1 in execution time, it could not find the exact solution because each subpopulation provided local optimal solutions rather than an exact solution.

The proposed method (GA3) found the exact solution. Table 2 shows that the proposed method requires approximately twice as many migrations to converge individuals to the rigorous solution.

Table II. Comparison of execution time and accuracy (Rastrigin function)

	GA1	GA2	GA3
Ave. Exec. Time	880.0 s	93.8 s	210.6 s
Ave. of f(x)	268.1	5.7	0.0
Min. of f(x)	259.6	1.2	0.0
The # of trials generating rigorous solutions	0.0	0.0	50
Ave. Migration count	N/A	N/A	2.3

Table III. Comparison of execution time and accuracy (Rosenbrock function)

	GA1	GA2	GA3
Ave. Exec. Time	6365 s	72.5 s	1754.1 s
Ave. of f(x)	0.068	0.082	6.6×10^{-5}
Min. of f(x)	0.041	0.060	5.6×10^{-5}
The # of trials generating rigorous solutions	0	0	50
Ave. Migration freq.	N/A	N/A	41.9

For the Rosenbrock function, we regarded individuals as a solution if their function value was less than 10^{-4}.

GA1 found a better individual than the method from a previous study. This is interesting because GA1 does not parallelize processes and seemed to be at a disadvantage when finding a good individual. This is likely caused by the greater diversity of individuals in GA1 than those of GA2, which helps improve search efficiency in this situation.

It should be noted that GA1 and GA2 provided results that did not satisfy our error limit condition, $f(x) < 10^{-4}$. Table 3 shows that the best individual obtained by the proposed method did satisfy this condition.

Table 3 also shows that GA3 appeared to require more execution time than GA2. The execution times listed in Table 3 are the times at which the system ended the judgment of convergence. This means that the time required to obtain an individual that satisfies $f(x) < 10^{-4}$ is much longer than the times shown in Table 3. In our trial, we could not find a case for which GA2 could obtain such individuals; therefore, we consider that GA2 definitely requires an unreasonable execution time.

The proposed method (GA3) demonstrated high accuracy compared to GA1 and GA2.

B. *Effects of population reduction*

We used the Rastrigin function and measured the execution time for cases with and without population reduction. Table IV. shows the execution time averaged over 50 trials and the minimum time for both cases.

Table IV. Comparison of reduction of individuals (Rastrigin function)

	Has reduced individuals	Has not reduced individuals
Ave. Exec. Time	210.6 s	222.5 s
Min. Exec. Time	139.2 s	146.5 s
Ave. of f(x)	0.0	0.0
Min. of f(x)	0.0	0.0
The # of trials generating rigorous solutions	50	50
Ave. Migration freq.	2.32	2.67
Ave. Exec. Time of the Shuffle tasks per one migration	36.1s	42.3s
Total file size of data (Ave.)	72.14 MB	157.17 MB

The results show that the average execution time for the population reduction case was approximately 5.3% less than the other cases. As for the minimum execution time, population reduction reduced the execution time by approximately 5.0%. A detailed investigation has revealed that the population reduction decreased the execution time of Shuffle tasks per one migration from 42.3 s to 36.1 s.

The estimated execution time of Shuffle tasks was 112.9(=42.3×2.67) s for the case without population reduction and 83.8(=36.1 × 2.32) s with population reduction. This indicates that the execution time of Shuffle tasks decreased by 26%.

The average migration frequency for both cases was less than 3. The iteration frequency of the proposed method is significant for system performance; therefore, population reduction has little influence on the performance.

V. CONCLUSION AND FUTURE WORK

A GA is a powerful way to solve combinatorial optimization problems. To improve GAs to be applicable to a problem whose solution space has a huge number of dimensions, we have proposed a method to parallelize RCGA

by means of a MapReduce processing model, which realizes effective parallelization and has strong fault tolerance.

In a previous study, Keco et al. proposed a method to parallelize GAs using MapReduce in a context similar to our proposed method. They adopted the island model and did not consider population migration; therefore, their method cannot secure sufficient population diversity to find a rigorous solution.

To solve this problem, we have proposed a method for population migration that is compatible with the MapReduce model.

We proposed a migration mechanism in Shuffle tasks. Shuffle tasks are the only tasks in which the MapReduce model allows modes to intercommunicate. To realize this, we have utilized the identification numbers of islands as keys to assign individuals to subpopulations. In addition, to reduce unnecessary network I/O in Shuffle tasks, we have also proposed a method to reduce the number of individuals during migrations.

We compared the performance of the proposed method with the original RCGA and its parallelized version without migration. The compared indices were the execution time and solution accuracy. To evaluate the effects of population reduction during migration, we compared the execution time for cases with and without population reduction. The results show a significant improvement in the solution accuracy and execution time.

In future, it will be necessary to discuss the optimum number of individuals reduced in Map tasks. Although we eliminated one-half of the individuals in each population, the ratio of elimination may be dependent upon the number of individuals. For example, assume that there are five subpopulations, each of which has one thousand individuals. Even though we eliminate one-half of these individuals in each population, many similar individuals could still be migrated to the same subpopulation. In this case, elimination of more individuals reduces the migration cost but will lead to the same solution.

REFERENCES

[1] J. H. Holland, "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence", MIT press, Cambridge, MA, (1992).

[2] D. E. Goldberg, "Genetic algorithms in search, optimization, and machine learning", Addison-Wesley Publishing Company Inc, (1989).

[3] M. Munetomo, "Genetic Algorithms, the theory and advanced technique" (Japanese), Morikita publishing Co. Ltd, (2008).

[4] E. Cantú-Paz, "A survey of parallel genetic algorithms", Calculateurs Parallèles, Reseaux et Systems Repartis 10(2), pp. 141-171, (1998).

[5] J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM, 51(1), pp. 107-113, (2008).

[6] Message Passing Interface Forum, http://www.mpi-forum.org, (2014).

[7] L. G. Valiant, "A bridging model for parallel computation", Communications of the ACM, 33(8), pp. 103-111, (1990).

[8] Apache Hadoop, http://hadoop.apache.org, (2014).

[9] T. White, "Hadoop: The definitive guide," O'Reilly Media, Yahoo! Press, (2009).

[10] D. Keco, A. Subasi, "Parallelization of genetic algorithms using Hadoop Map/Reduce", Southeast Europe Journal of Softcomputing, 1(2), pp. 56-59, (2012).

[11] H. Rasit, N. Erdogan, "Parallel Genetic Algorithm to Solve Traveling Salesman Problem on the MapReduce Framework using Hadoop Cluster", JSCSE, 3(3), pp. 380-386, (2013).

[12] T. Hiroyasu, M. Miki, M. Negami, "Parallel distributed genetic algorithm with randomized migration rate" (Japanese), The Science and Engineering Review of Doshisha University, 40(2), pp. 25-34, (1999).

[13] L. J. Eshelman, J. D. Schaffer, "Real-coded genetic algorithms and interval-schemata", Foundations of Genetic Algorithms 2, Morgan Kaufman Publishers, pp. 187-202, (1993).

[14] A. H. Wright, "Genetic algorithms for real parameter optimization", Foundations of Genetic Algorithms, Morgan Kaufman Publishers, pp. 205-218, (1991).

Method for Selecting Words in Japanese–English Translation Based on Ontology

Marina Naito
Shibaura Institute of Technology,
Department of Information Science and Engineering,
3-7-5 Toyosu, Koto Ward, Tokyo 135-8548, Japan
l10078@shibaura-it.ac.jp

Masaomi Kimura
Shibaura Institute of Technology,
Department of Information Science and Engineering,
3-7-5 Toyosu, Koto Ward, Tokyo 135-8548, Japan
masaomi@sic.shibaura-it.ac.jp

***Abstract*—In this study, we introduce a method of selecting translated words for translations of sentences from Japanese to English. We used topic maps to express relationships between words. For constructing topic maps, we extracted word data and categories from Japanese and English Wikipedia. We then used English–Japanese and Japanese–English dictionaries to find translation relationships between Japanese and English words. Our method selects translated words using dependency associations extracted from an input Japanese sentence and the categories that each word belongs to. In our experiment, we used a topic map that organizes database related words. The results suggest that our method can correctly translate words.**

Keywords—component, topic maps, ontology, translation, wikipedia

I. INTRODUCTION

In recent years, machine translation has often been used to translate Japanese sentences into English. However, because computers have difficulty comprehending the meaning of sentences, they often produce incorrect translations. For example, the Japanese sentence, “私はファイルを解凍する” (“I uncompress a file” in English), may be translated as “I defrost a file,” which is incorrect. The reason for the difficulty is that translation techniques consider only a word’s part of speech and the grammatical structure of the target sentence not homonym and inter-word semantic relationships. The homonym problem must be solved for an appropriate word to be selected.

To resolve this problem, we use an ontology [1] [2], which expresses three inter-word relationships, namely, word translation, category inclusion and potential-word-modification relationships. They were studied in the context of ontology [3], disambiguation [4] [5] [6]. To utilize those relationships in a computer system, we need to establish a method to create a data structure that expresses the relationships as an ontology. Topic Maps [7] [8] and Resource Description Framework (RDF) are well-known and general-purpose formats for describing the relationships between concepts. In this study, we implement the ontology as a topic map, which is suitable for expressing concepts and their relationships. Topic maps mainly consist of topics, associations, and association roles. A topic is an element that identifies a subject (*e.g.*, a person, a country, or a concept). A topic usually has a type, which is an attribution element of the topic. An association is an element that identifies the relationship between topics. An association role defines the role of a topic in a specific association. Although topic maps are essentially undirected graphs, the association role can define an association direction.

For example, the concepts, “コンピュータ” (computers), “解凍する” (uncompress/defrost), and “食べ物” (foods) can be expressed as topics. In the computer domain, the word “解凍する” should be translated as “uncompress” in English. In the food domain, it should be translated as “defrost.” We can, therefore, define *translation* associations between the Japanese word, “解凍する,” and the English words “uncompress” and “defrost.” *Inclusion* associations can be defined between “uncompress” and “computer” and between “defrost” and “foods.”

In this study, we propose a method of identifying the appropriate translated word based on the multilingual word semantic relationships defined in an ontology. A paragraph consists of sentences that coherently share a subject. Therefore, in order to find a domain, we use sentences in one paragraph as input sentences. Our method searches for a translated word in the ontological topic map based on the domain to which the input sentences belong. In this study, we focus on Japanese–English translation. Moreover, we focus on only nouns and verbs, which tend to be wrongly translated and are basic parts of speech.

II. METHODS

We focused not only on inter-lingual word relationships between Japanese and English words but also on the intra-lingual relationships of words. Inter-lingual word relationships are the relationships between translation pairs of words, which are usually listed in a dictionary. For intra-lingual relationships, we used (potential) modification relationships between words that appeared in sentences and the corresponding relationships between words and their domains.

We suggest these topics and associations:

- Topics of the *word* type are elements that correspond to each part of speech in English or Japanese: English nouns and verbs (referred to as *noun_e topics* and

verb_e topics, respectively) and Japanese nouns and verbs (*noun_j topics* and *verb_j topics*).

- Topics of the *domain* type are the categories of words.
- Associations of the *inclusion* type are relationships in which a word is included within the domain to which it is linked by the association.
- Associations of the *action* type are potential modification relationships between an English noun (*noun_e*) and an English verb (*verb_e*). The *potential* means that the noun can be used as a subject of the verb in sentences.
- Associations of the *translation* type are relationships between *noun_e* and *noun_j* topics or *verb_e* and *verb_j* topics, in which both words in the relationship have the same meaning.

Fig. **1**. illustrates an example of topic maps in this study.

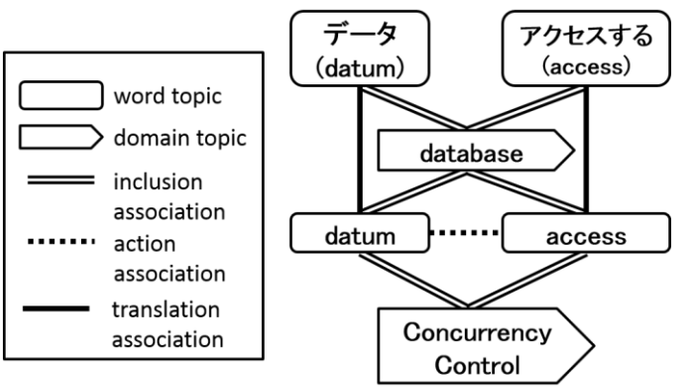


Fig. 1. Example of the topic map

A. Method of making a topic map

The abovementioned property of paragraphs suggests that the words in the sentences therein are used in a specific context and should belong to a common domain. As we mentioned in Introduction, domain information can help solve the homonym problem. Therefore, in our proposed topic map, we needed to associate words with domains.

In this study, we made a topic map from the description data in the Japanese version of Wikipedia [10] and the English version of Wikipedia [11]. We used these sources because they provide words and their domains. The pages contain title words and descriptions, and each of pages belong to categories. We used not only title words but also words in the descriptions as context for what categories the pages belonged to. Therefore, in order to include them in a topic map, we related words to the domains that correspond to the categories.

We used this procedure to construct the topic map:

1. We extracted the categories to which the articles belonged in the Japanese Wikipedia and English Wikipedia.
2. If a category in the English Wikipedia could be matched to a category in the Japanese Wikipedia, we integrate the English category to the Japanese category.
3. We assigned domain topics to the categories obtained in steps 1 and 2.
4. We extracted nouns and verbs from the main text in the pages of the Japanese Wikipedia articles. We did not include the Japanese verbs “ある(be)” and “できる(can)” in the topic map, because they are usually used as auxiliary verbs and do not carry information useful for improving translation.
5. We extracted nouns and verbs from the main text in the pages of the English Wikipedia articles.
6. We created *noun_j* and *verb_j* topics for Japanese nouns and verbs, and *noun_e* and *verb_e* topics for English nouns and verbs. In doing this, we changed nouns to their singular forms and verbs to their dictionary forms, and then assigned them to topics as topic names. We distinguished between topics for the same word in different categories, because topics can carry multiple meanings.
7. If there were dependency relationships between English words, we assigned action associations to the topics associated with those words.
8. We created inclusion associations between words and their domains based on articles and their categories in Wikipedia. Because any article in Japanese Wikipedia or English Wikipedia belongs to at least one category, we can assign at least one inclusion association to each topic.
9. If two words were listed as translations of each other in an English–Japanese dictionary and Japanese–English dictionary, we assigned a translation association between the topic of the English word and the topic of the Japanese word.

Fig. 2. illustrates how to create a word topic for the sentence “ビューはテーブルに対して利点を持つ。” Fig. 3. shows how to create a word topic for the sentence “Views can provide advantages over tables.”

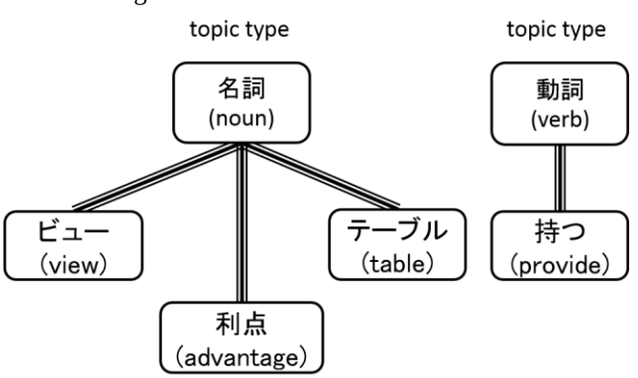


Fig. 2. Example of Japanese topics

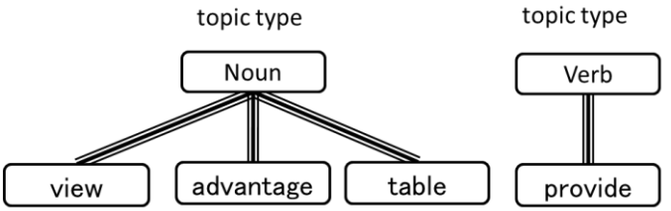


Fig. 3. Example of English topics

B. *Method for searching words*

We identified the domains to which input sentences belonged. We then searched the domains to which words belonged to obtain the top three domains in descendant order of the frequency with which the domain appeared. We limited the number of domains to narrow the search area of words to those that share a domain with the words to be translated. After identifying word domains, we searched for translated words that belonged to each domain. We followed these steps when searching for words:

1. Find words as pairs of nouns and verbs, which have dependency relationships when they appear in input sentences.
2. Search Japanese topics (*noun_j* and *verb_j* topics) that have the same topic names as the nouns and verbs found in step 1.
3. Obtain the *noun_e* topics associated with *noun_j* topics and *verb_e* topics associated with *verb_j* topics by translation association.
4. Extract the *noun_e* and *verb_e* topics that belong to the domain discussed above.
5. Judge whether the obtained *noun_e* and *verb_e* topics are connected by action association. If so, output the *noun_e* and *verb_e* topics as translated words.

III. SYSTEM IMPLEMENTATION

We implemented a topic map using the method discussed in Section A. To extract nouns and verbs from Japanese sentences, we used the CaboCha dependency parser [12]. To parse English sentences, we used Stanford Parser [13] and TreeTagger [14]. To add translation associations, we used the Eijiro [15] Japanese–English and English–Japanese dictionary database. We utilized TOME [16] as a topic map database.

Fig. **4**. schematically illustrates our implemented system. First, a user writes sentences for translation into a file that the system reads. Second, the system extracts nouns and verbs from the file and writes them to the file of words. Third, the system inquires with TOME and searches the domains of input sentences in the way discussed in section B. Fourth, the program searches translated words again, reads the file of words, gets results from the searched domains, and inquires with TOME to get the resultant words that belong to the domains. The program writes the results to the file for translated words. Finally, the user receives the translated words.

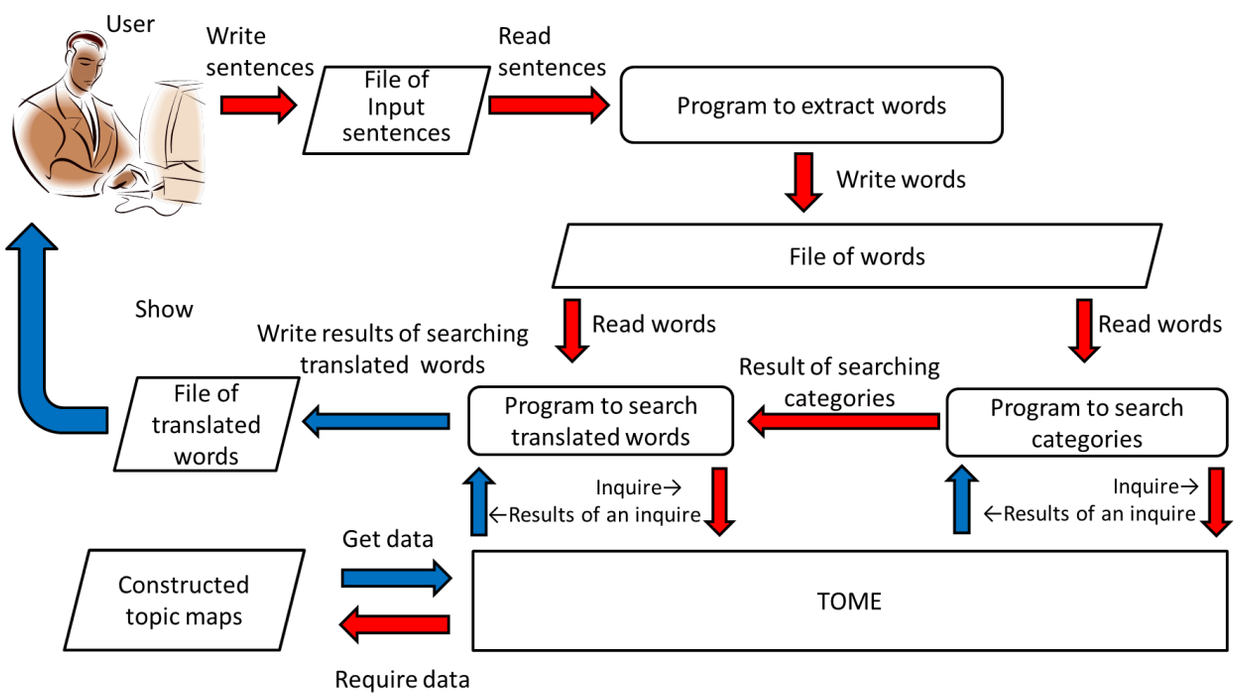


Fig. 4. System representation

IV. EXPERIMENTS

A. Purpose of experiments and target data

We conducted an experiment to see whether the proposed method selects the correct word translations. The sentences, including the words to be translated, were input into the system using the proposed method. We confirmed that the translated words produced by the system were correct.

The target topic map was constructed using the method in section A. Using all Wikipedia pages would produce too much data, so we limited the number of pages in the database category. The number of topics and the number of associations are shown in the following tables:

TABLE I. NUMBER OF TOPICS

Topic name	number
all topics	6,899
noun_j topics	3,615
verb_j topics	1,244
noun_e topics	1,007
verb_e topics	819
domain topics	190

TABLE II. NUMBER OF ASSOCIATIONS

Association name	number
all associations	44,703
action associations	662
inclusion associations	38,203
translation association	5,838

B. Experimental method

We entered sentences from 10 paragraphs in the Introduction sections of 9 papers focused primarily on databases. This is because our target topic map only contained topics in the database domain. The system searched the topic map and generated candidate pairs of translated nouns and verbs from the entered sentences.

To evaluate the candidates, we used the Google search engine to examine whether the obtained pairs of nouns and verbs were used in sentences written by native English authors.

C. Results

From the sentences in the input papers, we obtained 382 Japanese words. Among them, 155 words were not in the topic map. Among Japanese words in the topic map, 42 words did not have corresponding English words. Among words with translated words in the topic map, 165 translated candidates had no action associations. This shows that our method did not find translated words for 362 words out of 382 words. The 10 obtained pairs of Japanese words and translated English nouns and verbs are shown in TABLE **III**. TABLE **III** suggests that we obtained proper translated words for the pairs of No. 1, 2, 4, 5, 7, 8, 9, and 10. It is interesting that Japanese verbs No. 4 and 10 both generally mean *utilize*, their corresponding English words differed based on the corresponding nouns.

TABLE III: DETAILED LIST OF OBTAINED TRANSLATED WORDS

No.	Extracted words		Translated words	
	noun_j	verb_j	Noun	Verb
1	問題 (problem)	解決する (solve)	problem	solve
2	問題 (problem)	解決する (solve)	problem	solve
3	ファイル (file)	利用する (use)	file	take, use
4	知識 (idea)	利用する (use)	idea	use
5	情報 (information)	格納する (store)	information	store
6	概念 (concept)	表現する (express)	concept	put
7	コスト (cost)	減らす (reduce)	cost	reduce
8	データ (datum)	アクセスする (access)	datum	access
9	操作 (operation)	加える (add)	operation	add
10	技術 (technique)	用いる (employ)	technique	employ

D. Discussion

As shown in TABLE III, we obtained correct translated words for only 16 words (8 pairs). This is because our method focuses on the precision, not recall, of output. Our condition is severe, because the pairs to be translated need to be related to their translated words in dependency relationships in sentences that appear in Wikipedia. Therefore, words that appear in academic papers, but not necessarily in Wikipedia, may not obtain good results. However, the output pairs have nouns that actually modify the verbs in the sentences that appear in the English version Wikipedia. This guarantees that the results are highly precise. We focused on precision to prevent incorrect choices of translation candidates. Even if a method provides high recall, wrong choices leads users to wrong translation. To improve recall, it is preferable to add more action associations to the topic map based on the dependency relations contained in a much larger English corpus. In this study, we used the English version of Wikipedia as a corpus; in further studies, we should use a corpus that contains the lexicons used in input sentences.

As a side effect, the system generated different verbs depending on their corresponding nouns, even if the verbs had similar meanings. This illustrates that the topic map uses information about dependency relationships as action associations.

The reason why the system did not output the correct translated words for No. 3 is that our method does not consider idioms. For example, the verb *take* in No. 3 should form an

idiom *take advantage of* to express the meaning of *use*. Noticeably, the verb *take* alone does not have this meaning, which can lead users to incorrect translations.

As for No. 6, although the sentence, "the concept was put by him," makes sense, we categorized it as the wrong translation. The reason for this is that the original Japanese sentence intended to express "metadata to formalize a concept." The essential reason for this gap comes from the different purpose of the action specified by the verb. If we express a concept to people's mind, the verb *put* is adequate. But, in this case, we are not using *put* to express formalization of a concept in order to utilize it in a computer system. This can be due to a gap between multiple languages. This suggests that we need to store extra information to distinguish differences in the topic map.

Because the experiment showed the existence of words that were not contained in these topic maps, we must increase the number of word topics and translation associations. Because the experiment also showed a deficiency in information about translation relationships in this topic map, we need to use additional dictionaries to cover technical terms in order to increase the number of translation associations.

In addition, we did not consider synonyms in the topic map. If we considered them, we would get more translated words in this experiment. For example, we input the pair of words "技術, 使用する"(technique, use) into the system, but we could not obtain translations using our proposed method, because the Japanese word "使用する" is not directly related to its translation candidate words by a translation association. However, if we introduced a synonym association in the topic map, we could obtain translated words using this method: we find a *translation candidate* word of the synonym words of "使う," and regard it as a *translated* word if it is directly associated to one of the translated words of the Japanese noun "技術" by an action association. Note that the translated words of synonymous words are not necessarily the translated words of the original word. This is why we used the action association to identify the translated words. In this study, we did not use an association to express the (potential) dependency relationship of words in a Japanese sentence. A dependency relationship was implicit, because we assumed that input pairs of nouns and verbs had dependency relationships. To make the extension method for employing synonyms work, we would need to introduce a Japanese version action association to guarantee that the chosen synonymous words carry the same meanings as the original words.

As for the use of domains to narrow the number of translation candidate words, we could not confirm the effects on our data set in this experiment. To confirm that our topic map has the structure to enable search based on domain, we looked into it to find substructures that output different words depending on the domain. Fig. 5. illustrates one such substructure. The topic "保存する" is associated with the topics of the translation candidate words *preserve* and *store*. The *preserve* topic belongs to both the *database* domain and *data modeling* category, but the *store* topic belongs to not only the *database* domain but also the *artificial intelligence* domain, *transaction* domain, and so on. Because the latter topic does not belong to the *data modeling* domain, we can find *preserve* as the translated word of "保存する," if the words of the input sentence are in the *data modeling* domain. Conversely, if the domain is *artificial intelligence* or *transaction*, we can find *store* as the translated word.

Therefore, our method can adjust the translation of a verb depending on the noun word in a pair of search words and the domain to which the noun and verb belong.

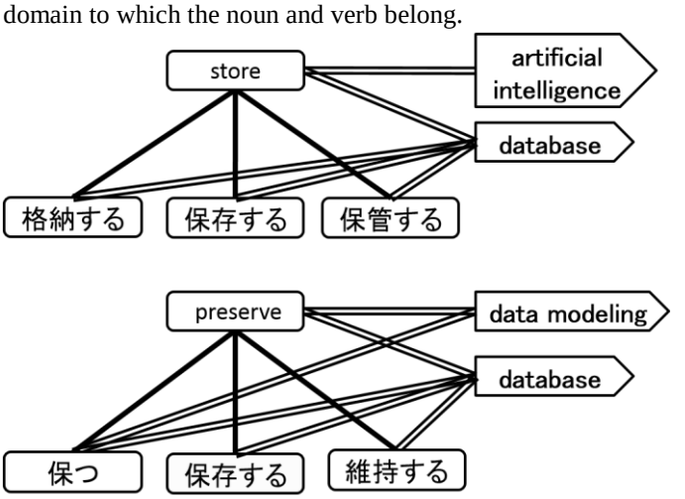


Fig. 5. Parts of "store" and "preserve" topics in the topic map

V. Conclusion and Future Works

In this study, we proposed a method of correctly translating words based on multilingual semantic relationships between words that were defined in an ontology. We implemented the ontology as a topic map.

We introduced two types of topics (*i.e.*, word and domain) and three types of associations (*i.e.*, inclusion, action, and translation). We utilized the Japanese and English versions of Wikipedia pages to find words and the domains to which they belonged, and we also assigned conclusion associations between them. We used the English version of Wikipedia to find the dependency relationships in sentences, to which we assigned action associations. We used English–Japanese and Japanese–English dictionaries to find translation relationships between Japanese and English words, to which we assigned translation associations.

We assumed that a paragraph consists of sentences coherently sharing a subject. To find a domain, we used sentences in one paragraph as input sentences. Our method finds words in input sentences as pairs of nouns verbs in dependency relationships. The method searches topics corresponding to the nouns or the verbs. Based on translation and conclusion associations, it searches translated nouns and verbs that share action associations. We focused on translations of nouns and verbs from Japanese to English, which tend to be incorrect despite being basic parts of speech in sentences.

To evaluate our method, we conducted an experiment. The inputs were sentences from 10 paragraphs in the Introduction sections of 9 papers focused primarily on databases. We

obtained 382 Japanese words, 155 of which were not in the topic map, and 42 of which had no corresponding English words. We obtained 10 pairs of translated nouns and verbs. The results also showed that some Japanese verbs that have similar meanings are translated into different English verbs depending on their corresponding nouns.

In the future, we plan to improve the coverage of words in the translation process by adding English and Japanese words to the topic map and adding action associations. To improve the accuracy of translation search, we will improve the searching method to consider idioms.

Though we used sentences in Wikipedia to find action associations, it is desirable to use sentences from documents that cover the words as those related to the input sentences. In our future study, we will combine Wikipedia data with data from other resources to increase the number of action associations in our topic map.

To suppress unwanted translation outputs, we need to increase the number of association types to make topic selection more precise.

Our method outputs only translated words. To make an effective automated translation system, we will apply our ontology-based method to an existing automated translation system based on grammatical techniques. We will also extend it to a system that can identify implied meaning.

REFERENCES

[1] R. Mizoguchi, K. Kozaki, Y. Kitamura, M. Sasajima, "An Introduction to Construct Ontology," Ohmsha, Ltd. Press, 2006.

[2] Y. Kitamura, "Spread and Apply Ontology," The Japanese Society for Artificial Intelligence, Ohmsha, Ltd., 2012.

[3] F. Kimura, A. Maeda, T. Koshida, J. Miyazaki, S. Uemura, "Construction of Bilingual Ontology using Web Directory," IPSJ, vol. 112, pp. 25–32, Nov. 2003.

[4] F. Kimura, A. Maeda, K. Hatano, J. Miyazaki, S. Uemura, "Analysis of Appropriate Category Level of Web Directory for Cross-Language Information Retrieval," Proceedings of IMECS, vol. 1, pp. 19–21, March 2008.

[5] F. Kimura, A. Maeda, J. Miyazaki, M. Yoshikawa, S. Uemura, "Cross-Language Information Retrieval Using Web Directories as a Linguistic Resource," IPSJ, vol. 45(SIG_7(TOD_22)), pp. 208–217, June. 2004.

[6] A. Maeda, F. Sadat, M. Yoshikawa, S. Uemura, "Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine," IRAL, Proceedings of the fifth international workshop on on Information retrieval with Asian languages, pp. 25–32, 2000.

[7] M. Naito, "An Introduction to Topic Maps," Tokyo Denki University Press, 2006.

[8] Lars Marius Garshol, "What Are Topic Maps," http://www.xml.com/pub/a/2002/09/11/topicmaps.html, 2002.

[9] RDF Primer, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/, 2014.

[10] Wikipedia(ja), http:///ja.wikipedia.org/, 2014.

[11] Wikipedia(en), http://en.wikipedia.org/, 2014.

[12] T. Kudo, Y. Matsumoto, "Japanese Dependency Analysis using Cascaded Chunking," CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops), pp. 63–69, 2002.

[13] Stanford Parser, http://nlp.stanford.edu/, 2014.

[14] TreeTagger, http://www.cis.uni-muenchen.de/~schmid/, 2014.

[15] Electronic Dictionary Project, Eijiro the fifth edition, 2010.

[16] Y. Kuribara, T. Hosoya, M. Kimura, "Tome: Topic maps database extended," in Proc. of SEATUC, pp. 245–248, 2010.

Point Cloud Generation for Ground Surface Modeling Employing MAV in Outdoor Environment

Shahmi Junoh* and Klaus-Dieter Kuhnert†
Bonn-Rhein-Sieg University of Applied Sciences, Germany
Institute of Real-Time Learning Systems, University of Siegen, Germany

ABSTRACT

Ground surface modeling has become an important research area during the past few years. While computer vision approach suffers from some inherent problems like lighting condition dependency, we propose a Micro Aerial Vehicle (MAV) equipped with Global Positioning System (GPS), barometer, Inertial Measurement Unit (IMU) and laser scanner as an alternative solution for generating ground surface models. We suggest a first method which make use of all those sensors except the laser scanner and a second method incorporating this laser scanner with a scan matching algorithm. We evaluate those two methods by generating point cloud in each case. We also devise and apply a simple yet useful algorithm called *BacktransformLaserScan* that is useful to evaluate the feasibility of any laser scan matching algorithm to certain application, hence, show its usefulness.

1 INTRODUCTION

Despite Autonomous Ground vehicles (AGVs) which are beneficial for many applications, MAVs have been found to be more advantageous over AGVs in 3D mapping. This is due to the fact that the degree of maneuverability and degree of dexterity of MAVs are higher than that of AGVs. In other words, an MAV is able of explore a wider space than an AGV can do particularly in the presence of obstacles.

We address the problem of 3D ground model generation. Some information which are known apriori like timestamp, position, orientation of the MAV and range data obtained from a laser scanner are used to build such a ground model in the form of a 3D point cloud. Upon successfully generating the point cloud, we need to optimize the output by applying a laser scan matching algorithm to get a more accurate generated point cloud. Hence, the problems we deal with are (1) generating a point cloud using MAV localization data and (2) improving the accuracy of the generated point cloud by employing laser scan matching algorithm.

The basic idea behind this research is about generating a 3D point cloud employing an MAV that utilizes GPS, IMU, barometer and 2D laser scanner. The data in the IMU have been fused using kalman filter to get the attitude of MAV[1]. The GPS gives the position value of the MAV. The frequency of this pose value is done at 10 Hz. The laser scanner gives raw data of distance which is updated at 30 Hz.

First of all, a point cloud is generated and then further improved using scan matching algorithm, which is eventually verified through several datasets in order to evaluate the approach on some different scenarios. This idea is illustrated in Figure 1.

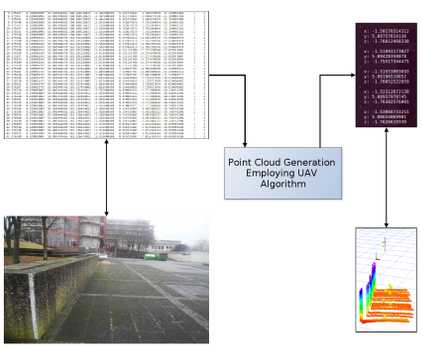


Figure 1: *This illustrates overview of the work. A dataset (see also the corresponding area where a flight has taken place by following the arrow) is taken as an input to our algorithm and our algorithm outputs a point cloud data from which the visualization data below is generated.*

The remainder of the paper is organized as follows. In Section 2, we introduce some of the related work in this domain. Section 3 presents our approach and later in Section 4, we evaluate its performance. Section 5 concludes this work.

2 RELATED WORK

The following gives an overview of some existing efforts in 2D mapping and 3D mapping.

2D Mapping A common sensor used in realizing 2D mapping is by using a single 2D laser scanner that results in 2D map as shown by Thrun in [1]. Thrun also analyzes in that survey about several algorithms like some Kalman filter

*shahmi.junoh@smail.inf.h-brs.de
†kuhnert@fb12.uni-siegen.de

[1]The sensor fusion is not included in the work package and is done by other developers.

approaches, expectation maximization (EM) and hybrid approach used in robotic mapping. Grzonka et al. [2] shows how a 2D map is created utilizing a quadrotor and a laser scanner.

3D Mapping There are a few ways of acquiring 3D maps. One popular way is realized by using 2 units of 2D laser scanners and the other is by utilizing a 3D laser scanner. The first method is done by mounting one laser scanner vertically and another horizontally. For instance, Früh et al. [3] have implemented this way to achieve the goal of 3D mapping. The latter uses a 3D laser scanner which means rotating or tilting 2D laser scanner is utilized to obtain 3D data. Surmann et al. [4] has built a low-cost 3D laser scanner and further employed it to improve the performance of robot navigation and recognition. The current trend is heading toward exploiting RGB-D data that a Kinect can offer as applied in [5], however, the use of this hardware is limited for indoor setup only.

3 Approach

3.1 Tools and Configurations

The md4-1000 quadcopter platform from *microdrones*[2] shown in Figure 3 is used. ROS[3] [6] is selected as a framework to ease and accelerate our development. Three cartesian coordinate frames — world frame, MAV's body frame and sensor frame — are defined as shown in Figure 2. To ensure reliable measurements, only laser scanner sensor readings of a maximum distance of 30 meters and field of view of 90 degrees of opening angle are regarded.

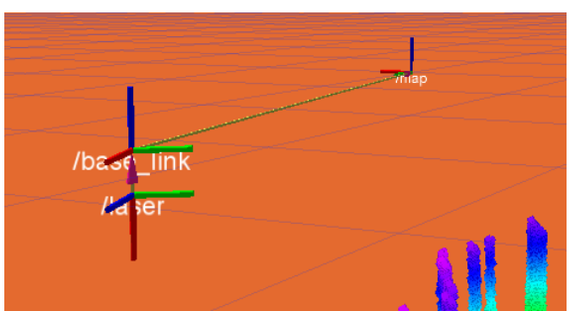


Figure 2: *A laser frame (*/laser*) is attached to a MAV's body frame (*/base_link*) by a fixed link.*

3.2 Scenarios

We briefly describe four scenarios we used to evaluate both the performance of the proposed approach and the quality of the employed laser scan matching algorithm. Each scenario varies with the way MAV is hovered (whether it is translational or rotational) and in the chosen type of ground surface.

[2]www.microdrones.com
[3]http://www.ros.org/wiki/

Figure 3: *This figure shows how the laser scanner is mounted on the MAV (It is noted that the laser scanner is mounted such that the beam will be scanning downwards).*

Scenario I The MAV is implementing forward flight going forward over a relatively flat surface for about 10 meters, returning to the place where it starts and this sequence is repeated for about 3 minutes.

Scenario II The ground surface is flat as well, same as in scenario I. The only difference is that the quadcopter is maneuvered in rotational movement about a relatively fixed position.

Scenario III The target ground surface is in the form of corner-like ground surface. Instead of hovering, we replicate the motion manually. Our purpose is to observe how the performance changes with this special type of target surface. The corresponding scenario is shown in Figure 4.

Figure 4: *Manual steering over a corner-like ground surface.*

Scenario IV We construct the test environment a structured environment two identical boxes as depicted in Figure 5. The quadcopter is then hovered over that small region by forward flight.

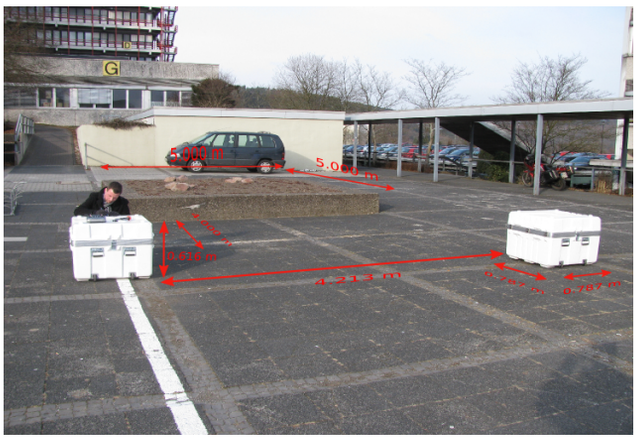


Figure 5: *This setup is used for evaluation using the mentioned quantitative method.*

3.3 Laser Scan Matching

Theoretically, given two independent scans or point clouds that correspond to a single shape, the goal of laser scan matching is to estimate the rotation **R** and translation **t** that reduces the following cost function:

$$E(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} w_{i,j} \|a_i - (\mathbf{R}b_j + \mathbf{t})\|^2 \qquad (1)$$

where the weight $w_{i,j}$ equals to 1 if the feature points a_i and b_j overlap or else, 0 is assigned.

It means that the correspondence must be first found before computing the transformation(**R**, **t**).

Several variants of laser scan matching algorithms which have been implemented in the past. Some of them have been wrapped in the context of ROS [6] and been packed as a ROS package. There are two packages that are already available in the ROS repository: the *polar_scan_matcher* and the *laser_scan_matcher* package. In terms of algorithm, the former is originally written by Diosi and Kleeman [7] while the latter is developed by Censi [8]. Both of them are wrapped by Dryanovski [9] in ROS framework. We opt for the *laser_scan_matcher* in this work.

3.4 Back-Transformation

We propose a back-transformation algorithm named *BacktransformLaserScan* as shown in Algorithm 3.1 to evaluate the output returned by *laser_scan_matcher* package. The developed algorithm detailed in the subsequent BACKTRANSFORMLASERSCAN pseudocode. Each variable associated with pose2D in the pseudocode corresponds to the data obtained from the *laser_scan_matcher* package. Variables are differentiated by the prefixes/suffixes of x (position of x-component), y (position of y-component), theta (angle), previous (value before the current one) and current (most recent value)

Algorithm
3.1: BACKTRANSFORMLASERSCAN(int row_number, dataset $scan$, dataset $pose2D$)

Input : int row_number, dataset$scan$, dataset $pose2D$
Output : backtransformed_scan $x\_backrotated$, $y\_backrotated$
for each $range\ data \in one\ scan\ line$
do
$x\_cartesian\_previous \leftarrow scan(row\_number) * sin(angle)$
$y\_cartesian\_previous \leftarrow -scan(row\_number) * cos(angle)$

$x\_cartesian\_current \leftarrow scan(row\_number + 1) * sin(angle)$
$y\_cartesian\_current \leftarrow -scan(row\_number + 1) * cos(angle)$

$x\_backtranslated \leftarrow x\_cartesian\_current -$
$(x\_pose2D\_current - x\_dataPose2D\_previous)$
$y\_backtranslated \leftarrow y\_cartesian\_current -$
$(y\_pose2D\_current - y\_dataPose2D\_previous)$

$x\_backrotated \leftarrow x\_backtranslated * cos(theta\_pose2D\_current -$
$theta\_pose2D\_previous) - y\_backtranslated *$
$sin(theta\_pose2D\_current - theta\_pose2D\_previous)$
$y\_backrotated \leftarrow x\_backtranslated * sin(theta\_pose2D\_current -$
$theta\_pose2D\_previous) + y\_backtranslated *$
$cos(theta\_pose2D\_current - theta\_pose2D\_previous)$
return $(x\_backrotated, y\_backrotated)$

3.5 Evaluation on laser_scan_matcher Package

The obtained information from the *laser_scan_matcher* package are a position in the x-axis, a position in the y-axis and an angle of orientation. These values are cumulative. In other words, the values are not absolute values, and that, in order to know the corresponding absolute value, one has to subtract the current measurement from the previous one.

In order to know whether the scan matching algorithm from the *laser_scan_matcher* package is reliable or at least the output of the package is accurate enough to be used in our case, we have applied the previously mentioned *back-transformation* algorithm.

Scenario I The performance of the *laser_scan_matcher* package is depicted in Figure 6. In Figure 6, we draw an ellipse in order to easily focus on particular regions and further evaluate the performance of the package. We observe that we have height error of about 0.1 m and orientation error of about less than 1 degree. In the ideal case, the blue and red markers should overlap. In this case, we can say that the *laser_scan_matcher* package gives a poor pose estimation, particularly in position estimation. Based on those results, we can conclude that this package is not sufficient enough for our application.

Scenario II For scenario II, the performance of the *laser_scan_matcher* is shown in Figure 7 using laser scan reading of 60 and 61. It can be deduced that with this kind of scenario, the *laser_scan_matcher* does not perform. The performance degrades rapidly in its performance when the MAV is tilted about roll angle (cf. Figure 7) because of the windy state at that moment.

Scenario III For this scenario, the performance of the *laser_scan_matcher* package is in Figure 8. We observe some regions that differentiate their performance. It can be divided into three groups. The first one which is bounded by the green

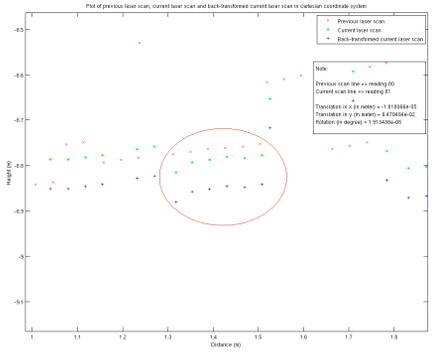


Figure 6: *The performance of the* laser_scan_matcher *package on scenario I using laser scan reading of 80 and 81.*

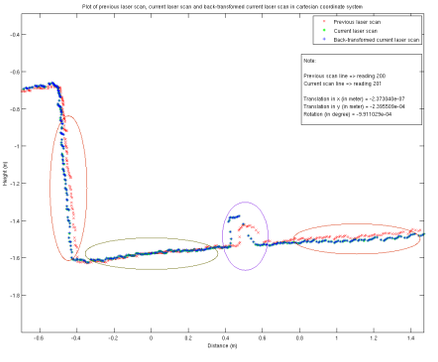


Figure 8: *The performance of the* laser_scan_matcher *package on scenario III using laser scan reading of 200 and 201.*

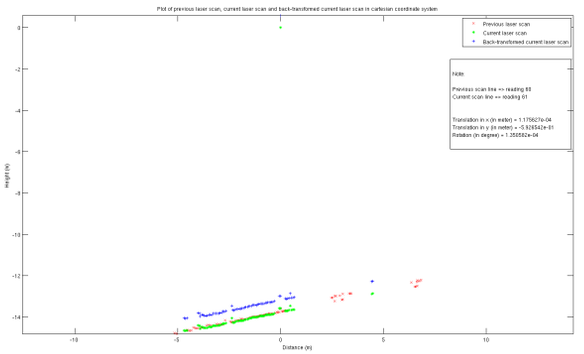


Figure 7: *The performance of the* laser_scan_matcher *package on scenario II using laser scan reading of 60 and 61.*

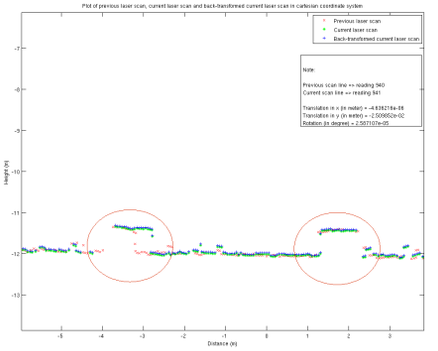


Figure 9: *The performance of the* laser_scan_matcher *package on scenario IV using laser scan reading of 940 and 941.*

ellipse is the one that performs best among those three. The one bounded by the red ellipse shows bad performance while the one with the purple ellipse is the worst among them.

Scenario IV Figure 9 shows the performance of the package on this scenario. It can be seen that the region which is bounded by two red ellipses are the top side of the two boxes. The performance shown in that graph is not very good.

3.6 ROS Nodes and Related Implementation

We design our software with ROS design standard in mind. In this regard, all nodes communicate via publisher and subscriber model. Effectively, there are nine nodes that communicate with each other. Out of these nine, there are mainly five nodes that we configure ourselves. These five nodes and their corresponding features are listed as follows:

- *publisher_node*: This node translates some relevant data from raw dataset by advertising on some topics. It also constructs and publishes pose information based on MAV localization data.
- *pose_estimation_node*: This node performs state estimation and publishes it.
- *robot_pose_broadcaster_node*: This node configures the transformation of the pose of *base_link* of the MAV with respect to the world and further publishes.
- *assembler_client_node*: This is a client node that uses a point cloud building service offered by *laser_assembler* package.
- *pointcloud_builder_node*: This node builds a point cloud[4].

These communicating nodes are shown in Figure 10.

Data format translation Particularly because we opt for ROS as our main software framework, we need to translate our existing dataset into ROS-compatible format. Thus, we translate all the data required to perform our point cloud generation. Those related are data obtained from laser, GPS,

[4]This node can be interchangeably used with assembler_client_node.

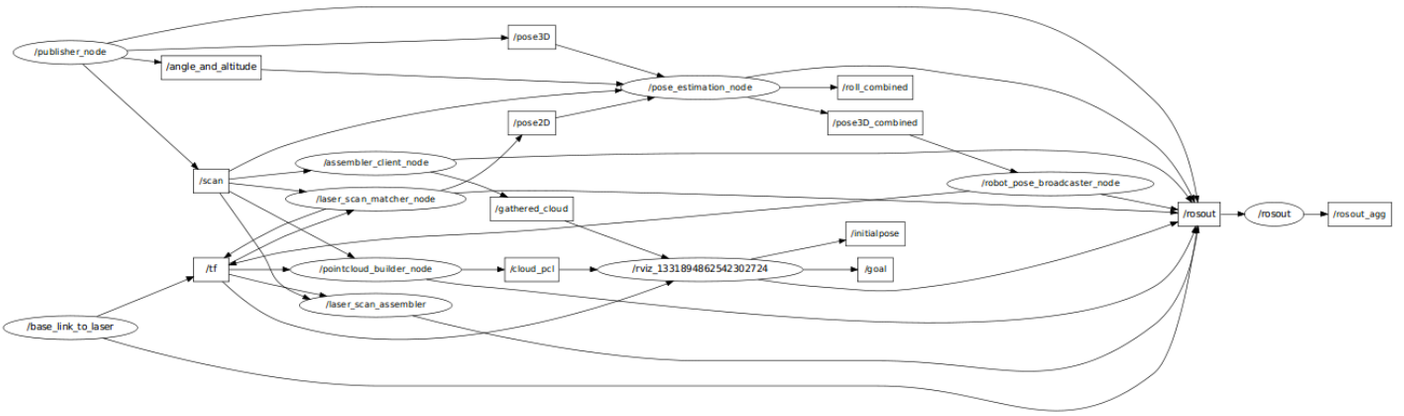


Figure 10: *ROS nodes communicate with each other in order to finally build a point cloud.*

barometer and IMU (roll, pitch and yaw). The translation is done by publishing them in order for other nodes to be subscribed when necessary. The translation is done in *publisher_node*.

Pose estimation The computation of pose estimation is done in *pose_estimation_node*. As already mentioned in Section 3.5 about the outputs by the *laser_scan_matcher* package, we utilize the position in x-axis, in particular. Thus, in this pose estimation, we are interested in using this x value to better *guess* the current height. We first take the barometer reading as height initialization value. By referring to the Figure 11, we take the difference between previous x-value (x1) and current x-value (x2) which gives d. This d is equivalent to the difference of height in two consecutive time steps. Conceptually, by integrating this d value with the current height over time, we should get a better height estimation. The output of the pose estimation is useful for robot_pose_broadcaster_node, as when it is used to broadcast the current transformation of the MAV's body with respect to the world frame.

Transformation Transformation in ROS is provided by *tf* library[5]. This library is very useful in order for computing any transformation. We utilize it in order to compute the static and non-static transformation. In static transformation, we use one of the nodes called *static_transform_publisher* to generate the transformation of *laser* with respect to *base_link*. This is considered as a static transform due to the fact that there exist a fixed link connecting the base of MAV and the laser scanner. This relation is shown in Figure 12. The *robot_pose_broadcaster_node* broadcasts the transformation of the *base_link* with respect to the *map*. Since, we have these two transformations done by the *tf*, we can infer or *listen* (in ROS terminology) the transformation of the laser with respect

[5]http://ros.org/doc/electric/api/tf/html/c++/

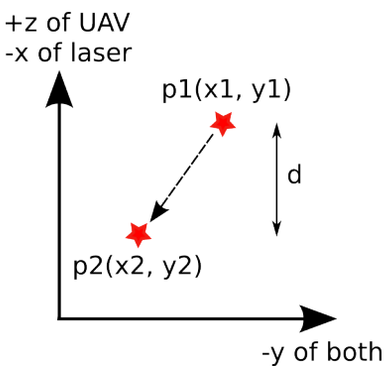


Figure 11: *The diagram shows the representation of the output of the* laser_scan_matcher. *The coordinates (x1, y1, x2, y2) correspond to the output of the package.*

to *map*. In other words, based on Figure 12, if we know a and b, then, we can get c.

Point cloud building There are two ways how we build the point cloud. The first one is by using an existing ROS package named *laser_assembler*. In order to use this one, we need to write a client node that defines the needed service parameters like start time, end time and so on. We name our client node as *assembler_client_node*.

The second method is done by building the cloud ourselves. The corresponding node is called *point-cloud_builder_node*. We first need to transform scan lines into point cloud datatype which we accumulate them in a buffer. The last accumulated point cloud in the buffer is the resulting point cloud.

Regarding the second method, essentially, the *point-cloud_builder_node* first needs to listen to a transformation of laser with respect to the map, given that the transformation of laser with respect to the base link is known. The transforma-

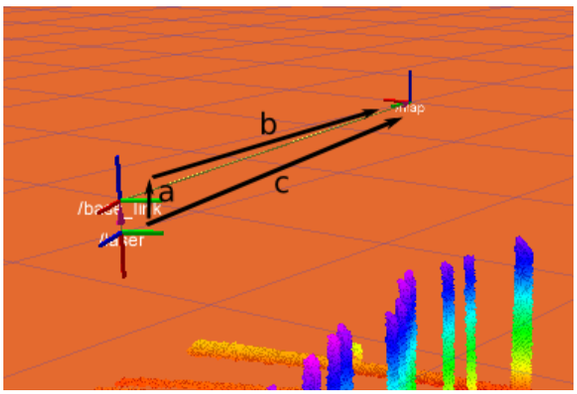


Figure 12: *We set-up transformation of a and b using* tf*, and hence, we can get the transformation of c by using the* tf.

tion of laser with respect to the base link is a static transform. When building the cloud, we first convert from *LaserScan* format to *PointCloud*. We then gather all the points in an accumulator.

Another difference between these two methods is that the former method is working on *sensor_msgs::PointCloud*[6] format while the latter is working on *sensor_msgs::PointCloud2*[7] format.

4 Evaluation

4.1 Experimental design

Nüchter has mentioned in [10] that there exist no standard evaluation method that can be adopted in order to evaluate a generated map. Consequently, we propose a naïve method to measure the performance of our approach which is adequate to our application. Thus, we distinguish our design of experiments into qualitative and quantitative method.

In qualitative method, we evaluate the performance of the developed approach by looking at the generated point cloud. We differentiate the output by assigning three classifications: GOOD, MEDIUM or BAD. This first approach is similar to the intent of Nüchter in [10].

The quantitative method can only be applied for the forth scenario since we found a way to measure such a performance. We have measured and collected some ground truth information regarding our environment. We place two boxes whose dimensions are identical on a flat ground surface as shown in Figure 5. Using this ground truth, we measure their ratio between width and length. Hence, the error E of the generated map can be calculated as follows:

$$E = \frac{|ratio_{measured} - ratio_{actual}|}{ratio_{actual}} \times 100\% \qquad (2)$$

$$E = \frac{|\frac{width_{measured}}{length_{measured}} - \frac{width_{actual}}{length_{actual}}|}{\frac{width_{actual}}{length_{actual}}} \times 100\% \qquad (3)$$

where

$ratio_{measured}$ = ratio between width and length of the box from generated point cloud

$ratio_{actual}$ = ground truth ratio between width and length of the box

$width_{measured}$ = width of the box output from generated point cloud

$length_{measured}$ = length of the box output from generated point cloud

$width_{actual}$ = width of the box output from the ground truth

$length_{actual}$ = length of the box output from the ground truth

and the performance P can be computed as,

$$P = 100\% - E \qquad (4)$$

4.2 Results

4.2.1 Scenario I

Without *laser_scan_matcher* The output of point cloud[8] that uses only data from localization of the MAV is shown in Figure 13.

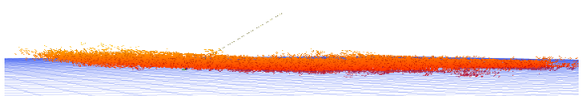


Figure 13: *The generated point cloud without using* laser_scan_matcher *information shown from the side view in scenario I.*

With *laser_scan_matcher* The output of point cloud that utilizes *laser_scan_matcher* information from localization of the MAV applied is shown in Figure 14.

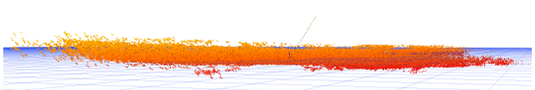


Figure 14: *The generated point cloud using* laser_scan_matcher *information shown from the side view in scenario I.*

[6] http://www.ros.org/doc/api/sensor_msgs/html/msg/PointCloud.html
[7] http://www.ros.org/doc/api/sensor_msgs/html/msg/PointCloud2.html

[8] All the generated point cloud outputs are in the form of RGB color-coded image in order to group points according to their heights.

4.2.2 Scenario II

Without *laser_scan_matcher* The output of point cloud that uses only data from localization of the MAV applied is shown in Figure 15.

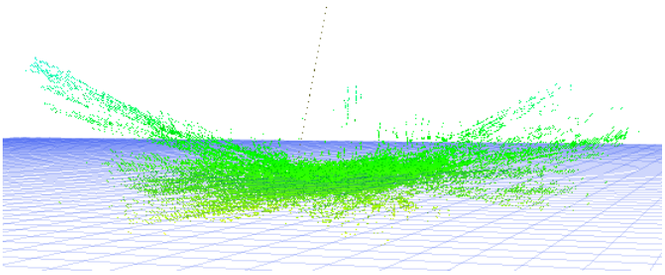


Figure 15: *The generated point cloud without using* laser_scan_matcher *information shown from the side view in scenario II.*

With *laser_scan_matcher* The output of point cloud that utilizes *laser_scan_matcher* information from localization of the MAV applied is shown in Figure 16.

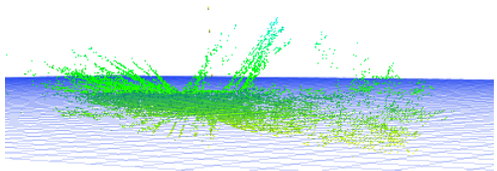


Figure 16: *The generated point cloud using* laser_scan_matcher *information shown from the side view in scenario II.*

4.2.3 Scenario III

Without *laser_scan_matcher* The output of point cloud that uses only data from localization of the MAV applied is shown in Figure 17.

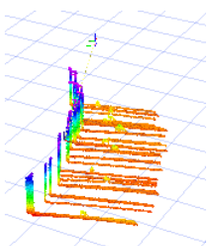


Figure 17: *The generated point cloud without using* laser_scan_matcher *information shown from the top view in scenario III.*

With *laser_scan_matcher* The output of point cloud that utilizes *laser_scan_matcher* information from localization of the MAV applied is shown in Figure 18.

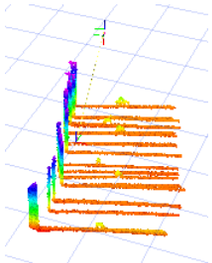


Figure 18: *The built point cloud using* laser_scan_matcher *information generated for scenario III.*

4.2.4 Scenario IV

Without *laser_scan_matcher* The output of point cloud that uses only data from localization of the MAV applied is shown in Figure 19 and 20.

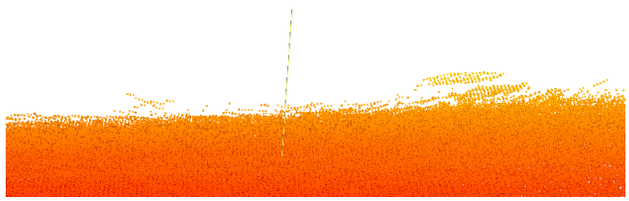


Figure 19: *The boxes in scenario IV without using* laser_scan_matcher *information.*

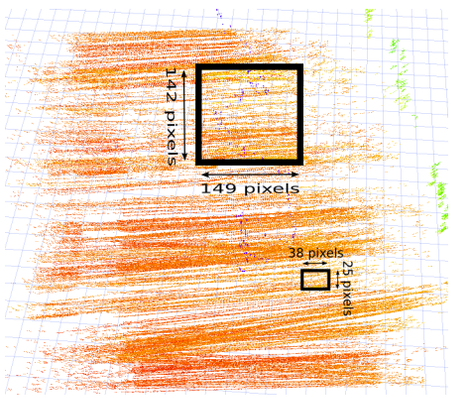


Figure 20: *The boxes are marked with their corresponding measurement shown from the top view in scenario IV without using* laser_scan_matcher *information.*

With *laser_scan_matcher* The output of point cloud that uses only data from localization of the MAV applied is shown in Figure 21 and 22.

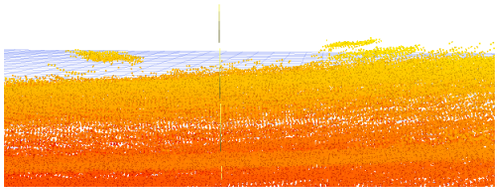


Figure 21: *The boxes in scenario IV using* laser_scan_matcher *information.*

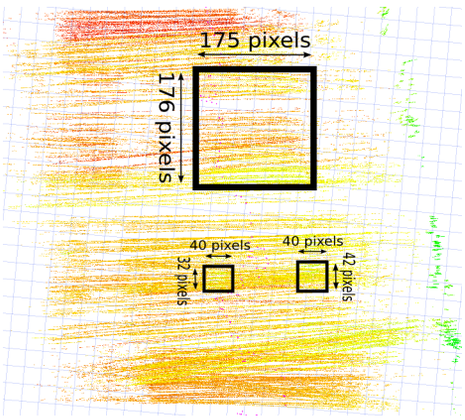


Figure 22: *The boxes are marked with their corresponding measurement shown from top view in scenario IV using* laser_scan_matcher *information.*

4.3 Analysis

From the output generated in Section 4.2, we can say that the 3D point cloud has successfully been generated. After analyzing the output, we summarize it in Table 1. While the generated point cloud in scenario III achieved our expectation, the problem remains challenging for other scenarios. There are some insights and considerations that may lead to any future betterment:

- As we had performed a qualitative evaluation - with an exception we did mention some figures in some extreme cases - on the feasibility of the laser scan matching algorithm, it can be evaluated more accurately if a quantitative evaluation that incorporates some statistical measures is taken into account.

- It can also be the case that *laser_scan_matcher* which was used for mobile robot application whose the coordinate frame of the sensor is exactly aligned with the robot coordinate frame is not suitable for our setting where the sensor frame has an offset with the robot(MAV) frame by 90 degrees.

- The dynamics e.g. vibration that imposes on the MAV much affecting the generated point cloud especially in scenario II. That is why in scenario II, the generated point cloud is rather very good as it was hovered manually with minimal dynamics.

Scenario	**Without LSM**	**With LSM**
Scenario I	MEDIUM*(-)	MEDIUM(-)
Scenario II	BAD(-)	BAD(-)
Scenario III	GOOD(-)	GOOD*(-)
Scenario IV	MEDIUM(71.54%)	MEDIUM(91.56%)

Table 1: *This table compares the performance resulting by using* laser_scan_matcher *information (written as With LSM) and without using it (written as Without LSM). The performance by qualitative measure is given in capital letters while the one following in bracket is the performance by quantitative measure. The ones with asterisk(*) implies it is a bit better than without it in performance.*

The values of 71.54% and 91.56% in the table 1 are obtained from the equation 4.

5 Conclusion

In this paper, we presented an approach for generating a point cloud for ground modeling application. We demonstrate how this can be pragmatically achieved. We also attempt to improve the accuracy of the generated point cloud by incorporating laser scan matching algorithm.

Furthermore, we devise a basic yet useful algorithm called *BacktransformLaserScan* that is beneficial in evaluating a laser scan matching algorithm for specific application. The proposed algorithm is not restricted to *laser_scan_matcher*, it is also applicable for any other variants of laser scan matching algorithm (or software packages) as well.

Acknowledgements

This work was a collaboration between Bonn-Aachen International Center for Information Technology (B-IT) and University of Siegen. The authors would like to thank Stefan Thamke, Rainer Herpers and Ibtissem Ben Makhlouf for their helpful insights and comments.

References

[1] S. Thrun. Robot mapping: A survey. *Exploring Artificial Intelligence in the New Millenium, Morgan Kaufmann*, 2002.

[2] Slawomir Grzonka, Giorgio Grisetti, and Wolfram Burgard. Towards a navigation system for autonomous indoor flying. In *ICRA*, pages 2878–2883, 2009.

[3] C. Fruh and A. Zakhor. 3d model generation for cities using aerial photographs and ground level laser scans. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society*

Conference on, volume 2, pages II–31 – II–38 vol.2, 2001.

[4] H. Surmann, K. Lingemann, A. Nüchter, and J. Hertzberg. A 3d laser range finder for autonomous mobile robots. In *Proceedings of the 32nd ISR (International Symposium on Robotics)*, volume 19, pages 153–158. Citeseer, 2001.

[5] Shaojie Shen, Nathan Michael, and Vijay Kumar. Autonomous indoor 3d exploration with a micro-aerial vehicle. In *ICRA*, pages 9–15, 2012.

[6] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.

[7] A. Diosi and L. Kleeman. Laser scan matching in polar coordinates with application to slam. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 3317–3322. IEEE, 2005.

[8] A. Censi. An icp variant using a point-to-line metric. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 19–25. IEEE, 2008.

[9] I. Dryanovski, W. Morris, and J. Xiao. An open-source pose estimation system for micro-air vehicles. In *Robotics and automation (ICRA), 2011 IEEE international conference on*, pages 4449–4454. IEEE, 2011.

[10] Andreas Nüchter. *3D Robotic Mapping - The Simultaneous Localization and Mapping Problem with Six Degrees of Freedom*, volume 52 of *Springer Tracts in Advanced Robotics*. Springer, 2009.

A Fuzzy Logic Controller for Thrust Level Control of Liquid Propellant Engines

Akbar Allahverdizadeh
School of Engineering-Emerging Technologies
University of Tabriz
Tabriz, Iran
allahverdizadeh@tabrizu.ac.ir

Behnam Dadashzadeh
School of Engineering-Emerging Technologies
University of Tabriz
Tabriz, Iran
b.dadashzadeh@tabrizu.ac.ir

***Abstract*— Thrust level control of liquid propellant engines is investigated in this paper. The dynamic equations of liquid propellant engines are formulated and a PID and a fuzzy controllers are designed to control its thrust level. Fuzzy logic deals with problems that have intrinsic or informational imprecision in definition of objective function or constraints. So fuzzy controller can be a good choice to control nonlinear systems like liquid propellant engine. Both PID and fuzzy controllers can control the engine thrust level well. Their performances are compared and investigated.**

Keywords— fuzzy control; liquid propellant engine; thrust level control.

I. INTRODUCTION

Liquid Propellant Rocket Engines (LPRE) are used in various operational Launch Vehicles all over the world. Based on the type of propellants used, they are classified as Earth-storable, Semi-cryogenic and Cryogenic rocket engines [1]. LPRE's are also classified as pump-fed engines and pressure–fed engines based on the propellant feed system. The Liquid Rocket Engine consists of Thrust chamber, propellant feed system, control components, ignition system, pre-conditioning system, control system etc. The control systems ensure proper functioning of the engine system with the desired performance. The basic Liquid-Propellant-Engine Control Systems are [3]:

- Engine start sequence control
- Engine cut off sequence control
- Engine duration control
- Engine safety control
- Propellant-tank pressurization control
- Engine-system Checkout and test controls
- Thrust vector control by gimballing the engine.
- Engine thrust-level control
- Propellant mixture ratio / Propellant utilization control

The control systems interconnect the components and logics designed to yield a desired response or output based on a command or reference input. Selection of the control method best suited for the propulsion system is influenced by the performance requirements, accuracy, and dynamic characteristics of the engine being controlled and particularly by engine reaction-time. This paper investigates mathematical modeling of combustion chamber and gas generator and then designing controllers for thrust level control of the engine.

Thrust vector control by Gimballing LPREs:

Steering of a vehicle over the desired trajectory employs thrust vector control systems (TVC). One of the methods of TVC is by gimballing either the main engine or by Gimballing vernier engines. Based on the vehicle trajectory the onboard computer generates the necessary error signal and gimbals the main engine using actuators. This topic is not in scope of this paper.

Thrust and Mixture Ratio Control (MRC) Systems:

As the Liquid Propellant engine and stage systems are configured and the propellant is loaded considering the optimum thrust and mixture ratio requirements, it is possible to achieve the safe engine operation, required vehicle performance and minimum propellant outage only if the engine is operated at the specified thrust and mixture ratio. Deviation from the requirements could be caused by factors such as engine tuning error, deviation in pressure and temperature of propellants at engine inlets etc. Deviation in thrust and mixture ratio can lead to either under performance and additional propellant outage or engine malfunctioning and failure. In order to ensure safe engine operation and optimum performance of the vehicle, it is essential to regulate the thrust and mixture ratio within the specified limits. Engine thrust and mixture ratio may be controlled by controlling the propellant flow to the engine, either in open loop mode or closed loop mode.

Thrust Control Schemes:

In pressure fed engine, pre-calibrated flow control devices such as orifices or venturies are used in the propellant feed circuits, to maintain the thrust within the specified limits in open loop mode and variable area flow control valves in the feed circuits or propellant tank pressure variation is used for controlling the thrust in close loop mode [2]. In pump-fed

Engines the thrust is regulated by controlling the power generated by the turbines by controlling the flow rate to the turbines. In the open-loop thrust control mode, pre-calibrated flow control elements are used for controlling the throughput to the turbines. In the case of GG cycle or staged combustion cycle engines, the propellant flow to the GG/ Pre Combusting Chamber (PCC) is controlled using fixed area orifices or venturies. Similarly in the case of engines working on other cycles, the thrust control is achieved by using pre-calibrated control elements in the turbine feed lines. In the closed loop control mode, flow to the turbine is controlled by variable area flow control valves. The thrust of the GG/SCC engines may be regulated either by controlling the propellant flow to the GG/PCC or by adjusting the hot gas flow from GG/PCC outlet to the turbines. The engine thrust is controlled in closed loop mode, by using either engine parameters (chamber pressure or propellant injection pressure) or vehicle parameters (vehicle acceleration or incremental velocity) as feedback signals [9].

Earth storable engines generally employ pneumo/hydraulic systems for thrust and mixture ratio control. Thrust is controlled by controlling the chamber pressure. The thrust control regulator uses a piston, balanced by the chamber pressure feedback on one side and the required chamber pressure fed as command pressure on the other side. Any unbalance will move the piston thereby changing the propellant flow rate to gas generator resulting in an increase or decrease in chamber pressure as required. Since the effect of propellant temperature on mixture ratio is negligible, the mixture ratio is controlled by controlling the thrust chamber inlet pressures. The mixture ratio control regulator equalizes the oxidizer and fuel pressures at thrust chamber inlet by means of a balancing piston. The required mixture ratio is ensured by suitably sizing the calibrated orifice mounted in the propellant line.

There are limited references in the open literature on the design of a combustion chamber control system. A class of literature on dynamic analysis and control system design for liquid propellant engines is limited to linear models of the engine utilizing a linear systems theory [5],[6]; therefore such designs would not be as robust as if the design were based on a nonlinear model of the engine. In the other class a nonlinear system design approach is employed utilizing nonlinear systems design techniques for use with reusable rocket engines and the design of the regulator loop is assumed to be based on a standard robust design that is used for linear systems [7]. The work presented herein for the design of the regulator control loop is based on a simplified mathematical model of the engine thrust force. The contribution of our work includes to show efficiency of fuzzy controller to adjust thrust level of LPREs.

II. MATHEMATICAL MODEL

LPRE is composed with combustor, turbopump, turbine, control valve, gas generator, pipes and so on as Fig. 1. In development phase, thrust control is one of the important requirements of LPRE. Also, mixture ratio control of propellants fed into combustor and gas generator is needed for safe operation of LPRE. For control of LPRE, 3 control valves are installed at the LOX line of gas generator, the main LOX line and the main fuel line of combustor.

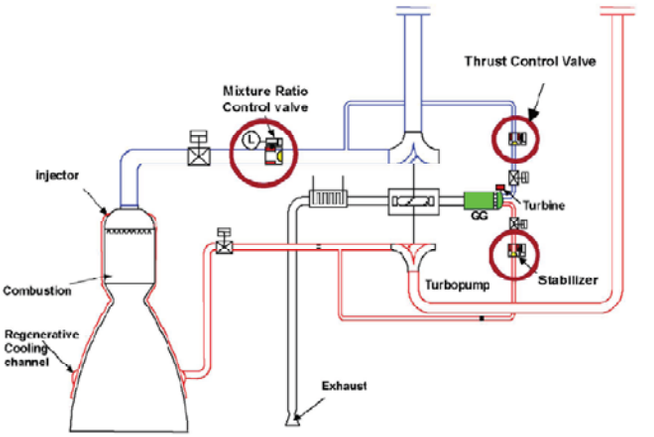


Fig. 1. Schematic of open type LPRE [4]

These simplifying assumptions have been considered in mathematical modeling:
The flow is incompressible,
Effects of temperature variations are ignored,
The output pressure of fuel and oxidizer tanks remain constant,
Flow density and viscosity is considered constant,
Gas flow in combustion chamber and gas generator is assumed to be adiabatic and non-viscous.

2.1 Main combustor and gas generator

In both main combustor and gas generator, combustion occurs and we take use of the outlet gas pressure, therefore their model are similar. We start with continuity equation for the gas inside combustion chamber or gas generator:

$$\dot{m}_{in}(t-\tau)=\dot{m}_{out}(t)+\frac{dm}{dt} \tag{1}$$

in which $\dot{m}_{in}$ is propellant flowrate into the chamber, $\dot{m}_{out}$ is outlet flowrate of propellant from the chamber, τ is combustion time constant and m is the mass of propellant inside the chamber. In this equation we have

$$\dot{m}_{in}(t-\tau)=\dot{m}_{ox}(t-\tau)+\dot{m}_{fu}(t-\tau) \tag{2}$$

So,

$$\dot{m}_{in}(t-\tau)=\dot{m}_{fu}(t-\tau)+\dot{m}_{ox}(t-\tau)=\dot{m}_{out}(t)+\frac{dm}{dt} \tag{3}$$

With the assumption of ideal gas

$$\mathrm{m}=\frac{\mathrm{PV}}{\mathrm{RT}} \tag{4}$$

If we assume that RT is constant, derivation of equation (4) yields to

$$\frac{dm}{dt} = \dot{m}_{in}(t-\tau) - \dot{m}_{out}(t-\tau) = \frac{V}{RT}\frac{dp}{dt} \quad (5)$$

In combustion chamber and gas generator, characteristic velocity is defined as

$$C^* = \frac{P.A_t}{\dot{m}_{out}} \quad (6)$$

and characteristic length is defined as

$$L^* = \frac{V}{A_t} \quad (7)$$

Therefore characteristic velocity is written as

$$C^* = \frac{\sqrt{R.T}}{\Gamma} \quad (8)$$

in which

$$\Gamma = \sqrt{k}\left[\frac{2}{k+1}\right]^{\frac{k+1}{2(k-1)}}. \quad (9)$$

Substituting these equations into (5) we will have

$$\frac{dm}{dt} = \frac{L^* A_t}{C^{*2}\Gamma^2}.\frac{dp}{dt} \quad (10)$$

Substituting equations (6) and (10) into (3) yields to

$$\frac{L^* A_t}{\Gamma^2 C^{*2}}\frac{dp}{dt} + \frac{A_t}{C^*} P_{cc} = \dot{m}_{in}(t-\tau) = \dot{m}_{fu}(t-\tau) + \dot{m}_{ox}(t-\tau) \quad (11)$$

This is the equation of combustion process. This equation is based on steady combustion and evaporation time is neglected.

2.2 Turbopump

In turbopump complex, the high pressure outlet gas from gas generator rotates turbine and this rotation is transferred to pumps of the main fuel and oxidizer lines.

Dynamic equation of torque is

$$TQ_{turbine} - TQ_{pump,fu} - TQ_{pump,ox} = J_{eq} \times \dot{\omega} \quad (12)$$

In which $TQ_{turbine}$ is turbine generated torque, $TQ_{pump,fu}$ and $TQ_{pump,ox}$ are consumed torques of fuel and oxidizer pumps, J_{eq} is equivalent moment of inertia of turbine and $\dot{\omega}$ is angular acceleration of its shaft.

The pump model using Avsianikeve equation is

$$\frac{H}{\omega^2} = (A + B(\frac{Q}{\omega}) - C(\frac{Q}{\omega})^2)/g \quad (13)$$

In which H is pump head, Q is flowrate, ω is angular velocity of pump and g is the gravity.

2.3 Pipelines and valves

Losses of flow is due to friction ΔP_f and resistances in flow path ΔP_R.

$$\Delta P = \Delta P_f + \Delta P_R \quad (14)$$

$$\Delta P_f = f\frac{L}{d}\frac{\rho V^2}{2} \quad (15)$$

$$\Delta P_R = k\frac{\rho V^2}{2} \quad (16)$$

In which ρ is density of fluid, V is its velocity, f is friction coefficient, L is length of pipe, d is its diameter and k is coefficient of minor losses.

2.4 Thrust control valve

The model of valve body can be simply written as

$$\begin{aligned} \dot{m} &= K_v\sqrt{2\rho\Delta P} \\ L_v &= \frac{1}{T_v s + 1}U_v \\ K_v &= A_v e^{0.04L_v} \end{aligned} \quad (17)$$

2.5 The overall model of LPRE

The overall model can be composed using above equations.

III. DESIGN OF PID CONTROL SYSTEM

The control system consists of pressure control of combustion chamber for thrust control of LPRE. The pressure of combustion chamber is controlled with thrust control valve operated by PI control or fuzzy control. Now having the simplified thrust model in Simulink we can design controllers. Fig. 2 shows the model and controller. Since there is restriction for flowrate of fuel and oxidizer we have used a saturation block after PID controller to restrict control signal in acceptable range.

Proportional–integral–derivative (PID) control of the engine was attempted first, since it is a popular closed-loop control approach that can be applied to a wide range of engineering problems.

Using automated PID tuning in Matlab SISO tool, this optimized PID controller was obtained:

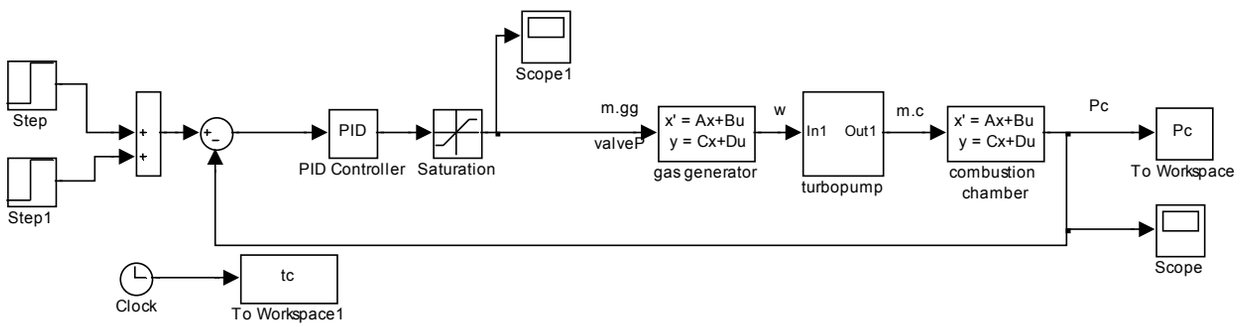


Fig. 2. PID control system of thrust level

$$C = 0.00016 \times \frac{\left(1 + 2.6\times10^{-5}\, s\right)\left(1 + 4.2\times10^{-4}\, s\right)}{s}$$

Figs. 3,4,5 show root locus, bode diagram and step response of this controller.

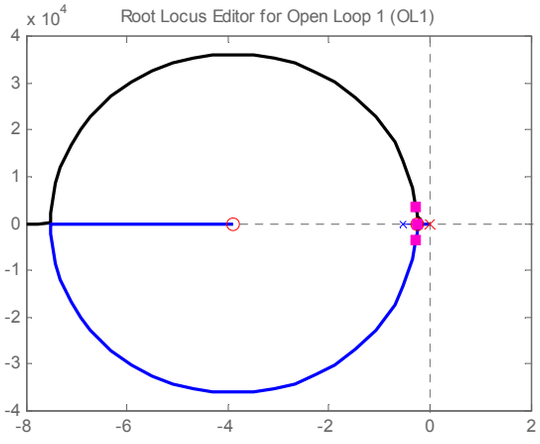


Fig. 3. Open loop root locus with optimized PID controller

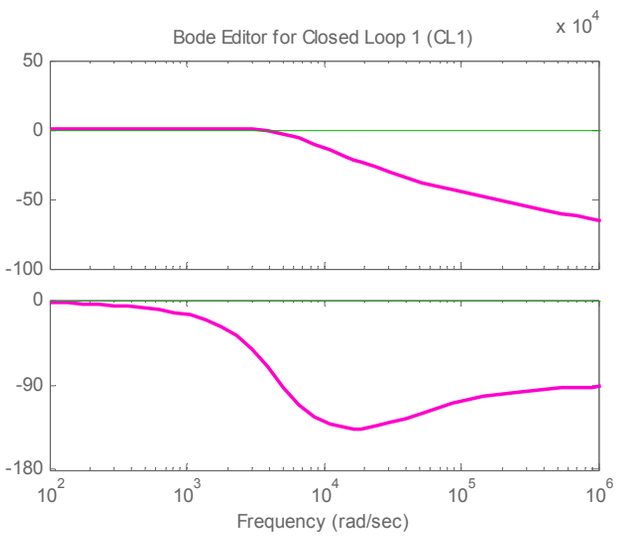


Fig. 4. Closed loop bode diagram with optimized PID controller

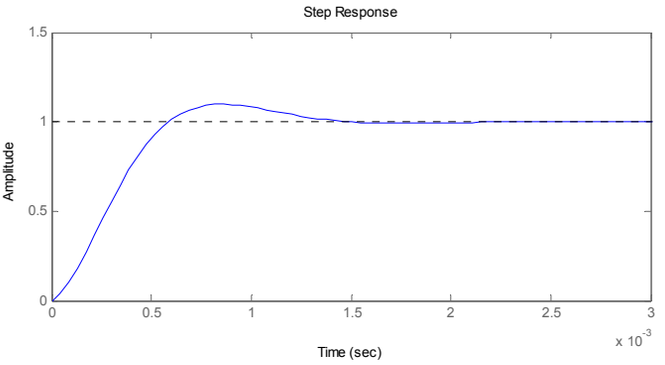


Fig. 5. Step response with optimized PID controller

IV. FLC DESIGN AND APPLICATION

The employment of fuzzy techniques belongs to the "soft computing'' family of algorithms. Here "soft computing'' refers to computational mechanisms that can determine suitable relationships (in a system data set) to assess and determine a quantitative opinion(s) based on future conditions. Within MSFC, such computational mechanisms are viewed as a collection of algorithms that can achieve optimal or near-optimal results in the presence of imprecise data, uncertainty [8], unknown physics, and probabilistic outcomes. The central goal in soft computing is to obtain greater robustness to these and other uncertainties.

Similar to the PID design, for main thrust level control, a FLC is designed and the response of the engine to a step input using the PID and fuzzy controllers is compared. The use of fuzzy logic is seen to be suitable since it accommodates the uncertainties associated with the engine. The technology of fuzzy logic enables a computer to make decisions based on vagueness or imprecision intrinsic in most physical systems. Fuzzy logic also provides a convenient way to introduce useful nonlinearities into the control law to achieve specific effects, such as reducing large overshoots.

A fuzzy controller was designed with two inputs and one output. The input variables used in the design of the controller were thrust error (e) and thrust error rate (ei). The thrust error is defined as the desired thrust minus the actual thrust in pounds. The thrust error rate is the change in thrust error in one sampling interval. The defined membership functions

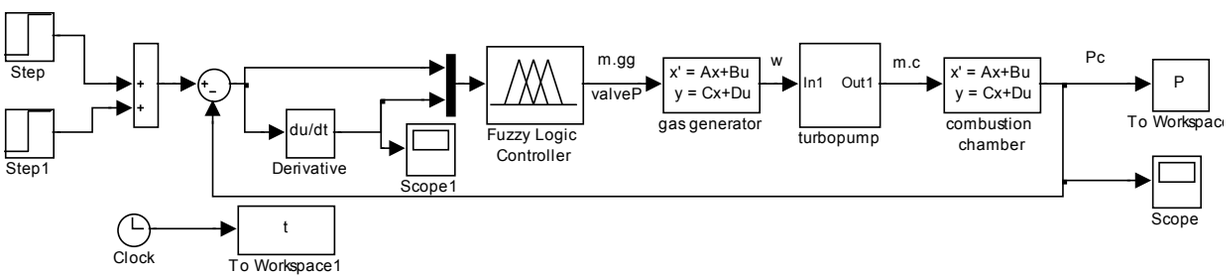


Fig. 6. Fuzzy control system of thrust level

have been shown in Figs. 7-9. We have used Gaussian membership functions for e and triangular membership functions for $\dot{e}$ and control valve position.

Our rule-base consists of ten fuzzy rules that have been constructed using heuristics and experience, as follows:

- 1. If (e is pF) then (V is pH) (1)
- 2. If (e is nF) then (V is nH) (1)
- 3. If (e is pC) then (V is pM) (1)
- 4. If (e is nC) then (V is nM) (1)
- 5. If (e is Z) then (V is Z) (1)
- 6. If (e is pN) then (V is pL) (1)
- 7. If (e is nN) then (V is nL) (1)
- 8. If (ei is nL) then (V is nM) (1)
- 9. If (ei is Z) then (V is Z) (1)
- 10. If (ei is pL) then (V is pM) (1)

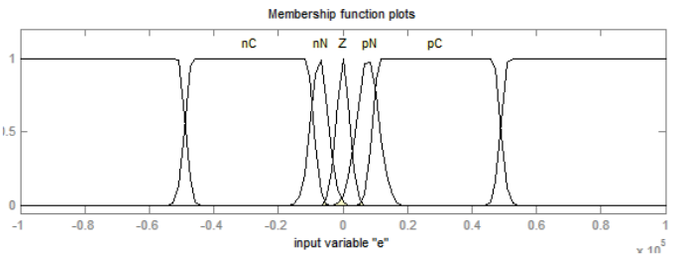


Fig. 7. membership functions for e

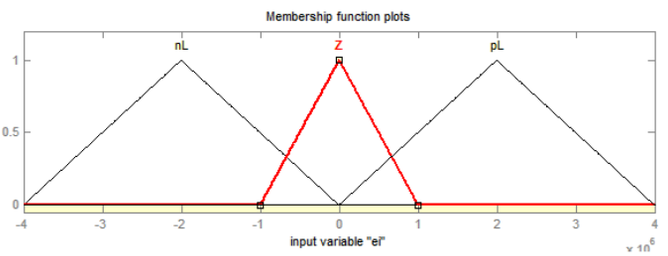


Fig. 8. membership functions for $\dot{e}$

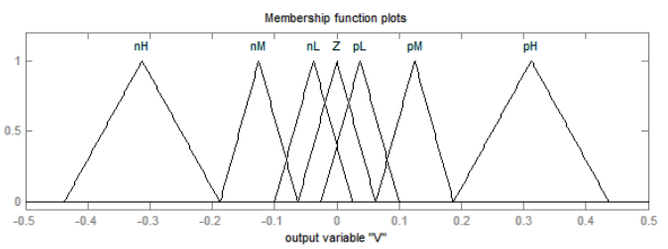


Fig. 9. membership functions for valve position

We define a desired P_c with two steps and apply our optimized PID controller and designed fuzzy controller to this input. The desired chamber pressure values are 40bar, 43bar and 45bar. Results are shown in Figs. 10-13. It can be seen that the controlled thrust of optimized fuzzy controller has less oscillations than fuzzy controller, but both of the controllers have an acceptable performance.

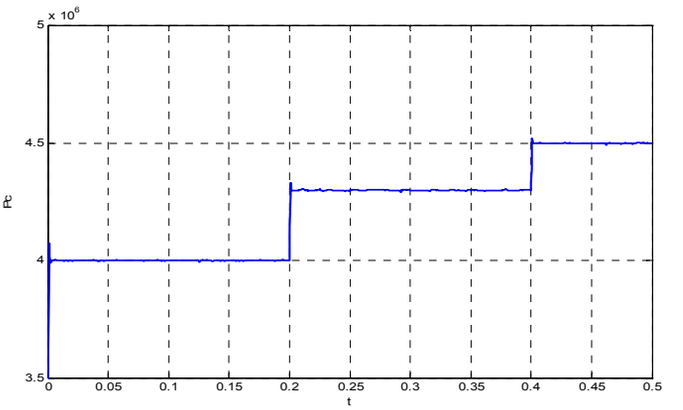


Fig. 10. Controlled chamber pressure using PID controller

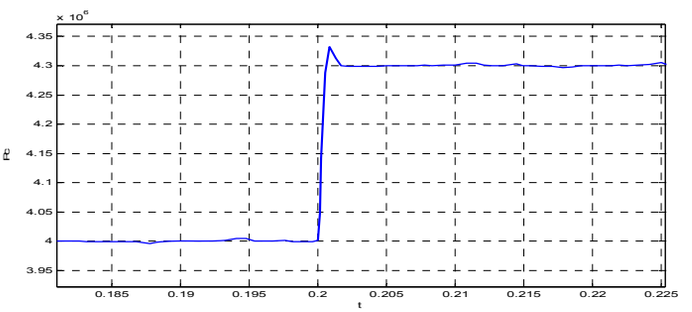


Fig. 11. Controlled chamber pressure using PID controller (zoomed)

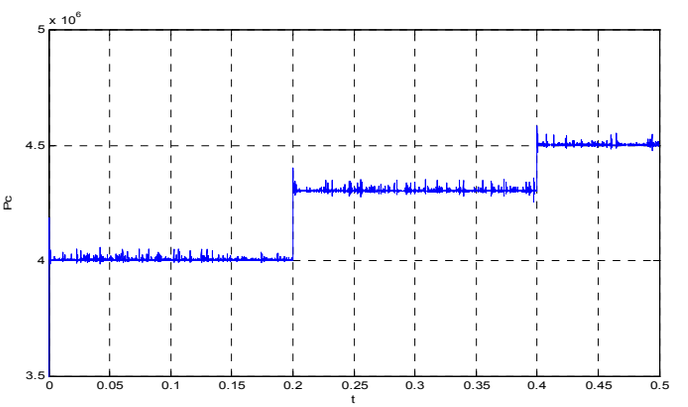


Fig. 12. Controlled chamber pressure using fuzzy controller

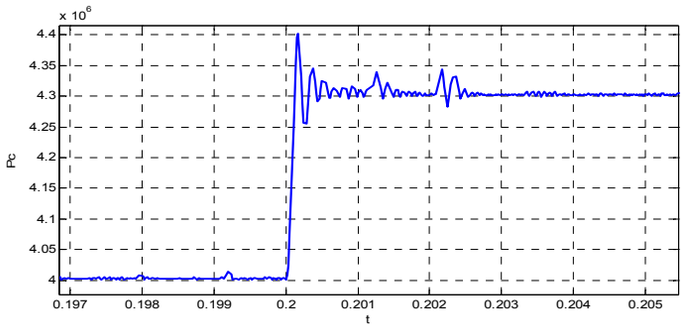


Fig. 13. Controlled chamber pressure using fuzzy controller (zoomed)

V. Concluding Remarks

The mathematical model of LPRE was established with the 1st order ordinary differential equations. For the control of thrust level of LPRE, we established two control systems, an optimized PID controller with saturators and a fuzzy controller using the parameters of the optimized PID controller to choose the values of fuzzy membership functions. We defined a desired chamber pressure with two constant levels and applied the controllers to adjust the pressure. Both of the controllers have an acceptable performance but the optimized PID controller had overall better results. This has two reasons; first, the parameters of PID controller was optimized but parameters of fuzzy controller was chosen using the optimized PID controller results but they were not directly optimized; second, we have used simplifying assumptions to model the motor with linear differential equations. Actual LPRE is a nonlinear system and probably the designed fuzzy controller will has better results if implemented on real motor, because fuzzy controller is robust against uncertainties and nonlinearities.

VI. References

[1] Huu P. Trinh, William Neill Myers, "Injector element which maintains a constant mean spray angle and optimum pressure drop during throttling by varying the geometry of tangential inlets," US8763362 B1 Grant US 13/452,303, Jul 1, 2014.

[2] Zachary W. Peterson, Closed-Loop Thrust and Pressure Profile Throttling of a Nitrous Oxide/Hydroxyl-Terminated Polybutadiene Hybrid Rocket Motor, Master of Science (MS) Thesis, Mechanical and Aerospace Engineering Department, Utah State University, 12-2012.

[3] V. Gnanagandhi, "Liquid Propellant Rocket Engine Control System," Workshop on engine control system technology IIT, MUMBAI November 2004.

[4] Young-Suk Jung, Seung-Hyub Oh, "Thrust and Propellant Mixture Ratio Control of Open type LPRE using Q-ILC," International Conference on Control, Automation and Systems, in COEX, Seoul, Korea, 2007.

[5] Santana, A. Jr. Barbosa, F.I. and Niwa M., "Modeling and robust analysis of a liquid rocket engine," 36th AIAA Joint Propulsion Conference Exhibit, Huntsville, Alabama, 2000.

[6] Schinstock, D.E., Scot, D.A. and Haskew, T.A., "Modeling and estimation for electromechanical thrust vector control of rocket engines," AIAA J. of Propulsion and Power, 1998.

[7] Lorenzo C.F., Ray A. and Holmes M.S., "Nonlinear control of a reusable rocket engine for life extension," AIAA J. of Propulsion and Power, 2001.

[8] J. Steincamp, First Annual Report, Marshall Space Flight Center, Huntsville, AL, September 2001.

[9] Hui Tian, Peng Zeng, Nanjia Yu, Guobiao Cai, "Application of variable area cavitating venturi as a dynamic flow controller," Flow Measurement and Instrumentation 38, 21-26, 2014.

Hopping Gait Generation for a Biped Robot with Hill-Type Muscles

Behnam dadashzadeh, Mohammad Esmaeili, Behrooz Koohestanim, M.R. Seyed Noorani
School of Engineering-Emerging Technologies
University of Tabriz
Tabriz, Iran
b.dadashzadeh@tabrizu.ac.ir, mohammad.esmaeili91@ms.tabrizu.ac.ir

***Abstract*—This work presents a novel gait generation method for biped hopping with point feet. The investigated biped model consists of a kneed massless leg and a trunk with two Hill-type muscles at the hip and knee joints. The dynamic equations of the system are derived using Lagrange method. Since the most important phase to stabilize bipedal running and hopping motion is stance phase, this paper deals only with the stance phase and develops a control law for the actuators to generate an arbitrary trajectory for hip of the underactuated biped robot. Without loss of generality, a forth order curve with properly chosen parameters has been used as the desired robot trajectory in stance phase. This curve has similarities to the Spring Loaded Inverted Pendulum gait, and can generate the desired initial and final position and velocity of the stance phase. Hill-type muscles are used as the actuators of our model, because being simple it includes fundamental elements that are necessary for biped running efficiency. The proposed control law calculates the needed actuators inputs and gets the robot to undergo the desired trajectory. The designed control law is verified in simulation.**

Keywords—biped hopping; gait generation; Hill-type muscle

I. INTRODUCTION

Similarity to humans and having high performance and maneuverability has fascinated researchers to investigate bipedal locomotion more deeply in recent years. Energy efficiency makes it necessary to switch from walking to running as the progression speed increases [1]. To use compliant elements in legs reduces touch-down impact transferred to robot links, increases energy efficiency of bipedal running, and generates more natural looking gaits. However, compliant structure of the robot makes controller design more difficult. Biped running motion consists of stance phase, take-off, flight phase, and touch-down. In stance phase one leg is pivoted to the ground and another leg is swinging. For our biped model with point feet, the robot has one degree of underactuation in stance phase. In flight phase the robot has no contact point with the ground and undergoes a ballistic motion, so it has three degrees of underactuation. Most of the control effort to reject disturbances and stabilize biped running is done in the stance phase and so in this paper we concentrate to control this phase only. Touch-down is an instantaneous phase in which a leg contacts to the surface and the robot switches from flight phase to stance phase. Touch-down impact causes an instantaneous change in robot links velocities and can damage motors if no springy elements are used in legs of the robot.

Spring Loaded Inverted Pendulum (SLIP) is a simple and useful compliant model that was proposed by Blickhan [2] in 1989 to describe biped walking and running dynamics. This theoretical model consists of a point mass at the hip and two massless springs as robot legs. This model generates human's center of mass (CoM) trajectory and ground reaction force (GRF) profiles for walking and running [3]. The dynamic model of SLIP has no closed form solution but it has been solved approximately with small angles hypothesis [4]. In the general case it should be solved numerically. Although SLIP running gaits cannot be implemented directly on multibody robots, we use it to plan the desired trajectory for the biped robot investigated in this paper.

Animals and human running takes advantages of viscoelastic properties of their muscles. Hill [5] in 1949 proposed a mechanical model for muscle consisting of: (1) a main contraction part, (2) a passive elastic part series to 1, and (3) a passive elastic part parallel to 1 or parallel to 1 and 2. The part 3 prevents the actuator reaction to small initial loads. The passive elastic part series to the main active contraction part plays an important role in mechanical behavior of muscles. This part as an spring accumulates energy when a high tension is exerted to the muscle during its sudden change from rest to active state [6]. Ahmadi et al. [7] in 1997 presented a control method for a one legged hopper robot with compliant elements series to hip motor. They found unstable passive hopping gaits for their model and designed a controller to stabilize the hopping motion around its passive gait. Hyon et al. [8] in 2004 proposed an energy preserving controller for planar biped hopping with torso. Their controller aimed to preserve energy during touch-down. Sato et al. [9] in 2004 modeled a robot with one springy leg and a motor at hip to study SLIP model. They generated running motion in simulation and experiment with the velocity of 0.8 m/s. Meghdari et al. [10] in 2008 proposed a feedback linearization controller to follow SLIP trajectory by a three link rigid hopper. Their controller could generate stable hopping motion in simulation starting from Poincare map fixed point initial condition. Iida et al. [11] in 2008 designed a biped robot with only one motor at the hip and passive springs as muscles at the knees and ankles. Their robot could generate walking and running like motions in experiments. Eilenberg et al. [12] in 2009 proposed an adaptive

muscle-reflex controller using Hill-type muscle model for flexor-plantar muscles of ankle joint. Oyama et al. [13] in 2009 designed and fabricated KenkenII robot which had springs similar to tendons of organisms. This robot was designed to run with 0.7 to 0.9 m/s using a robust controller to ground disturbances. Andrews et al. [14] in 2011 made successful experiments based on this controller for a one legged planar hopper. Their controller could stabilize hopping on unknown ground ups and downs less than %25 of the free leg length. Dadashzadeh et al. [15] in 2014 generated biped running gaits with constant motor torques for a biped model having springs parallel to motors. Then they stabilized it using a state feedback controller on the Poincare map.

In this paper we consider a one legged hopper with three links and two Hill-type muscles as actuators. The trunk angle is assumed to be constrained as vertical any time. The dynamic equations are derived using Lagrange equation. A symmetric trajectory is planned for the stance phase consistent with initial condition of the robot. Then the necessary angles, velocities and accelerations of the robot links are calculated using inverse kinematics and inverse dynamics to make the robot undergo the desired trajectory. Using the calculated parameters, the control inputs of the robot can be found at any time. The designed control law is verified in simulation. The main contribution of this paper is developing a control law for underactuated biped hopper to make the robot to undergo any arbitrary trajectory. The other contribution is generating efficient hopping gaits using Hill-type muscle model.

II. Trajectory Planning

The first step in our controller design procedure for stance phase is to plan a desired trajectory. Inspired by SLIP trajectory in stance phase, without loss of generality, we choose a 4th order curve

$$y = -\frac{a^2 x^4}{12} + \frac{b^2 x^2}{2} + c\,, \tag{1}$$

which is symmetric around y axis and has two points of inflection during stance phase as shown in Fig. 1. The velocity and acceleration equations are derived using derivatives of (1) as

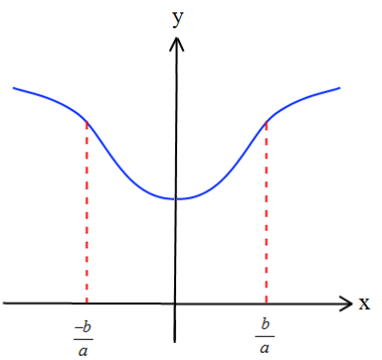


Fig. 1. Schematic view of 4th order curve chosen as the desired trajectory for the stance phase

$$\dot{y} = \frac{-a^2 \dot{x} x^3}{3} + b^2 \dot{x} x\,, \tag{2}$$

$$\ddot{y} = \frac{-a^2 \ddot{x} x^3}{3} - a^2 \dot{x}^2 x^2 + b^2 x \ddot{x} + b^2 \dot{x}^2\,. \tag{3}$$

Parameters a, b, and c are calculated using stance phase initial condition of the robot CoM. Assuming massless feet and constrained torso angle, the GRF passes through the robot's hip and using Newton method the dynamic equations can be written as:

$$\sum F_x = m\ddot{x} \rightarrow -F_r \sin\theta = m\ddot{x}\,, \tag{4}$$

$$\sum F_y = m\ddot{y} \rightarrow F_r \cos\theta - mg = m\ddot{y}\,. \tag{5}$$

Using trigonometric equations sin and cos can be written in terms of x and y:

$$\sin\theta = \frac{-x}{\sqrt{x^2 + y^2}}\,, \tag{6}$$

$$\cos\theta = \frac{y}{\sqrt{x^2 + y^2}}\,. \tag{7}$$

Equations (1-7) are combined to obtain acceleration equation as

$$\ddot{x} = \frac{-a^2 \dot{x}^2 x^3 + b^2 x \dot{x}^2 + xg}{\frac{a^2 x^4}{4} - b^2 x^2 + \frac{b^2 x^2}{2} + d}\,. \tag{8}$$

By solving (8) numerically, the velocity and acceleration on the robot CoM for the chosen trajectory are found over time.

III. Dynamic Model of the System

The one legged robot model in this paper consists of a kneed massless leg and a constrained vertical trunk with mass. According to Fig. 3 the robot has 2 degrees of freedom (DOF) in stance phase with variables θ_1 and θ_2.

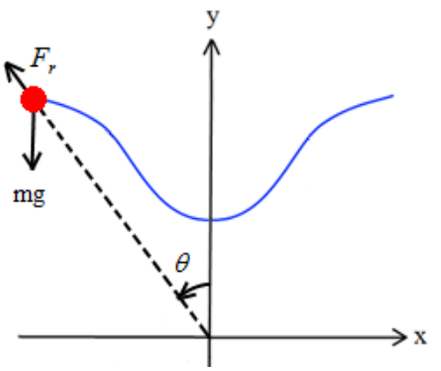


Fig. 2. The free diagram of the system

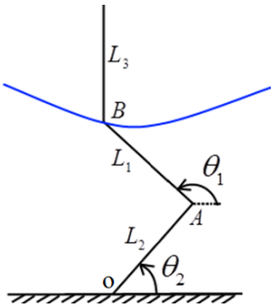


Fig. 3. Robot parameters and stance phase generalized coordinates

Using hip velocity $\vec{V}_B$ and acceleration $\vec{a}_B$ calculated from desired trajectory and relative velocity and acceleration equations, the angular velocity and acceleration of the links can be calculated as

$$\begin{cases} \vec{V}_A = \vec{V}_B + \vec{V}_{A/B} \\ \vec{V}_A = \vec{\omega}_{OA} \times \vec{r}_A \\ \vec{V}_{A/B} = \vec{\omega}_{AB} \times \vec{r}_{A/B} \end{cases}, \tag{9}$$

$$\begin{cases} \vec{a}_A = \vec{a}_B + (\vec{a}_{A/B})_n + (\vec{a}_{A/B})_t \\ \vec{a}_A = \vec{\alpha}_{OA} \times \vec{r}_A + \vec{\omega}_{OA} \times (\vec{\omega}_{OA} \times \vec{r}_A) \\ (\vec{a}_{A/B})_n = \vec{\omega}_{AB} \times (\vec{\omega}_{AB} \times \vec{r}_{A/B}) \\ (\vec{a}_{A/B})_t = \vec{\alpha}_{AB} \times \vec{r}_{A/B} \end{cases}. \tag{10}$$

The Lagrange equations corresponding to θ_1 and θ_2 are written as

$$\frac{d}{dt}\left(\frac{dT}{d\dot{\theta}_1}\right) - \frac{dT}{d\theta_1} + \frac{dV}{d\theta_1} = Q_1, \tag{11}$$

$$\frac{d}{dt}\left(\frac{dT}{d\dot{\theta}_2}\right) - \frac{dT}{d\theta_2} + \frac{dV}{d\theta_2} = Q_2, \tag{12}$$

in which kinetic energy consists of trunk translational energy

$$T = \frac{1}{2} m(\dot{\theta}_2^2 L_2^2 + \dot{\theta}_1^2 L_1^2 + 2\dot{\theta}_1 \dot{\theta}_2 L_1 L_2 \cos(\theta_1 - \theta_2)), \tag{13}$$

and the potential energy consists of gravitational energy

$$V = mg(L_1 \sin(\theta_1) + L_2 \sin(\theta_2) + \frac{L_3}{2}). \tag{14}$$

Similar to organisms and to generate torque, the muscles of our robot are attached to the leg segments using two little arms L_4 and L_5 which are fixed to L_3 and L_1 respectively, as shown in Fig. 4. Therefore θ_3 and θ_4 are fixed angles. To find the generalized forces Q_1 and Q_2, we use virtual work of muscle forces F_1 and F_2 as:

$$\begin{aligned} \delta W &= F_1\left(\frac{\partial AD}{\partial q_1}\delta q_1 + \frac{\partial AD}{\partial q_2}\delta q_2\right) \\ &+ F_2\left(\frac{\partial OC}{\partial q_1}\delta q_1 + \frac{\partial OC}{\partial q_2}\delta q_2\right). \\ &= \sum Q_i \delta q_i \end{aligned} \tag{15}$$

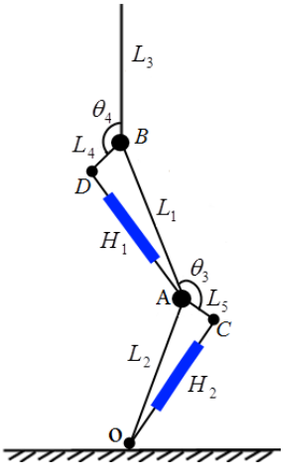


Fig. 4. Muscles configurations on the leg

The parameters values of the robot model are shown in table 1. By solving Lagrange equations for the velocities and accelerations of the desired trajectory, the needed muscle forces are found as a function of time.

TABLE I. THE ROBOT NOMENCLATURE

Parameter	Value (in SI units)
m	58 (kg)
L_1	0.5 (m)
L_2	0.5 (m)
L_3	1 (m)
L_4	0.1 (m)
L_5	0.1 (m)
θ_3	0.75π (rad)
θ_4	0.75π (rad)

IV. CONTROL INPUT OF THE MUSCLES

In the previous section we calculated the needed overall muscle force. In this section, considering muscle elements, we aim to calculate the needed force for the main active part of the muscle. It is noticeable that to avoid unnecessary complications in simulation, we assume that the main active part of the muscle can generate both tensile and compressive forces. In muscles of organisms this part is contractive and just exerts tensile force, so each joint has biarticular muscles. With the mentioned assumption we can actuate each joint by a single muscle that is feasible in robotics. Components of the muscle are shown in Fig. 5. The dynamic equations of the muscle are written as:

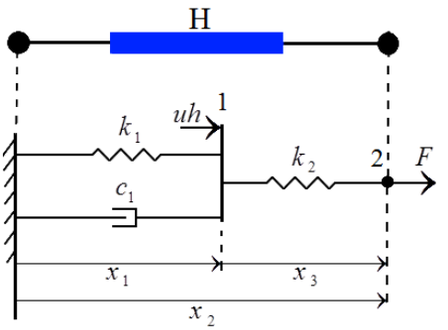


Fig. 5. Hill-type muscle model used as actuation system of our robot

$$-F_1 = k_2(x_3 - x_{3,0}) \rightarrow x_1 = x_2 - x_{3,0} + \frac{F}{k_2}, \tag{16}$$

$$uh = k_1(x_2 - x_{1,0}) + c_1\dot{x}_1 + k_2(x_2 - x_1 - x_{3,0}), \tag{17}$$

which its parameters are shown in Fig. 5, and subscript 0 indicates the free lengths of springs. The springs are assumed to be in their free length at touch-down. The used muscle parameters in simulations are shown in table 2. We have defined these parameters using trial and error to generate feasible motion for the robot.

TABLE II. THE MUSCLE MODEL PARAMETERS

Parameter		Value (in SI units)
Muscle 1	k_1	20
	k_2	30
	c_1	0.001
Muscle 2	k_1	30000
	k_2	40000
	c_1	0.001

V. SIMULATION RESULTS AND DISCUSSION

Starting from an appropriate initial condition and using control law (17), the robot follows the desired trajectory and generates a periodic running gait as shown in Fig. 6. In this figure, the hip trajectory is shown by solid line in stance phase and by dashed line in flight phase. This figure verifies the validity of our control law, because it shows that the dynamic response of the system with the designed controller is coincident with the desired trajectory.

Fig. 8 depicts the overall force (solid line) and active part force (dashed line) of muscle 1 in stance phase which shows an almost zero force for this muscle. This is because the trunk angle is constrained to be vertical and shows that for hopping motion of this configuration the muscle 1 in hip joint can be removed. The overall force (solid line) and active part force (dashed line) of muscle 2 in stance phase are depicted in Fig. 8. According to this figure, the active part of the muscle needs to generate smaller forces than the coverall muscle. This is a desirable matter and shows usefulness of hill-type muscle actuation system. Also the muscle force starts from zero and ends with zero in stance phase which is again desirable in biped robots.

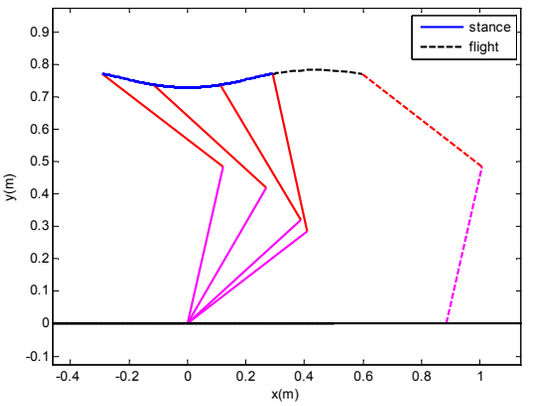


Fig. 6. One step of running using the proposed control law

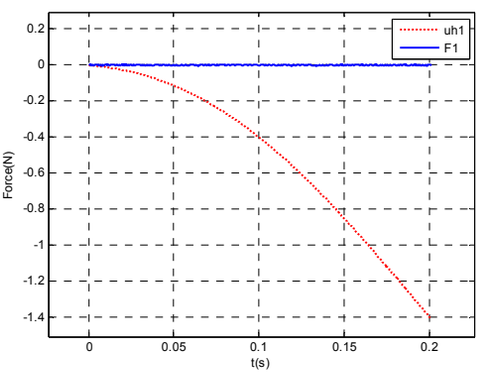


Fig. 7. The overall force (solid line) and active part force (dashed line) of muscle 1 in stance phase

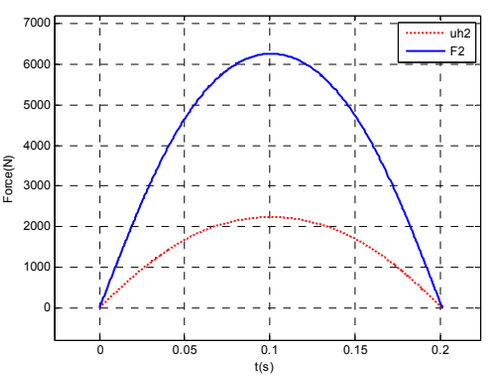


Fig. 8. The overall force (solid line) and active part force (dashed line) of muscle 2 in stance phase

To calculate ground reaction force components we use acceleration components of the robot CoM:

$$\sum F_x = m_G \bar{a}_x , \tag{18}$$

$$\sum F_y = m_G \bar{a}_y . \tag{19}$$

The horizontal (solid line) and vertical (dashed line) GRF profiles in stance phase are shown in Fig. 9. These profiles are qualitatively similar to SLIP running force profiles.

Cost of Transport (COT) is defined as the consumed energy per weight of the robot per traveled distance:

$$COT = \frac{\int_0^t |F.\dot{x}|\,dt}{m_G g L} , \tag{20}$$

in which t is stance time and L is range of one complete step. COTs of hopping with and without Hill-type muscles are shown in table 3. According to this table COT using muscles is greater than COT without muscle. This is undesirable and we aim to reach better COTs using muscles. COT is proportional to the product of the applied force by its velocity. Although the active part of the muscle had smaller values of forces than the overall muscle (Fig. 8), it has greater values of velocity than the overall muscle (Fig. 10) and their product causes bigger

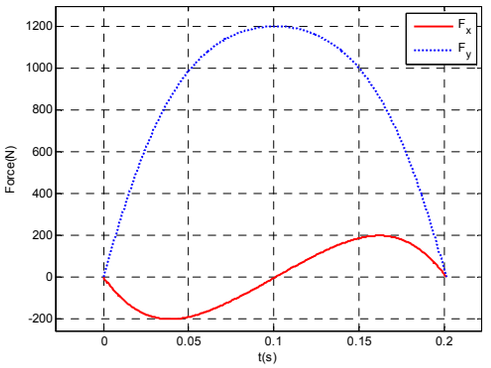


Fig. 9. Ground Reaction Force Components in stance phase

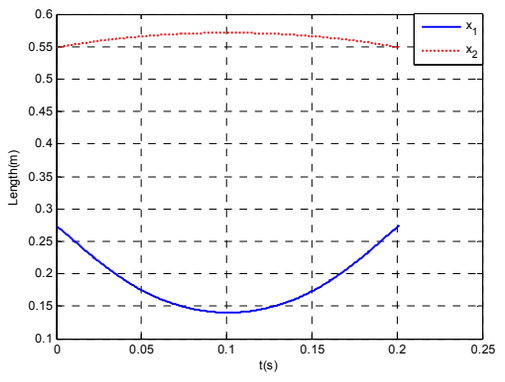


Fig. 10. The overall muscle length displacement (dashed line) and active part displacement (solid line) of muscle 2 in stance phase

TABLE III. HOPPING COT VALUES

Parameter	Value (in SI units)
COT of the overall muscle	0.3057
COT of the active part of the muscle	0.6064

COT for the active part. This is because the muscles parameters shown in table 2 are not optimized for minimum energy consumption and they were chosen just to generate a feasible hopping motion. By optimizing these parameters we would be able to reach more efficient biped hopping and running using muscles.

VI. CONCLUSIONS

A novel control strategy was proposed in this work to generate an arbitrary trajectory for underactuated biped robots running and hopping in stance phase. The desired trajectory should be consistent with stance phase initial conditions. A 4th order trajectory was chosen due to its similarity to SLIP trajectory. This strategy was applied to a kneed three link hopper. At first the necessary velocities and accelerations were found and then using dynamic equations its necessary actuators forces were found to undergo the desired trajectory. Then Hill-type muscles were considered as actuators of the robot and their necessary active parts forces were found. Simulation results showed that the proposed control law gets the robot to undergo the desired trajectory very well. The corresponding force of the active part of the muscle was smaller than the overall muscle force which shows a positive effect of using Hill-type muscle. But the COT of the active part was greater than the COT of the overall muscle which is not desirable.

As future works we are going to apply this method to more general biped running with unlocked torso. Also optimizing muscles parameters to reach lower active part forces and lower COTs seems to be very promising. This would guarantee the optimal use of muscles in robotic actuating systems.

REFERENCES

[1] Hasaneini SJ, Macnab CJB, Bertram JEA, Leung H., "The dynamic optimization approach to locomotion dynamics: human-like gaits from a minimally-constrained biped model," Adv. Rob. 2013;27:845–859.

[2] R. Blickhan, "The spring-mass model for running and hopping," Journal of biomechanics, vol. 22, pp. 1217-1227, 1989.

[3] H. Geyer, A. Seyfarth, and R. Blickhan, "Compliant leg behaviour explains basic dynamics of walking and running," Proceedings of the Royal Society, B, 273:2861-2867, 2006.

[4] Yvonne Blum, Susanne W. Lipfert, Andre Seyfarth, "Effective leg stiffness in running", Journal of Biomechanics 42 pp.2400–2405, 2009.

[5] A. V. Hill, "The Abrupt Transition from Rest to Activity in Muscle", Series B, Biological Sciences, Vol. 136, No.884, pp. 399-420, Oct. 19, 1949.

[6] A. V. Hill, "The Series Elastic Component of Muscle", Series B, Biological Sciences, Vol. 137, No.887, pp. 273-280, Jul. 24, 1950.

[7] Mojtaba Ahmadi and Martin Buehler, "Stable Control of a Simulated One-Legged Running Robot with Hip and Leg Compliance", IEEE

TRANSACTIONS ON ROBOTICS AND AUTOMATION, VOL. 13, NO. 1, FEBRUARY 1997.

[8] Sang-Ho Hyon, Takashi Emura, "Running Control of a Planar Biped Robot based on Energy-Preserving Strategy", International Conference on Robotics & Automation, New Orleans, LA April 2004.

[9] Akihiro Sato, Martin Buehler, "A Planar Hopping Robot with One Actuator, "Design, Simulation, and Experimental Results/ Conf. Intelligent Robots and Systems (IROS), Sendai, Japan, Sept/Oct. 2004.

[10] A. Meghdari & S. Sohrabpour & D. Naderi & S. H. Tamaddoni & F. Jafari & H. Salarieh, "A Novel Method of Gait Synthesis for Bipedal Fast Locomotion," J Intell Robot Syst, 53:101–118, 2008.

[11] Fumiya Iida, Ju¨rgen Rummel, Andre´ Seyfarth, "Bipedal walking and running with spring-like biarticular muscles", Journal of Biomechanics 41, 656–667, 2008.

[12] Michael F. Eilenberg, Hartmut Geyer, and Hugh Herr, "Control of a Powered Ankle-Foot Prosthesis Based on a Neuromuscular Model," TNSRE 00034, 2009

[13] Hiroyuki Oyama, Masaki Yamakita, Sang-Ho Hyon, Susumu Ohtake/ Control of underactuated biped running robot via CPG/ ICROS-SICE International Joint Conference, Fukuoka International Congress Center, Japan, August 2009.

[14] Ben Andrews, Bruce Miller, John Schmitt and Jonathan E Clark/ Running over unknown rough terrain with a one-legged planar robot/ Bioinsp. Biomim. 6, 026009 (15pp), 2011

[15] Behnam Dadashzadeh, M.J. Mahjoob, M. Nikkhah Bahrami and Chris Macnab, "Stable Active Running of a Planar Biped Robot Using Poincare Map Control", *Advanced Robotics*, vol.28, Issue 4, 2014.

Development of an Automatic TEMPEST Test and Analysis System

Cihan Ulaş, Serhat Şahin, Emir Memişoğlu
TUBITAK BILGEM,Gebze, Kocaeli Turkey
{cihan.ulas, serhat.sahin, emir.memisoglu}
@tubitak.gov.tr

Ulaş Aşık, Cantürk Karadeniz, Bilal Kılıç, Uğur Saraç
TUBITAK BILGEM, Gebze, Kocaeli Turkey
{ulas.asik, canturk.karadeniz, bilal.kilic, ugur.sarac}
@tubitak.gov.tr

***Abstract*— Today, it is clearly known that the electronic devices generate electromagnetic radiations unintentionally, which may contain critical information called compromising emanations (CE). CE is also known as TEMPEST radiation, which is a code name firstly used by an U.S government program. Every developed country has a TEMPEST Test Laboratory (TTL) connected to their National Security Agency (NSA). The main objective of these laboratories is to investigate equipment, systems, and platforms processing cryptographic information in terms of CE. TEMPEST tests might take very long time depending on the item under test. In this paper, a complete Automatic TEMPEST Test and Analysis System (ATTAS) developed in TUBITAK, BILGEM TTL is introduced. The system has the following properties, which are automatic system calibration unit, automatic test matrix generator based on the SDIP-27/1 standard, implementation of tunable and nontunable tests, automatic CE investigations, rendering of the CE of video display units, playing of the CE of audio signals, measurement of detection system sensitivity, zoning of TEMPEST equipment based on SDIP-28 standard, and generation of graphical results.**

Keywords—Comprimising Emanations, TEMPEST Test System

I. INTRODUCTION

Military history of exploitation of compromising emanations began as early as 1914s during World War I. The concept of information intercept prevention came into existence when the German army successfully eavesdropped on enemy voice communication from the earth loop current of allied battlefield phone lines. US Army engaged Herbert Yardley and his staff of the Black Chamber to develop methods to detect, intercept and exploit combat telephones and covert radio transmitters [1]. During researches, US Army discovered that equipment without modifications was vulnerable to enemy attacks and started a classified program to develop methods to prevent leakage of the classified information. The standards and works on emission security are kept secret all over the world. Only the US government declassifies some part of its emission security program but the revealed material can also be found in the open computing, security, and electromagnetic-compatibility literature. The only known is the name of US national compromising-emanations test standards name and their publication year. "NAG1A" and "FS222" were the first defined compromising-emanations test standards published in the 1950s and 1960s. In 1970 a new version called "National Communications Security Information Memorandum 5100: Compromising Emanations Laboratory Test Standard, Electromagnetics" was released. "NACSIM 5100" was replaced with "NACSIM 5100A" in 1981. The last known revision is "NSTISSAM TEMPEST/1-92" declassified in 1999 after a Freedom-of-Information-Act request made by John Young [2]. All these standards and their NATO equivalent "SDIP-27/1" are still classified documents [3].

Academic research on compromising electromagnetic emanations started in the mid 1980's and there have been significant recent progresses. The first open publication about compromising emanation risks was an 18-page booklet released by a Swedish government committee in 1984 to inform business community [4]. The risk was brought to general attention by van Eck in 1985 [5]. In the paper, van Eck reconstructed the Cathode Ray Tubes (CRT) screen content at a distance by using low-cost home built equipment. Furthermore, attacks to recover information from RS232 cable [6], Liquid Crystal Display (LCD) [7], laser printer [8], keyboard [9], [10], and [11] have been carried out.

In this paper, Automatic TEMPEST Test and Analysis System (ATTAS) has been developed to meet the overall TEMPEST test requirements and evaluation procedures defined in SDIP-27/1 [3]. ATTAS provides several advantages with respect to manual test system and conventional CE investigations. First, the system has an automatic validation unit providing a very fast and practical system-check capability. Second, the test matrix can be generated based on the rules and procedures defined in the standard SDIP-27/1 or can be imported to the system if it is already prepared by the test engineer. Third, the traditional tunable and nontunable tests can be performed with the generated or imported test matrix. Fourth, the CE investigations are carried out automatically, where the RED signal, which represents the classified information, is required for correlation between the RED and BLACK signal, which represents the unclassified or encrypted information. The suspected CE signals can be analyzed based on the signal type, which can be a video signal, audio signal, and a digital signal. A video signal, which is assumed that their screen resolution is known, is displayed within a user-friendly panel. Similarly, if the signal is a voice signal it can be easily played by the speakers. Moreover, the detection system sensitivity (DSS) measurements can be performed automatically in a few minutes. TEMPEST Device

Zoning (TDZ) described in the SDIP-28/1 [12] standard can be carried out successfully. Finally, the TEMPEST test report builder can be used to combine the graphical results.

II. Related Works

A. Types of Compromising Emanations

CE of the Video Display Units (VDU) has a high importance in the concept of TEMPEST. Van Eck showed that CRT screen content could be obtained by affordable equipment [5]. Nowadays, the CRT monitors is almost out of date, and instead, LCD and LED monitors are in fashion. Screens, represented in 2D, are composed of streaming frames represented in 1D. The vertical representation is called as a monitor screen (S_t) at time *t* and its frame representation (F_t) is given in (1).

$$S_t = \begin{bmatrix} row_1 \\ row_2 \\ . \\ . \\ row_N \end{bmatrix}, \qquad F_t = \begin{bmatrix} row_1 & row_2 & . & . & row_N \end{bmatrix} \qquad (1)$$

There are redundant blanks behind and in front of the rows and frames. These blanks provide the time for the scanning and are not visible in the monitor screen. To be able to reconstruct the information of monitor screen from an obtained frame, the Horizontal Synchronization (HS) is very crucial since it is used as conversion from 1D data to 2D screen. Both the horizontal and vertical (refresh rate) synchronization signals exist in the actual transmission between monitor and PC. In the case of CE of VDUs, an attacker can only obtain the frames; however, Furkan et al showed that the synchronization signals could be obtained from the frames by using signal-processing techniques [13]. It is also known that the synchronization signals are constant and sufficient to solve them once; moreover, VESA standards provide synchronization signals if the screen resolution is known. For instance, the HS for 1280x1024 @60 Hz resolution is given as 63.981 kHz. Frame data including frame blanks and row blanks is converted to 2D screen as shown in Fig. 1.

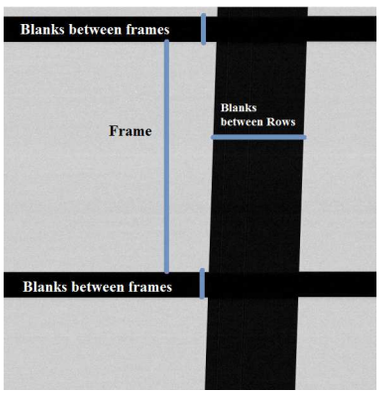


Fig. 1. A whole frame representation in 2D.

The existing ATTAS video rendering module requires the row frequency (equal to HS), which is obtained by VESA standard. An extensive report is published by Kuhn on eavesdropping risk of computer displays for analog and digital video platforms [4]. Signal processing applications for information extraction from the VDUs emissions is presented by Koksaldi [14].

Another type of the CE is composed of keyboard leakages. Keyboard is an important input device of computer processing confidential information. Although Han Fang did some analysis on CE of the keyboards [15], a breakthrough has been achieved by Vuagnoux et al, who proposed a method to recover keyboard emanations with 95% success and a distance around 20 meters. Wang and Yu analyzed a representative control circuit of keyboards by focusing on the PS/2 keyboard protocol in order to find the sensitive signals causing CE [11]. A recent study on CE of keyboard, which investigates the information leakage on the ground line of the PS/2 serial cable, is presented in 2013 [10]. It has been shown that the keystroke signals might leak to ground line network which then be recovered on the other power outlets sharing the same electric line.

B. TEMPEST Test Procedures

For any equipment under test (EUT), it is necessary to prepare a document called as TEMPEST test plan that gives the detailed information about test procedures to be performed. A TEMPEST test plan document should contain the purpose of the test, technical information about the EUT, general representation of RED/BLACK signals, potential emanations, test environment, exercising equipment, operation modes, test media, test setups, and test matrix. TEMPEST tests start with the approval of TEMPEST test plan by NTA. A test process consists of three steps, which are verification of measurement system, measurement of emissions, and advanced signal analysis. After tests, a TEMPEST test report document, containing equipment of test and their dates of calibration, test setups, TEMPEST test procedures, and graphs of the test results, is prepared. Thus, one can say that TEMPEST tests consist of following three steps, preparing a TEMPEST test plan document, executing tests and finally preparing a TEMPEST test report document.

In TEMPEST evaluation procedures, examination and classification of detected emanations is the most difficult step because of the dependency on many parameters such as detection system capabilities, techniques used in signal analyses, personnel experiences, qualification of test environment etc. CE mostly appears in three signal types, which are in baseband, modulated by a carrier signal, and impulsive emanations; therefore, it is not easy to classify detected emanations unless implemented with a detailed examination. In addition, in some cases, the electromagnetic emissions can be data related but one cannot prove whether they are compromising emanations or not. Therefore, the difficulties in the search of compromising emanations require

a fast and reliable TEMPEST test and analysis system in addition to save cost and time.

III. TEMPEST AUTOMATIC TEST AND ANALYSIS SYSTEM

The ATTAS block diagram is given in Fig. 2. The system has three main equipment, which is pre-amplifier, test receiver (FSET), and oscilloscope, connected to a control personal computer (PC). A low noise pre-amplifier is very important to increase the SNR, and in ATTAS, a custom (switchable) pre-amplifier, including two pre-amplifiers working in two different frequency bands, is used depending on the working frequency band. A pre-amplifier is used in 100 Hz - 1 GHz frequency band, which has 32 dB gain, and another amplifier is used in 1-20 GHz frequency band, which has a 26 dB gain. The second equipment, connected to control PC via GPIB/USB converter, is the TEMPEST Test Receiver (FSET) produced by the company of Rohde & Schwarz. This test receiver is devoted to the TEMPEST tests, working in the frequency range of 20 Hz - 22 GHz with at most 500 MHz bandwidth, which is the largest bandwidth in the marketplace. The last equipment is the oscilloscope, where the model of either Lecroy Wave Runner 640Zi or Lecroy WavePro 7300A, which has 4 GHz and 3 GHz bandwidths, respectively, can be used.

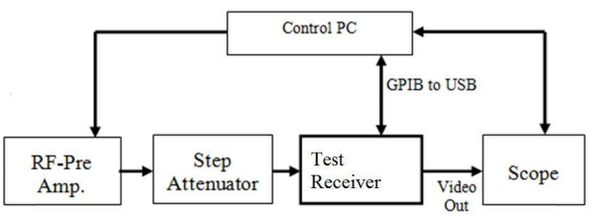


Fig. 2. ATTAS block diagram.

In Fig. 3, a sample test setup in fully anechoic chamber and the control room including ATTAS is shown. The properties of the ATTAS are introduced in the following subsections.

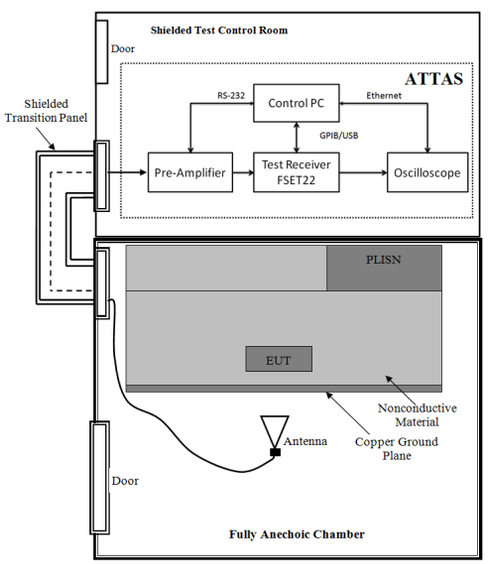


Fig. 3. Test setup and control room including ATTAS.

A. Automatic System Calibration/Validation

According to the Laboratory standards like ISO 17025, before starting tests, the test infrastructure should be checked and validated if everything is all right. For this reason, a test engineer has to start a test day by validating that the devices used in the measurement system are working properly. Every component or transducer, like cables, connectors, antennas, probes, amplifiers, RF limiters in the measurement system has a correction factor. Therefore, this fact has to be considered when evaluating the measurement results. The conventional calibration procedure is applied in three steps. First, the correction factor of the transducer or set, which may include more than one transducer like cable, antenna, and amplifier, is selected in the Test Receiver (TR). Second, the Signal Generator (SG) is tuned to a starting frequency, which might be the lowest frequency of the test receiver system, with constant amplitude. Third, the generated signal is measured in the TR while tuning it to the frequency set by the SG with the proper span and bandwidth. This procedure is repeated with frequency increments up to the highest frequency of the measurement systems. In traditional system calibration, the test engineer checks about 20 points by finding the errors between the generated and measured signal amplitudes. This error is also considered as an overall system correction factor and has to be added to the measurement results. This boring procedure takes about more than an hour for every measurement day. The aim of the automatic calibration unit is to carry out this procedure automatically by controlling the SG, TR, and amplifiers in a remote mode. With this way, the overall system correction factor is computed in much more points and loaded to the system automatically in a few minutes. Automatic calibration user interface is shown in Fig. 4.

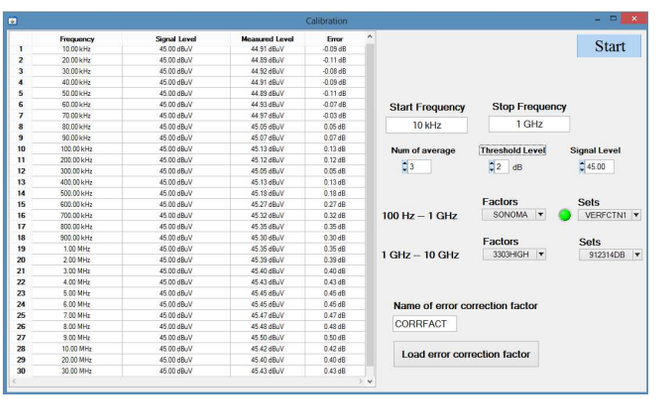


Fig. 4 Automatic calibration user interface.

B. Automatic Test Matrix Generator and Importer

The TEMPEST Test plan preparation procedure defined by the SDIP-27/1 consists of straightforward rules. These rules are based on the signal properties like speed, bandwidth, type, and the test medium, which can be Electrical Radiations (ER), Magnetic Radiations (MR), and BLACK Line Conducted (BLC). The automatic test matrix generator provides the test matrix based on the rules defined by the standard. However,

the test engineers usually prefer to prepare the test matrix manually due to the traditional reflexes. For this purpose, a user interface is developed to import a prepared test matrix from an excel file. Thus, the test engineer can start the tests easily and quickly.

C. Implementation of Tunable and Nontunable Tests

In the TEMPEST standard, tests are divided into two main parts, which are tunable and nontunable tests. Tunable tests are based on a test receiver, where FSET 22 produced by Rohde & Schwarz (RS) is used. Tunable tests are performed based on generated or imported test matrix, which contains the start frequency, stop frequency, resolution bandwidth, and transducer. In the second part, Nontunable tests are carried out based on an oscilloscope, where Lecroy Wave Runner 640zi or WavePro 7300A is used.

D. Automatic Compromising Emanation Investigations

The heart of the analysis system is the investigations of the CE automatically. In the conventional method, the CE are searched in the frequency points where the emissions pass above the limit line, which is defined by SDIP-27/1 [3]. During the tests, it is assumed that the RED signal, is known and applied periodically. In frequency points looked for CE, the test receiver system is operated in the zero-span mode, which produces AM-demodulated signal at tuned center frequency. The video out of the test receiver system is connected to the oscilloscope with a high storage capability. The test engineer constantly looks for the similarities and correlations between the RED signal and demodulated signal. The searching of CE in such a way is very problematic, especially when the RED signal is a type of audio signal. The audio signals are narrowband and mostly checked by the test engineers' ear. If we assume that the test span is wideband spectrum, such as from DC to 1GHz, and the audio signal bandwidth is 5 kHz, the test engineer has to check 200.000 points. This process obviously takes weeks or months depending on the emission levels above the limit lines. In addition, the tests are carried out with at least two people switching in every 30 minutes to save their ear and keep the tests reliable. The aim of the automatic CE investigations is to solve this problem by controlling test receiver, amplifiers, and oscilloscope remotely. Before starting the search of automatic compromising emanations, the RED signal, which we are looking for correlation, is saved with a sufficient number of samples and sampling rate. Then the spectrum is swept with a proper bandwidth, span, and transducer. The related limit line defined in the standard is added to the spectrum graph, and the frequency points above the limit line are determined. The correlation, given by (2), between the RED and demodulated signal is calculated for each frequency point. The results are listed in a table.

$$\text{corr} = \max \int_{-\infty}^{+\infty} r(t-\tau)b(\tau)\, d\tau \qquad (2)$$

In ATTAS, the digitized signals can be interpreted as follows.

1. Displaying and Correlating CE with Zooming

In the analyses, the signal digitized by the oscilloscope, which either might be the BLACK or the RED signal, can be displayed in the control PC's screen with zooming property. The correlation between two signals can be computed and they can be shifted on each other to see the similarities on the zoom panel.

2. Playing CE of Audio Signals

In the analyses, the digitized signal by the oscilloscope, which might be either the BLACK or the RED signal, can be played on the speakers connected to control PC.

3. Rastering CE of Video Signal

In the analyses, the signal digitized by the oscilloscope, which might be either the BLACK or the RED signal, can be rendered and visualized by the system. To be able to visualize the signal, the vertical synchronization frequency has to be known, which is found from the VESA standards [16] for the time being.

E. The Detection System Sensitivity Measurements

The Detection System Sensitivity (DSS) measurements are carried out for tunable and nontunable detection systems and should apply to all signal classes as appropriate for the test to be performed. All DSS measurements should be made using correct calibration source. These methods are specified using sine wave substitution source and given as [3]:

Method 1 requires a calibrated unmodulated carrier as the substitution signal, and is applicable when measuring the DSS at the pre-detection (e.g. IF) output of tunable detection systems and at the output of nontunable detection systems.

Method 2 requires a calibrated sine wave carrier modulated at 30% by a sine wave at any suitable frequency less than or equal to the repetition rate as the substitution signal, and is applicable when measuring the DSS at the AC or DC coupled post-detection output.

Method 3 is applicable when measuring the DSS at the DC coupled post-detection output possessing technical limitations preventing the use of a modulated sine wave carrier as the substitution signal. The required substitution signal for Method 3 is a calibrated unmodulated carrier.

In ATTAS, it is considered that the Method A is more convenient than other methods for DSS measurements. In this method, the signal amplitude level is set to the minimum amplitude level of the signal generator and increased gradually until the detection system has 10 dB SNR. The applied signal level (L) is read and DSS measurement is specified as L-10.

F. Equipment TEMPEST Zoning

Equipment TEMPEST zoning (ETZ) procedures is given by SDIP-28 [12] standard as well as facility TEMPEST zoning procedures. To assign a TEMPEST zone to an equipment or system, ER test procedures given in SDIP-27/1 is applied [12].

One has to know that ETZ cannot be applied to crypto equipment and transmitters. This means that the Level A tests are out of scope of ETZ procedures. ETZ assignment can be evaluated using the following two methods.

Method 1 compares CE levels of each RED Signal of equipment with the limits and is in practice identical to a SDIP-27/1 Level B/C (ER) test. If CE levels are below the Level B ER limits defined in SDIP-27/1, the equipment is assigned an equipment zone 1. However, if CE levels exceed Level B ER limits defined in SDIP-27/1, but are below the Level C ER limits, the equipment is assigned an equipment zone 2.

Method 2 does not analyze the modulations to be compromising or not, and in worst-case assessment, all peak levels are assumed to be compromising and have to be below the limits. The bandwidth is selected based on the highest data rate of the respective category within the "Bounds on Tunable Overall Detection System Bandwidth" defined in SDIP-27/1. As a result, if peak signal levels are below the Level B ER limits, equipment is assigned equipment zone 1. If peak signal levels exceed Level B ER limits, but are below the Level C ER limits, the equipment is assigned an equipment zone 2. ATTAS implements Method 2 for convenience.

G. Automatic Test Report Builder

Automatic Test Report Builder (ATRB) is developed to generate the report of graphical representation of TEMPEST tests results. The report generation process is carried out in two steps. In the first step, each test result graph is prepared as a report page whenever a related part of the test finished. The information about the test is entered through a user interface and a single page Word report is obtained. In the second step, these single page reports are combined to obtain the report of graphical representation of TEMPEST test results.

IV. EXPERIMENTAL STUDIES

In this section, first, automatic investigations of compromising emanations (CE), a powerful part of the system, is implemented. Second, CE of the LCD monitors are captured and displayed in a user interface.

A. Investigation of CE

In order to show the system performance, we performed two different experiments. First, 300 Hz-3.5 kHz chirp signal, which represents the human voice, is amplitude modulated (Double sideband) with a 10 MHz carrier frequency by a signal generator. The spectrum of the signal is given in Fig. 5.

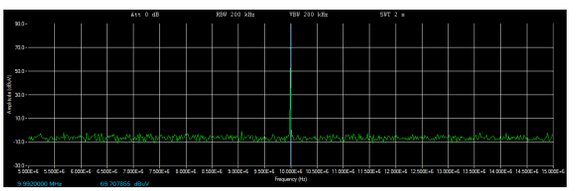
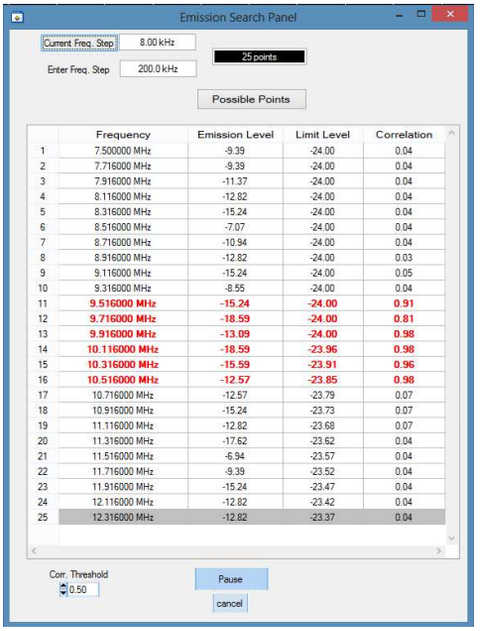


Fig. 5. AM-modulated chirp signal in 10 MHz.



	Frequency	Emission Level	Limit Level	Correlation
1	7.500000 MHz	-9.39	-24.00	0.04
2	7.716000 MHz	-9.39	-24.00	0.04
3	7.916000 MHz	-11.37	-24.00	0.04
4	8.116000 MHz	-12.82	-24.00	0.04
5	8.316000 MHz	-15.24	-24.00	0.04
6	8.516000 MHz	-7.07	-24.00	0.04
7	8.716000 MHz	-10.94	-24.00	0.04
8	8.916000 MHz	-12.82	-24.00	0.03
9	9.116000 MHz	-15.24	-24.00	0.05
10	9.316000 MHz	-8.55	-24.00	0.04
11	9.516000 MHz	-15.24	-24.00	0.91
12	9.716000 MHz	-18.59	-24.00	0.81
13	9.916000 MHz	-13.09	-24.00	0.98
14	10.116000 MHz	-18.59	-23.96	0.98
15	10.316000 MHz	-15.59	-23.91	0.96
16	10.516000 MHz	-12.57	-23.85	0.98
17	10.716000 MHz	-12.57	-23.79	0.07
18	10.916000 MHz	-15.24	-23.73	0.07
19	11.116000 MHz	-12.82	-23.68	0.07
20	11.316000 MHz	-17.62	-23.62	0.04
21	11.516000 MHz	-6.94	-23.57	0.04
22	11.716000 MHz	-9.39	-23.52	0.04
23	11.916000 MHz	-15.24	-23.47	0.04
24	12.116000 MHz	-12.82	-23.42	0.04
25	12.316000 MHz	-12.82	-23.37	0.04

Fig. 6. Automatic CE search table of a chirp signal.

Then this signal is investigated with 200 kHz frequency steps and the result is given in Fig. 6. It is seen that the correlations are very high around 10 MHz and these points are highlighted and shown by red color in the search table. The demodulated signal is digitized with 50 kHz sampling rate for 0.5 second and is shown in scope panel as given in Fig. 7.

Here, the critical point is that the data acquisition time of the demodulated signal should be at least two times the duration of the RED signal. In this case, it is about 2.5 times. In the zoom panel, two signals can be aligned and their correlation can be computed in real time.

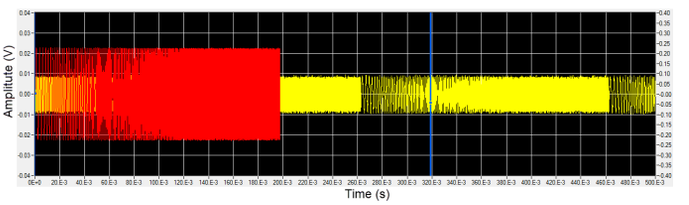


Fig. 7 Capturing the demodulated chirp signal by oscilloscope.

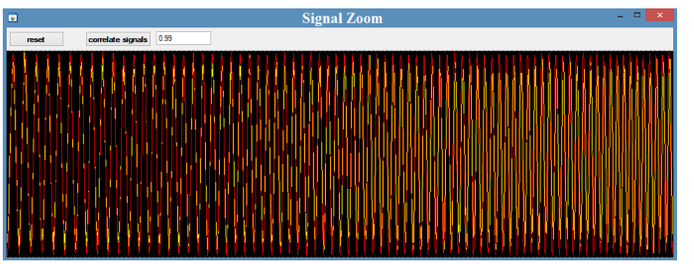


Fig. 8. Zoom Panel and real-time correlation computation.

B. Displaying CE of Video Display Units

The similar CE search procedure is applied to an LCD monitor and correlation result is computed as 66%, which is shown in zoom panel in Fig. 9. In this test setup, log-periodic

antenna is used since the CE of the monitor occurs around 776 MHz. The demodulation bandwidth of the receiver is set to 10 MHz.

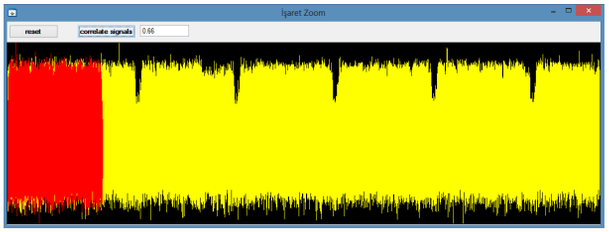


Fig. 9. An LCD Monitor compromising emanation.

In Fig. 10, the one-dimensional data of a frame is transformed to two-dimensional data with the row frequency, which is known by VESA standards [16], and the result is shown in video rendering panel. For more information about compromising emanations of video display units, we refer to [4].

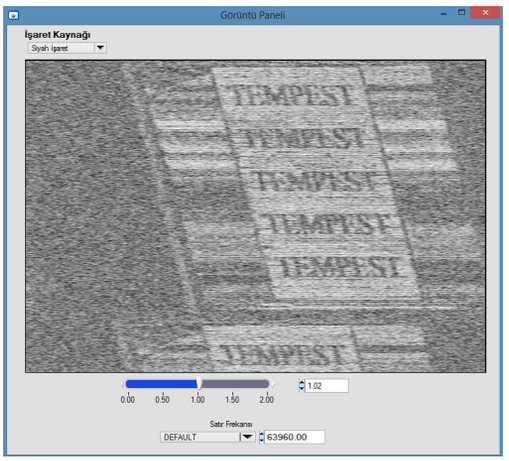


Fig. 10. Video Rendering Panel.

V. CONCLUSION

In this paper, a complete TEMPEST Automatic Test and Analysis System (ATTAS) is developed to improve the test reliability by reducing the testing time. ATTAS includes automatic system calibration unit, test matrix generator and importer, implementation of tunable and nontunable tests, automatic compromising emanations (CE) search, interpretation of the CE with displaying, zooming, rendering, and playing panels. In addition, the measurement of detection system sensitivity, device zoning based on SDIP-27/1, and a report builder of graphical results is achieved by automatically by the system. The system and the software is designed in a modular manner and suitable to update and upgrade devices used. ATTAS has been used successfully in TEMPEST Test Laboratory for almost a year.

Acknowledgments

This study is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under the project TEMPEST Tests.

Reference

[1] H. O. Yardley, *The American black chamber*: Naval Institute Press, 1931.

[2] N. TEMPEST, "1-92," *Compromising Emanations Laboratory Test Requirements*, 1992.

[3] Standard, "SDIP-27/1: NATO TEMPEST Requirements and Evaluation Procedures [confidential]," 2009.

[4] M. G. Kuhn, "Compromising emanations: eavesdropping risks of computer displays," *University of Cambridge Computer Laboratory, Technical Report, UCAM-CL-TR-577*, 2003.

[5] W. Van Eck, "Electromagnetic radiation from video display units: an eavesdropping risk?," *Computers & Security*, vol. 4, pp. 269-286, 1985.

[6] P. Smulders, "The threat of information theft by reception of electromagnetic radiation from RS-232 cables," *Computers & Security*, vol. 9, pp. 53-58, 1990.

[7] M. G. Kuhn, "Electromagnetic eavesdropping risks of flat-panel displays," in *Privacy Enhancing Technologies*, 2005, pp. 88-107.

[8] T. Tosaka, K. Taira, Y. Yamanaka, A. Nishikata, and M. Hattori, "Feasibility study for reconstruction of information from near field observations of the magnetic field of laser printer," in *Electromagnetic Compatibility, 2006. EMC-Zurich 2006. 17th International Zurich Symposium on*, 2006, pp. 630-633.

[9] M. Vuagnoux and S. Pasini, "Compromising Electromagnetic Emanations of Wired and Wireless Keyboards," in *USENIX Security Symposium*, 2009, pp. 1-16.

[10] Y. Du, Y. Lu, and J. Zhang, "NOVEL METHOD TO DETECT AND RECOVER THE KEYSTROKES OF PS/2 KEYBOARD," *Progress In Electromagnetics Research C*, vol. 41, 2013.

[11] W. Litao and Y. Bin, "Analysis and Measurement on the Electromagnetic Compromising Emanations of Computer Keyboards," in *Computational Intelligence and Security (CIS), 2011 Seventh International Conference on*, 2011, pp. 640-643.

[12] Standard, "SDIP-28/1: NATO Zoning Procedures [confidential]," 2005.

[13] F. Elibol, U. Sarac, and I. Erer, "Realistic eavesdropping attacks on computer displays with low-cost and mobile receiver system," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1767-1771.

[14] N. E. Koksaldi, I. Olcer, U. Yapanel, and U. Sarac, "SIGNAL PROCESSING APPLICATIONS FOR INFORMATION EXTRACTION FROM THE RADIATION OF VDUs," *National Institute Of Electonics & Cryptology, Gebze, Kocaeli*, vol. 41470.

[15] H. Fang, "Electromagnetic Information Leakage and its Protection of Computer," presented at the Science Press, Beijing, 1993.

[16] V. M. T. Standard, "Monitor timing specifications," ed: San Jose: Video Electronics Standards Association, 1998, 1998.

Smartphone's Embedded Sensors Performance Analytics

Yasmin BARZAJ
Institute of Fondamental Electronics
University of Paris-SUD X1
91405, Orsay Cedex, Paris -France
yasmin.barzaj@u-psud.fr

Stéphane ESPIÉ
IFSTTAR
(ex INRETS/LCPC)
277447 Marne la Vallée Cedex 2, Paris –France
Stephane.espie@ifsttar.fr

Abderrahmane Boubezoul
IFSTTAR
(ex INRETS/LCPC)
277447 Marne la Vallée Cedex 2, Paris –France
abderrahmane.boubezoul @ifsttar.fr

Jean- Michel DOUIN
CNAM
75003 Paris- France
Jean-michel.douin@cnam.fr

***Abstract*— This paper aims to identify a hybrid solution for Data Acquisition, by using recent Smartphone's embedded sensors. The assumption is that Smartphone's sensors will reduce the complexity and the high cost of instrumentations. The objective is to achieve a sub-meter accuracy of the collected trajectories and to allow a large-scale deployment of the system's instrumentation, such as a helpful system in the domain of transport. Various parameters have been taken into account to identify and characterize the performance of the sensors under different Android Smartphones and tablets. Two experiments have been conducted and Android software has been implemented in cooperation of PC software, developed to retrieve the data via Wi-Fi, and to store the data via USB by a mbed microcontroller in a SD-Card. Some of devices respond better than others depending on the used mode or method.**

Keywords*— *Applications of Signal Processing, Signal Sensing-Radar-Sonar and Sensor, Array Signal Processing, Signal Processing for Sceuirty, Signal Processing Theory and Methods.

I. INTRODUCTION

The number of Smartphone applications increases rapidly due to the increasing use of Smartphone in various aspects of the life. The use smartphones for controlling transportation behavior itself is not new, previous work has primarily focused on elaborate use of the phone's integrated GPS receiver. While GPS-based systems can be very efficient when GPS signals are available, they suffer from some important limitations. The ability to capture the behavior of transport accurately on smartphones would have a positive impact on many research fields. For example, human mobility tracking would benefit from an ability to automatically control the behavior of individuals in the domain of transport [9, 10]. It can be considered as a field of activity recognition and a widely studied field within the wearable, the mode detection of transport [16]. As a Location detection has been increasingly explored on smartphones. A number of newly systems used the embedded accelerometer for detecting different pedestrian and non-motorised modalities, such as walking and running [14, 8], ascending or descending stairs [12] or cycling [11]. Some studies are focused for detection by using sensor-fusion output of both accelerometer, gyroscope and also extracting information from GPS measurements [15][14]. Recent work has also focused on the external sensors that installed on vehicle to analysis and reconstruct the accident situation on the real-time location for the movement of the vehicle [16]. Most of these applications use the Smartphone's sensors to transfer data from and to the devices. With the development of their operating systems, such as Google's Android platform [1, 2, 3], it is imperative that these systems already have several sensors (e.g., GPS, accelerometer, compass and microphone), to acquire data from sensors, by using a Personal Area Network (PAN) technology [6,7], such as Wi-Fi and USB.

We have developed a hybrid solution for Data acquisition by using Smartphone's embedded accelerometer and gyroscope and combining GPS data that can provide a huge database of detection different cases of accident situation in the domain of transport. We have implemented our approach on android smartphones and integrated it as a part of a mobile application that aims at sending an alarm when detect an ordinary situation (like accidents, crash, etc.).

II. RELATED WORK

A. Overview

The latest smartphone are equipped with many inputs of research and not limited:

- Camera.
- 3-axis accelerometer.

- 3-axis gyroscope.
- GPS.
- etc.

Smartphone driving detection systems have become increasingly common since they only need to use sensor data acquired from GPS and sensor-fusion (accelerometer, gyroscope) that are available in a model smartphone. These devices are powerful, inexpensive and research platforms that make instrumenting vehicle for Data acquiring. Most of Smartphone's platform has Android platform that is flexible for building real-world systems, and is especially useful for those who develop applications using Java programming language. It is also useful for creating applications exploiting the hardware sensors. Actually, the inability to access the hardware underlying the device has been frustrating to the mobile platform developers. While the Android Java environment is still the intermediary between the device and the human, and the Android developers bring much of the hardware's capability to the surface. This presents a great opportunity to write some code to accomplish our tasks using an open source platform.

B. Android Application

Every Android application consists of one or more components which are defined in the application's manifest file. The Android platform allows the usage of all normal Java concurrency constructs.

Android platform provides three types of classes:

Service: The Service doesn't have a visual user interface, but rather it runs in the background for an indefinite period of time. It is possible to be connected with an ongoing service (and start the service if it is not already running).While running, it can communicate with another service through an interface of the service.

The *UI Thread* is a concurrent unit of execution, which has its own call stack for methods being invoked. At least one main thread is running when it is started for each virtual machine (VM) instance. The application might decide to start additional threads for specific purposes.

AsyncTask enables a proper and easy use of the UI Thread, it allows performing background operations and publishing results on the UI Thread without having to manipulate threads at a low-level.

III. HYPOTHESIS AND METHODOLOGY

In order to check the ability and the limit of these embedded sensors of smartphone, we suppose that the performance of sensors is approximately invariant at different acquisition rates, at various modes of operation classes and methods of communication, for different source of data. In order to prove this hypothesis, we have to acquire the sensor's data, and then store them to be analyzed. Three types of sensor's data will be exploited for our objective: GPS data (Longitude, Latitude, Altitude, and Speed), 3-axis accelerometer [8] (ax, ay, az), 3-axis gyroscope (rx, ry, rz) for characterizing their performance with different rates: 100Hz and 200Hz, and verify if sensors remain responsive or not.

Each mode and method is mentioned in above, will be tested over our sensor's data. A single software implementation will be applied on different devices.

In our implementation, we take into account three operating modes:

AsyncTask and UI Thread Classes: We apply each one separately and compare their responsive on each device.

Communication: We study the frequency of Data Acquisition vs. modes of communication (local database, Wi-Fi and USB).

Acquisition Rate: We estimate the approximated frequency of Data Acquisition.

IV. EXPERIMENTATION

In this section, for our experiments, in order to determine the accuracy of Data Acquisition, we used two experiments. The first one is acquiring data with a frequency 200Hz and storing it. The second one is Acquisition frame-rate of the embedded sensors (especially accelerometers and gyrometers). Each experiment consists of three parts (software, test, and result analysis).

Software: We realize a software that allows us to acquire the data by two classes (AsyncTask and UI Thread),using Wi-Fi and USB communication , then to store data in a local database at a configurable frequency [2 Hz.. 200 Hz], to give the frames of the Data sensors.

Test: Here we select the frequency 200 Hz by the software, the Smartphone communicates with the PC and then the PC receives the collected data which will be stored in a file/local database.

Result Analysis: To analyze the results of each experiment, we study the number of received frames according to the Data Acquiring method to estimate the performance of the acquisition mode and the method of communication.

During this test, we have to take into account enough time, about fifteen seconds for every experiment, within which we are supposed to shake the phone in three dimensions (X-axe, Y-axe and Z-axe), and ensuring that no activity is running on the phone at the same time of data acquisition. In these experiments, there is no comparison between sensor values, but the comparison will be performed for the frequencies. These values are not reproducible, so the data does not represent the same phenomenon.

A. Acquired Data with a frequency at 200Hz and Data Storage

1. Acquired Data with a frequency at 200Hz

The devices used for acquiring the sensors data are Galaxy Tablet 10, Galaxy Tablet2 7.0, Galaxy S2 and HTC (see Table I). For this purpose, we developed an application to collect the required data. It manages the communication with the sensors. This data is then previewed as variable signals to be distinguished by the naked eye. Figure 1 shows a screen shot of the application running on the devices.

TABLE I. THE TYPES OF SMARTPHONES USED.

Device	Android Version	Kernel Version	Processor
Galaxy Tab 10	3.2	2.6.36.3	P7510 dual core
Galaxy Tab 2 7.0	4.0.3	3.08-379370	GT-P3110 dual core
Galaxy S2	2.3.6	3.0.15	GT-I9100G
HTC	4.0.3	3.28.401.1	5GHz NVIDIA Tegra3quad core

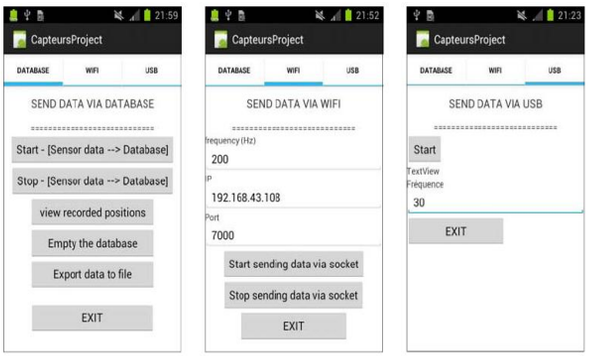


Fig. 1. The application of data acquisition running on the devices.

As we referred, the main objective of our study will be on the sampled sensors. They will be acquired at a frequency of 200 Hz. We will be able to check whether their values change to 100 Hz. Firstly, we started the Data Acquisition by the communication channel (Wi-Fi). This operation need about fifteen seconds for the Data Acquisition and we repeat it for all kinds of Smartphone (Galaxy tablet 10, Galaxy tablet 2 7.0, Galaxy S2 and HTC). The results of sensor's data values via Wi-Fi and via local database are presented in Figure 2.

Note that the results of using the class UI Thread are more efficient for acquiring data sensors than the results of the class AsyncTask; because we receive many more packets for the first class than the second during a limited time, for both Wi-Fi and the local database in all of the Smartphone (see Table II), but we were not able to send the data via Wi-Fi for Galaxy Tab2 7.0 because this version cannot be a host to communicate with PC (see Figure 3), but via Bluetooth port will create a hotspot the same as a Wi-Fi; this problem is already well known[2].

By analyzing the results of Acquired Data with a frequency at 200Hz via Wi-Fi (see Table III), and via local database (see Table IV), we find that the best estimated frequency via Wi-Fi is achieved by using Galaxy Tablet 10, then HTC and finally Galaxy S2. This order is inversed via Local database, the best result for the fixed frequency is found for Galaxy Tablet2 7.0 then HTC, Galaxy S2 and finally Galaxy Tablet 10. We conclude that the speed of acquiring data from the Smartphone's sensors is not associated with its software versions but instead depends on the hardware specifications.

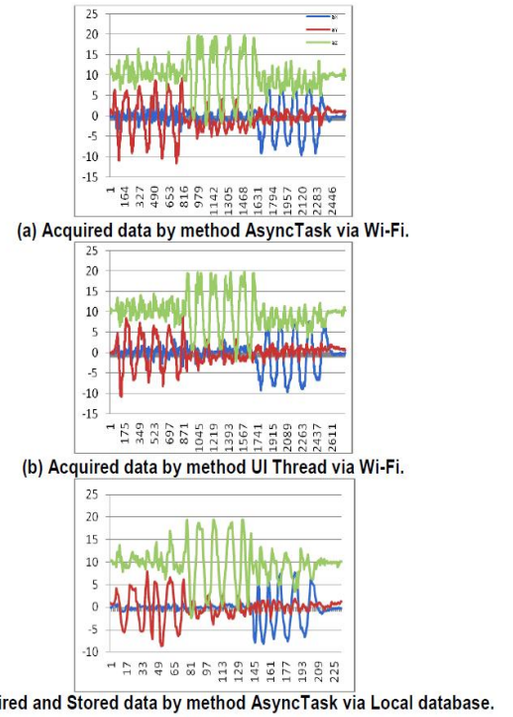


Fig. 2. Acquired data results for Galaxy S2.

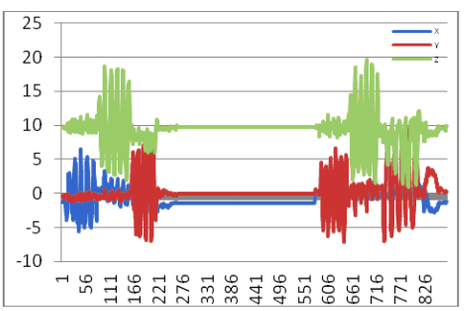


Fig. 3. Acquired data results for Galaxy Tablet2 7.0 via mode Local database.

TABLE II. THE ESTIMATED FREQUENCY OF ACQUIRING PACKETS FOR TABLET10, TABLET 7, GALAXY S2 AND HTC.

Smartphone	**Tablet 10**	
	Number of Packets	***Estimated Frequency***
UI Thread via Wi-Fi	2899	193.26
AsyncTask via Wi-Fi	2880	192
AsyncTask via Local database	225	15
Smartphone	**Tablet 7**	
	Number of Packets	***Estimated Frequency***
UI Thread via Wi-Fi	-	-
AsyncTask via Wi-Fi	-	-
AsyncTask via Local database	826	55.06
Smartphone	**Galaxy S2**	
	Number of Packets	***Estimated Frequency***
UI Thread via Wi-Fi	2655	183.99
AsyncTask via Wi-Fi	2598	180.04
AsyncTask via Local database	603	41.78

Smartphone	HTC	
	Number of Packets	*Estimated Frequency*
UI Thread via Wi-Fi	2848	189.48
AsyncTask via Wi-Fi	2643	175.84
AsyncTask via Local database	804	53.49

TABLE III. THE PERCENTAGE OF THE ESTIMATED FREQUENCY VIA WI-FI.

Smartphone	Estimated Frequency	Percentage
Galaxy Tab 10	193.26 Hz	96.63%
Galaxy Tab 2 7.0	-	-
Galaxy S2	183.99 Hz	91.99%
HTC	189.48 Hz	94.74%

TABLE IV. THE PERCENTAGE OF THE ESTIMATED FREQUENCY VIA LOCAL DATABASE.

Smartphone	Estimated Frequency	Percentage
Galaxy Tab 10	15 Hz	7.5%
Galaxy Tab 2 7.0	55.06 Hz	27.53%
Galaxy S2	41.78 Hz	20.89%
HTC	53.49 Hz	26.74%

2. *Acquired Data with a frequency at 30HZ*

For this experiment, we designed a special mbed microcontroller for our test (see Figure 4). As Known, the mbed Microcontrollers are a series of ARM microcontroller development boards designed for rapid prototyping. We design [5] a model of the mbed NXP LPC1768 Microcontroller in particular for prototyping all sorts of devices, especially those including Ethernet, USB, and the flexibility of lots of peripheral interfaces and FLASH memory. It is packaged as a small DIP form-factor for prototyping with stripboard and breadboard, and includes a built-in USB FLASH programmer. It is based on the NXP LPC1768, with a 32-bit ARM Cortex-M3 core running at 96MHz. It includes lots of interfaces including built-in Ethernet, USB Host and Device, CAN, SPI, I2C, ADC, DAC, PWM and other I/O interfaces. The mbed NXP LPC1768 includes a built-in USB programming interface that is as simple as using a USB Flash Drive.

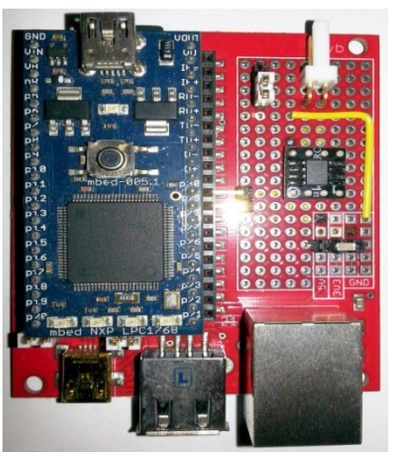


Fig. 4. Design of mbed NXP LPC1768 microcontroller.

By analyzing the results of Acquired Data with a frequency at 200Hz via USB; we find after the calculation of the real frequency for all Smartphones is 2.62 ≈ 3Hz. Maybe the reason is the core of mbed which is running at 96MHz and by acquiring the data by a frequency 200 Hz, we have just 3Hz and it is equal to 0.000003MHz and it is not efficient to acquire the data at this limit of frequency.

To ensure the purpose of the performance of Smartphone data reception, we decrease gradually the frequency until 30Hz. we evaluate the frequency of the sampled values during a limited of acquired Data at 30Hz. Table V presents the frequency results of each Smartphone.

TABLE V. THE ESTIMATED FREQUENCY VIA USB.

Smartphone	Tablet 10	Tablet2 7.0	GalaxyS2
Time	***Frequency***		
(T+32)-T	18.35	18.35	18.35
(T+64)-(T+32)	20.97	20.97	18.35
(T+96)-(T+64)	18.35	18.35	20.97
(T+128)-(T+96)	20.97	20.97	20.98

By analyzing the frequency of Data acquired via USB at 30Hz, we found it is more efficient than the Wi-Fi one at the same frequency (Table VI).

With Wi-Fi communications, it can save cabling costs and installation time. But the USB mode is low-cost multifunction and it is well suited for purposes due to its small size, easy USB connection and an energy source for the Smartphone installed in a vehicle.

TABLE VI. THE ESTIMATED FREQUENCY VIA WI-FI AND VIA USB

Smartphone	Galaxy Tablet10	Galaxy S2
Frequency via Wi-Fi	12.00 Hz	3.6 Hz
Frequency via USB	19.66 Hz	19.66Hz

3. *Data storage*

We performed an additional experiment aiming to evaluate the speed of the data storage. We designed an application which examines the speed of writing this data to a file by using two methods UI thread and AsyncTask. This operation has been tested for fifteen seconds. For each second we received the information and then wrote it to the file (see Table VII). The objective was to know which way is more speed than other and is more efficient for saving the sensor's data to the file.

In general, the mode of local database for Smartphone takes more time to interact with the device's memory and to work on the Database engine. We couldn't write the received information, and we lost much information for the Acquired Data with a frequency at 200Hz because the high speed of acquisition data.

The results of acquiring data using a timer are faster and more efficient to write data to the database via Wi-Fi than the results of acquiring data via local database. Probably the reason is the type of the Smartphone's hardware and the software layers.

TABLE VII. EXPORTING DATA TO FILE PER 1000 SECOND VIA LOCAL DATABASE.

Smartphone	Local Database (per 1000 sec)	
	Method UI Thread	*Method AsyncTask*
Galaxy Tab 10	17 packets	18
Galaxy Tab 2 7.0	20	21
Galaxy S2	15	17
HTC	39	25

B. The Acquisition framerate of the embedded sensors

All Here, the Data Acquisition uses the sensor's frequency. The purpose is to check the performance of Smartphone data reception and compare different devices at this frequency in order to determine its impact. The aim of this experiment is to determine the slope during a limited moment, which permits us to check the frequency of a sampled value of Data acquisition. We should ensure that the movement of each device (the shaking) is fast enough to observe the sampled data, and by the movement, it will have a frequency that is higher than the frequency of sampled value of the sensor; for that reason we will observe the crenels.

For this operation, we had to shake each Smartphone in several directions, five times up and down, five times forward and backward and five times left and right. The objective was to analyze the signals of the Smartphone obtained by the applied movement, then to check the performance of its frequency by acquiring data to test the quantity of received data. These experiments are performed using the method *UI Thread* via Wi-Fi, because we had good results for the first experiment. Figure 6 presents the sampled values for the 3-axis accelerometer (ax, ay and az), over a 15 seconds period where the device was shaken in three dimensions (X-axe, Y-axe and Z-axe).

In order to bring out the significant changes in the sensor's data, corresponding to the shaking moments, we take only a sample of the signals to analyze the frequency of data acquisition during milliseconds (Figure 6).

To estimate the frequency of these sampled values during a time of acquired Data. Mean (μ) and Sd (σ) are calculated on these sampled values to analyze statistically the significant changes. Table VIII, IX and X show the results of these statistics for all type of Smartphones.

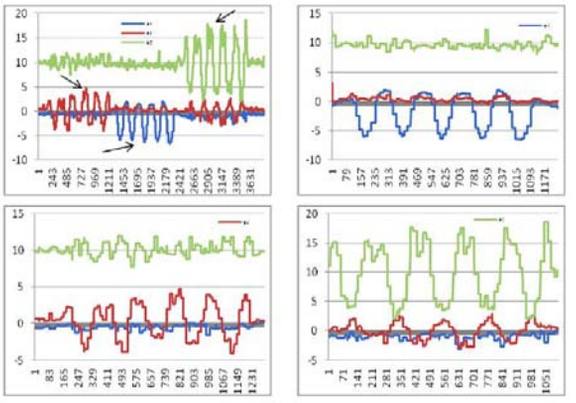


Fig. 5. Sensor data of 3-axis accelerometer.

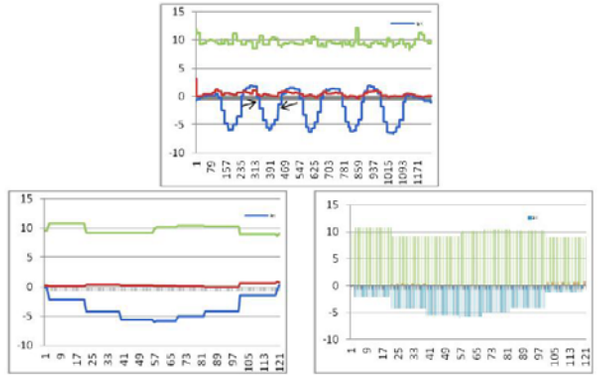


Fig. 6. Sampled values for Tablet 10.

TABLE VIII. CALCULATED Σ, M AND ESTIMATED FREQUENCY FOR RANDOM SAMPLED VALUES (AX).

Δx(x)	0,14	2,109	4,278	-5,562	6,017	5,764	4,193	1,352	0,499
F(x)	2	19	18	17	1	11	14	18	19

Sampled Values(ax)	Mean μ	Standard Deviation σ	Frequency (F)	Time
336-456	-3,7819	1,671	12	0.08s

TABLE IX. CALCULATED Σ, M AND ESTIMATED FREQUENCY FOR RANDOM SAMPLED VALUES (AY).

Δy(x)	-5,899	5,057	-2,49	-1,417	-0,766	-0,421	1,034	2,873	4,559
F(x)	6	5	2	2	10	3	3	5	3

Sampled Values(ax)	Mean μ	Standard Deviation σ	Frequency (F)	Time
32-81	1,134	4,673	3,6	0.27s

TABLE X. CALCULATED Σ, M AND ESTIMATED FREQUENCY FOR RANDOM SAMPLED VALUES (AZ).

Δz(x)	9,485	9,935	9,61	8,336	6,612	5,728	5,358	4,875	4,768
F(x)	2	5	4	- 2	6	1	6	4	2

Sampled Values(ax)	Mean μ	Standard Deviation σ	Frequency (F)	Time
199-289	6,3199	1,8548	3,826	0.26s

We conclude, it is not relevant to acquire and store the data at more than 12Hz for Galaxy Tablet 10, and the estimated frequency for HTC and Galaxy S2 is about 4Hz. We consider that this difference in the frequency is not related to the current version of Android for each Smartphone but to its processor. The best results are achieved by Galaxy Tablet 10 then HTC and lastly Galaxy S2 (see Table XI).

We found that the used method is sufficient to observe the frequency (ex. The crenels) and the frequency depends on the puissance of the smartphone.

Smartphone	Galaxy Tab 10	Galaxy S2	HTC
Estimated Frequency	12 Hz	3.6 Hz	3.8 Hz
Time	0.08s	0.27s	0.26s

V. CONCLUSION

main objective of this study is to propose a method to identify the performance of various Smartphone's embedded sensors, like Galaxy Tablet10, Galaxy Tablet7, Galaxy S2 and HTC for our experiments. The technical tools of Android platform have two classes UI Thread and AsyncTask, which are exploited to achieve our purpose. Two experiments have been conducted and we have ensured a sufficient time: The first one is to develop a method to store the acquired data with a frequency 200Hz. The second one is to identify the acquisition frame-rate of the embedded sensors (accelerometers and gyrometers). We found that the closer for the frequency is the better for the performance.

Three communication modes are elaborated for acquiring data via Wi-Fi, via USB and via local database. The local database mode is not considered in this study, for the frequency of acquired data because it consumes more time to deal with the memory of the device and Database engine at the same time. By a high speed of data, the received data was not written. For the pre-defined frequency of 200 Hz via Wi-Fi, both classes AsyncTask and UI Thread are efficient to send and store all acquired data completely. For the second experiment *Acquisition framerate of the embedded sensors*, is performed to check the performance of Smartphone data reception and compare different devices at this frequency, in order to determine its impact during a period of time.

To ensure that the performance of the sensor is determined by the processors, not by the Android system, we will take into account another Smartphone (Galaxy S5) to compare and confirm the additional results. In additional, by the knowledge of the Smartphone's context hardware, this allows us to estimate and identify a hybrid solution for data acquiring by their embedded sensors. For our future works, we will evaluate the performance of the Smartphone's sensors for failure detection (sensors faults , false alarms, transport accident case, etc.) to be able to distinguish between the sensors that correctly measure the structural data, and sensors that may fail to correctly measure the structural ones in real-time.

REFERENCES

[1] L;Bao and S.S.Intille. Activity recognition from user-annotated acceleration data. In *proceedings of the 2nd International Conference on Pervasive Computing* (PERVASIVE), volume 3001 of *Lecture Notes in Computer science*, pages 1-17. Springer-Verlag, 2004.

[2] S.Consolvo, D.W.McDonald, T.Toscos, M.Y.Chen,J.Frohlich, B.Harrison, P.Klasnja, A.LaMarca, L.LeGrand,R.Libby, I.Smith, and J.A. Landay. Activity sensing in the wild: a field trail of ubifit garden. In *CHI' 08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1797-1806, New York, NY, USA, 2008.ACM.

[3] http://www.droidforums.net/forum/htc-rezound-general-discussions/207230-how-do-i-share-data-mysamsung-tab-2-7-0-a.html.

[4] A what is google android? android news - android google phone forums, July 2010. http://www.talkandroid.com/google-android-faq.

[5] https://mbed.org/platforms/mbed-LPC1768.

[6] J. Krumm and E. Horvitz. LOCADIO: Inferring motion and location from Wi-Fi signal strengths. In *Proceedings of the 1st International Conference on Mobile and Ubiquitous Systems (Mobiquitous)*, pages 4 - 14. IEEE, 2004.

[7] K. Muthukrishnan, M. Lijding, N. Meratnia, and P. Havinga. Sensing Motion Using Spectral and Spatial Analysis of WLAN RSSI. In *Proceedingds of the 2nd European Conference on Smart Sensing andContext (EuroSSC)*, pages 62-76,. Springer, 2007.

[8] T. Iso and K. Yamazaki. Gait analyzer based on a cell phone with a single three-axis accelerometer. *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 141-144, 2006.

[9] D. Lazer, A. P. L. Adamic, S. Aral, A.-L. Barab_asi,D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. R. 2, and M. V. Alstyne. Computational social science. Science, 323(5915):721-723, 2009.

[10] C. Song, Z. Qu, N. Blumm, and A.-L. Barab_asi. Limits of predictability in human mobility. Science, 19(5968):1018-1021, 2010.

[11] G. Bieber, J. Voskamp, and B. Urban. Activity recognition for everyday life on mobile phones. In Proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction (UAHCI), pages 289-296, 2009.

[12] T. Brezmes, J.-L. Gorricho, and J. Cotrina. Activity recognition from accelerometer data on a mobile phone. In Workshop Proceedings of the 10th International Work-Conference on Arti_cial Neural Networks (IWANN), pages 796-799, 2009.

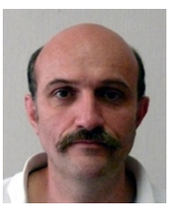
[13] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In Proceedings of the 6th ACM conference on Embedded network sensor systems (SenSys), pages 337-350, New York, NY, USA, 2008. ACM.

[14] Y. Zheng, Y. Chen, Q. Li, and W.-Y. Xie, X. Ma.Understanding transportation modes based on gps data for web applications. ACM Transactions on the Web, 4,1, 2010.

[15] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In Proceedings of the 10th international conference on Ubiquitous computing, pages 312-321, 2008.

[16] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. Transportation mode detection using mobile phones and gis information. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 54-63, 2011.

Yasmin BARZAJ PhD student in laboratory IEF option : Software engineering and data fusion, University Paris-SUD XI
Paris, France (2009-2014), and her Master's degree in Master 2 research informatics from University Paris-SUD XI (2007-2008).



Stéphane ESPIÉ is a senior researcher at IFSTTAR. His main research areas are behavioural traffic simulation (MAS based), and the design of tools to study road user behaviours (driving/riding simulators and instrumented vehicles). He was the scientific director of the 7th FP collaborative project 2BESAFE aiming at a better understanding of the motorcyclists behaviour, and of the French collaborative projects DAMOTO and SIM2CO+ focussing on safety systems for motorcyclists and on the design of new training modules for riders, including riding simulators.

Abderrahmane Boubezoul received his Ph.D. in Computer Science and Mathematics from University Paul CÃl'zanne (Aix-Marseille III), France in 2008 and his Master's degree in Virtual Reality and Complex Systems from Evry Val d'Essone University, France. Since 2008, he is a researcher at IFSTTAR institute. His current work is about statistical signal processing and machine learning applied to road transport systems.