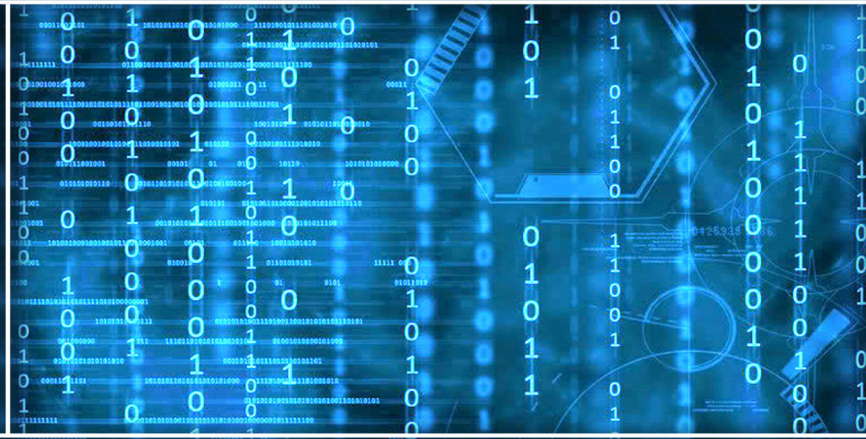


Volume 7 Issue 9

September 2016



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor...

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 7 Issue 9 September 2016
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)
©2013 The Science and Information (SAI) Organization

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University

Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

Associate Editors

Chao-Tung Yang

Department of Computer Science, Tunghai University, Taiwan

Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing

Elena SCUTELNICU

"Dunarea de Jos" University of Galati, Romania

Domain of Research: e-Learning, e-Learning Tools, Simulation

Krassen Stefanov

Professor at Sofia University St. Kliment Ohridski, Bulgaria

Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications

Maria-Angeles Grado-Caffaro

Scientific Consultant, Italy

Domain of Research: Electronics, Sensing and Sensor Networks

Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Intelligent Systems, Data Mining, Databases

T. V. Prasad

Lingaya's University, India

Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics

Reviewer Board Members

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

St. Xaviers College(Autonomous), 30 Park Street,
Kolkata-700 016

- **Athanasios Koutras**
- **Ayad Ismaeel**
Department of Information Systems Engineering-
Technical Engineering College-Erbil Polytechnic
University, Erbil-Kurdistan Region- IRAQ
- **Ayman Shehata**
Department of Mathematics, Faculty of Science,
Assiut University, Assiut 71516, Egypt.
- **Ayman EL-SAYED**
Computer Science and Eng. Dept., Faculty of
Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Bae Bossoufi**
University of Liege
- **BALAMURUGAN RAJAMANICKAM**
Anna university
- **Balasubramanie Palanisamy**
- **BASANT VERMA**
RAJEEV GANDHI MEMORIAL COLLEGE, HYDERABAD
- **Basil Hamed**
Islamic University of Gaza
- **Basil Hamed**
Islamic University of Gaza
- **Bhanu Prasad Pinnamaneni**
Rajalakshmi Engineering College; Matrix Vision
GmbH
- **Bharti Waman Gawali**
Department of Computer Science & information T
- **Bilian Song**
LinkedIn
- **Binod Kumar**
JSPM's Jayawant Technical Campus, Pune, India
- **Bogdan Belean**
- **Bohumil Brtnik**
University of Pardubice, Department of Electrical
Engineering
- **Bouchaib CHERRADI**
CRMEF
- **Brahim Raouyane**
FSAC
- **Branko Karan**
- **Bright Keswani**
Department of Computer Applications, Suresh Gyan
Vihar University, Jaipur (Rajasthan) INDIA
- **Brij Gupta**

University of New Brunswick

- **C Venkateswarlu Sonagiri**
JNTU
- **Chanashekhhar Meshram**
Chhattisgarh Swami Vivekananda Technical
University
- **Chao Wang**
- **Chao-Tung Yang**
Department of Computer Science, Tunghai
University
- **Charlie Obimbo**
University of Guelph
- **Chee Hon Lew**
- **Chien-Peng Ho**
Information and Communications Research
Laboratories, Industrial Technology Research
Institute of Taiwan
- **Chun-Kit (Ben) Ngan**
The Pennsylvania State University
- **Ciprian Dobre**
University Politehnica of Bucharest
- **Constantin POPESCU**
Department of Mathematics and Computer
Science, University of Oradea
- **Constantin Filote**
Stefan cel Mare University of Suceava
- **CORNELIA AURORA Gyorödi**
University of Oradea
- **Cosmina Ivan**
- **Cristina Turcu**
- **Dana PETCU**
West University of Timisoara
- **Daniel Albuquerque**
- **Dariusz Jakóbczak**
Technical University of Koszalin
- **Deepak Garg**
Thapar University
- **Devena Prasad**
- **DHAYA R**
- **Dheyaa Kadhim**
University of Baghdad
- **Djilali IDOUGH**
University A.. Mira of Bejaia
- **Dong-Han Ham**
Chonnam National University
- **Dr. Arvind Sharma**

- Aryan College of Technology, Rajasthan Technology University, Kota
- **Duck Hee Lee**
Medical Engineering R&D Center/Asan Institute for Life Sciences/Asan Medical Center
 - **Elena SCUTELNICU**
"Dunarea de Jos" University of Galati
 - **Elena Camossi**
Joint Research Centre
 - **Eui Lee**
Sangmyung University
 - **Evgeny Nikulchev**
Moscow Technological Institute
 - **Ezekiel OKIKE**
UNIVERSITY OF BOTSWANA, GABORONE
 - **Fahim Akhter**
King Saud University
 - **FANGYONG HOU**
School of IT, Deakin University
 - **Faris Al-Salem**
GCET
 - **Firkhan Ali Hamid Ali**
UTHM
 - **Fokrul Alom Mazarbhuiya**
King Khalid University
 - **Frank Ibikunle**
Botswana Int'l University of Science & Technology (BIUST), Botswana
 - **Fu-Chien Kao**
Da-Y eh University
 - **Gamil Abdel Azim**
Suez Canal University
 - **Ganesh Sahoo**
RMRIMS
 - **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh
 - **George Pecherle**
University of Oradea
 - **George Mastorakis**
Technological Educational Institute of Crete
 - **Georgios Galatas**
The University of Texas at Arlington
 - **Gerard Dumancas**
Oklahoma Baptist University
 - **Ghalem Belalem**
University of Oran 1, Ahmed Ben Bella
 - **gherabi noreddine**
 - **Giacomo Veneri**
University of Siena
 - **Giri Babu**
Indian Space Research Organisation
 - **Govindarajulu Salendra**
 - **Grebenisan Gavril**
University of Oradea
 - **Gufan Ahmad Ansari**
Qassim University
 - **Gunaseelan Devaraj**
Jazan University, Kingdom of Saudi Arabia
 - **GYÖRÖDI ROBERT STEFAN**
University of Oradea
 - **Hadj Tadjine**
IAV GmbH
 - **Haewon Byeon**
Nambu University
 - **Haiguang Chen**
ShangHai Normal University
 - **Hamid Alinejad-Rokny**
The University of New South Wales
 - **Hamid AL-Asadi**
Department of Computer Science, Faculty of Education for Pure Science, Basra University
 - **Hamid Mukhtar**
National University of Sciences and Technology
 - **Hany Hassan**
EPF
 - **Harco Leslie Henic SPITS WARNARS**
Bina Nusantara University
 - **Hariharan Shanmugasundaram**
Associate Professor, SRM
 - **Harish Garg**
Thapar University Patiala
 - **Hazem I. El Shekh Ahmed**
Pure mathematics
 - **Hemalatha SenthilMahesh**
 - **Hesham Ibrahim**
Faculty of Marine Resources, Al-Mergheb University
 - **Himanshu Aggarwal**
Department of Computer Engineering
 - **Hongda Mao**
Hossam Faris
 - **Huda K. AL-Jobori**
Ahlia University
 - **Imed JABRI**

- **iss EL OUADGHIRI**
- **Iwan Setyawan**
Satya Wacana Christian University
- **Jacek M. Czerniak**
Casimir the Great University in Bydgoszcz
- **Jai Singh W**
- **JAMAIAH HAJI YAHAYA**
NORTHERN UNIVERSITY OF MALAYSIA (UUM)
- **James Coleman**
Edge Hill University
- **Jatinderkumar Saini**
Narmada College of Computer Application, Bharuch
- **Javed Sheikh**
University of Lahore, Pakistan
- **Jayaram A**
Siddaganga Institute of Technology
- **Ji Zhu**
University of Illinois at Urbana Champaign
- **Jia Uddin Jia**
Assistant Professor
- **Jim Wang**
The State University of New York at Buffalo,
Buffalo, NY
- **John Sahlin**
George Washington University
- **JOHN MANOHAR**
VTU, Belgaum
- **JOSE PASTRANA**
University of Malaga
- **Jui-Pin Yang**
Shih Chien University
- **Jyoti Chaudhary**
high performance computing research lab
- **K V.L.N.Acharyulu**
Bapatla Engineering college
- **Ka-Chun Wong**
- **Kamatchi R**
- **Kamran Kowsari**
The George Washington University
- **KANNADHASAN SURIYAN**
- **Kashif Nisar**
Universiti Utara Malaysia
- **Kato Mivule**
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Kennedy Okafor**
Federal University of Technology, Owerri
- **Khalid Mahmood**
IEEE
- **Khalid Sattar Abdul**
Assistant Professor
- **Khin Wee Lai**
Biomedical Engineering Department, University
Malaya
- **Khurram Khurshid**
Institute of Space Technology
- **KIRAN SREE POKKULURI**
Professor, Sri Vishnu Engineering College for
Women
- **KITIMAPORN CHOOCHOTE**
Prince of Songkla University, Phuket Campus
- **Krasimir Yordzhev**
South-West University, Faculty of Mathematics and
Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov**
Professor at Sofia University St. Kliment Ohridski
- **Labib Gergis**
Misr Academy for Engineering and Technology
- **LATHA RAJAGOPAL**
- **Lazar Stošić**
College for professional studies educators
Aleksinac, Serbia
- **Leanos Maglaras**
De Montfort University
- **Leon Abdillah**
Bina Darma University
- **Lijian Sun**
Chinese Academy of Surveying and
- **Ljubomir Jerinic**
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Computer Science
- **Lokesh Sharma**
Indian Council of Medical Research
- **Long Chen**
Qualcomm Incorporated
- **M. Reza Mashinchi**
Research Fellow
- **M. Tariq Banday**
University of Kashmir
- **madjid khalilian**
- **majzoob omer**
- **Mallikarjuna Doodipala**
Department of Engineering Mathematics, GITAM
University, Hyderabad Campus, Telangana, INDIA

- **Manas deep**
Masters in Cyber Law & Information Security
- **Manju Kaushik**
- **Manoharan P.S.**
Associate Professor
- **Manoj Wadhwa**
Echelon Institute of Technology Faridabad
- **Manpreet Manna**
Director, All India Council for Technical Education,
Ministry of HRD, Govt. of India
- **Manuj Darbari**
BBD University
- **Marcellin Julius Nkenlifack**
University of Dschang
- **Maria-Angeles Grado-Caffaro**
Scientific Consultant
- **Marwan Alseid**
Applied Science Private University
- **Mazin Al-Hakeem**
LFU (Lebanese French University) - Erbil, IRAQ
- **Md Islam**
sikkim manipal university
- **Md. Bhuiyan**
King Faisal University
- **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
- **Mehdi Bahrami**
University of California, Merced
- **Messaouda AZZOUZI**
Ziane Achour University of Djelfa
- **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in Vranje
- **Miriampally Venkata Raghavendra**
Adama Science & Technology University, Ethiopia
- **Mirjana Popovic**
School of Electrical Engineering, Belgrade University
- **Miroslav Baca**
University of Zagreb, Faculty of organization and
informatics / Center for biometrics
- **Moeiz Miraoui**
University of Gafsa
- **Mohamed Eldosoky**
- **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
- **Mohamed Kaloup**
- **Mohamed El-Sayed**
Faculty of Science, Fayoum University, Egypt
- **Mohamed Najeh LAKHOUA**
ESTI, University of Carthage
- **Mohammad Ali Badamchizadeh**
University of Tabriz
- **Mohammad Jannati**
- **Mohammad Alomari**
Applied Science University
- **Mohammad Haghghat**
University of Miami
- **Mohammad Azzeh**
Applied Science university
- **Mohammed Akour**
Yarmouk University
- **Mohammed Sadgal**
Cadi Ayyad University
- **Mohammed Al-shabi**
Associate Professor
- **Mohammed Hussein**
- **Mohammed Kaiser**
Institute of Information Technology
- **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**
University Tun Hussein Onn Malaysia
- **Mokhtar Beldjehem**
University of Ottawa
- **Mona Elshinawy**
Howard University
- **Mostafa Ezziyyani**
FSTT
- **Mouhammd sharari alkasassbeh**
- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**
University Malaysia Pahang
- **MUNTASIR AL-ASFOOR**
University of Al-Qadisiyah
- **Murphy Choy**
- **Murthy Dasika**
Geethanjali College of Engineering & Technology
- **Mustapha OUJAOURA**
Faculty of Science and Technology Béni-Mellal
- **MUTHUKUMAR SUBRAMANYAM**
DGCT, ANNA UNIVERSITY
- **N.Ch. Iyengar**
VIT University
- **Nagy Darwish**

Department of Computer and Information Sciences,
Institute of Statistical Studies and Researches, Cairo
University

- **Najib Kofahi**
Yarmouk University
- **Nan Wang**
LinkedIn
- **Natarajan Subramanyam**
PES Institute of Technology
- **Natheer Gharaibeh**
College of Computer Science & Engineering at
Yanbu - Taibah University
- **Nazeeh Ghatasheh**
The University of Jordan
- **Nazeeruddin Mohammad**
Prince Mohammad Bin Fahd University
- **NEERAJ SHUKLA**
ITM UNiversity, Gurgaon, (Haryana) Inida
- **Neeraj Tiwari**
- **Nestor Velasco-Bermeo**
UPFIM, Mexican Society of Artificial Intelligence
- **Nidhi Arora**
M.C.A. Institute, Ganpat University
- **Nilanjan Dey**
- **Ning Cai**
Northwest University for Nationalities
- **Nithyanandam Subramanian**
Professor & Dean
- **Noura Aknin**
University Abdelamlek Essaadi
- **Obaida Al-Hazaimeh**
Al- Balqa' Applied University (BAU)
- **Oliviu Matei**
Technical University of Cluj-Napoca
- **Om Sangwan**
- **Omaima Al-Allaf**
Asesstant Professor
- **Osama Omer**
Aswan University
- **Ouchtati Salim**
- **Ousmane THIARE**
Associate Professor University Gaston Berger of
Saint-Louis SENEGAL
- **Paresh V Virparia**
Sardar Patel University
- **Peng Xia**
Microsoft

- **Ping Zhang**
IBM
- **Poonam Garg**
Institute of Management Technology, Ghaziabad
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **PROF DURGA SHARMA (PHD)**
AMUIT, MOEFDRE & External Consultant (IT) &
Technology Tansfer Research under ILO & UNDP,
Academic Ambassador for Cloud Offering IBM-USA
- **Purwanto Purwanto**
Faculty of Computer Science, Dian Nuswantoro
University
- **Qifeng Qiao**
University of Virginia
- **Rachid Saadane**
EE departement EHTP
- **Radwan Tahboub**
Palestine Polytechnic University
- **raed Kanaan**
Amman Arab University
- **Raghuraj Singh**
Harcourt Butler Technological Institute
- **Rahul Malik**
- **raja boddu**
LENORA COLLEGE OF ENGINEERNG
- **Raja Ramachandran**
- **Rajesh Kumar**
National University of Singapore
- **Rakesh Dr.**
Madan Mohan Malviya University of Technology
- **Rakesh Balabantaray**
IIIT Bhubaneswar
- **Ramani Kannan**
Universiti Teknologi PETRONAS, Bandar Seri
Iskandar, 31750, Tronoh, Perak, Malaysia
- **Rashad Al-Jawfi**
Ibb university
- **Rashid Sheikh**
Shri Aurobindo Institute of Technology, Indore
- **Ravi Prakash**
University of Mumbai
- **RAVINA CHANGALA**
- **Ravisankar Hari**
CENTRAL TOBACCO RESEARCH INSTITUE
- **Rawya Rizk**
Port Said University

- **Reshmy Krishnan**
Muscat College affiliated to Stirling University.U
- **Ricardo Vardasca**
Faculty of Engineering of University of Porto
- **Ritaban Dutta**
ISSL, CSIRO, Tasmania, Australia
- **Rowayda Sadek**
- **Ruchika Malhotra**
Delhi Technological University
- **Rutvij Jhaveri**
Gujarat
- **SAADI Slami**
University of Djelfa
- **Sachin Kumar Agrawal**
University of Limerick
- **Sagarmay Deb**
Central Queensland University, Australia
- **Said Ghoniemy**
Taif University
- **Sandeep Reddivari**
University of North Florida
- **Sanskriti Patel**
Charotar University of Science & Technology,
Changa, Gujarat, India
- **Santosh Kumar**
Graphic Era University, Dehradun (UK)
- **Sasan Adibi**
Research In Motion (RIM)
- **Satyena Singh**
Professor
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Seema Shah**
Vidyalankar Institute of Technology Mumbai
- **Seifedine Kadry**
American University of the Middle East
- **Selem Charfi**
HD Technology
- **SENGOTTUVELAN P**
Anna University, Chennai
- **Senol Piskin**
Istanbul Technical University, Informatics Institute
- **Sérgio Ferreira**
School of Education and Psychology, Portuguese
Catholic University
- **Seyed Hamidreza Mohades Kasaei**
University of Isfahan
- **Shafiqul Abidin**
HMR Institute of Technology & Management
(Affiliated to GGSIP University), Hamidpur, Delhi -
110036
- **Shahanawaj Ahamad**
The University of Al-Kharj
- **Shaidah Jusoh**
- **Shaiful Bakri Ismail**
- **Shakir Khan**
Al-Imam Muhammad Ibn Saud Islamic University
- **Shawki Al-Dubae**
Assistant Professor
- **Sherif Hussein**
Mansoura University
- **Shriram Vasudevan**
Amrita University
- **Siddhartha Jonnalagadda**
Mayo Clinic
- **Sim-Hui Tee**
Multimedia University
- **Simon Ewedafe**
The University of the West Indies
- **Siniša Opic**
University of Zagreb, Faculty of Teacher Education
- **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
- **Slim BEN SAOUD**
National Institute of Applied Sciences and
Technology
- **Sofien Mhatli**
- **sofyan Hayajneh**
- **Sohail Jabbar**
Bahria University
- **Sri Devi Ravana**
University of Malaya
- **Sudarson Jena**
GITAM University, Hyderabad
- **Suhail Sami Owais Owais**
- **Suhas J Manangi**
Microsoft
- **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
- **Süleyman Eken**
Kocaeli University
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia

- **Sumit Goyal**
National Dairy Research Institute
- **Supareerk Janjarasjitt**
Ubon Ratchathani University
- **Suresh Sankaranarayanan**
Institut Teknologi Brunei
- **Susarla Sastry**
JNTUK, Kakinada
- **Suseendran G**
Vels University, Chennai
- **Suxing Liu**
Arkansas State University
- **Syed Ali**
SMI University Karachi Pakistan
- **T C.Manjunath**
HKBK College of Engg
- **T V Narayana rao Rao**
SNIST
- **T. V. Prasad**
Lingaya's University
- **Taiwo Ayodele**
Infonetmedia/University of Portsmouth
- **Talal Bonny**
Department of Electrical and Computer Engineering, Sharjah University, UAE
- **Tamara Zhukabayeva**
- **Tarek Gharib**
Ain Shams University
- **thabet slimani**
College of Computer Science and Information Technology
- **Totok Biyanto**
Engineering Physics, ITS Surabaya
- **Touati Youcef**
Computer sce Lab LIASD - University of Paris 8
- **Tran Sang**
IT Faculty - Vinh University - Vietnam
- **Tsvetanka Georgieva-Trifonova**
University of Veliko Tarnovo
- **Uchechukwu Awada**
Dalian University of Technology
- **Udai Pratap Rao**
- **Urmila Shrawankar**
GHRCE, Nagpur, India
- **Vaka MOHAN**
TRR COLLEGE OF ENGINEERING
- **VENKATESH JAGANATHAN**
- **ANNA UNIVERSITY**
- **Vinayak Bairagi**
AISSMS Institute of Information Technology, Pune
- **Vishnu Mishra**
SVNIT, Surat
- **Vitus Lam**
The University of Hong Kong
- **VUDA SREENIVASARAO**
PROFESSOR AND DEAN, St.Mary's Integrated Campus, Hyderabad
- **Wali Mashwani**
Kohat University of Science & Technology (KUST)
- **Wei Wei**
Xi'an Univ. of Tech.
- **Wenbin Chen**
360Fly
- **Xi Zhang**
illinois Institute of Technology
- **Xiaojing Xiang**
AT&T Labs
- **Xiaolong Wang**
University of Delaware
- **Yanping Huang**
- **Yao-Chin Wang**
- **Yasser Albagory**
College of Computers and Information Technology, Taif University, Saudi Arabia
- **Yasser Alginahi**
- **Yi Fei Wang**
The University of British Columbia
- **Yihong Yuan**
University of California Santa Barbara
- **Yilun Shang**
Tongji University
- **Yu Qi**
Mesh Capital LLC
- **Zacchaeus Omogbadegun**
Covenant University
- **Zairi Rizman**
Universiti Teknologi MARA
- **Zarul Zaaba**
Universiti Sains Malaysia
- **Zenzo Ncube**
North West University
- **Zhao Zhang**
Deptment of EE, City University of Hong Kong
- **Zhihan Lv**

Chinese Academy of Science

- **Zhixin Chen**
ILX Lightwave Corporation
- **Ziyue Xu**
National Institutes of Health, Bethesda, MD

- **Zlatko Stacic**
University of Zagreb, Faculty of Organization and
Informatics Varazdin
- **Zuraini Ismail**
Universiti Teknologi Malaysia

CONTENTS

Paper 1: An Approach for Energy Efficient Dynamic Virtual Machine Consolidation in Cloud Environment

Authors: Sara Nikzad, Seyed EnayatOllah Alavi, Mohammad Reza Soltanaghaei

PAGE 1 – 9

Paper 2: Maneuverability of an Inverted Pendulum Vehicle According to the Handle Operation Methods

Authors: Chihiro NAKAGAWA, Takuya CHIKAYAMA, Akikazu OKAMOTO, Atsuhiko SHINTANI, Tomohiro ITO

PAGE 10 – 16

Paper 3: Gaussian Mixture Model and Deep Neural Network based Vehicle Detection and Classification

Authors: S Sri Harsha, K. R. Anne

PAGE 17 – 25

Paper 4: Designing and Implementing of Intelligent Emotional Speech Recognition with Wavelet and Neural Network

Authors: Bibi Zahra Mansouri, Hamid Mirvaziri, Faramarz Sadeghi

PAGE 26 – 30

Paper 5: An IoT Middleware Framework for Industrial Applications

Authors: Nicoleta-Cristina Gaitan, Vasile Gheorghita Gaitan, Ioan Ungurean

PAGE 31 – 41

Paper 6: A Survey of IPv6 Deployment

Authors: Manal M. Alhassoun, Sara R. Alghunaim

PAGE 42 – 46

Paper 7: Intelligent Image Watermarking based on Handwritten Signature

Authors: Saeid Shahmoradi, Nasrollah Sahragard, Ahmad Hatam

PAGE 47 – 53

Paper 8: Fuzzy Risk-based Decision Method for Vehicular Ad Hoc Networks

Authors: Riaz Ahmed Shaikh

PAGE 54 – 62

Paper 9: Good Quasi-Cyclic Codes from Circulant Matrices Concatenation using a Heuristic Method

Authors: Bouchaib AYLALJ, Said NOUH, Mostafa BELKASMI, Hamid ZOUAKI

PAGE 63 – 68

Paper 10: Balanced Distribution of Load on Grid Resources using Cellular Automata

Authors: Amir Akbarian Sadeghi, Ahmad Khademzadeh, Mohammad Reza Salehnamadi

PAGE 69 – 76

Paper 11: Camera Self-Calibration with Varying Intrinsic Parameters by an Unknown Three-Dimensional Scene

Authors: B. SATOURI, A. EL ABDERRAHMANI, H. TAIRI, K. SATORI

PAGE 77 – 87

Paper 12: On the Internal Multi-Model Control of Uncertain Discrete-Time Systems

Authors: Chakra Othman, Ikbel Ben Cheikh, Dhaou Soudani

PAGE 88 – 98

Paper 13: High Performance Computing Over Parallel Mobile Systems

Authors: Doha Ehab Attia, Abeer Mohamed ElKorany, Ahmed Shawky Moussa

PAGE 99 – 103

Paper 14: ROHDIP: Resource Oriented Heterogeneous Data Integration Platform

Authors: Wael Shehab, Sherin M. ElGokhy, ElSayed Sallam

PAGE 104 – 109

Paper 15: Improving the Emergency Services for Accident Care in Saudi Arabia

Authors: Amr Jadi

PAGE 110 – 115

Paper 16: Analysis of Purchasing Tendency using ID-POS Data of Social Login User

Authors: Kohei Otake, Takashi Namatame

PAGE 116 – 123

Paper 17: Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus

Authors: Issa Atoum, Ahmed Otoom

PAGE 124 – 130

Paper 18: Trends of Recent Secure Communication System and its Effectiveness in Wireless Sensor Network

Authors: Manjunath B E, P.V. Rao

PAGE 131 – 139

Paper 19: Estimation Medicine for Diseases System to Support Medical Diagnosis by Expert System

Authors: Noor T. Mahmood

PAGE 140 – 144

Paper 20: Context-Sensitive Opinion Mining using Polarity Patterns

Authors: Saeedeh Sadat Sadidpour, Hossein Shirazi, Nurfadhline Mohd Sharef, Behrouz Minaei-Bidgoli, Mohammad Ebrahim Sanjaghi

PAGE 145 – 150

Paper 21: Application of Intelligent Data Mining Approach in Securing the Cloud Computing

Authors: Hanna M. Said, Ibrahim El Emary, Bader A. Alyoubi, Adel A. Alyoubi

PAGE 151 – 159

Paper 22: Identifying Green Services using GSA Model for Achieving Sustainability in Industries

Authors: Iqbal Ahmed, Hiroshi Okumura, Kohei Arai

PAGE 160 – 167

Paper 23: Using a Cluster for Securing Embedded Systems

Authors: Mohamed Salim LMIMOUNI, Khalid BOUKHDIR, Hicham MEDROMI, Siham BENHADOU

PAGE 168 – 172

Paper 24: Developing a Transition Parser for the Arabic Language

Authors: Aref abu Awad, Essam Hanandeh

PAGE 173 – 175

Paper 25: Multi- Spectrum Bands Allocation for Time-Varying Traffic in the Flexible Optical Network

Authors: KAMAGATE Beman Hamidja, Michel BABRI, GOORE Bi Tra, Souleymane OUMTANAGA

PAGE 176 – 183

Paper 26: Robust Image Watermarking using Fractional Sinc Transformation

Authors: Almas Abbasi, Chaw Seng Woo

PAGE 184 – 189

Paper 27: A Semantic Approach for Mathematical Expression Retrieval

Authors: Zahra Asebriy, Soulaïmane Kaloun, Said Raghay, Omar Bencharef

PAGE 190 – 194

Paper 28: Variability of Acoustic Features of Hypernasality and it's Assessment

Authors: Shahina Haque, Md. Hanif Ali, A.K.M. Fazlul Haque

PAGE 195 – 201

Paper 29: Optimization of Dynamic Virtual Machine Consolidation in Cloud Computing Data Centers

Authors: Alireza Najari, Seyed EnayatOllah Alavi, Mohammad Reza Noorimehr

PAGE 202 – 208

Paper 30: A Light Weight Service Oriented Architecture for the Internet of Things

Authors: Omar Aldabbas

PAGE 209 – 216

Paper 31: Fingerprint Gender Classification using Univariate Decision Tree (J48)

Authors: S. F. Abdullah, A.F.N.A. Rahman, Z.A. Abas, W.H.M. Saad

PAGE 217 – 221

Paper 32: Enhancing Wireless Sensor Network Security using Artificial Neural Network based Trust Model

Authors: Adwan Yasin, Kefaya Sabaneh

PAGE 222 – 228

Paper 33: Security and Privacy Issues in Ehealthcare Systems: Towards Trusted Services

Authors: Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi

PAGE 229 – 236

Paper 34: Estimation of Trajectory and Location for Mobile Sound Source

Authors: Mehmet Cem Catalbas, Merve Yildirim, Arif Gulden, Hasan Kurum, Simon Dobrišek

PAGE 237 – 241

Paper 35: Proposed Bilingual Model for Right to Left Language Applications

Authors: Farhan M Al Obisat, Zaid T Alhalhouli, Hazim S. AlRawashdeh

PAGE 242 – 247

Paper 36: Between Transition from IPv4 and IPv6 Adaption: The Case of Jordanian Government

Authors: Iman Akour

PAGE 248 – 252

Paper 37: A Machine Vision System for Quality Inspection of Pine Nuts

Authors: Ikramullah Khosa, Eros Pasero

PAGE 253 – 267

Paper 38: Predicting CO2 Emissions from Farm Inputs in Wheat Production using Artificial Neural Networks and Linear Regression Models

Authors: Majeed Safa, Mohammadali Nejat, Peter Nuthall, Bruce Greig

PAGE 268 – 274

Paper 39: E-Learning for Secondary and Higher Education Sectors: A Survey

Authors: Sadia Ashraf, Tamim Ahmed Khan, Inayat ur Rehman

PAGE 275 – 283

Paper 40: Design of a Prediction System for Hydrate Formation in Gas Pipelines using Wireless Sensor Network

Authors: Ahmed Raed Moukhtar, Alaa M. Hamdy, Sameh A. Salem

PAGE 284 – 292

Paper 41: The Role of Image Enhancement in Citrus Canker Disease Detection

Authors: K. Padmavathi, K. Thangadurai

PAGE 293 – 296

Paper 42: Analysis of Compensation Network in a Correlated-based Channel using Angle of Arrivals

Authors: Affum Emmanuel Ampoma, Paul Oswald Kwasi Anane, Obour Agyekum Kwame O.-B, Maxwell Opong Afriyie

PAGE 297 – 303

Paper 43: Differential Evolution based SHEPWM for Seven-Level Inverter with Non-Equal DC Sources

Authors: Fayçal CHABNI, Rachid TALEB, M'hamed HELAIMI

PAGE 304 – 311

Paper 44: Human Face Classification using Genetic Algorithm

Authors: Tania Akter Setu, Md. Mijanur Rahman

PAGE 312 – 317

Paper 45: An Example-based Super-Resolution Algorithm for Multi-Spectral Remote Sensing Images

Authors: W. Jino Hans, Lysiya Merlin.S, Venkateswaran N, Divya Priya T

PAGE 318 – 323

Paper 46: Fitness Proportionate Random Vector Selection based DE Algorithm (FPRVDE)

Authors: Qamar Abbas, Jamil Ahmad, Hajira Jabeen

PAGE 324 – 340

Paper 47: Internet of Things based Expert System for Smart Agriculture

Authors: Raheela Shahzadi, Javed Ferzund, Muhammad Tausif, Muhammad Asif Suryani

PAGE 341 – 350

Paper 48: Dependency Test: Portraying Pearson's Correlation Coefficient Targeting Activities in Project Scheduling

Authors: Jana Shafi, Amtul Waheed, Sumaya Sanober

PAGE 351 – 356

Paper 49: Comparison of Digital Signature Algorithm and Authentication Schemes for H.264 Compressed Video

Authors: Ramzi Haddaji, Samia Bouaziz, Raouf Ouni, Abdellatif Mtibaa

PAGE 357 – 363

Paper 50: A Novel Information Retrieval Approach using Query Expansion and Spectral-based

Authors: Sara Alnofaie, Mohammed Dahab, Mahmoud Kamal

PAGE 364 – 373

Paper 51: A Hybrid Steganography System based on LSB Matching and Replacement

Authors: Hazem Hiary, Khair Eddin Sabri, Mohammed S. Mohammed, Ahlam Al-Dhamari

PAGE 374 – 380

Paper 52: A Novel High Dimensional and High Speed Data Streams Algorithm: HSDStream

Authors: Irshad Ahmed, Irfan Ahmed, Waseem Shahzad

PAGE 381 – 392

Paper 53: Anti-noise Capability Improvement of Minimum Energy Combination Method for SSVEP Detection

Authors: Omar Trigui, Wassim Zouch, Mohamed Ben Messaoud

PAGE 393 – 401

Paper 54: An Analysis on Natural Image Small Patches

Authors: Shengxiang Xia, Wen Wang, Di Liang

PAGE 402 – 407

Paper 55: Intelligent Pedestrian Detection using Optical Flow and HOG

Authors: Huma Ramzan, Bahjat Fatima, Ahmad R. Shahid, Sheikh Ziauddin, Asad Ali Safi

PAGE 408 – 417

Paper 56: Modeling and Analyzing Anycast and Geocast Routing in Wireless Mesh Networks

Authors: Fazle Hadi, Sheeraz Ahmed, Abid Ali Minhas, Atif Naseer

PAGE 418 – 423

Paper 57: MOSIC: Mobility-Aware Single-Hop Clustering Scheme for Vehicular Ad hoc Networks on Highways

Authors: Amin Ziagham Ahwazi, MohammadReza NooriMehr

PAGE 424 – 431

Paper 58: Peak-to-Average Power Ratio Reduction based Varied Phase for MIMO-OFDM Systems

Authors: Lahcen Amhaimar, Saida Ahyoud, Adel Asselman, Elkhaldi Said

PAGE 432 – 437

Paper 59: Solving Nonlinear Eigenvalue Problems using an Improved Newton Method

Authors: S.A Shahzadeh Fazeli, F. Rabiei

PAGE 438 – 441

Paper 60: Automatic Generation of Model for Building Energy Management

Authors: Quoc-Dung Ngo, Yanis Hadj-Said, St´ephane Ploix, Ujjwal Maulik

PAGE 442 – 454

An Approach for Energy Efficient Dynamic Virtual Machine Consolidation in Cloud Environment

Sara Nikzad¹

Department of Computer Engineering
Isfahan (Khorasgan) Branch
Islamic Azad University
Isfahan, Iran

Seyed EnayatOllah Alavi^{2,*}

Department of Computer Engineering
Shahid Chamran University of Ahvaz
Ahvaz, Iran

Mohammad Reza Soltanaghaei³

Department of Computer Engineering
Isfahan (Khorasgan) Branch
Islamic Azad University
Isfahan, Iran

Abstract—Nowadays, as the use of cloud computing service becomes more extensive and the customers welcome this service, an increasing trend in energy consumption and operational costs of these centers may be seen. To reduce operational costs, the providers should decrease energy consumption to an extent that Service Level Agreement (SLA) maintains at a desirable level. This paper adopts the virtual machine consolidation problem in cloud computing data centers as a solution to achieve this goal, putting forward solutions to make the decision regarding the necessity of migration from hosts and finding appropriate hosts as destinations of migration. Using time-series forecasting method and Double Exponential Smoothing (DES) technique, the proposed algorithm predicts CPU utilization in near future. It also proposes an optimal equation for the dynamic lower threshold. Comparing current and predicted CPU utilization with dynamic upper and lower thresholds, this algorithm identifies and categorizes underloaded and overloaded hosts. According to this categorization, migration then occurs from the hosts that meet the necessary conditions for migration. This paper identifies a certain type of hosts as “troublemaker hosts”. Most probably, the process of prediction and decision making regarding the necessity of migration will be disrupted in the case of these hosts. Upon encountering this type of hosts, the algorithm adopts policies to modify them or switch them to sleep mode, thereby preventing the adverse effects caused by their existence. The researchers excluded all overloaded, prone-to-be-overloaded, underloaded, and prone-to-be-underloaded hosts from the list of suitable hosts to find suitable hosts as destinations of migration. An average improvement of 86.2%, 28.4%, and 87.2% respectively for the number of migrations of virtual machines, energy consumption, and SLA violation is among the simulation achievements of this algorithm using Clouds tool.

Keywords—Cloud Computing; Service Level Agreement; Energy Consumption; Virtualization; Dynamic Consolidation; Data Center

I. INTRODUCTION

Cloud computing is a model that provides access to infrastructure including a set of configurable computing resources such as servers, storages, applications, services, etc. This model provides them for applicants via available connection infrastructure such as network and the Internet in an easy, rapid and on-demand manner, while taking into account quality of service. IaaS, PaaS, and SaaS are the three major types of cloud computing services [1, 2]. IaaS presents data centre infrastructures, servers and storage spaces as well as hardware policies independently from location and

geographical limitations and under computer networks. Instead of purchasing IT infrastructure and getting involved with equipment maintenance and enhancement, organizations fulfill their computational needs using cloud computing on a pay-as-you-go basis [3, 4].

In recent years, in light of the ever increasing expansion of the use of cloud computing services and due to the fact that customers welcomed this service, cloud computing service providers have increased the number and volume of greedy data centres that consume huge amounts of energy [5]. This has incurred enormous operational costs. Quality of service assurance, included in SLA, and is agreed upon between customers and providers, is a necessity for the cloud computing environment. Hence, cloud computing service providers tend to bring about a trade-off between energy and performance and should reduce energy consumption because it would not disrupt or decrease quality of service to reduce operational costs [3].

The major energy loss mainly occurs in hardware infrastructure of cloud computing data centres. Research has shown that although the power consumed by hardware equipment is idle, it is almost equal to that at the peak of consumption. Thus, failure to utilize them in a perfect manner will result in a huge energy loss [6]. In this respect, Forrester research team observed that when a server is idle for 70% of the time, it consumes a power of almost 30% of the consumption peak power [7]; therefore, what mainly accounts for energy loss in cloud computing data centres is the use of equipment while their utilization is at low levels [6]. Virtualization, is the key feature and main basis of cloud computing, making possible the establishment of several VMs on a host as well as the migration of VMs [8].

The optimal consolidation problem of VMs using virtualization technology is an effective approach to achieve energy efficient cloud computing data centres [9-11]. Because it allows VMs on hosts to migrate to other suitable hosts when the work load of hosts is low, and the hosts that have become idle to switch to sleep mode [5].

The study employed live migration of VM to transfer VM without suspension and with minimum downtime. However, any VM migration involves certain performance degradation and consequently potential SLA violation [12]. On the other hand, unnecessary VM migration will lead to extra management costs (such as VM reconfiguration, VM creation

and destruction, etc.), resulting in additional energy consumption [13]. Hence, the researchers avoided unnecessary VM migrations to reduce SLA violation and energy consumption to the extent possible, i.e. they should minimize the number of migrations. This paper addressed VM dynamic consolidation problem in cloud computing data centres as a solution to tackle the mentioned problems.

In some related studies, the authors made decisions just based on current utilization of hosts. On the other hand, in some other related studies, the researchers made decisions just based on upper threshold. However, In this study, Proposed Algorithm makes decision based on dynamic upper and lower threshold as well as current and predicted CPU utilization. Solutions are put forward in this regard that are briefly described as follows:

- Proposing an optimal equation to calculate the dynamic lower threshold and presenting a technique to identify and categorize underloaded hosts.
- Decision making regarding the necessity of migration from hosts using the comparison of current and predicted CPU utilization with dynamic upper and lower thresholds as well as the identification and categorization of overloaded and underloaded hosts.
- Identifying a certain type of hosts called troublemaker hosts and adopting policies to modify them or switch them to sleep mode.
- Presenting a method to find hosts that are appropriate destinations by excluding existing hosts in all of the considered categories.

Further in this paper, section 2 investigates previous research. Section 3 describes the proposed algorithm. Section 4 determines the capability of running the proposed algorithm using Clouds tool. Finally, Section 5 presents the conclusion and looks into future works.

II. PREVIOUS RESEARCH

In [14], Wu et al. used GA to solve the consolidation problem. They investigated the energy consumed by physical machines and inter-connection networks in data centers. They found out that compared with FFD [15], their technique generates better solutions. However, FFD is faster in calculation compared with their method.

In [11], the authors put forward MU, MMT, MC, and RC policies for the VM selection problem. They proposed MAD and IQR techniques to find the dynamic upper threshold. In their study, a host is considered to be an overloaded one provided that its current CPU utilization is greater than the dynamic upper threshold.

In [16], Tang and Pan proposed an Hybrid Genetic Algorithm (HGA) to solve the consolidation problem. To rapidly improve solutions, they adopted a local optimization procedure. To gradually work out the violations of conditions in infeasible solutions and convert an infeasible solution to a feasible one, they employed an infeasible solution repairing procedure. They realized that HGA converges faster than GA

and also exhibits remarkably better results in terms of performance and utilization.

In [17], the authors divided up the entire population to a number of families and performed genetic operations on these families in parallel in order to generate an optimal mapping between the set of hosts and VMs. Thus, they presented Family GA, which is a model of Parallel GA. They made use of a self-adjusting mutation operator to prevent untimely convergence in the people of the population.

In [18], Farahnakian et al. adopted the regression forecasting technique k-nearest Neighbor, proposed in [19], to forecast resource utilization. They solved the consolidation problem using the current utilization of resources and prediction of resource utilization in future.

In [6], the authors proposed policies to determine underloaded hosts as well as a policy for the placement of migratable VMs, where they used a multi-criteria decision making technique. They also put forward a novel and comprehensive procedure for cloud resource management called Enhanced Optimization (EO) that offers an all-embracing outlook on the resource management procedure.

In [12], Singh and Shaw proposed an algorithm for decision making about the necessity of migration and finding the appropriate destination host using time-series forecasting technique as well as dynamic upper threshold and moving average, SES, and DES techniques. In their algorithm, a host is determined to be overloaded whose current and predicted CPU utilization is greater than the upper threshold.

In [20], the authors made changes to certain parts of the original ant colony algorithm, namely pheromone updating, definition, and aggregation, in a way that it would be fit to be used in multi-objective problems. They adopted the mentioned algorithm to allocate resources to VMs in order to reduce energy consumption and waste of resources. The results indicated a better performance of this algorithm compared with GA in terms of both aspects.

In [13], Fu and Zhou presented two new policies, namely MP and MCC. The first policy functions with the aid of satisfaction from resources and CPU utilization and selects VMs for migration using dynamic upper and lower thresholds. The second policy functions using a correlation coefficient and finds the suitable destination host.

III. THE PROPOSED ALGORITHM

In light of [21], the authors divided the VM dynamic consolidation problem in cloud computing data centres into the following three parts:

- Part 1: Decision making regarding the necessity of migration from hosts.
- Part 2: VM selection for migration.
- Part 3: Finding suitable destination hosts.

The authors answered the first and the third parts of the consolidation problem using the proposed algorithm. To answer to the VM selection problem, the study adopted the

minimum utilization (MU) policy presented in [11], as in [12]. From among the VMs existing on one host, this policy selects the VM with minimum CPU utilization.

A. Decision Making regarding the Necessity of Migration from Hosts

In the proposed algorithm, to make decision regarding the necessity of migration from hosts, current CPU utilization is not the mere criterion to take action; rather, as in [12], the authors also used the CPU utilization history of the host in several recent periods and forecasted CPU utilization in near future by taking advantage of time-series method and DES technique. [22-24] provided further explanation about the forecasting method. In contrast with [12], where it merely employed dynamic upper threshold, this algorithm also makes use of the dynamic lower threshold. Comparison of forecasted and current CPU utilization with dynamic upper and lower thresholds determined the status of each host. The authors, then categorized hosts and made decisions regarding the necessity of migration using the categorization mentioned above. The following section describes the proposed algorithm.

This algorithm receives a host as input, examines its status, and makes a decision for its migration. As in [12], the method presented in [11] (Step 4) calculated the upper threshold, where the authors considered the value of parameter s_1 to be 2.5 in conformity with [11]. To calculate the dynamic lower threshold, they proposed and adopted equation (1), (Step 6) inspired by the upper threshold method presented in [11] and using median absolute deviation (MAD) method. It is noteworthy that MAD technique [11] also proposed. In (1), they considered the value of s_2 to be 2.5 empirically and through conducting numerous experiments.

$$\text{Lower Threshold} = 0.3 + s_2 \times \text{MAD} \quad (1)$$

The study included a great number of experiments to identify a certain type of hosts that are prone to the emergence of undesirable conditions. These hosts are called “troublemaker hosts” in this paper. To prevent the emergence of undesirable conditions and improve results, Step 6 of the proposed algorithm adopts policies to identify troublemaker hosts given the status of the hosts, and then modifies or removes them.

The first type of troublemaker hosts are those whose MAD is greater than 0.065 (this number is obtained empirically). Considering that MAD shows the strength of the deviation of the host CPU utilization, the experiments demonstrated that when the value of MAD exceeds 0.065, the deviations of CPU utilization increase and algorithm accuracy in predicting CPU utilization decreases.

On the other hand, as MAD increases, the accuracy of the calculation of dynamic upper and lower thresholds decreases, in a way that it may consider a host with a normal load to be overloaded or underloaded and this may lead to unnecessary migrations; or it may consider an overloaded or underloaded host to be a host with normal load, and this may lead to SLA violation in the host mentioned above. It is notable that the bigger the MAD, the smaller the upper threshold. Thus, CPU utilization is not used desirably. On the other hand, as MAD increases, it is more likely for CPU utilization to reach 100% and SLA to violate [11]. In the proposed algorithm, to prevent

potential problems from arising regarding this type of troublemaker hosts, it considered the lower threshold to be equal to 0.9. Using this technique, this type of hosts most probably become underloaded and all of the VMs thereon are transferred to suitable hosts. Then the idle hosts will switch to sleep mode.

Proposed Algorithm:

Decision making regarding the necessity of migration

Input: host

Output: migration decision (true/false)

- 1) $\text{flagPO} = \text{flagFO} = \text{flagPU} = \text{flagFU} = \text{false}$
 - 2) Find current Utilization of host h
 $\text{Utilization} = \text{total requested MIPS}/h.\text{getTotalMips}()$
 - 3) $\text{data}[] = h.\text{getUtilizationHistory}()$
 - 4) Find UpperThreshold using MAD
 $\text{UpperThreshold} = 1 - s_1 \times \text{MAD}$
 - 5) Find LowerThreshold using MAD
 $\text{LowerThreshold} = 0.3 + s_2 \times \text{MAD}$
 - 6) **if** ($\text{MAD} > 0.065$) **then**
 $\text{LowerThreshold} = 0.9$
else
 if ($\text{UpperThreshold} < 0.85$) **then**
 $\text{UpperThreshold} = 0.9$
 if ($\text{LowerThreshold} > 0.35$) **then**
 $\text{LowerThreshold} = 0.3$
 - 7) **if** ($\text{Utilization} > \text{UpperThreshold}$) **then**
 $\text{flapPO} = \text{true}$
 - 8) **if** ($\text{Utilization} < \text{LowerThreshold}$) **then**
 $\text{flapPU} = \text{true}$
 - 9) **if** ($\text{data.lenght} < 10$ and $\text{flagPO} == \text{true}$) **then**
 $\text{OverUtilizedHosts.add}(h)$
 Return True
 - 10) Find future Utilization using DES Technique
 $\text{future_Utilization} = \text{getHostFutureLoad}(\text{data})$
 - 11) **if** ($\text{future_Utilization} > \text{UpperThreshold}$) **then**
 $\text{flagFO} = \text{true}$
 - 12) **if** ($\text{future_Utilization} < \text{LowerThreshold}$) **then**
 $\text{flagFU} = \text{true}$
 - 13) **if** ($\text{flagFO} == \text{false}$ and $\text{flagPO} == \text{true}$) **then**
 $\text{currentOverUtilizedHosts.add}(h)$
 - 14) **if** ($\text{flagFO} == \text{true}$ and $\text{flagPO} == \text{false}$) **then**
 $\text{predictedOverUtilizedHosts.add}(h)$
 - 15) **if** ($\text{flagFO} == \text{true}$ and $\text{flagPO} == \text{true}$) **then**
 $\text{overUtilizedHosts.add}(h)$
 - 16) **if** ($\text{flagFU} == \text{false}$ and $\text{flagPU} == \text{true}$) **then**
 $\text{currentUnderUtilizedHosts.add}(h)$
 - 17) **if** ($\text{flagFU} == \text{true}$ and $\text{flagPU} == \text{false}$) **then**
 $\text{predictedUnderUtilizedHosts.add}(h)$
 - 18) **if** ($\text{flagFU} == \text{true}$ and $\text{flagPU} == \text{true}$) **then**
 $\text{undertilizedHosts.add}(h)$
-

A MAD empirically below 0.065 and various experiments identified the second type of troublemaker hosts. In this case, if

the upper threshold is smaller than 0.85, it is equal to 0.9 because authors determined overloaded hosts incorrectly and used their capacity improperly. On the other hand, if the lower threshold is greater than 0.35, the lower threshold is equal to 0.3 because they determined underloaded hosts incorrectly and unnecessary migrations increase.

Contrary to [12] that uses two flags, the proposed algorithm uses four flags. Each of the flags flagFO, flagPO, flagFU, and flagPU being true respectively shows the host's being potentially overloaded in near future. The host's being overloaded at present, the host's being potentially underloaded in near future, and the host's being underloaded at present. In this algorithm, if current CPU utilization is greater than the upper threshold, flagPO will be true (Step 7). If current CPU utilization is below the lower threshold, flagPU will be true (Step 8).

As in [12], the length of data array is considered to be 10; that is, to forecast a host's being overloaded or underloaded required at least 10 data from CPU utilization history. If fewer than 10 data are available and flagPO is true, that host will enter the list of overloaded hosts and the algorithm ends; otherwise, the algorithm continues and runs the subsequent steps (Step 9). Authors used DES technique (Step 10) to forecast CPU utilization in near future. They compared the resulting value with dynamic upper and lower thresholds and flagFO and flagFU values are then set (Step 11 and 12). Step [12] considered three categories to categorize overloaded hosts. In the proposed algorithm, three other categories will add to the previous three categories to categorize underloaded hosts. Thus, they categorized overloaded and underloaded hosts in 6 different categories. Further, section below describes each category:

- First category: FlagFO is false and flagPO is true. In this case, the host overloads at present; however, the authors forecasted that it will not overload in near future. In these circumstances, they adds the respective host to the list of currently overloaded hosts (currentOverUtilizedHosts). However, the migration of VMs does not take place from this category of hosts since they forecasted that this host will not overload in future and to decrease unnecessary migration (Step 13).
- Second category: FlagFO is true and flagPO is false. In this case, the host does not overload at present; however, the authors forecasted that it will overload in near future. In these circumstances, they added the respective host to the list of hosts that they forecasted to overload in future (predictedOverUtilizedHosts). However, the migration of VMs does not take place from this category of hosts since they will not overload at present (Step 14).
- Third category: FlagFO and flagPO are both true. In this case, the host overloads at present and the forecast is that it will overload also in near future. In these circumstances, the respective host will add to the list of overloaded hosts (overUtilizedHosts). To reduce the load of this category of hosts, they selected and migrated a number of VMs (Step 15).

- Fourth category: FlagFU is false and flagPU is true. In this case, the host underloads at present; however, the forecast is that it will not remain underloaded in near future. In these circumstances, the respective host will add to the list of currently underloaded hosts (currentUnderUtilizedHosts). However, the migration of VMs does not take place from this category of hosts since the forecast is that this host will not underload in future and in order to decrease unnecessary migration (Step 16).
- Fifth category: FlagFU is true and flagPU is false. In this case, the host does not underload at present; however, the forecast is that it will underload in near future. In these circumstances, the respective host will add to the list of hosts that based on forecast, will underload in future (predictedUnderUtilizedHosts). However, the migration of VMs does not take place from this category of hosts since they does not underload at present (Step 17).
- Sixth category: FlagFU and flagPU are both true. In this case, the host underloads at present and the forecast is that it will underload in near future as well. In these circumstances, the respective host will add to the list of underloaded hosts (underUtilizedHosts). To reduce energy consumption, all of the VMs on these hosts will migrate. Afterward, idle hosts will switch to sleep mode (Step 18).

In the above categorization, migration is necessary merely for the hosts in the third and sixth categories. The hosts in the third category will definitely be overloaded. In order for their load to become normal and their CPU utilization to be below the upper threshold, one VM or more should migrate therefrom. The hosts in the sixth category will definitely be underloaded. All of the VMs thereon should migrate to hosts that meet the necessary conditions as migration destinations. If the migration of all of the VMs on the source host succeeds, that host will switch to sleep mode because it becomes idle. Table (1) shows a summary of the above categorization.

TABLE I. CATEGORIZATION OF OVERLOADED AND UNDERLOADED HOSTS

NO.	flagFO	flagPO	flagFU	flagPU	List Name
1	False	True	-	-	currentOverUtilizedHosts
2	True	False	-	-	predictedOverUtilizedHosts
3	True	True	-	-	overUtilizedHosts
4	-	-	False	True	currentUnderUtilizedHosts
5	-	-	True	False	predictedUnderUtilizedHosts
6	-	-	True	True	underUtilizedHosts

As mentioned in [12] and in light of the mechanism of finding underloaded hosts in [12], it is noteworthy that at load peak time, when the utilization of all hosts is at a high level, the hosts that have lower utilizations compared with other hosts will identify as underloaded hosts and the VMs thereon will migrate to other hosts. This may increase the rate of

unnecessary migration. In the proposed algorithm, as stated earlier, this problem will resolve using the dynamic lower threshold and a proper method to find underloaded hosts.

B. Finding Suitable Destination Hosts

To find appropriate destination hosts for migration, the authors eliminate a number of hosts that do not fit to be a destination host from the list of suitable destination hosts. They consider a total of 6 different categories for overloaded and underloaded hosts in this paper. In [12], they exclude only the first three categories from the mentioned categorization from the list of suitable destination hosts. In addition to the first three categories, they also excluded the second three categories that contain underloaded hosts or are prone to be underloaded from the list of suitable destination hosts in this paper. In fact, this prevents the hosts that are underloaded or prone to become underloaded from remaining on. A considerable reduction in the energy consumption of the data center may be brought about by turning them off. In contrast with [12] that tries to select the destination host from among underloaded hosts and those with normal loads, efforts are made in this paper to select those hosts as destination hosts that have normal loads. With this policy adopted, they optimized the selection of destination hosts. As a result, they eliminated unnecessary migrations and energy consumption decreases substantially.

IV. INTEGRATION OF PARTS OF PROPOSED ALGORITHM

In the first part of the proposed algorithm, the authors have classified the overloaded hosts, those hosts prone to be overloaded, also underloaded hosts, and those prone to be underloaded into six different categories.

To make the load normal on the overloaded hosts by utilizing MU policy, the authors have selected certain number of virtual machines to migrate from these hosts. They have also added all the virtual machines available on the underloaded hosts, for migration, to the migration list.

In the second part of the proposed algorithm with the aim of preventing from unnecessary migrations of virtual machine, authors have excluded the six identified categories in the first part from destination host list. Therefore, they can create an optimized list of destination hosts.

Finally, the authors have selected a mapping from among virtual machines for migration (from overloaded and underloaded hosts) and made a suitable list of destination hosts. Then, each virtual machine will migrate to a host that has the minimum increasing power after migration of that virtual machine.

V. PERFORMANCE ANALYSIS

Authors selected Clouds 3.0.3 tool for the simulation of this paper. Further explanation about this tool may be found in [25-27]. To investigate the performance of the proposed algorithm, they compared this algorithm, MAD-MU algorithm presented in [11], and the algorithm presented in [12] in terms of different metrics. They analyzed and examined the results obtained from the comparison of these algorithms.

MAD-MU algorithm, henceforth called MM in this paper for brevity, is implemented in Cloudsim tool by the authors of

[11]. To select a VM for migration, this algorithm makes use of MU policy and takes advantage of MAD technique to calculate dynamic upper threshold. The algorithm presented in [12], henceforth called MMD (MAD-MU-DES) in this paper for brevity, functions similarly to MM upon selecting VM and calculating dynamic upper threshold. It makes use of DES technique for the best results obtained to predict host CPU utilization in future. Since the algorithm proposed in this paper strives to optimize MMD, it is henceforth called OMMD (Optimized MMD) for brevity.

A. Performance Metrics

This paper adopted 6 metrics to compare the proposed algorithm with MM and MMD algorithms, namely energy consumption, the number of VM migrations, PDM, SLATAH, SLAV, and ESV.

PDM metric demonstrates performance degradation due to VM migration. SLATAH metric shows the percentage of time that the host has a CPU utilization of 100%. SLAV metric specifies in what percentage of time the resources allocated to the host are less than the resources demanded by that host and it determines the rate of SLA violation. ESV metric is obtained by multiplying energy consumption by SLAV. It indicates the simultaneous improvement of these two metrics and reveals a trade-off between them. section [11] includes further explanations about each of these metric.

B. Experiment Settings

Since the algorithm presented in this paper attempts to improve the performance of MM and MMD, the authors employed experiment settings similar to these algorithms, which are available in [11, 12]. They simulated a data centre including 800 heterogeneous physical hosts. Half of these hosts are of type HP ProLiant ML110 G4 (Intel Xeon 3040, 2 cores \times 1860 MHz, 4 GB) and the other half is of type HP ProLiant ML110 G5 (Intel Xeon 3075, 2 cores \times 2660 MHz, 4 GB). This data centre has several types of VMs including High-CPU Medium Instance (2500 MIPS, 0.85 GB), Extra Large Instance (2000 MIPS, 3.75 GB), Small Instance (1000 MIPS, 1.7 GB), and Micro Instance (500 MIPS, 613 MB).

C. Workload Data

Today, research projects that require the work load of real data centres for simulation make use of the data pertaining to a 10-day workload from CoMon project [28], which is a monitoring infrastructure for PlanetLab and collected in March and April 2011. This data comprises CPU utilization data collected at 5-minute intervals from over thousands of operational VMs relating to service providers in more than 500 locations around the world. They will embed as defaults in Clouds simulator. This paper adopts the same data to evaluate the performance of the proposed algorithm and compare it with MM and MMD.

D. Simulation Results

This section will compare the proposed algorithm with MM and MMD algorithms according to the mentioned metrics and using the data of 10 workdays. Fig. 1 to Fig. 6 depict the comparison results.

Fig. 1 shows the comparison of the performances of MM, MMD, and OMMD in terms of the number of migrations metric. On average, OMMD has achieved 89.16% and 83.25% decrease respectively in comparison with MM and MMD.

In OMMD, the number of migrations has considerably decreased using the method presented for finding underloaded hosts and adding 3 new categories for categorizing this sort of hosts and also by eliminating unnecessary migrations from the hosts that are not really underloaded. Another reason why OMMD shows improved results regarding this metric is the identification of troublemaker hosts and adopting policies to modify or eliminate them.

By modifying the aforesaid hosts, the accuracy of identifying overloaded and underloaded hosts increases. This fact causes the number of migrations stemming from the existence of this type of hosts to decrease. On the other hand, in OMMD, in addition to overloaded hosts and those that are prone to be overloaded, the authors excluded underloaded hosts and those that are prone to be underloaded from the list of destination hosts. Consequently, they select the destination hosts with higher accuracy and quality. Thus, this will prevent the repeated migration of VMs as a result of migration to inappropriate destinations.

Fig. 2 shows the comparison of the performance of MM, MMD, and OMMD with respect to the energy consumption metric. On average, OMMD has achieved 35.09% and 21.63% decrease respectively in comparison with MM and MMD. What mainly accounts for the reduced energy consumption is the fact that OMMD obtains underloaded hosts optimally and with greater accuracy in comparison with the other two algorithms. As a result, it prevents energy loss in data centers to a great extent by turning off hosts where their utilization is at a low level.

On the other hand, since the authors selected the destination hosts more accurately in OMMD, this will prevent from the migration of VMs to hosts that are underloaded or prone to be

underloaded. Thus they provided more appropriate conditions to switch to sleep mode. In addition to the above, the policies adopted to manage the problems of troublemaker hosts exerted favorable effects on the quality of selecting underloaded and overloaded hosts and, hence, reduced energy consumption.

Fig. 3 exhibits a comparison of the performance of the three mentioned algorithms with regard to PDM metric. On average, OMMD has achieved 90.65% and 87.54% decrease respectively in comparison with MM and MMD. What mainly accounts for this remarkable improvement is the substantial decrease in the number of migrations in OMM

Fig. 4 depicts a comparison of the performance of MM, MMD, and OMMD with respect to SLATAH metric. In comparison with MM and MMD, OMMD has shown a poorer performance in the majority of cases. Efforts made to maximize the utilization of the hosts perhaps have caused this. Since SLAV metric is the multiplication of PDM and SLATAH metrics, in light of the remarkable results of PDM metric, somewhat poor results regarding SLATAH metric is negligible in the proposed algorithm. This may be clearly seen upon investigating and analyzing the figure pertaining to SLAV metric.

Fig. 5 shows a comparison of the performance of the three mentioned algorithms with regard to SLAV metric. On average, OMMD has achieved 89.46% and 84.86% decrease respectively in comparison with MM and MMD. What mainly accounts for this substantial improvement is the improvement of PDM metric.

Fig. 6 shows a comparison of the performance of MM, MMD, and OMMD with respect to ESV metric. On average, OMMD has achieved 93.17% and 88% decrease respectively in comparison with MM and MMD. The reason behind this considerable decrease is the decreases in energy consumption and SLA violation rate. As a matter of fact, these results suggest that there has been a successful trade-off in this paper between these two metrics.

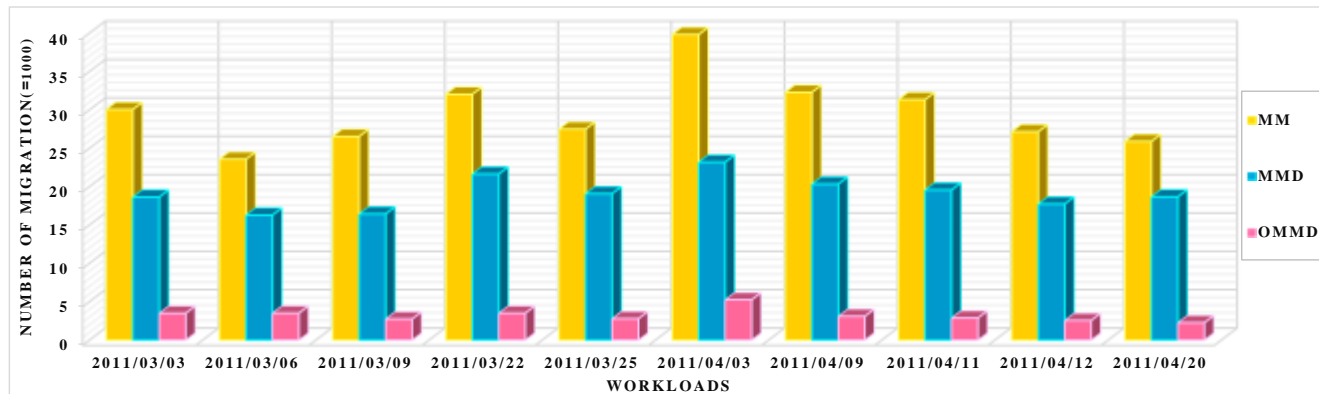


Fig. 1. Comparison of algorithms with regard to number of migration for 10 workdays

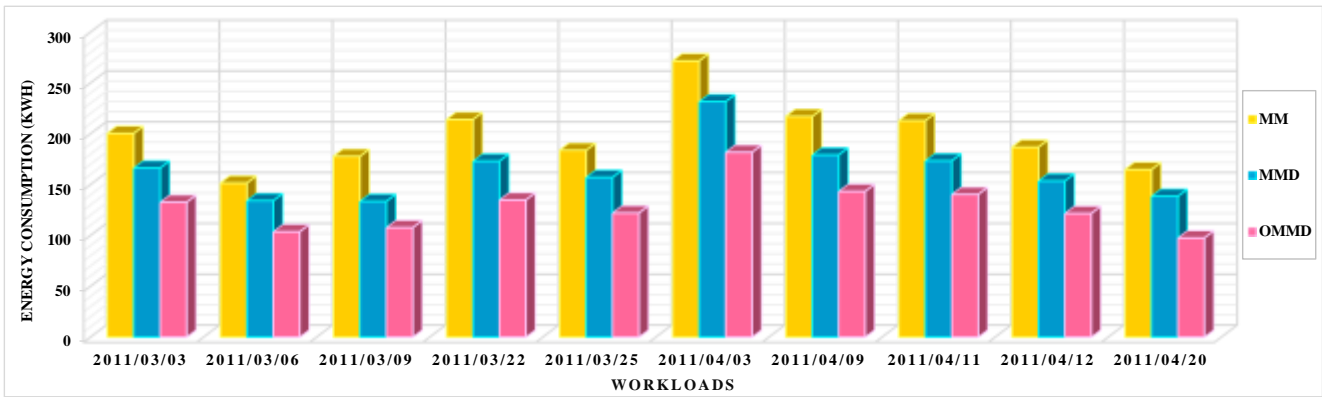


Fig. 2. Comparison of algorithms with regard to energy consumption for 10 workdays

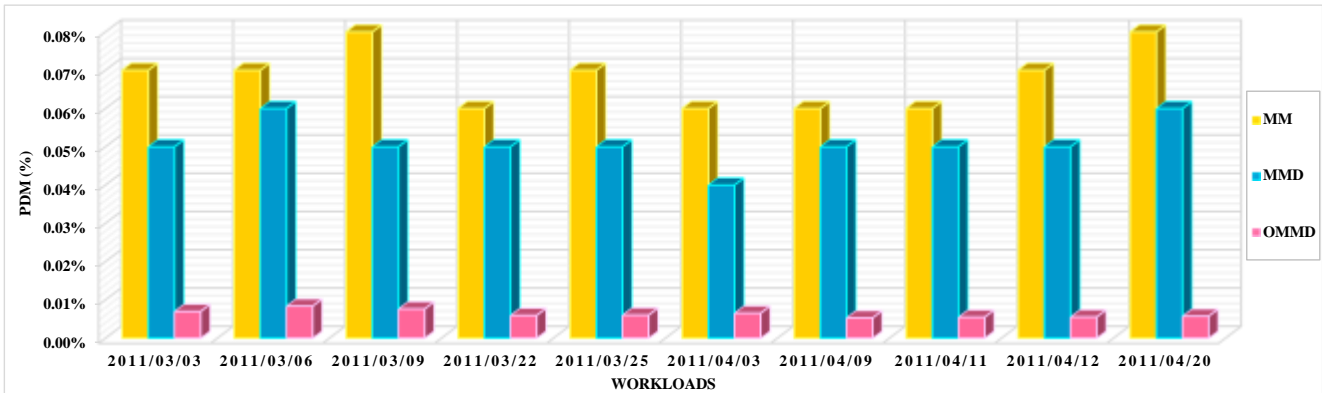


Fig. 3. Comparison of algorithms with regard to PDM for 10 workdays

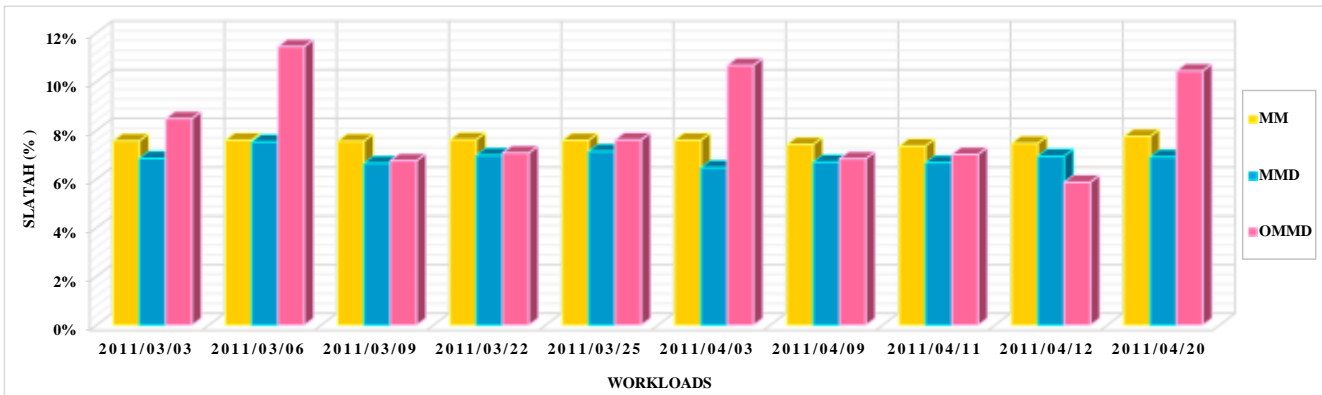


Fig. 4. Comparison of algorithms with regard to SLATAH for 10 workdays

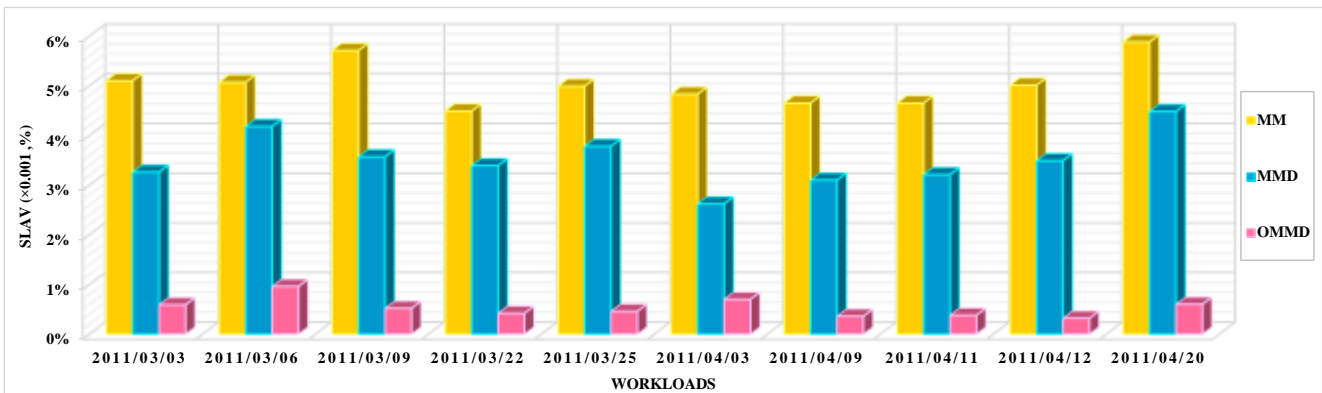


Fig. 5. Comparison of algorithms with regard to SLAV for 10 workdays

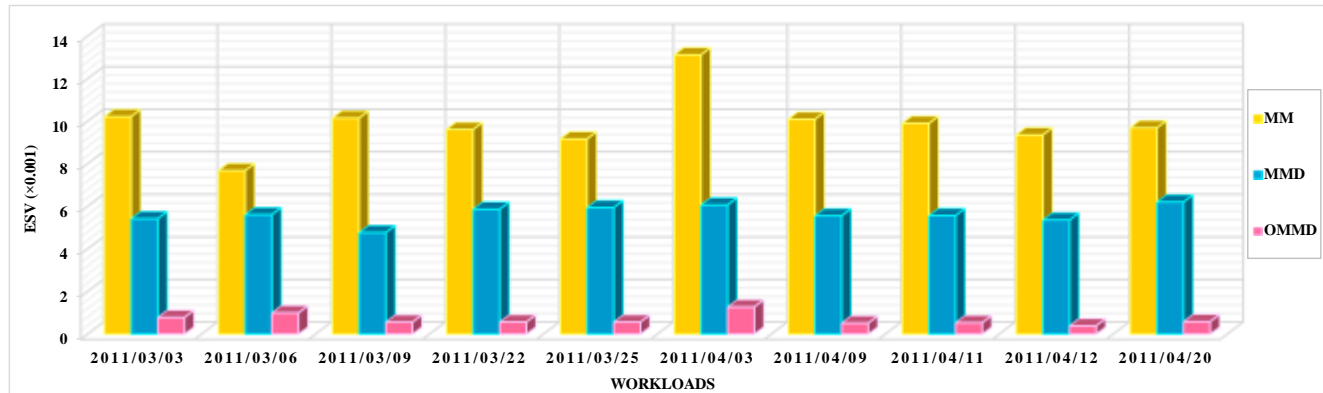


Fig. 6. Comparison of algorithms with regard to ESV for 10 workdays

VI. CONCLUSION AND FUTURE RESEARCH OPPORTUNITIES

The major concern of cloud computing data centres is the decrease in energy consumption and, consequently, reduced operational costs and increased profitability of these centres.

In OMMD, the VM dynamic consolidation problem in cloud computing data centers was brought into spotlight as a solution to battle this problem. In this respect, the authors have provided solutions to make decision about the necessity of migration from hosts and finding suitable destination hosts.

To make decision about the necessity of migration, they have compared current and predicted CPU utilization with dynamic upper and lower thresholds. Thereby, they have identified and categorized underloaded and overloaded hosts.

According to the categorization and the identity of each category, migration took place from the hosts that met necessary conditions for migration. The number of migrations and, as a result, SLA violation rate decreased remarkably using the method proposed for calculating the dynamic lower threshold and finding underloaded hosts and adding 3 new categories to categorize these hosts and also by eliminating unnecessary migrations from hosts that are not really underloaded.

On the other hand, as the accuracy in identifying underloaded hosts increased and by turning them off, they prevented from energy loss in the data center to a considerable extent.

To encounter and prevent the disruptions and adverse effects stemming from the existence of troublemaker hosts, OMMD adopted the policy of modifying them or switching them to sleep mode given the status of those hosts. Thus, the accuracy of identifying overloaded and underloaded hosts increased. This fact had a substantial effect on reduced number of migrations, SLA violation rate, and energy consumption.

OMMD managed to establish a proper trade-off between energy consumption and SLA violation. The results of comparing OMMD with MM and MMD are as follows: respectively 89.16% and 83.25% improvement in the number of migration metric, respectively 35.09% and 21.63% improvement in the energy consumption metric, respectively

90.65% and 87.54% improvement in PDM metric, respectively 89.46% and 84.86% improvement in SLAV metric, and respectively 93.17% and 88% improvement in ESV metric.

Proposed future works:

- OMMD have adopted MU technique for the VM selection problem and no new algorithm was put forward for that. Therefore, the authors recommended that this technique be optimized or a new method be adopted to improve the results even more.
- Given that the improvement of SLA metric can substantially affect the quality improvement of the results of the proposed algorithm, efforts should be made in future studies to alleviate the defect of the SLATAH metric.
- Even though OMMD exhibited remarkable results in the simulation environment, the effect of this algorithm in a real cloud infrastructure is not clearly obvious. Hence, in order to evaluate the performance of the proposed algorithm, it can develop in a real cloud environment such as OpenStack, which is a free open-source software, for future works.
- In addition to physical hosts energy consumption, energy consumption can be Investigate, examine and take into consideration in the communication infrastructures.

REFERENCES

- [1] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future generation computer systems*, vol. 28, no. 5, pp. 755-768, 2012.
- [2] R. Jeyarani, N. Nagaveni, and R. V. Ram, "Design and implementation of adaptive power-aware virtual machine provisioner (APA-VMP) using swarm intelligence," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 811-821, 2012.
- [3] A. Beloglazov, and R. Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers.", *MGC '10 Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science*, article no. 4, 2010.
- [4] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities." pp. 5-13, 2008.

- [5] Y. Gao, H. Guan, Z. Qi, T. Song, F. Huan, and L. Liu, "Service level agreement based energy-efficient resource management in cloud data centers," *Computers & Electrical Engineering*, vol. 40, no. 5, pp. 1621-1633, 2014.
- [6] E. Arianyan, H. Taheri, and S. Sharifian, "Novel energy and SLA efficient resource management heuristics for consolidation of virtual machines in cloud data centers," *Computers & Electrical Engineering*, vol. 47, pp. 222-240, 2015.
- [7] M. Poess, and R. O. Nambiar, "Energy cost, the key challenge of today's data centers: a power consumption analysis of TPC-C results," *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1229-1240, 2008.
- [8] A. Horri, M. S. Mozafari, and G. Dastghaibiyfard, "Novel resource allocation algorithms to performance and energy efficiency in cloud computing," *The Journal of Supercomputing*, vol. 69, no. 3, pp. 1445-1461, 2014.
- [9] S. Esfandiarpour, A. Pahlavan, and M. Goudarzi, "Structure-aware online virtual machine consolidation for datacenter energy improvement in cloud computing," *Computers & Electrical Engineering*, vol. 42, pp. 74-89, 2015.
- [10] A. Beloglazov, and R. Buyya, "Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 7, pp. 1366-1379, 2013.
- [11] A. Beloglazov, and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397-1420, 2012.
- [12] S. B. Shaw, and A. K. Singh, "Use of proactive and reactive hotspot detection technique to reduce the number of virtual machine migration and energy consumption in cloud data center," *Computers & Electrical Engineering*, vol. 47, pp. 241-254, 2015.
- [13] X. Fu, and C. Zhou, "Virtual machine selection and placement for dynamic consolidation in Cloud computing environment," *Frontiers of Computer Science*, vol. 9, no. 2, pp. 322-330, 2015.
- [14] G. Wu, M. Tang, Y.-C. Tian, and W. Li, "Energy-Efficient Virtual Machine Placement in Data Centers by Genetic Algorithm," vol. 7665, pp. 315-323, 2012.
- [15] J. Xu, and J. A. Fortes, "Multi-Objective Virtual Machine Placement in Virtualized Data Center Environments." pp. 179-188.
- [16] M. Tang, and S. Pan, "A Hybrid Genetic Algorithm for the Energy-Efficient Virtual Machine Placement Problem in Data Centers," *Neural Processing Letters*, vol. 41, no. 2, pp. 211-221, 2014.
- [17] C. T. Joseph, K. Chandrasekaran, and R. Cyriac, "A Novel Family Genetic Approach for Virtual Machine Allocation," *Procedia Computer Science*, vol. 46, pp. 558-565, 2015.
- [18] F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, and H. Tenhunen, "Utilization Prediction Aware VM Consolidation Approach for Green Cloud Computing." pp. 381-388, 2010.
- [19] F. Farahnakian, P. Liljeberg, and J. Plosila, "LiRCUP: Linear Regression Based CPU Usage Prediction Algorithm for Live Migration of Virtual Machines in Data Centers." pp. 357-364, 2013.
- [20] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230-1242, 2013.
- [21] S. B. Shaw, and A. Singh, "A survey on scheduling and load balancing techniques in cloud computing environment." pp. 87-95, 2014.
- [22] M. Natrella, "NIST/SEMATECH e-handbook of statistical methods", 2010.
- [23] [C. Chatfield, *The analysis of time series*. Boca Raton, FL: Chapman & Hall/CRC, 2016.
- [24] C. Chatfield, *Time-series forecasting*. Boca Raton: Chapman & Hall/CRC, 2001.
- [25] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities." pp. 1-11, 2009.
- [26] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, vol. 41, pp. 23-50, 2011.
- [27] R. N. Calheiros, R. Ranjan, C. A. De Rose, and R. Buyya, "CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services," *arXiv preprint arXiv:0903.2525*, 2009.
- [28] K. Park, and V. S. Pai, "CoMon: a mostly-scalable monitoring system for PlanetLab," *SIGOPS Oper. Syst. Rev.*, vol. 40, no. 1, pp. 65-74, 2006.

Maneuverability of an Inverted Pendulum Vehicle According to the Handle Operation Methods

Chihiro NAKAGAWA, Takuya CHIKAYAMA, Akikazu OKAMOTO, Atsuhiko SHINTANI and Tomohiro ITO
Graduate School of Engineering, Osaka Prefecture University
1-1, Gakuen, Naka, Sakai, Osaka, 599-8531, Japan

Abstract—This study investigated what handle operation and turning gain is comfortable for people using an inverted pendulum vehicle that is changeable the handle operation. Experimental conditions were three conditions. First is a slalom course with two cones placed at an interval of 1.8 m. Second is a slalom course with five cones placed at an interval of 1.4 m. Third is a slalom course with six cones placed at 1.8m, 1.4m, 1.8m, 1.4m, 1.8m, and 1.8m interval. The first condition considered the difference of handle operation between subjects who were used to ride and not used to ride. The second condition considered the difference of maneuverability due to gains. The third condition considered the difference of maneuverability between two handle operations in real running space in a condition of 10 gains. In a result of the first condition, a subject who was used to ride run effectively and running time is short compared with a subject who was used to ride. However, in handle yaw rotation, the difference of maneuverability was small. In a result of the second condition, running mileage about the same in two handle operation, but running time of handle yaw rotation is shorter than that of handle roll rotation. In a result of the third condition, like the second condition, running time of handle yaw rotation is shorter than that of handle roll rotation. In questionnaire evaluation, the best gain is the lower gain, 0.02. At last, An experiment was carried out by 14 subjects in the best gain, 0.02 that is best both handle operation. In the result of this experiment, 12 subjects answered that handle yaw rotation is better than handle roll rotation.

Keywords—personal mobility vehicle; inverted pendulum vehicle; maneuverability; handle operation; number of operations; questionnaire evaluation

I. INTRODUCTION

Environmental and energy problems are of increasing interest, and alternative means of transportation have been considered to solve this problem. Daily transportation involves walking, bicycles, and cars depending on the time and situation. Recently, personal mobility vehicles (PMVs)—which are environmentally friendly and useful for short- to mid-range transportation—have attracted some attention.

Some examples of PMVs include inverted pendulum vehicles such as the Segway and the Winglet. In addition to their usefulness over short- to mid-range distances, inverted pendulum vehicles are either human-powered or electrically powered, which emit no exhaust gas and are thus environmentally friendly. For this reason, the use of such vehicles has been spreading for security and recreational activities. For ensuring their convenient use, verification

experiments have been made for the implementation of PMVs as efficient and user-friendly transportation systems.

The studies of PMVs include those focused on the improvements and usage of common PMVs, as well as those on developing new PMVs such as high-performance wheel chairs and amphibious bicycles and tricycles. Studies on inverted pendulum vehicles include those focusing on the stability and relationships between the vehicle and the controller and between the vehicle and pedestrians. Controllability studies on PMVs involve the investigation of the stable driving of rides using bicycle steering, model analysis, and braking. However, few studies have focused on the vehicle's handle operation and controllability.

This study focused on two types of operations:

1) Turns by holding down the handle axis (handle roll rotation) and

2) Turns by rotating the handle (handle yaw rotation).

The purpose of this study is to experimentally compare the controllability of the two types of handle operation.

II. TEST VEHICLE

The handle of the test vehicle used in this study was replaceable. The handle of the test vehicle could be either of the handle roll rotation (Fig. 1) or the handle yaw rotation (Fig. 2) variety. The maximum rotation of the handle roll rotation was 20° to the left and the right, and that of the handle yaw rotation was 30° to the left and the right. The vehicle equipped with a Bluetooth communication function for measuring the number of wheel rotations and handling angle. The data sampling period for the measurements was 0.1 s.

III. CONTROLLABILITY EXPERIMENT

The experiments were carried out under the three conditions listed below.



Fig. 1. Handle roll rotation



Fig. 2. Handle yaw rotation

Condition 1: Using both experienced and inexperienced drivers for the comparison

Condition 2: Selection of steering gain

Condition 3: Comparison of the controllability of the handle roll rotation and the handle yaw rotation

The condition 1 studied the behaviors of experienced drivers (who had driven 10 times or more) and completely inexperienced drivers. The condition 2 selected the most appropriate steering gain for each handle operation. The condition 3 investigated which handle operation had better controllability.

A. Difference Between The Experienced And Inexperienced Drivers

This experiment was carried out under the conditions shown in Table 1. The drivers were six adult males, five of whom were experienced and one of whom was inexperienced. The test course was a slalom with two cones with a distance of 1.8 m between them, as shown in Fig. 3. The driving trajectory was measured using a three-dimensional operation analysis device (VICON). Each driver rode around the slalom for three cycles with each of the two handle operations and three different gains, i.e., large (0.02), medium (0.012), and small (0.008), making a total of six patterns. After the test rides, the drivers answered a five-level controllability evaluation survey.

TABLE I. EXPERIMENTAL CONDITION 1

Subjects	6 (5 experienced, 1 inexperienced)
Course	Two cones with a distance of 1.8 m
Gain	0.02, 0.012, and 0.008
Cycles	3
Order	0.02, 0.012, and 0.008, each with 3 cycles

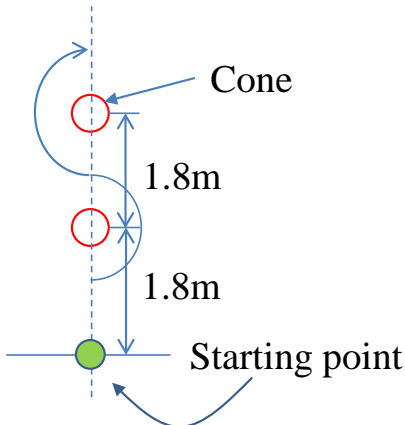


Fig. 3. Experimental Course

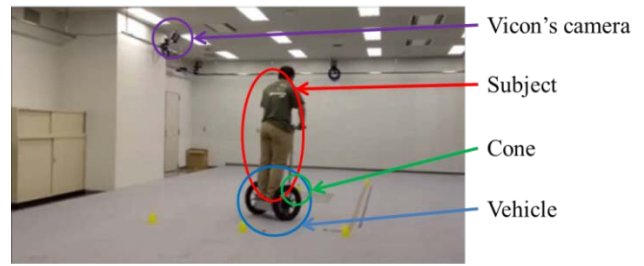


Fig. 4. Experimental Course

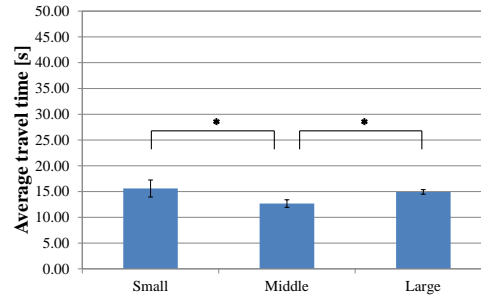


Fig. 5. Average travel time (roll, experienced driver)

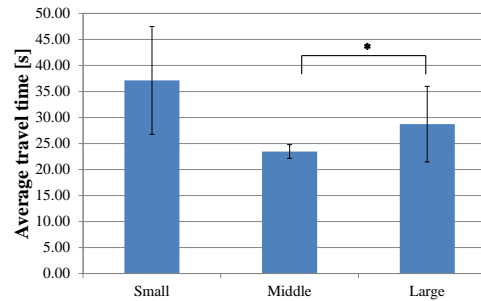


Fig. 6. Average travel time (roll, inexperienced driver)

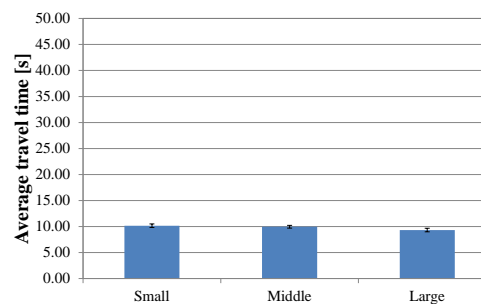


Fig. 7. Average travel time (yaw, experienced driver)

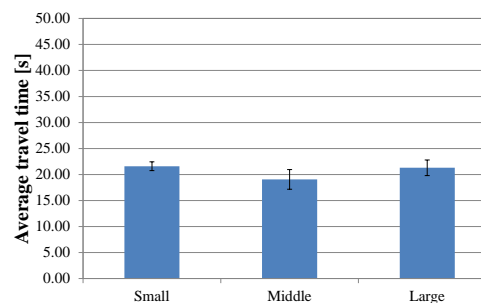


Fig. 8. Average travel time (yaw, inexperienced driver)

Fig. 5 and Fig. 6 show the ride times of the experienced and inexperienced drivers, respectively, with handle roll rotation. Fig. 7 and Fig. 8 show these ride times with handle yaw rotation. Comparing the rides of the experienced and inexperienced drivers revealed that the both types of the driver had smaller handle angles with increasing steering gain. The inexperienced driver's ride time was significantly longer than those of the experienced drivers in both handle roll rotation and handle yaw rotation. The difference in the ride time between the experienced and inexperienced drivers was smaller for the handle yaw rotation than that for the handle roll rotation.

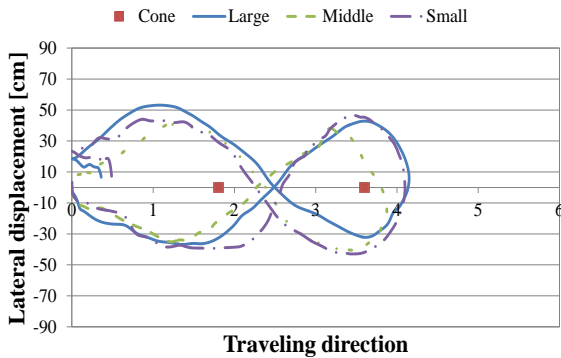


Fig. 9. Trajectory of the experienced driver (handle roll rotation)

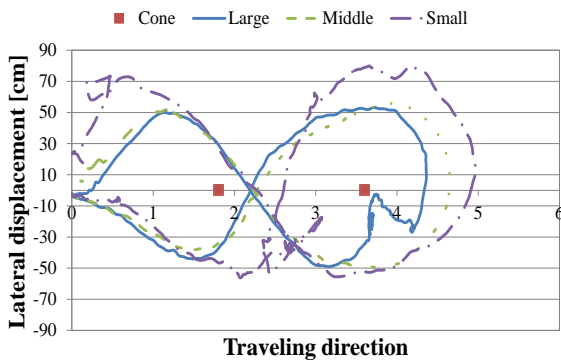


Fig. 10. Trajectory of the inexperienced driver (handle roll rotation)

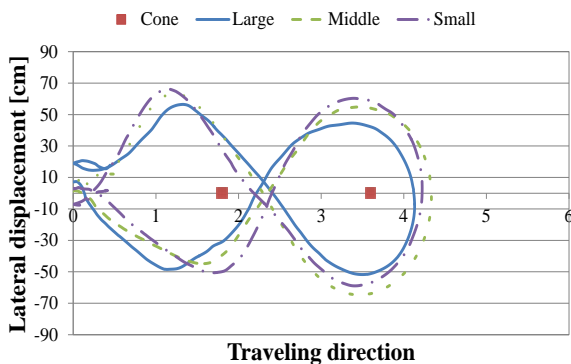


Fig. 11. Trajectory of the experienced driver (handle yaw rotation)

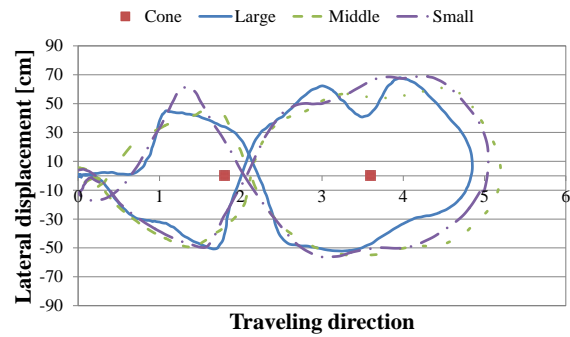


Fig. 12. Trajectory of the inexperienced driver (handle yaw rotation)

Fig. 9 and Fig. 10 are the drive trajectories of the experienced and inexperienced drivers, respectively, using handle roll rotation; similarly, Figs. 11 and Fig. 12 are those with handle yaw rotation. With both handle operations, the inexperienced driver had trajectories that were compared with those of the experienced drivers. Moreover, Fig. 13 and Fig. 14 show that the inexperienced driver cut the wheel more than the experienced drivers. Here, “cutting the wheel” means the change of sign of the handle angular velocity.

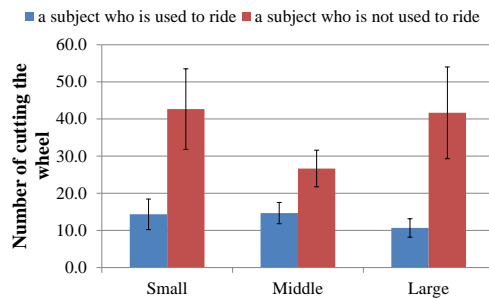


Fig. 13. Amount of cutting of the wheel (handle roll rotation)

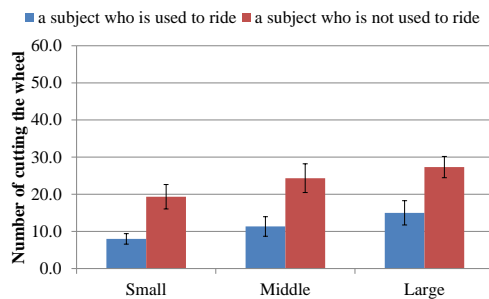


Fig. 14. Amount of cutting of the wheel (handle yaw rotation)

B. Selection of Steering Gain

This experiment was conducted under the conditions shown in Table 2. Five male adult subjects drove on a slalom course with combinations of large (1.8 m) and small (1.4 m) distances between the cones. Each subject made two cycles around the course for each of the five steering gains; 0.012, 0.02, 0.03, 0.04, and 0.05. This study included gains larger than 0.05;

0.06, 0.07, 0.08, 0.09, and 0.1 in the survey. Considering the order effect, the steering gains were measured in the order 0.012, 0.03, 0.02, 0.04, and 0.05. For the survey, the gains were placed in the order 0.06, 0.08, 0.07, 0.1, and 0.09 in addition to the former gain order. The subjects answered an evaluation survey after the test drives.

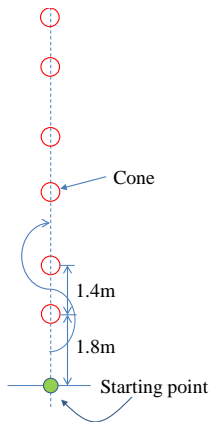


Fig. 15. Experimental course

TABLE II. EXPERIMENTAL CONDITION 2

Subjects	5 (All experienced)
Course	At an interval combination of 1.8 m and 1.4 m
Gain	Five gains for data collection: 0.012, 0.02, 0.03, 0.04 and 0.05.
	Ten gains for survey: 0.06, 0.07, 0.08, 0.09, and 0.1 in addition to the above five gains
Cycles	2
Orders	Data collection: 0.012, 0.03, 0.02, 0.04, and 0.05, each two cycles
	For survey: 0.06, 0.08, 0.07, 0.1, and 0.09, each 2 cycles

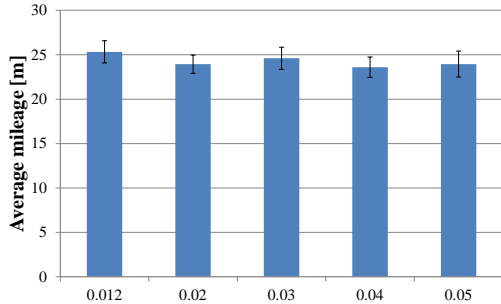


Fig. 16. Average mileage (handle roll rotation)

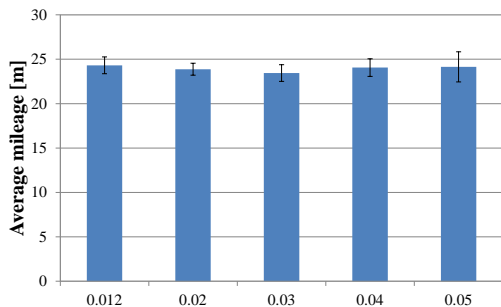


Fig. 17. Average mileage (handle yaw rotation)

Fig. 16 and Fig. 17 show the distances traveled; Fig. 18 and Fig. 19 show the ride times; Fig. 20 and Fig. 21 show the numbers of cuttings of the handle, in other words, changes in

the turning direction of the handle. Similar to the section 3.1, the handling angles become smaller as the steering gain increase for both handle operations. There was almost no difference in the distance traveled between the two handle operations. On the other hand, the ride time with handle yaw rotation was approximately 6 s shorter than that with handle roll rotation. Moreover, the amount of cutting of the handle was less for the handle yaw rotation for all the steering gains. These comparison implied a higher efficiency of the handle yaw rotation.

The survey results were evaluated from two aspects: the handling performance, which pertained to the controllability, and riding performance, which was related to the balance of the ride. Fig. 22 and Fig. 23 show the results of the handling performance of the handle roll rotation and handle yaw rotation, respectively. Fig. 24 and Fig. 25 are the results of the controlling performance. The vertical axis shows the rating with 5 being the most and 1 the least desirable riding conditions. The error bars show the standard deviations. In both surveys, the steering gain of 0.02 had the highest ratings.

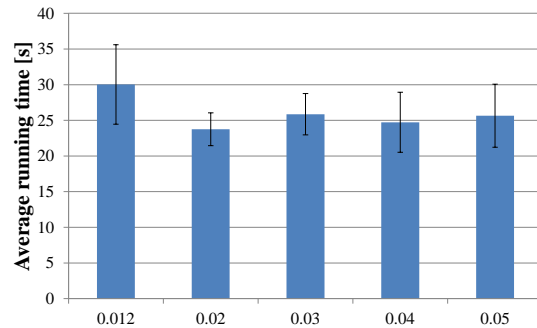


Fig. 18. Average running time (handle roll rotation)

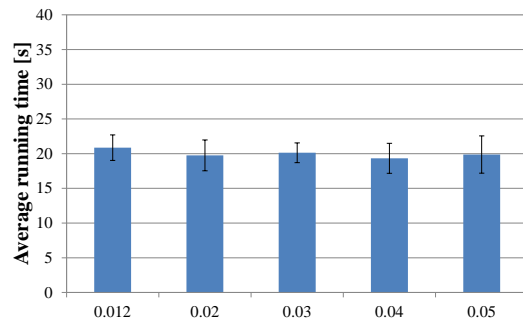


Fig. 19. Average running time (handle yaw rotation)

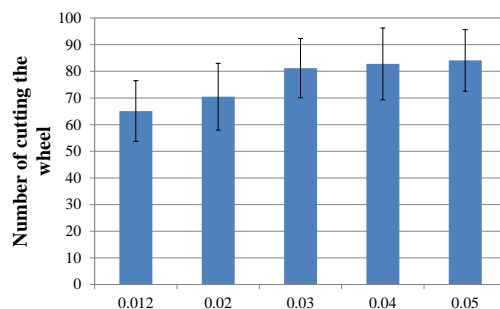


Fig. 20. Amount of cutting of the wheel (handle roll rotation)

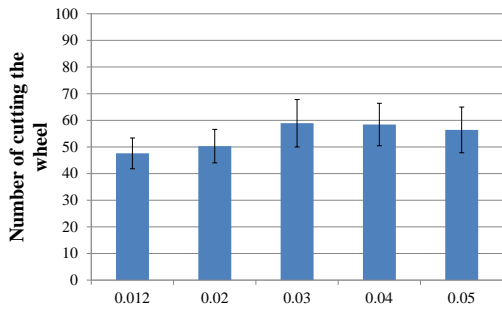


Fig. 21. Amount of cutting of the wheel (handle yaw rotation)

C. Comparison Of The Controllability Of Handle Roll Rotation And Handle Yaw Rotation

In this section, it is examined which handle operation—handle roll or handle yaw—was better based on the experiment and survey results. By using the gain of the highest rating (0.02 for both handle roll rotation and handle yaw rotation) as shown in section B, answers to a survey were obtained from 14 subjects. The test course was a slalom with three cones with an equal distance of 1.8 m. The subjects were adult males and females, and none were experienced in driving these vehicles. The method of answering the survey was the same five-level rating as in the previous section. Fig. 26 and Fig. 27 show the five-level survey results for the respective handle operations. Of the 14 subjects, two answered that the handle roll rotation had better controllability than did the handle yaw rotation, and 12 answered the other way around. As a result of a sign test, there was a significant difference between the two handle operations, as shown in Fig. 28, which indicates that the handle yaw rotation has better controllability. Moreover, there was no subject who indicated that the handle yaw rotation was difficult to ride. The above results thus suggest that handle yaw rotation has better controllability.

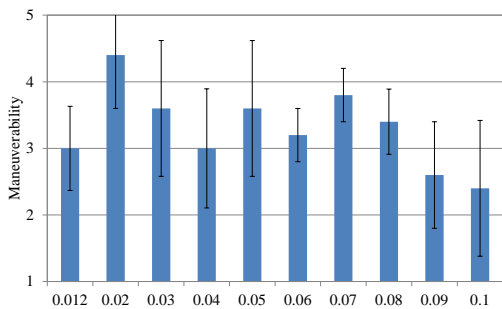


Fig. 22. Questionnaire result of handle roll rotation (handle operation)

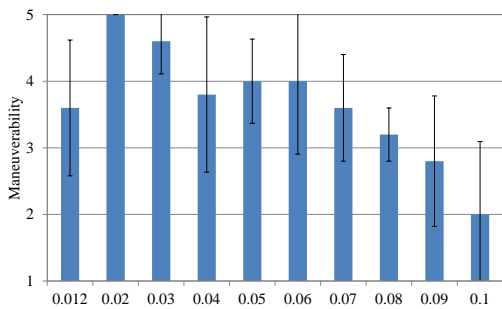


Fig. 23. Questionnaire result of handle yaw rotation (handle operation)

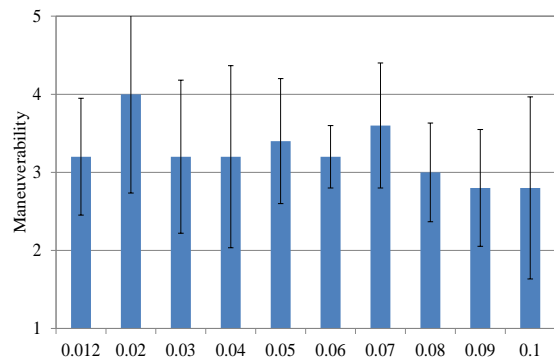


Fig. 24. Questionnaire result of handle roll rotation (running operation)

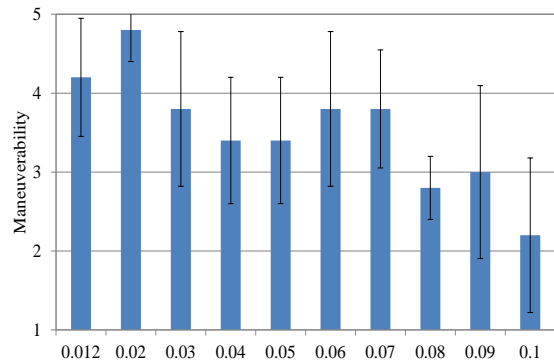


Fig. 25. Questionnaire result of handle yaw rotation (running operation)

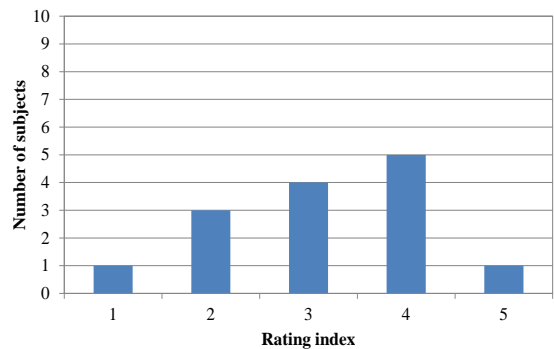


Fig. 26. Questionnaire result of handle yaw rotation

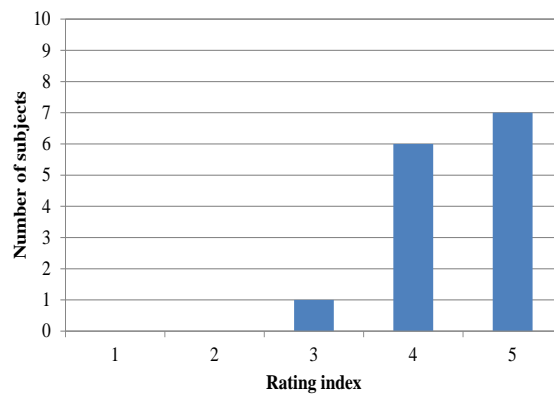


Fig. 27. Questionnaire result of handle roll rotation

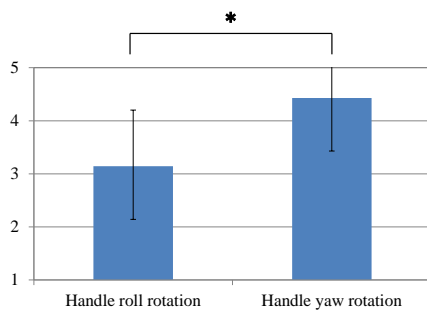


Fig. 28. Comparison of evaluation

IV. DISCUSSIONS

This section discusses the characteristic differences between the handle roll rotation and the handle yaw rotation and these effect. In handle roll rotation, the driver's body tends to tilt as the handle is held down from left to right and vice versa. Consequently, the ride easily goes off course on a slalom as the driver's body tilts frequently. In handle yaw rotation, on the other hand, the handle axis is fixed, and the steering is performed by rotating the handle in the direction of the yaw. Unlike handle roll rotation, this method allows steering in the upright position, leading to stability without tilting the driver's body to the right or the left, even during sharp turns. The unfavorable results for handle roll rotation obtained in section C were because inexperienced drivers were likely to feel instability while controlling the vehicle with their bodies tilted.



Fig. 29. State of handle roll rotation



Fig. 30. State of handle yaw rotation

V. CONCLUSION

As a result of this study, it is discovered that both experienced and inexperienced drivers found the handle yaw rotation easier to control. This is likely because handle roll rotation involves tilting one's body while handle yaw rotation does not. As the driver experiences centrifugal force while steering, the force is larger for handle roll rotation than for handle yaw rotation. Thus, it is likely that the inexperienced drivers felt instability upon steering, leading to a higher rating for handle yaw rotation than that for handle roll rotation.

REFERENCES

- [1] Segway <http://www.segway-japan.net/index.html> (accessed on January 26, 2015)
- [2] Winglet http://www.toyota.co.jp/jpn/tech/personal_mobility/winglet.html (accessed on January 26, 2015)
- [3] Segway Implementation results available from <http://www.segway-japan.net/casestudy/police/> (accessed on January 26, 2015)
- [4] Robot special district validation and experiment promotion council available from <http://council.rt-tsukuba.jp/> (accessed on January 26, 2015)
- [5] Start of road test of Winglet, Toyota's standing ride robot, available from <http://ascii.jp/elem/000/000/811/811044/> (accessed on January 26, 2015)
- [6] Miura, M. and Takahashi, Y., Proposal for Personal Mobility Vehicle for People with Limited Mobility, the Japan Society of Mechanical Engineers, the Transportation and Logistics Conference 2010(19), 2010, pp. 299–302.
- [7] Mizokami, S., Kawashima, H and Yaguchi, T., A Study of Potential for QOL Improvements through the Use of Personal Mobility in an 'Aged Society', Internal Association of Traffic and Safety Science, IATSS review 36(3), 2012, pp. 44–51.
- [8] Ono, H., Yoshiko, K., Ohsaki, H., Hirahara, H., Watanabe, K., Iwase, M. and Hatakeyama, S., Design of structure and control for bicycle operation assist system, Proceedings of the Japan Joint Automatic Control Conference 50(0), 2007, pp. 194–194.
- [9] Satoh, Y. and Sakaue, K., A secure and reliable next generation mobility: an intelligent electric wheelchair with a stereo omnidirectional camera system, National Institute of Advanced Industrial Science and Technology, Synthesiology 2(2), 2009, pp. 113–126.
- [10] Toda, M., Design and production of amphibious bicycle of a removable system, Polytechnic College Kinki Proceedings, (18), 2014, pp. 6–10.
- [11] Akita, H., The Structure and its Development of an Electric Power Assisted Cycle, Japanese Society of Automotive Engineers, JSAE, 63(9), 2009, pp. 50–54.
- [12] Ishikawa, T., Takafumi, F., Terauchi, F., Kubo, M. and Aoki, H., A Front-Wheel-Drive Tricycle by using the Elastic Behavior of the Rear Frames, Proceedings of the Annual Conference of JSSD (51), 2004, pp. 70–71.
- [13] Nakano, K., Nakamori, D., Zheng, R., Ohori, M. and Suda, Y., Stability Analysis on a Two-Wheeled Inverted Pendulum Type Personal Mobility Vehicle Considering Human Motion, The Japan Society of Mechanical Engineers, TRANSACTIONS OF THE JAPAN SOCIETY OF MECHANICAL ENGINEERS Series C 79(801), 2013, pp. 1427–1440.
- [14] Nakamori, D., Nakano, K., Ohori, M. and Suda, Y., Relations Between Stability and Ride Comfort of the Two-wheeled Inverted Pendulum Vehicle, The Japan Society of Mechanical Engineers, The Transportation and Logistics Conference 2011, pp. 163–166.
- [15] Suda, Y., Hirayama, Y., Yamaguchi, D. and Aki, M., Communication with Personal Mobility Vehicle and Driver, SEISAN KENKYU 64(2), 2012, pp. 71–74.
- [16] Arakawa, S., Nakagawa, C., Shintani, A. and Ito, T., Basic Study on the Behavior of Inverted Pendulum Vehicle and Driver Using Multibody Dynamics, The Japan Society of Mechanical Engineers, TRANSACTIONS OF THE JAPAN SOCIETY OF MECHANICAL ENGINEERS Series C 78(789), 2012, pp. 1497–1506.

- [17] Nakagawa, C., Imamura, K., Shintani, A. and Ito, T., Experimental Study on the Influence of the Size of Personal Mobility Vehicle on Pedestrians, the Japan Society of Mechanical Engineers, TRANSACTIONS OF THE JAPAN SOCIETY OF MECHANICAL ENGINEERS Series C 78(794), 2012, pp. 3332–3342.
- [18] Maeda, K. and Yamada, M., Stable running by the steering of two-wheel self-propelled bicycle, The Japan Joint Automatic Control Conference, Proceedings of the Japan Joint Automatic Control Conference 49(0), 2006, pp. 175–175.
- [19] Katayama, T., Kinetic model for stability and safety analyses of bicycles, Japan Bicycle Promotion Institute, Bicycle Technology Information (77), 2000–02, pp. 1–16.
- [20] Matsui, H., et al., Bicycle braking by college students, adults and seniors, Japan Bicycle Promotion Institute, Bicycle Technology Information (20), 1983–2007, pp. 25–30.
- [21] Oguro, H. and Raksincharoensak, P., Study on Driving Characteristics of Two-Wheel Inverted Pendulum Electric Vehicle in Pedestrian Walking Space, The Japan Society of Mechanical Engineers, Robotics/mechatronics conference proceedings, 2011, “2A1-Q13 (1)”–“2A1-Q13 (2).”
- [22] Suda, Y., Hirayama, Y. and Takagi, M. A. T., Yaw Control Experiment of Tricycle-type Personal Mobility Vehicle: Maneuvering Experiment by Different Rotation Speed Control in Right and Left Wheel. The Japan Society of Mechanical Engineers, Dynamics & Design Conference 2011, 2011, “148-1”–“148-8.”
- [23] Nakagawa, C., Suda, Y., Nakano, K. and Hirayama, Y., Maneuvering Experiment of Parallel Two-Wheeled Personal Mobility Vehicle with Human Pedaling, Institute of Industrial Science The University of Tokyo, SEISAN KENKYU 62(1), 2010, pp. 119–122.
- [24] Horiuchi, E., Matsumoto, O., Takei, T., Koyachi, N., Hashimoto, T., Ando, T. and Iwao, K., A Steering Assist Method for Formation Control of Personal Vehicles, The Robotics Society of Japan, JRSJ 28(10), 2010, pp. 1243–1250.
- [25] Nakagawa, C., Imamura, K., Shintani, A. and Ito, T., Experimental Study on the Influence of the Size of Personal Mobility Vehicle on Pedestrians, the Japan Society of Mechanical Engineers, TRANSACTIONS OF THE JAPAN SOCIETY OF MECHANICAL ENGINEERS Series C 78(794), 2012, pp. 3332–3342.
- [26] Maeda, K. and Yamada, M., Stable running by the steering of two-wheel self-propelled bicycle, The Japan Joint Automatic Control Conference, Proceedings of the Japan Joint Automatic Control Conference 49(0), 2006, pp. 175–175.
- [27] Katayama, T., Kinetic model for stability and safety analyses of bicycles, Japan Bicycle Promotion Institute, Bicycle Technology Information (77), 2000–02, pp. 1–16.
- [28] Matsui, H., et al., Bicycle braking by college students, adults and seniors, Japan Bicycle Promotion Institute, Bicycle Technology Information (20), 1983–07, pp. 25–30.
- [29] Oguro, H. and Raksincharoensak, P., Study on Driving Characteristics of Two-Wheel Inverted Pendulum Electric Vehicle in Pedestrian Walking Space, The Japan Society of Mechanical Engineers, Robotics/mechatronics conference proceedings, 2011, “2A1-Q13 (1)”–“2A1-Q13 (2).”
- [30] Suda, Y., Hirayama, Y. and Takagi, M. A. T., Yaw Control Experiment of Tricycle-type Personal Mobility Vehicle: Maneuvering Experiment by Different Rotation Speed Control in Right and Left Wheel. The Japan Society of Mechanical Engineers, Dynamics & Design Conference 2011, 2011, “148-1”–“148-8.”
- [31] Nakagawa, C., Suda, Y., Nakano, K. and Hirayama, Y., Maneuvering Experiment of Parallel Two-Wheeled Personal Mobility Vehicle with Human Pedaling, Institute of Industrial Science The University of Tokyo, SEISAN KENKYU 62(1), 2010, pp. 119–122.
- [32] Horiuchi, E., Matsumoto, O., Takei, T., Koyachi, N., Hashimoto, T., Ando, T. and Iwao, K., A Steering Assist Method for Formation Control of Personal Vehicles, The Robotics Society of Japan.

Gaussian Mixture Model and Deep Neural Network based Vehicle Detection and Classification

S Sri Harsha

Assistant Professor, Department of IT,
VR Siddhartha Engineering College, Vijayawada
Andhra Pradesh, India

K. R. Anne

Professor, Director Academics,
Veltech University, Chennai,
Tamil Nadu, India

Abstract—The exponential rise in the demand of vision based traffic surveillance systems have motivated academia-industries to develop optimal vehicle detection and classification scheme. In this paper, an adaptive learning rate based Gaussian mixture model (GMM) algorithm has been developed for background subtraction of multilane traffic data. Here, vehicle rear information and road dash-markings have been used for vehicle detection. Performing background subtraction, connected component analysis has been applied to retrieve vehicle region. A multilayered AlexNet deep neural network (DNN) has been applied to extract higher layer features. Furthermore, scale invariant feature transform (SIFT) based vehicle feature extraction has been performed. The extracted 4096-dimensional features have been processed for dimensional reduction using principle component analysis (PCA) and linear discriminant analysis (LDA). The features have been mapped for SVM-based classification. The classification results have exhibited that AlexNet-FC6 features with LDA give the accuracy of 97.80%, followed by AlexNet-FC6 with PCA (96.75%). AlexNet-FC7 feature with LDA and PCA algorithms has exhibited classification accuracy of 91.40% and 96.30%, respectively. On the contrary, SIFT features with LDA algorithm has exhibited 96.46% classification accuracy. The results revealed that enhanced GMM with AlexNet DNN at FC6 and FC7 can be significant for optimal vehicle detection and classification.

Keywords—Vehicle detection and classification; deep neural network; AlexNet; SIFT; Gaussian Mixture Model; LDA

I. INTRODUCTION

The high pace development of technologies predominantly image or video processing techniques have enabled a number of application scenarios. Visual traffic surveillance (VTS) based intelligent transport system (ITS) is one of the most sought and attractive application and research domains, which has attracted academia-industries to enable better efficiency. The significant application prospects of ITS systems have motivated researchers to achieve a certain effective solution. An efficient vehicle detection and localization scheme can enable ITS to make efficient surveillance, monitoring and control by incorporating semantic results, like “X-Vehicle crossed Y location in Z direction and overtaking A Vehicle with Speed B”. Considering these needs, in previous works [1,2], vehicle detection, tracking, and speed estimation model were developed. However, the further optimization could enable more efficient ITS solution. Developing a novel and robust system to detect and classify the vehicle simultaneously can be of paramount significance. Vehicle detection features such as size, shape, color, stopped or moving object and their

type can be vital for ITS decision systems [3]. The type of the detected vehicle can provide significant information that may lead ITS administrators to ensure that certain type of vehicle doesn't appear in a certain region. Implementation of the multi-camera infrastructures [4] might enable ITSs to identify and detect the targeted vehicle or vehicle class by matching it from traffic data from different functional data acquisition cameras.

Recently, some efforts have been made for vehicle detection and tracking, however, very few efforts have been made towards its classification. Especially, not much effort has been made on developing a simultaneous vehicle detection and classification system. There are numerous issues like cluttered image scene, occlusion, the exceptionally higher number of classes and features, etc. that make classification highly intricate. Background segmentation based object detection can be beneficial as it can remove clutter [5]. The images retrieved through surveillance cameras are used to be of low resolution, different lighting conditions, and more importantly, size of the vehicle is very small in complete traffic video frame that makes classification too tedious task. In practice, vision-based surveillance applications require dealing with huge unlabelled data elements, features, occlusion, unannotated images, localization and classification under different lighting or background conditions, etc. To deal with such issues, a system with effective background subtraction, feature extraction and vehicle region or ROI localization and classification can be of paramount significance. Also, to provide time effective solution reduced data processing and computationally efficient approach is required. Considering such requirements and motivations, in this paper, a multilevel optimization measure has been proposed. In this paper, an enhanced Gaussian Mixture Model (GMM) algorithm and connected component analysis (CCA) scheme has been developed for optimal vehicle region or ROI identification and localization. Further, to enable accurate and swift vehicle classification an enhanced multilayered deep convolutional neural network (DNN) was developed that functions on AlexNet DNN model. An additional feature extraction model, space invariant feature transform (SIFT) were prepared to extract ROI features. Implementing dimensional reduction schemes over extracted features support vector machine (SVM) based classification was performed that classifies vehicles into different classes.

The other sections are divided as follows: Section II presents the related work. Section III discusses the proposed research or contribution. In Section IV, the algorithmic development and its discussion are presented and the results

obtained are given in Section V. Section 6 presents conclusion and future work. References used are presented at the last of manuscripts.

II. RELATED WORK

Recently a vision based model for vehicle detection, feature extraction and classification were developed in [6], where researchers applied GMM with Hole Filling algorithm for vehicle detection, Gabor kernel based feature extraction and Multi-Class classification. To deal with dense vehicle classification, a vector sparse coding scheme with SVM was proposed in [7]. Applying sparse coding technique, they projected features to the high dimensional vector that assisted SVM to perform better classification. The combined shape and gradient feature based classification was proposed in [8][9]. To perform shape-based classification, at first they performed background subtractions and obtained shape features from silhouettes in the omnidirectional video frames. Similarly, for gradient based classified Histogram of Oriented Gradients (HOG) Features were obtained, where researchers found that the combined features based classification can be more useful than the individual feature based classification. The features like geometry, number plate location and shape was used as input of dynamic Bayesian network (DBN) for vehicle classification [10]. Researchers applied GMM to calculate the probability distribution of features. However, they could not address the detection issues under varying illuminations and frame dynamicity. A sparse learning based vehicle detection and classification model were proposed in [9]. Later, in [11] sparse coding and spatial pyramid matching scheme were used for vehicle classification, where they extracted the patch based sparse features using a discriminate dictionary. The extracted features were classified using histogram intersection kernel based SVM classifier. An integrated vehicle detection and classification model was proposed in [12] where multi-resolution vehicle recognition (MRVR) scheme was introduced to support cascade boosted classifiers for vehicle classification. The combined feature including HAAR and HOG was used for vehicle detection and classification [13]. The concept of multi-feature fusion was proposed in [14], where authors combined local as well as global feature of the detected vehicle region or ROI. In their work, they applied SIFT for local feature extraction and PCA based global feature extraction process. The combined features were used for classification using SVM [14]. To increase accuracy, researchers [15][16] used higher layer features of the deep neural network (DNN). Researchers [15] extracted PHOG and LBP-EOH using DNN. They combined these features for classification. An appearance based vehicle classification scheme has been developed in [17], where vehicle front features has been applied for classification using semi-supervised CNN algorithm. On the contrary, in this paper, the rear information and lane dash line information have been applied to perform multi-lane vehicle detection. Also, it deals with occlusion issues. A shape-based multi-class classification scheme has been proposed in [18] where the concavity property of vehicles such as buses and sedans was used for classification. Authors in [19] applied a Deep Belief Networks (DBN) based vehicle classification. They have used key features such as image pixel value, HOG features and Eigen features to perform classification. An approach named

cascade classifier ensemble has been suggested in [19] for vehicle classification. As the first ensemble, they applied classifiers such as SVM, K-NN, random forest and multiple-layer perceptrons (MLPs) for vehicle classification.

Recently, real-time vision-based vehicle detection and the classification system were proposed in [20], where a simple morphology-based approach has been formulated for ROI detection. To deal with vehicle occlusion issues, they applied the ROI accumulative curve method and Fuzzy Constraints Satisfaction Propagation (FCSP). Retrieving the Time-Spatial Images (TSI) from the surveillance video, they eliminated shadowed region using SVM and Deterministic Non-Model Scheme (DNMS). A combined model to perform vehicle detection, tracking, classification, counting has been proposed in [21]. In [16], researchers applied conventional median filter and Otsu method based background subtraction for vehicle detection. However, they could not address the problems introduced due to illumination change and background features variations. To deal with these issues, GMM scheme can be a potential alternative for background subtraction [6][10], however, traditional GMM scheme remains questionable especially with dynamic frame movement and varying illumination conditions because of its fixed learning rate and pixel saturation issues. To deal with this in this paper, an adaptive learning rate based GMM model has been developed for vehicle ROI detection. On the other hand, the direct deep neural network (DNN) implementation for vehicle detection and classification is highly intricate and almost impractical. Therefore, in this paper an enhanced AlexNet DNN with CaffeNet model [22] has been developed that enables optimal vehicle detection and classification, even with huge dataset. Considering the effectiveness of the SVM classifier, in this paper, 10-fold cross-validation scheme has been applied to achieve accurate classification performance.

III. CONTRIBUTION

In this paper, a robust vehicle detection and classification system has been developed for vision-based surveillance system to be used for ITS purposes. In fact, the presented work is a multilevel optimization effort where numerous optimization efforts have been introduced on a different phase of vehicle detection and classifying. The proposed approach includes enhanced GMM (adaptive learning rate) based background subtraction and vehicle detection, CCA based ROI identification or localization, DNN model; AlexNet and CaffeNet based feature extraction, dimensional reduction and SVM based efficient vehicle classification. To perform vehicle localization in image and occlusion avoidance, the vehicle's rear features along with lane dash markings have been applied. Once performing background subtraction, to reduce irrelevant blob presence, CCA has been applied that eventually achieves precise vehicle region or ROI. To extract ROI features, an enhanced DNN algorithm has been applied based on convolutional neural network (CNN) principle. Here, AlexNet DNN model [23] extracts multidimensional features at the higher DNN layers (Fig. 3). In existing works [23], DNN has been used for vehicle classification using different datasets [24]. However, AlexNet can't be applied directly as in practical situations the labeled data used to be smaller than the DNN parameters. In generic DNN based approaches the probability

of degraded accuracy and over-fitting can't be ignored. To deal with this issue, in this paper, CaffeNet [22] with AlexNet DNN has been used that enables optimal performance even with general purpose computing systems. In practice, due to higher unannotated data, performing DNN learning and classification is a tedious task. To deal with such issues, multilayered DNN has been implemented and trained over large scale labeled vehicle dataset that enables swift and accurate data classification. In this work, the ROI features have been retrieved at each layer of the trained DNN (Convolutional Layer-1 to Layer-5 and Fully Connected Layer 6 and Layer 7). Since, features at the higher layers (fully connected 6, 7 and 8) of DNN used to be more informative [16] and therefore a set of 4096-dimensional features have been retrieved for individual vehicle image at FC6 and FC7 (Fig. 3). Recently, researchers [25] suggested that SIFT features can also enable accurate classification; therefore in this paper, 4096 SIFT feature descriptors have been obtained from each image, which is equivalent to AlexNet FC6 and FC7 features. The extracted features have been projected to the dimensional reduction schemes, the principle component analysis (PCA) and linear discriminant analysis (LDA). After dimensional reduction with PCA and LDA individually, the retrieved AlexNet features have been projected to the polynomial kernel based SVM classifier for vehicle classification. Similarly, SIFT feature vectors have been used as input of SVM for classification. The detailed discussion of the proposed vehicle detection and classification system is presented in the following sections.

IV. SYSTEM MODEL

This section discusses the overall development and implementation of the proposed enhanced GMM and DNN based vehicle detection and classification system (Fig. 1).

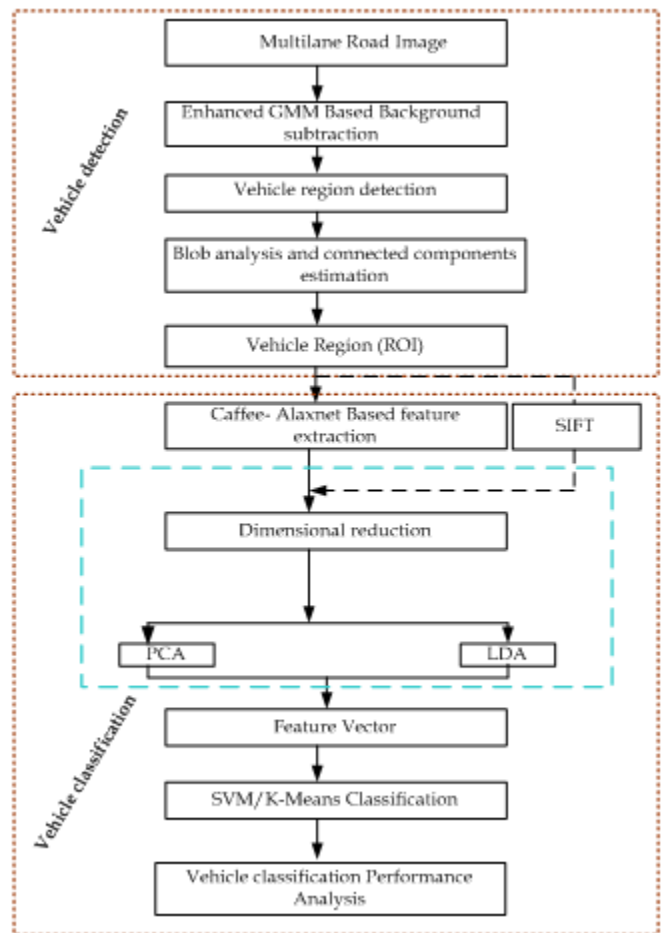


Fig. 1. Proposed vehicle detection and classification system

The overall proposed model comprises four sequential phases. These are:

- A. Vehicle detection,
- B. Feature extraction,
- C. Dimensional reduction and
- D. Classification

The discussion of the proposed methodology is presented as follows

A. Vehicle Detection

This section discusses the proposed vehicle detection mechanism.

1) *Multilane road image retrieval*: In this work, the vehicle image data has been obtained using static a camera placed on the road side. In real-time vision based surveillance applications, occlusion plays a significant role for limiting the efficiency. To deal with such issue, vehicle's images with rear information including lane dash line marking have been collected. It enables the proposed approach to detect and classify multilane vehicles. The dash line detection makes it feasible to detect occluded vehicles and their exact location. The camera has been placed in such a way that it takes the rear view of vehicles images on multiple lanes of the highway. To detect or localize the vehicle on image, background subtraction scheme has been applied.

2) *Background subtraction*: Considering the significance of Gaussian Mixture Model (GMM) algorithm for background subtraction [6] [10], in this paper an enhanced GMM scheme has been employed for background subtraction and vehicle detection. The proposed GMM model is discussed as follows:

a) *Enhanced gaussian mixture model based background subtraction*: Unlike conventional threshold-based approaches [16], proposed model applies an enhanced GMM scheme for background subtraction. GMM based background subtraction is nothing else but a pixel-based approach. Consider x be a pixel value at certain time instant. A flexible measure to estimate the probability density function (PDF) of x can be the GMM, in which the PDF comprises the sum of Gaussians. With K component densities the PDF of the Gaussian mixture $p(x)$ can be estimated as

$$p(x) = \sum_{k=1}^K w_k N(x; \mu_k, \sigma_k) \quad (1)$$

Where w_k represents the weight factor, and $N(x; \mu_k, \sigma_k)$ gives the normal density of mean μ_k and the covariance matrix $\Sigma_k = \sigma_k I$. GMM as suggested in [26] calculates these parameters to obtain the background. Initially, these parameters are initialized with zero, (i.e., $w_k = w_0, \mu_k = \mu_0, \sigma_k = \sigma_0$). In the case of any similarity, i.e., $\|x - \mu_j\| / \sigma_j < \tau$, with $j \in [1, \dots, K]$ and $\tau (> 0)$ as a certain threshold level, the GMM parameters are updated as follows:

$$w_k(t) = (1 - \alpha)w_k(t - 1) + \alpha M_k(t) \quad (2)$$

$$\mu_k(t) = (1 - \beta)\mu_k(t - 1) + \beta x \quad (3)$$

$$\sigma_k^2(t) = (1 - \beta)\sigma_k^2(t - 1) + \beta \|(x - \mu_k(t))\|^2 \quad (4)$$

Where $M_k(t) = 1$ in the case of the matching element j otherwise is considered as 0.

In case of zero similarity or non-matching elements, the component with minimum w_k is re-initialized, i.e., $w_k = w_0, \mu_k = \mu_0, \sigma_k = \sigma_0$. In above equations (2-4), α represents the learning rate, and β is obtained as

$$\beta = \alpha N(x; \mu_k, \sigma_k) \quad (5)$$

Here, the weight parameter w_k is normalized iteratively so as to increase to 1. In [26], researchers sorted Gaussians w_k / σ_k in decreasing order so as to perform background subtraction. In background subtraction, GMM applies a threshold value λ which is used to the cumulative sum of weights so as to obtain the set $\{1, \dots, B\}$. Mathematically, background subtraction is performed using equation (6).

$$B = \arg \min_{K_B} \left(\sum_{k=1}^{K_B \leq K} w_k > \lambda \right) \quad (6)$$

In this approach, the Gaussians with the maximum w_k and minimum standard deviation σ_k represent the background region. In major GMM models μ_k and σ_k are updated with certain constant learning rate β [26]. However, it can't be effective for dynamic application scenarios such as traffic movement, background changes, and varying lighting or illumination conditions. To deal with such issues, a modification was made in [27]. In [27] the learning rate β was assigned in an initial learning process that enabled adaptation under dynamic surface change. In real time applications, there can be pixels which might neither be a foreground nor a background object. However, such pixel is classified either as foreground or background. It leads inaccurate vehicle detection and classification. As proposed in [27], increasing β might cause extremely high rate pixel feature variations such as illumination that may make the system vulnerable. Similarly, with the square of the difference between mean and the pixel values might lead higher variance, resulting in continuous increase in illuminations till the saturation of Gaussian mixture over entire pixel color range. Observing both these approaches [26][27], it can be found the earlier [26] lacks dealing with dynamic surface variation, while later [27] suffers from pixel saturation caused due to fast variations (in variance). To deal with such issues, in this paper, an adaptive learning rate based enhanced GMM model has been developed that alleviates such degeneracy, especially in variance by introducing an optimal parameter update paradigm. In the proposed approach, the learning rate has been decoupled for μ_k and σ_k . Unlike conventional approaches, an adaptive learning rate $\gamma_k(t)$ has been applied for updating μ_k that comprises a relative probability factor $R_k = N(x; \mu_k, \sigma_k)$ that signifies whether a pixel belongs to the k th Gaussian component or not.

$$\gamma_k(t) = \gamma_k(t-1) + \frac{K+1}{K} R_k - \frac{1}{K} \sum_{i=1}^K R_i \quad (7)$$

The implementation of the proposed adaptive learning rate can provide fast Gaussian component mean update as suggested in [27]. It can also enable coping up with fast illumination changes that can ensure precise ROI identification and localization. Now, substituting γ_k as β in (3), it can be found that the self-governing update of the variance can avoid pixel saturation; however, a fast update might result into degeneracy situation. To alleviate this issue, a semi-parametric model has been applied for variance calculation that can significantly enable quasi-linear adaptation, particularly in the case of small changes from the mean and a degraded response for significantly higher deviations. To achieve this, a sigmoid function has been derived as follows:

$$f_{a,b}(x, \mu_k) = a + \frac{b-a}{1 + e^{-SE(x, \mu_k)}} \quad (8)$$

Where, $E(x, \mu_k) = (x - \mu_k)^T(x - \mu_k)$. Here, S plays the role of sigmoid slope controller. Now, substituting (8) in (4), the variance update is obtained as

$$\sigma_k^2(t) = (1 - \rho)\sigma_k^2(t-1) + \rho f_{a,b}(x, \mu_k(t-1)) \quad (9)$$

Where, $\eta = 0.6$ and $f_{a,b}(x, \mu)_{\mathcal{R}^+}$ limits σ_k to the region $\mathcal{R} \in \left[\frac{a+b}{2}, b\right]$. Here, the values of a and b are selected in such way that \mathcal{R} spans over one kth of the pixel range. Thus, applying the proposed adaptive learning rate based GMM model, background subtraction has been performed. The evaluation of the proposed scheme revealed that $\gamma_k(0) = 0.05$ can give better performance for background subtraction. Once performing background subtraction, a connected component analysis (CCA) mechanism has been implemented so as to remove irrelevant connected pixels or blobs so as to enable accurate ROI localization.

3) *Vehicle region localization*: To enhance the vehicle region detection, CCA scheme has been applied that considers valid region, size, and location on the image to remove irrelevant components. Here, a hypothesis that the connected region signifies the Gaussian components belonging to the single lane has been taken into consideration. In the proposed approach, CCA has been performed based on the centroid position. To use the lane information, the width of the individual connected components based on the allied lane has been normalized. The normalized width has been used as the width of the connected component region divided by the width of the lane at the centroid of the connected region. Using the normalized width, it becomes flexible to compare the vehicle size at distinct locations. Thus, employing the enhanced GMM and CCA approaches the exact vehicle regions or the ROI have been localized, which has been followed by its feature extraction.

B. Feature Extraction

Once estimating the vehicle region or the ROI, features have been extracted to execute further vehicle classification. In

a practical scenario, the vehicles of different categories such as sedan, SUV, MPV, van, truck, etc. would have different features. These high differences in features make classification intricate. As depicted below (Fig. 2) the vehicle (a) represents a MPV, (b) taxi, (c) van and (d) is the other commercial vehicles. These vehicles have different shape, size and color and therefore would have different features too. Considering a broad view of classification where these vehicles have to be classified into two categories, passenger and commercial or other types, to distinguish these vehicles correctly would be highly intricate because these vehicles can have same color, size etc. To enable efficient classification there is the need of certain robust image feature extraction and semantic learning paradigm.

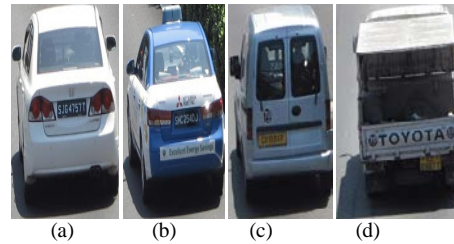


Fig. 2. Vehicle images with significantly higher different features

In this paper, the deep learning approach has been applied to perform vehicle or ROI feature extraction. Here, a well-known and robust image feature extraction model based on convolutional neural network (CNN) named AlexNet has been applied to extract ROI features. AlexNet is a multilayered DNN that functions based on convolutional neural network concept and works on ImageNet data. Ironically, the direct implementation of AlexNet DNN with generic computing systems and data elements is not feasible; therefore we have applied a parallel DNN model called CaffeNet [28] with AlexNet. It enabled AlexNet function on general purpose computers. The brief of the AlexNet DNN scheme is presented as follows:

1) *AlexNet DNN based feature extraction*: In this paper, CaffeNet based AlexNet feature extraction has been performed on vehicle dataset LSVRC-2012. The developed feature extraction model has been trained over the localized vehicle ROI data. To enable ROI data for feature extraction with multilayered AlexNet DNN, each vehicle region image has been resized to 256×256 dimension. As depicted in Fig. 3, AlexNet comprises five CONVOLUTIONAL LAYERS (CONV1-CONV5) and three FULLY CONNECTED LAYERS (FC6-FC8). The initial layer of this model can have general features resembling Gabor information and blob features. On the contrary, the higher layers comprise significant information for classification; therefore in AlexNet (Fig. 3) five CONVOLUTIONAL LAYERS and two FULLY CONNECTED LAYERS (FC6 and FC7) have been applied to extract features at different layers. Here, each convolutional layer comprises multiple kernels where each kernel signifies a 3D filter connected to the outputs of the preceding layer. In case of fully-connected layers (FC6-FC8), the individual layer comprises multiple neurons containing a real positive value.

The individual neuron is connected to all the neurons of the previous layer. In this paper, features have been obtained at the two fully connected layers, FC6 and FC7. To achieve better performance, 4096-dimensional features have been obtained at the higher layers of the DNN, FC6 and FC7. These extracted features have been presented in terms of a feature vector $F_V = (f_1, f_2, f_3, \dots, f_{4096})$ which has been later processed for dimensional reduction and feature selection. Once retrieving the features, the implementation of dimensional reduction schemes can enable swift and accurate vehicle classification. In this work, two-dimensional reduction algorithms, principle component analysis (PCA) and linear Discriminant analysis (LDA) have been applied to perform dimensional reduction and feature selection. Similar to the AlexNet DNN based feature extraction, SIFT approach has been applied to examine relative performance efficacy.

2) *SIFT based feature extraction*: This is the matter of fact that feature extraction, selection and its mapping plays a significant role to perform classification. The majority of classification systems are still insignificant because of lower inter-class scatter, particularly with vehicle's multiclass classification. In practice, the vehicle region or ROI in the image might be very small in size than the overall image size and even the change in lighting can introduce additional intricacies and the insignificant feature that eventually might impact classification accuracy. Here an effort has been made to enhance vehicle detection by applying an enhanced GMM background subtraction model. However, considering existing work and suggestions [25], in this paper, SIFT approach has also been applied to extract ROI features. To retrieve SIFT-based features, four directional filtering 128 SIFT feature descriptors of the each image have been obtained, i.e., 128-dimensional vectors. Similar to AlexNet features, SIFT features has been processed for dimensional reduction using PCA and LDA. It has been followed by SVM-based classification. The retrieved vectors have been projected to PCA algorithm for dimensional reduction. In this paper, the first 64 dimensional vectors have been considered and employing 32 Gaussian components distribution; fisher encoding has been done that eventually generates 4096-dimensional feature vector, which is equivalent to the AlexNet-FC6/FC7.

The discussion of the proposed dimensional reduction approach is presented as follows:

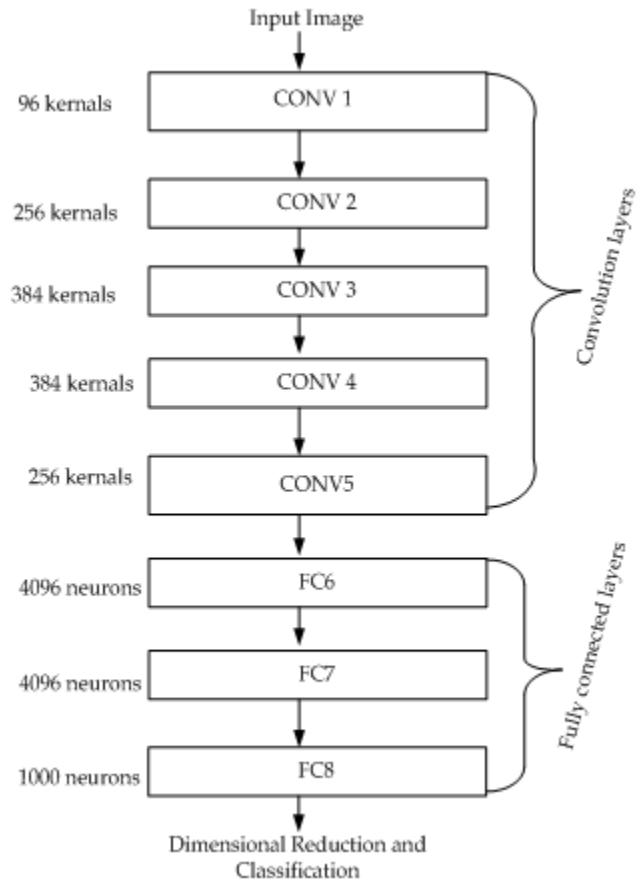


Fig. 3. AlexNet DNN Architecture

C. Dimensional Reduction and Classification

As discussed above, in feature extraction AlexNet as well as SIFT feature descriptors retrieved 4096-dimensional features for each image and therefore to achieve computation and time efficient classification, two predominant dimensional reduction and feature selection approaches, PCA and LDA have been applied. A brief of the applied dimensional reduction approaches is given as follows:

1) *PRINCIPLE COMPONENT ANALYSIS*: In this work, it is intended to classify vehicles in multiple classes. In general, the feature components extracted from PCA algorithm used to be the most expressive features (MEF), while LDA employs the most discriminating features (MDF) function. In PCA-based approach distinct principle component (PCS) is

generated for an individual class. However, despite of retrieving the distance from the average principal component of each class, the PCA vectors have been trained using SVM classifier. Here, radial basis function (RBF) kernel has been applied for SVM training. SVM has been trained to retrieve the largest feasible classification margin that signifies the lowest value of w in

$$\frac{1}{2}w^T w + E \sum \varepsilon_i \quad (10)$$

Where $\varepsilon_i \geq 0$ and E is the error tolerance level.

To perform classification, the training vectors have been categorized in labeled pairs $L_i(x_i, y_i)$ where x_i states the training vector, while the class label of x_i is given by $y_i \in \{-1, 1\}$. In classification, the hyperplane groups highest feasible points of the same class on the same side, while increasing the distance of either class from it. To achieve optimal classification accuracy 10-fold cross validation has been performed. To perform testing, a test image data has been processed for PCS estimation which has been followed by its principle component classification using trained SVM.

2) *Linear discriminant analysis:* As discussed above, PCA-based schemes employ MEFs to perform classification. However, MEFs can't be the MDFs all the time. On the contrary, LDA can perform automatic feature selection that can enable efficient feature space for further classification. To alleviate the issue of high dimensionality, LDA has been initiated by employing PCA, where all the vehicle region data or ROI irrespective of the class label has been projected onto a single PCS. The dimension of the PCS has been confined by the total training image minus the number of classes. In this model, two distinct metrics have been estimated, intra-class scatter matrix I_{ICW} and inter-class scatter matrix I_{IOS} . Mathematically these matrixes have been estimated as

$$I_{ICW} = \sum_{i=1}^C \sum_{j=1}^{M_i} (y_j - \mu_i)(y_j - \mu_i)^T \quad (11)$$

$$I_{IOS} = \sum_{i=1}^C (\mu_i - \mu)(\mu_i - \mu)^T, \quad (12)$$

Where C represents the total number of classes, μ_i states the average vector of a class i , and M_i signifies the number of samples within i . Thus, the average of the average vectors is obtained as

$$\mu = \frac{1}{C} \sum_{i=1}^C \mu_i \quad (13)$$

LDA approach focuses on maximizing the inter-class scatter while reducing the intra-class scatter by increasing the ratio $\det[S_B]/\det[S_w]$. The significance of applying this ratio is that in the case of a non-singular I_{IOS} matrix, the ratio can be increased when the column vectors of the projection matrix W can be the eigenvectors of $I_{ICW}^{-1}I_{IOS}$. Here, the projection matrix W with $C - 1$ dimension assigns the training data onto a new space, usually called fisher vector. Thus, W is applied for projecting all the training samples onto the fisher vector. The

retrieved feature vector $F_{VR} = (f_{1R}, f_{2R}, f_{3R}, \dots, f_{4096R})$ has been further used for classification.

In the proposed approach, the obtained vectors have been used to form a know discovery-tree that in the later stage has estimated the nearest neighbors during classification.

In addition to the AlexNet DNN based feature extraction, in this research SIFT has been applied for feature selection, which has been further processed for dimensional reduction using PCA and PLA (Fig. 1).

D. Classification

In this paper, a polynomial kernel based support vector machine (SVM) has been applied to perform vehicle classification. The extracted and dimensionally reduced features from LDA and PCA (Table 1) have been projected and mapped for SVM- based classification. To achieve optimal classification accuracy, 10-fold cross validation has been done. The vehicles have been classified into two broad classes, passenger and other, where passenger class contains vehicle types SUV, van, bus, and cars.

TABLE I. DIMENSIONAL REDUCTION AND CLASSIFICATION SCHEMES

Data Feature	Dimensional Reduction	Classification
AlexNet	PCA	SVM
AlexNet	LDA	SVM
SIFT-FV	PCA	SVM
SIFT-FV	LDA	SVM

Thus, the overall research implementation of the presented work is depicted in Fig. 4

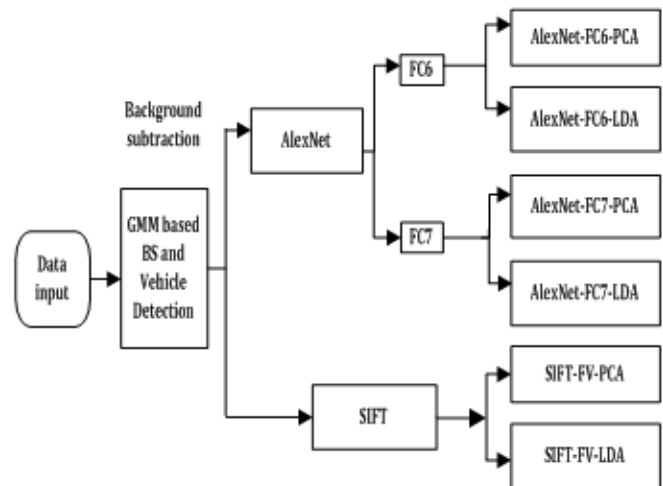


Fig. 4. Implementation Model

The performance evaluation of the proposed vehicle detection and classification algorithm has been discussed in following section.

V. RESULTS AND ANALYSIS

The results obtained are discussed in this section. To perform vehicle detection and classification, a total of 400 images of the vehicles with rear information were used for analysis. Among these images 200 images were from the vehicle category sedan, SUV, etc. or passenger category while

remaining 200 images were other types including bus, cab, etc. The data was equally selected so as to maintain the similar size of both classes. The initial or the normal size of the images was 4184×3108 that was later resized to 1046×777 for vehicle detection purpose. Once performing background subtraction using Gaussian Mixture Model (GMM), the localized ROI or vehicle region was mapped to the original image with natural resolution. To initiate classification process, the mapped image data was resized to 256×256 size and was fed as input to the AlexNet DNN. With AlexNet DNN based feature extraction, from each image the high layer features FC6 and FC7 were obtained, where each layer possesses 4096 dimensional features. Retrieving the overall features, it was passed to dimensional reduction schemes, PCA and LDA. The dimensionally reduced features were then projected to polynomial kernel function based SVM for multi-class classification. Also, as a parallel model was developed to retrieve features using SIFT scheme. In this approach, the SIFT descriptors of the images were obtained in 128-dimensional vectors which was then projected to PCA for dimensional reduction. The initial 50% of the PCA were selected for analysis, i.e., 64 dimensions. Finally applying 32 Gaussian components distribution, fisher encoding was performed that eventually provided 4096-dimensional feature vector, equivalent to the AlexNet-FC6/FC7. The overall algorithm was developed using MATLAB 2015b tool. Also, VIFeat-0.9.20 toolbox was used to enable swift and easy implementation and processing. The two-class classification results for passenger vehicles and others are presented in Table 2.

TABLE II. VEHICLE CLASSIFICATION ACCURACY

Features	Classification Accuracy (%)	
FC-6	GAX6-PCA	96.75
	GAX6-LDA	97.80
FC-7	GAX7-PCA	91.40
	GAX-7LDA	96.30
SIFT	GSIFT-FV-PCA	96.25
	GSIFT-FV-LDA	96.45

*GAX signifies proposed GMM detection preceded AlexNet DNN

TABLE III. RELATIVE PERFORMANCE ANALYSIS

Techniques	Classification Accuracy (%)
PCA+DFVS[25]	95.85
PCA+DIVS [25]	94.15
PCA+DIVS+DFVS [25]	96.42
SIFT-FV-PCA [29]	92.30
SIFT-FV-LDA [29]	91.30
AlexNet F6-PCA [16]	96.45
AlexNet F6-LDA [16]	97.00
AlexNet F7-PCA [16]	96.80
AlexNet F7LDA [16]	96.10
*GAX6-PCA	96.75
*GAX6-LDA	97.80
*GAX7-PCA	91.40
*GAX-7LDA	96.30
*GSIFT-FV-PCA	96.25
*GSIFT-FV-LDA	96.45

*GAX signifies proposed GMM detection preceded AlexNet DNN

Considering comparative performance of the proposed model and others [16], the impact the enhanced GMM for vehicle detection can be easily visualized. The proposed adaptive learning rate based GMM has enabled more accurate vehicle detection under different background and illumination conditions. Now, for classification performance analysis, it can be found that AlexNet FC6 with LDA outperforms other combinations or approaches for two-class classification. Here (Table 2), it can be seen that FC6 feature with LDA dimensional reduction gives classification accuracy of 97.80%, which is higher than FC7 features with PCA (91.40%) and LDA (96.30%). Considering an alternative of DNN, SIFT-based features have also exhibited better with LDA (96.45%) than PCA (96.25%). It states that higher layer DNN features with LDA can give efficient mapped features to perform vehicle classification even with low annotated data.

VI. CONCLUSION

In this paper, a multilevel optimization measure has been proposed for vehicle detection and classification. Considering the limitations of traditional threshold-based background subtraction schemes, an enhanced adaptive learning rate based GMM algorithm has been developed, which has enabled precise vehicle detection under varying frame background frame features and illumination. To avoid occlusion, in multilane traffic conditions, vehicle's rear features and lane dash markings have been taken into consideration. The application of connected component analysis (CCA) has enabled efficient vehicle region or ROI localization. An enhanced deep convolutional neural network (DNN), named AlexNet has been applied for ROI feature extraction. The implementation of AlexNet-DNN's higher layer features (FC6 and FC7) has exhibited better accuracy, because of higher feature informative contents. As a comparative model, SIFT feature descriptors have been obtained for the ROI. The retrieved 4096-dimensional features from AlexNet-FC6, FC7 and SIFT has been processed for dimensional reduction using PCA and LDA. To perform classification, in this paper polynomial kernel based SVM classifier has been applied that classifies vehicle data into passenger (car, taxi, sedan, SUV) and other types. Results exhibit that AlexNet FC6 features with LDA gives highest classification accuracy of 97.80%, followed by AlexNet-FC6 with PCA (96.75%). The highest accuracy with AlexNet-FC7 has been found lower than AlexNet FC6. Similarly, SIFT features with PCA and LDA (SVM with 10-fold cross validation) has exhibited classification accuracy of 96.25% and 96.45% respectively. The proposed scheme has outperformed other approach because of enhancements introduced regarding adaptive learning rate based GMM. This work has exhibited that adaptive learning rate based GMM with higher layers DNN features can lead optimal vehicle detection and classification. In general, DNN suffers from weight estimation and learning complexity issues, and hence to make this system more effective and time efficient, in future efforts can be made to enhance DNN learning. Concepts such as, shared weight estimation based CNN learning can also be explored to make the proposed system time efficient. In future,

the efficiency of the proposed system could be examined with multiple camera based real time vehicle detection and classification process.

REFERENCE

- [1] S. S. Harsha and A. Koteswara Rao, "A Highly Robust Vehicle Detection, Tracking and Speed Measurement Model for Intelligent Transport Systems", *International Journal of Applied Engineering Research*, Vol. 11, No. 5, pp 3731-3740, 2016.
- [2] S. S. Harsha and K. R. Anne, "A Robust and Efficient Optical Flow Analysis Based Vehicle Detection and Tracking System for Intelligent Transport System," *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 14, No. 6, pp 605-13, 2016.
- [3] "USDOT, USDOT intelligent transportation systems research", 2011, <http://www.fhwa.dot.gov/research/>, Jan 2013.
- [4] D. Ang, et al., "Video analytics for multi-camera traffic surveillance," second international workshop on computational transportation science, Seattle, WA, USA, pp. 25-30, 2009.
- [5] A. Ambardekar, et al., "Efficient vehicle tracking and classification for an automated traffic surveillance system," international conference on signal and image processing, Kailua-Kona, HI, USA, pp. 1-6, 2008.
- [6] A. Nurhadiyatna, et al., "Gabor filtering for feature extraction in real time vehicle classification system," 9th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, pp. 19-24, 2015.
- [7] T. Ma, et al., "Urban vehicle classification based on linear SVM with efficient vector sparse coding," Information and Automation (ICIA), IEEE International Conference on, Yinchuan, pp. 527-532, 2013.
- [8] H. C. Karaimmer, et al., "Combining Shape-Based and Gradient-Based Classifiers for Vehicle Classification," 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, pp. 800-805, 2015.
- [9] X. Chen, et al., "Vehicle representation and classification of surveillance video based on sparse learning," in China Communications, vol. 11, no. 13, pp. 135-141, 2014.
- [10] Y. Liu and K. Wang, "Vehicle classification system based on dynamic Bayesian network," Service Operations and Logistics, and Informatics (SOLI), 2014 IEEE International Conference on, Qingdao, pp. 22-26, 2014.
- [11] Y. Peng, et al., "Vehicle classification using sparse coding and spatial pyramid matching," 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, pp. 259-263, 2014.
- [12] B. Zhang, et al., "Vehicle Classification with Confidence by Classified Vector Quantization," in IEEE Intelligent Transportation Systems Magazine, vol. 5, no. 3, pp. 8-20, 2013.
- [13] Y. Hu et al., "Algorithm for vision-based vehicle detection and classification," 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO), Shenzhen, pp. 568-572, 2013.
- [14] W. Ma, et al., "Vehicle classification based on multi-feature fusion," Wireless, Mobile and Multimedia Networks (ICWMMN 2013), 5th IET International Conference on, Beijing, pp. 215-219, 2013.
- [15] H. Qian, et al., "Vehicle classification based on the fusion of deep network features and traditional features," Advanced Computational Intelligence (ICACI), 2015 Seventh International Conference on, Wuyi, pp. 257-262, 2015.
- [16] Y. Zhou, et al., "Vehicle Classification using Transferable Deep Neural Network Features," arXiv:1601.01145v1 [cs.CV] 6 Jan 2016.
- [17] Z. Dong, et al., "Vehicle Type Classification Using Unsupervised Convolutional Neural Network," Pattern Recognition (ICPR), 2014, 22nd International Conference on, Stockholm, pp. 172-177, 2014.
- [18] H. Gu, et al., "Vehicle size classification for real time intelligent transportation system," Conference Anthology, IEEE, China, pp. 1-5, 2013.
- [19] B. Zhang, "Reliable Classification of Vehicle Types Based on Cascade Classifier Ensembles," in IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 1, pp. 322-332, March 2013.
- [20] C. Y. Chen, et al., "Vehicle classification and counting system," Audio, Language and Image Processing (ICALIP), 2014 International Conference on, Shanghai, pp. 485-490, 2014.
- [21] R. H. Peña-González and M. A. Nuño-Maganda, "Computer vision based real-time vehicle tracking and classification system," 2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS), College Station, TX, pp. 679-682, 2014.
- [22] Y. Jia, et al., "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the ACM International Conference on Multimedia. ACM, pp. 675-678, 2014.
- [23] A. Krizhevsky, et al., "ImageNet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097-1105, 2012.
- [24] A. S. Razavian, et al., "CNN features off-the-shelf: an astounding baseline for recognition," arXiv preprint arXiv:1403.6382, 2014.
- [25] A. Ambardekar, et al., "Vehicle classification framework: a comparative study," EURASIP Journal on Image and Video Processing, vol. 2014, no. 1, p. 29, 2014.
- [26] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246-252, 1999.
- [27] D. S. Lee, "Effective gaussian mixture learning for video background subtraction," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 827-832, 2005.
- [28] Y. Jia, et al., "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the ACM International Conference on Multimedia. ACM, pp. 675-678, 2014.
- [29] J. Sánchez, et al., "Image classification with the fisher vector: Theory and practice," International journal of computer vision, vol. 105, no. 3, pp. 222-245, 2013.

Designing and Implementing of Intelligent Emotional Speech Recognition with Wavelet and Neural Network

Bibi Zahra Mansouri

Department of Computer
Engineering, Kerman Branch,
Islamic Azad University
Kerman, Iran

Hamid Mirvaziri*

Department of Computer
Engineering, University of
Shahid Bahonar
Kerman, Iran

Faramarz Sadeghi

Department of Computer Science,
University of Shahid Bahonar,
Kerman, Iran

Abstract—Recognition of emotion from speech is a significant subject in man-machine fields. In this study, speech signal has analyzed in order to create a recognition system which is able to recognize human emotion and a new set of characteristic has proposed in time, frequency and time-frequency domain in order to increase the accuracy. After extracting features of Pitch, MFCC, Wavelet, ZCR and Energy, neural networks classify four emotions of EMO-DB and SAVEE databases. Combination of features for two emotions in EMO-DB database is 100%, for three emotions is 98.48% and for four emotions is 90% due to the variety of speech, existing more spoken words and distinguishing male and female which is better than the result of SAVEE database. In SAVEE database, accuracy is 97.83% for two emotions of happy and sad, 84.75% for three emotions of angry, normal and sad and 77.78% for four emotions of happy, angry, sad and normal

Keywords—Recognition of emotion from speech; feature extraction; MFCC; Artificial neural network; Wavelet

I. INTRODUCTION

Speech is a communicative process among humans. One of the most significant characteristics of speech is transferring of internal emotion to the audiences. When the speech is presented by the speaker, the speech includes the individual's emotion. In this study, the researcher is going to recognize individual emotion. Recognizing emotions mean understanding speaker's emotion by speech's samples. It is better to use suitable parameters for improving the result of emotional speech. Firozshah et al have used MFCC and ANN to recognize four emotions as angry, happy, neutral and sad which have the accuracy of recognition 72.05, 66.05 and 71.25 for

women, men and mixtures of them respectively [1]. Javidi et al have used MFCC, ZCR, Pitch, Energy and combination of the CHAID decision Tree, Regression, SVM, C5.0 and ANN to recognize as angry, happy, neutral, sadness, disgust, fear and boredom emotions, and the accuracy of recognition using ANN was 71.70 [2]. Dai et al [3], have presented neural network and combination of feature as a landmark, Pitch, energy to recognize speech's emotions as angry, happy, neutral and sadness and the accuracy of 90% was obtained for recognizing angry and neutral and more than 80% accuracy was obtained for sad and happy and more than 49% was obtained for classifying four emotions. Ayadi et al have worked by feature extraction ANN and HMM. The number of emotions was 7 for them. The accuracy rate was 71% for HMM and 50% for ANN which has shown better operation of HMM. [4]. Haq et al have used 7 emotions of angry, disgust, fear, happy, neutral, sad, surprised and energy feature extraction, duration, MFCC, pitch and MLB which has obtained 53% accuracy rate [5]. Ververidis et al have used angry, happy, neutral and sad emotions. They extract the features of energy, formant and pitch and their accuracy was 53.7% [6]. Gharavian et al have used GMM model and used four emotions and its accuracy was 65.1%. In this study, they used modular neural- SVM and applied happy, angry and neutral's emotions. The accuracy rate was 76.3%. In this study, the accuracy rate for C5.0 was 56.3% [7]. Table 1 shows the previous works in this field. This article organized as follows: In section two emotion speech recognition system is introduced. In section three feature extraction is stated. proposed method is in section four and it is evaluated with different dataset in section five and six and finally there is a conclusion and future works in the last two sections.

TABLE I. PREVIOUS MODELS AND THEIR RESULTS IN RECENT YEARS

Previous work	Emotion	Feature	Classifier	Database	Recognition rate
Ververidis D, Kotropoulos c (2006) [6]	Anger, happiness, neutral, sad, surprise	Pitch, Energy, Formant	MLB	DES-SUSAS	53.7%
Ayadi M, Kamel S, Karray F (2007) [4]	Anger, disgust, happiness, neutral, anxiety, tiredness	Energy, MFCC	ANN, HMM	EMO-DB	71% 55%
Dai, K, Fell, H.J, MacAuslan, J(2008)[3]	Hot-Anger, happiness, sadness, neutral, cold anger	Pitch, energy, landmark	ANN	Emotional Prosody Speech and Transcripts corpus	49%
Haq S, Jackson PJB, Edge J (2008) [5]	Happiness, anger, disgust, fear, sadness, surprise, neutral	Pitch, Energy, MFCC	MLB	EMO-DB	53%
Firoz Shah. A, RajiSukumar, A,BabuAnto . P(2010)[1]	Anger, happiness, neutral, sadness	DWT	ANN	INDIAN - DB	72,05% 66,05% 71,25%
Javidi M, Roshan E(2013)[2]	Anger, disgust, happiness, neutral, sadness, fear, boredom	Energy, Pitch, MFCC, ZCR,	SVM, ANN, C5.0,	EMO-DB	71.7%
Gharavian D, Sheikhan M, (2013)[7]	Happiness, anger, neutral, sad	MFCC, formant, pitch	GMM,C5.0, MLP, MODULAR NEURAL - SVM	PERSIAN - DB	65.1% 56.3% 76.3%

II. EMOTION SPEECH RECOGNITION SYSTEM

The emotion recognition system includes four main parts. "Fig. 1" shows the information of emotion recognition system.

The Pattern and emotion recognition system include four main processes, which are as the following: speech input, feature extraction, classifier, emotion speech output.

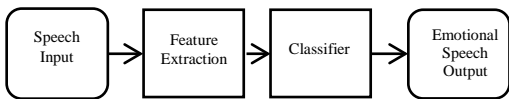


Fig. 1. Emotion speech recognition system

III. ANALYSIS AND FEATURE EXTRACTION

Extraction and selection of the best parameters of the speech signal are the most significant duty in designing a speech recognition system. Fig. 2 shows the preprocessing steps of speech analysis and feature extraction.

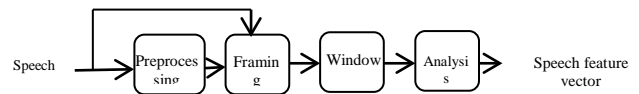


Fig. 2. Block diagram of speech signal analysis

A. Framing

When an audio vector is analyzed, the features are divided into two parts, half of them are in audio frame and the rest is in the frames. It is probable that the features are not achieved completely in each window analysis and they probably are hidden. Since, after converting analog signals to digital one, the speech samples are divided into frames in order to overlap each other. The new frame includes part of the previous frame and the next one [8].

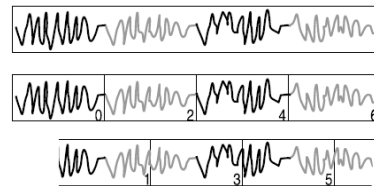


Fig. 3. Frame speech signal with overlapping frames

B. Windowing

Due to the non-static hood of speech signal and variable statistical features during time, the speech signal is divided into short times about 4-20ms and analyzing is conducted during mentioned time, these speech are called window.

After framing all frames at the beginning and end of each frame include interruption that is spectral distortion are reached to the least by framing at the beginning and end of each frame [9].

N is the number of symbols in each frame and n is the number of frames. Then the result of framing is the equation (1)

$$y(n) = x(n) w(n) \tag{1}$$

X(n) is the input signal of speech and y(n) is the output of the framing. Windows include varied models and some of them are introduced in the following equations [2].

$$W(n) = \left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] 0 \leq n \leq N-1 \tag{2-a}$$

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) 0 \leq n \leq N-1 \tag{2-b}$$

$$W(n) = 1 0 \leq n \leq N-1 \tag{2-c}$$

C. Energy

Energy is the most and significant and basic features in speech signal which recognize the boundary between speech and silence. The energy can be obtained as follows [2].

The energy (E) of a signal frame of length N is obtained by

$$E = \sum_{n=0}^{N-1} y(n) * y(n) \tag{3}$$

D. Zero Crossing (ZCR)

The calculation of ZCR is done with audio signal which recognize the speech signals from silence [10].

The ZCR Crossing of a signal frame of length N is obtained by

$$ZCR = \frac{1}{N} \sum_{n=0}^{N-1} \frac{|sgn[y(n)] - sgn[y(n-1)]|}{2} \quad (4)$$

$$\begin{aligned} Sgn[y(n)] &= 1 && \text{if } y(n) \geq 0 \\ Sgn[y(n)] &= -1 && \text{if } y(n) < 0 \end{aligned}$$

E. Mel Frequency Cepstral Coefficient (MFCC)

The main aim of using MFCC is inspiring from human's ears feature in receiving and understanding speech. The operation of human's ear is in a way that understanding frequency is varied from real frequency. One Mel is the measurement unit of heard frequency of a phoneme. It doesn't rely on pitch frequency linearly, since the operation of human's ear is in a way that it doesn't understand more than 1 kHz frequency auditory system of human doesn't understand the frequencies ever scale linear, since the researcher has presented Mel scale for developing human understands. Mel frequency is a logarithmic mapping of physical frequency to understand frequencies. MFCC's coefficient considers certain coefficient for each frequency and since varied emotions considering different morality and mood have different frequency; therefore, anger is different from happiness and using these features increase the strength of emotion recognition.

F. Pitch

The periodic information of thin and thick speech is recognized mainly by pitch frequency. The more pitch the more thin sound and the less pitch frequency the more thick sound.

This frequency, which is called base frequency and it is shown by F_0 , is about 50 to 150 hertz in men. In women it is about 150 to 450 hertz and in children it is about 300 to 700 hertz.

One of the oldest ways of estimating pitch in speech is autocorrelation. In this method autocorrelation changes of the function $r(\eta)$ are plotted with respect to η (sample frame) [11].

$$r(\eta) = \sum_{n=0}^{N-1-\eta} y(n) * y(n - \eta) \quad (5)$$

G. Discrete wavelet transform(DWT)

The most usual signal analysis is Fourier transform which break up signals to different frequencies and keep the information of frequencies and lose the information of time, while the wavelet includes both of them that is, frequency information and time - oriented information. Equation of wavelet transform has presented in the following[1].

$$W_{j,k} = \sum_j \sum_k s(k). 2^{-\frac{j}{2}} \Psi(2^{-j}n - k)k, j \in z \quad (6)$$

$\Psi(t)$ is the main wavelet of analysis function and $S(k)$ is speech signal, j is time measurement, k is the amount of movement in each measurement (transform parameter) and $w_{j,k}$ is wavelet coefficients.

In discrete wavelet transform, the main signal passes through the low-pass and high-pass filters which are appeared as approximation and detailed coefficients. In speech signals, low frequency is known as approximation $h(n)$, and high frequency is known as details $g(n)$ which has shown in Fig.4.

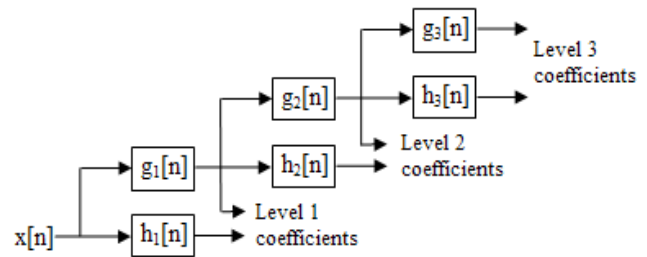


Fig. 4. Analysis of discrete wavelet on three level[12]

IV. CLASSIFYING MODEL OF ARTIFICIAL NEURAL NETWORK

A multilayer neural network (MLP) shows a non-linear relationship between input and output vectors. It operates through connecting neurons of each node to the previous and next layers. The output of each neuron is multiplied by the weighting coefficients and is given as input to non-linear functions. In training phase, educational information is given to perceptron. Then weights are adjusted so that error between the output current and target reduces to minimum or the number of training reaches the predetermined one. Then to evaluate the accuracy of the training process, a series of inexperienced input is applied to the network.

MLP architecture consists of an input layer, hidden layer and output layer each of them includes specific neurons. The numbers of neurons of input layers is equal to the vector's features plus bias neuron and the number of output neurons is equal to the defined class for classifier. It is possible to change the number of middle neurons till the best accuracy is obtained. In this research the number of middle neurons is changed from one to five and the best result has been obtained when the network has just had one hidden layer. Therefore a hidden layer of 12 neurons were considered.

V. DATABASE

Emotional database in emotion recognition is usually applied for studying acoustic, phonetic and research and development in the field of emotion speech recognition system. SAVEE and EMO-DB have been studied for this research.

A. Berlin Database of Emotional Speech (EMO-DB)

This database is produced in technical university of Berlin. Seven emotional moods have used in this database which are angry, happy, boredom, sadness, fear, disgust and neutral. Ten artists who were five men and five women run this program. The number of speeches is about 535. The number of audio files for seven emotions is classified as follows [13]: angry (127), boredom (81), disgust (69), happy (71), sadness (62) and neutral (79).

B. Surrey Audio-Visual Expressed Emotion Database (SAVEE)

The database consists of the recorded voice of four made actor in 7 different emotions and 480 British speeches which has chosen amongst TIMIT database. These databases have recorded on forms of audio, video and audio-video. The database has recorded four British speakers of surrey’s postgraduate student between the ages of 27 to 31. The samples are 44.1 KHz for audio and 60 fps for video. The number of audio files for seven emotions classified as follows [14]: angry (60), surprise (60), disgust (60), fear (60), happy (60), sadness (60) and neutral (120).

VI. IMPLEMENTATION METHOD AND ANALYSIS OF RESULTS

In this study data mining is done by ANN classifier and IBM SPSS software.

IBM SPSS Modeler is one of the best data mining tools and professional software to perform complex calculations and statistical analyzes for server and client.

Our data include 60 features extracted from a Berlin database speech and SAVEE. The outputs were angry, happy, sad and neutral. In this study, 340 speeches are chosen from Berlin database and 300 speeches are chosen from SAVEE. 20 percent of data were used for testing and 80 percent were used for training. This process has been repeated many times and the accuracy of classification is obtained based on the samples which recognized rightly to the all samples. Then, average of accuracy values, calculates for all repetition and presented as final accuracy (5 fold cross validation).

In the first experiment, recognizing of two emotions was performed. Our emotional moods were happy and sad. The whole number of speeches in Berlin database were 132 which include happy (70), sadness (62). The number of speeches was 120 in SAVEE database which include happy (60) and sadness (60). The result of each test for all features and features combination has done and their accuracy is presented in table 2. As you see in the table, the wavelet features in EMO-DB is 85.29% and in SAVEE is 53.57% by using the feature combination as wavelet, MFCC, energy, ZCR, pitch, energy Fourier, ZCR Fourier, pitch Fourier accuracy is obtained 100% in EMO-DB and 97.83% in SAVEE database, and for the Berlin database, accuracy is 2.7% better than SAVEE.

In table 2 the accuracy of emotions happy and sad in IBM SPSS Modeler Software in Berlin database obtained as 91.43%, and after removing lost data in MFCC features, it changes to 100%.

In the second test, it is tried for three emotions of angry, sad and neutral. All of 69 features were used in the test. The whole number of speech in Berlin database was 267; 126 for angry, 79 for neutral and 62 for sadness. The speech numbers were 240 in SAVEE database, 60 for angry, 120 for neutral and 60 for sadness. The result of each test has done for all features and features combination which their accuracy is shown in table 3. As you see in the table, for the wavelet feature is 63.64% in Berlin database and 54.24% in SAVEE database and by using feature combination as wavelet, MFCC, energy, ZCR, pitch, energy Fourier, ZCR Fourier, and pitch Fourier, it

is obtained as 98.48% in Berlin database and 84.75 in SAVEE that is, in Berlin database 13.73% is better than SAVEE.

TABLE II. THE ACCURACY OF HAPPY AND SAD

Table with 3 columns: Dataset, SAVEE, EMO-DB. Rows include various feature combinations like WAVELET+FFTZCR, WAVLET+FFTPITCH, etc., and a 'Mixture of features' row.

TABLE III. THE ACCURACY OF ANGRY, SAD, NEUTRAL

Table with 3 columns: Dataset, SAVEE, EMO-DB. Rows include various feature combinations like WAVELET+FFTZCR, WAVLET+FFTPITCH, etc., and a 'Mixture of features' row.

In table 3, the accuracy of angry, sad and neutral emotions in IBM SPSS Modeler software in Berlin database firstly obtained 96.97%, and after removing lost data in MFCC feature it changes to 98.48%

In the third test, it is tried to classify angry, happy, sad, neutral emotions include 60 features. The whole number of speeches in Berlin database were 337 which include 126 angry, 79 neutral, 62 sadness and 70 happy. In SAVEE database, the speech number was 300 which include 60 angry, 120 neutral, 60 sadness and 60 happy. The result of each test for all features and feature combination has done and their accuracy has presented in table 4. As you see in the table, for wavelet feature, it is 56.25% in Berlin database and 37.5% in SAVEE, and by using the feature combination as wavelet, MFCC, energy, ZCR, pitch, energy Fourier, ZCR Fourier, pitch Fourier, it is obtained as 90% in Berlin database and 77.78% in SAVEE and Berlin database accuracy is 12.22% more than SAVEE.

TABLE IV. THE ACCURACY OF ANGRY, HAPPY, SAD, NEUTRAL

Dataset	SAVEE	EMO-DB
WAVELET+FFTZCR	50	62.5
WAVELET+FFTPITCH	61.11	56.25
WAVELET+FFTENERGY	65.28	67.5
WAVELET+ZCR	61.11	86.25
WAVELET+ENERGY	56.94	65
WAVELET+PITCH	65.28	70
WAVELET+MFCC	43.06	41.25
FFT ZCR	55.56	62.5
FFT pitch	63.89	47.25
FFT energy	76.39	63.75
Energy	70.83	70
ZCR	63.89	88.75
Pitch	79.17	67.5
MFCC	47.22	34
Wavelet	37.5	56.25
Mixture of features	77.78	90

In table 4, the accuracy of happy, sad, angry and neutral emotions in IBM SPSS Modeler software in Berlin database firstly obtained as 76.5%, and after removing lost data in MFCC feature changes to 90%.

VII. CONCLUSION

In this study, EMO-DB emotional speech is used which was made at the technical Berlin University and SAVEE emotional speech, made in Surrey University in England. In this database, the data have high quality. The next step in recognizing emotional speech is feature extraction, that wavelet, MFCC, energy, pitches, ZCR. In this study, the feature combination of time – frequency, time, frequency domain and ANN classifier is used and feature combination in EMO-DB for two emotion is 100%, for three emotion is 98.48%, and for four emotion is 90%, which all are better than SAVEE accuracy of recognizing emotional speech as 97.83% for happiness and sadness, 84.75%, for angry, sad and neutral and 77.78%, for happiness, sadness, angry, neutral.

VIII. SUGGESTION FOR FUTURE WORKS

It can be concluded from the mentioned studies that by a combination of features, the accuracy increases, which is significant, but it is not the best method since it is time consuming. Regarding this reason, the researcher has chosen the combination of neural network with Evolutionary Algorithm which improves the quality of speech emotion recognition.

REFERENCES

- [1] F.Shah, A.R. Sukumar, and B.Anto, "Discrete wavelet transforms and artificial neural networks for speech emotion recognition," International Journal of Computer Theory and Engineering, 2(3), 2010, pp. 1793-8201.
- [2] M.M. Javidi, and E.F. Roshan, "Speech Emotion Recognition by Using Combinations of C5.0, Neural Network (NN), and Support Vector Machines (SVM) Classification Methods," Journal of Mathematics and Computer Science, 6, 2013, pp. 191-200.
- [3] K. Dai, H.J.Fell, and J.MacAuslan, "Recognizing emotion in speech using neural networks," Telehealth and Assistive Technologies, 2008, pp. 31-38.
- [4] E.Ayadi, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in: The proceedings of the international conference on Acoustics, Speech, and Signal Processing, vol 5, 2007, pp. 957-960.
- [5] S.Haq, P.J.Jackson, and J.Edge, "Audio-visual feature selection and reduction for emotion classification," In: The proceedings of international conference on Auditory-Visual Speech Processing, 2008, pp. 185-190.
- [6] D.Ververidis, and C.Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," In: The proceedings of European signal processing conference, 2006, pp. 1-5.
- [7] M.Sheikhan, M.Bejani, and D.Gharavian, "Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method," Neural Computing and Applications, 23(1), 2013, pp. 215-227.
- [8] I.McLoughlin, "Applied speech and audio processing: with Matlab examples," Cambridge University Press, 2009
- [9] J.S.Devi, Y. Srinivas, and S.P. Nandyala, "Automatic Speech Emotion and Speaker Recognition Based on Hybrid GMM and FFNN," International Journal, 2014.
- [10] L.R. Rabiner, and R.W.Schafer, "Introduction to digital speech processing, Foundations and trends in signal processing," 1(1), 2007, pp. 1-194.
- [11] X.Li, "SPEECH Feature Toolbox (SPEFT) Design and Emotional Speech Feature Extraction," Faculty of Graduate School, Marquette University, 2007.
- [12] S. Sunny, S.David Peter, and K.P. Jacob, "Performance Analysis of Different Wavelet Families in Recognizing Speech," International Journal of Engineering Trends and Technology, 4, 2013, pp. 512-517.
- [13] <http://emodb.bilderbar.info/docu/#emodb>
- [14] <http://kahlan.eps.surrey.ac.uk/savee/Evaluation.html>

An IoT Middleware Framework for Industrial Applications

Nicoleta-Cristina Gaitan^{1,2}, Vasile Gheorghita Gaitan^{1,2}, Ioan Ungurean^{1,2}

¹Faculty of Electrical Engineering and Computer Science, ²Integrated Centre for Research, Development and Innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for Fabrication and Control (MANSiD) Stefan cel Mare University of Suceava, Romania

Abstract—Starting from the RFID and the wireless sensor networks, the Internet of connected things has attracted the attention of major IT companies and later, of the industrial environment that recognized the concept as one of their key axes for future growth and development. The implementation of IoT in the industrial environment raises some significant issues related to the diversity of fieldbuses, the large number of devices and their configuration. The requirements related to reliability, security and real-time are very important. This paper proposes an industrial IoT and communications at the edge framework which has some outstanding features related to: the easy integration of fieldbuses and devices used in industrial environments with automatic configuration features, integration of multiple middleware technologies (CORBA, OPC and DDS), the uncoupling of the industrial activity from the publishing data on the Internet, security at different levels of the framework. Another important feature of the proposed framework is that it is based on mature standards and on open source or public implementations of these standards. The framework is modular, allowing the easy integration of new fieldbus protocols, middleware technologies and new objects in the client application. This paper is focused mainly on CORBA and DDS approaches.

Keywords—Internet of Things; Middleware; CORBA; ACE ORB (TAO); Data Distribution Service

I. INTRODUCTION

KEVIN Ashton, from the MIT Auto-ID Center, was the first who proposed the term "Internet of Things" (IoT), referring to the connection between the information provided by radio frequency identifiers (RFID) and the Internet [1]. Quickly, the interest in the Internet of connected things caught the attention of governments and IT companies which have recognized the concept as one of the key axes for their future growth and development [2]. An increasingly accepted definition of the IoT was provided in [3]. In this definition, the emphasis is placed on virtual and physical "things" which: use intelligent interfaces; are fully integrated into an information network; have identifiers, physical attributes, and virtual personalities using a global infrastructure network with dynamic configuration (mobile), auto-configuration facilities, and interoperable communication protocols.

The potential growth of IoT technologies has led to increased interest in their use in various industries, where devices, machines, sensors, or simple things communicate with each other using standard Internet technologies [4]. It can be stated that the real value of the Industrial IoT (IIoT) is the

availability of ubiquitous information and consequently, the decisions that can be made from it. An IIoT platform must validate the sharing of dispersed and ubiquitous data in an efficient and timely way for the web, cloud, desktop, embedded, and mobile applications. Therefore, IIoT can be defined [5] as the connection between the sensors from the physical world, devices and machines on the Internet and, by applying a thorough analysis using the software, the transformation of massive data into powerful insight, and intelligence.

It is becoming increasingly clear that the industry needs a functional and useful architecture for the Industrial Internet of Things (IIoT), which should include the recent progresses and novelty technologies in the field. Such an architecture should be easily understood and, at the same time, complete. Most projects and specialized literature are focused on how "things" can be converted so that they can be connected to the Internet through the addition of intelligence and connectivity, for instance by using the RFID technology for things/objects in everyday life [6][7]. Beside the RFID technology, they also take into account the wireless sensor networks. These architectures can be found in [8][9][10][11][12]. An important issue of this solution is security [13].

Although the issues listed above are essential for the IoT, it can be considered that in addition of RFID, wireless communication, sensors and actuators as IoT things, it can be added devices and machines with wired communication in order to define IIoT things. Furthermore, it can be pointed out that industrial automation involves difficult requirements regarding communication and the ensuring security and reliability. These requirements must be met by IIoT from the beginning. Currently, the implementation and operation of the complex production processes or of the Internet applications (Internet-enabled) requires time and a manual network setup that is susceptible to errors. This situation is generated by the need to ensure a high level of determinism, safety, and security during the production process and to avoid both critical security failures and costly production interruptions. These objectives should be IIoT-specific, including a high level of automation for the network configuration processes (including the fieldbuses pertaining to the industrial environment).

In this paper, it is proposed an IIoT framework organized on three levels, based on the three observations outlined above (italic): the device that integrates the hardware (sensors, actuators, RFID) in order to sense/control the physical world and to acquire data, middleware for data transport and an

application which provides the means to interact with the user and other IoT applications [14]. The proposed framework can be the edge that bridges the information technologies and world of things, where the available resources in the cloud cannot be directly accessed [14]. In this case, the operational technologies are the fieldbuses with their features that represents additional challenges. At the low level, the framework understands different network topology, and data protocols that will be found into the world of things. This contains solutions for automatically discovery and identification of the real industrial things, data associated and to be able to perform storage at high-frequency updates. At the high level, the framework collects the data and sent it to the cloud via IIoT standards. In the CISCO visions [15], the framework represents the Edge Computing (that is also called Fog Computing).

The framework is in accordance with the IIoT definition which was presented previously. The solution uses OPC (Open Platform Communication) [16], OPC .NET [16], OPC UA [16], TAO [17] and DDS (Data Distribution Service) standard [18][19] are used as middleware (an important component in IIoT) in order to show the data at enterprise level and DDS for the external enterprise interoperability. This article mainly takes into account the implementations based on TAO and DDS. In the process of defining the framework, three great challenges arise: (i) the large number of fieldbuses, description of devices and automatic configuration; (ii) middleware choice and provision of real-time services; and (iii) separation of the industrial activity from the operations, for the sake of data publication and subscription on the Internet, and incorporation of different types of technology. The proposed framework can be used in smart factory but the utilisation can be extended for smart home, smart buildings, smart living, and smart city.

Furthermore, this paper is organized as follows: Section II briefly presents different architectures proposed for the use of IoT in industry. Section III presents our proposal for an IoT based on TAO for the industry field. Section IV presents the test performed in order to compare the bandwidth used by a TAO-based server with one based on OPC DA, OPC UA and OPC.NET in a local network. Section V presents a comparison between TAO/OpenDDS and OPC UA as support for IIoT. The final conclusions are drawn in Section VI.

II. RELATED WORK OF THE INDUSTRIAL IOT ARCHITECTURE

When a new IIoT architecture and a practical implementation are proposed, a natural question which arises is: what are the existing solutions? The literature specialised in the field is very poor in such solutions because IIoT is at the beginning. A courageous attempt is made in [20]. The authors, relying on a rich bibliography, tried to understand the current status and the future research opportunities related to the use of the IoT concept in industry. Only Section V strictly refers to the applications of IoT in industry, fields such as healthcare service, food supply chain, transport and logistics, and firefighting, which are more in the field of services and infrastructure and not industry, are being taken into account. The only industrial sector already addressed is mining production [21][22]. Our bibliographic studies have led to

similar conclusions. There are few articles related to IIoT and those are strictly focused on specific applications. In what follows, we will briefly present some concerns present at institutional level or which are covered by research projects.

In Germany, the IoT is associated with the field of production and logistics through the term "Industry 4.0"[23], and the grounds are being prepared for a new social and technological revolution which will drastically change the whole industrial environment. Industry 4.0 is a sophisticated change of the entire chain of values: communication, planning, logistics, and production. Due to the success it recorded in the fields of information and communication technologies (ICT) (currently 90% of all manufacturing processes are already supported by ICT) and embedded systems, (strong autonomous microcomputers) either connected to each other or to the Internet, wired or wireless, it will lead to a convergence between the physical and the virtual (cyberspace) world. This convergence takes the form of a Cyber- Physical Systems (CPS), term used international to describe Industry 4.0 concept. With the development of IPv6 standards, there are now enough addresses to allow, for the first time, the networking of resources, information, objects, and people, in order to create the Internet of Things and Services. The proposed architecture is set on four levels (from bottom to top): Internet of Things, Internet-based System & Service Platforms, Internet of Services and Applications. More details can be found in [23].

Another interesting research project in the IIoT field is the IoT@Work [24]. The project focuses on the exploitation of IoT technologies in the industrial and automation sectors. The architecture proposed in this project has five horizontal levels and three vertical planes. The horizontal levels refer to: Field/Control Infrastructure & Network, Device and Network Embedded Services (auto-configuration, device semantic, network management), Device Resource Creation & Management Services (abstraction, context/dependencies), Application Level Middleware Services (commissioning, composition, adaptation), and Automation Applications. The vertical planes are the following: communication plane, security plane and management plane. The project proposes the following technologies for the IIoT: Directory Service, Auto-Configuration of Real-Time Ethernet, Event Dispatching (Event Notification Service), Capability-based Access Control, Complex Event Processing, Network Slices, and Embedded Access Control. More details on the proposed architecture and technologies can be found in [24].

An interesting discussion is launched by Herman Storey (co-chair ISA 100), Rick Bullota and Daniel Drolet in [25]. The discussion begins with the observation that IIoT should primarily provide security, robustness, and punctuality as far as the requirements of automation networks are concerned and, secondly, remote access. The IIoT proposed architecture has four horizontal levels: multiple physical media and link layer, IPv6/6LoWPAN common network layer, more communication stack layers and multiple applications layer. Vertically, the architecture has two levels: Common time and Common network management and security. As an essential element, the IIoT must provide a way to integrate multiple physical environments and multiple applications in a single industrial network system using common technologies. To integrate such

a variety of communication and application environments, the IIoT must use IPv6 as a network protocol. IPv6 has an extension called 6LoWPAN which allows it, as a network layer, to be used for low power networks or limited bandwidth. Although it was designed for battery-powered wireless devices, it may be used for wired networks as well. ISA 100.15 published a document which provides models and concepts for architectures adequate for IIoT.

Following the analysis of the three architectures presented above, it can be said that currently, there is a low degree of standardization. Efforts are being made to achieve an IIoT standard (Industry 4.0, ISA 100). The IIoT is a different IoT from the non-industrial ones due to the special characteristics of the production processes. Except for Industry 4.0, IIoT architecture is based on ground level devices, which are interconnected via fieldbuses and which have access points to local networks and the Internet, while on the upper level it has specific applications. Intermediate levels ensure services for the safe transport of information. In addition to the horizontal levels, there may also be vertical planes, able to ensure management and security, time management, and so on. The expectations of IIoT refer to the possibility that devices, machines, and other objects could interact with each other without relying on human intervention to achieve added value. Among the most important requirements for IIoT [26], we can mention: reliability, robustness, reasonable cost, security and safety, easy use, low/no maintenance, optimal and adaptive set of features, standardization, integration capabilities, reach sensing and data capabilities, industry degree support, and services. The challenges faced by IIoT refer to IoT devices, lifetime and energy, data and information, humans and business.

III. THE IOT FRAMEWORK PROPOSED FOR INDUSTRY

In this section, it is presented the new proposed IoT framework for industry where devices, machines, sensors, or simple things must communicate with each other. This IIoT framework is composed of three levels (device, middleware, and application). The first level is the device level. It is composed of three elements, namely: the device which acquires data directly from the environment and can transfer this information using a wired or wireless network/fieldbus, the gateway which adapts the specific network protocol to the specific computer protocol used by the middleware in order to connect to the IIoT environments (which can also add real time facilities) and the software driver for the gateway device which adapts the information sent or received to/from the gateway in order for it to be compatible with the middleware. The middleware level is designed to provide data transportation inside the IIoT and it is based on the OPC, CORBA (with TAO implementation (The ACE ORB)) and DDS. The application level provides support for the implementation of the basic applications pertaining to the proposed framework and the level's middleware objects which can be embedded in other IIoT applications [27]. The specific interoperability model is provided by the OPC and TAO, while the global interoperability is ensured by the DDS middleware standard [28].

A. The motivation of the proposed framework

In order to motivate the proposed framework, we can begin from the question: is it a new technology? The answer is that it is a new vision related to the reorganisation of a sum of existing technologies in order to satisfy new requirements concerning the future development of the industry.

Regarding the device level, the following major problems were considered: there are different physical and data link layers which respond to different requirements of specific applications in the industry field; at the extremity of the global network, there are fieldbuses that are intended to acquire information from sensors and transducers, and to emit commands via actuators; and that all these fieldbuses must have common support for IPv4/IPv6. For this level, a gateway device is defined, one which must implement the gateway function [29] in order to transfer the information to the higher level. It must transform the process-specific information into information useful for the higher level [30] and it must provide real-time behaviour at fieldbus level. Furthermore, a description method for devices, recognised by all partners who require information about devices, must be developed. Network/fieldbus configuration for acquisition of information from the process is a time-consuming and expensive operation which means that tools capable of automating this operation must be created. In the fieldbuses area field, there is currently a multitude of standards (and perhaps new standards will appear in the future) which means that, consequently, the framework must support the integration of new protocols.

The middleware level has the important task of transporting information between different nodes placed in the Intranet, Extranet, and the Internet. This level implies important design decisions. Standard-based middleware's were taken into account due to their stability and impact on the industry. Since the OPC specifications are specially designed for industrial applications, a first major question is: why TAO and DDS? A second question may be: why not just DDS? The short answer to the first question is: the OPC specifications have no explicit real-time requirements and use the client-server paradigm, which is less suitable for data centre frameworks of the publisher/subscriber type; and answer to the second question is: TAO is better prepared for real time. Further, these two answers are expanded.

A very interesting discussion on the utilization of standards for real-time distribution middleware is presented in [31]. The authors, out of several distribution models, chose those which are based on the standard, are mature, stable and with impact on the industry; namely: CORBA/RT-CORBA, Distributed System Ada Annex (DSA), Data Distribution Service for Real-time System (DDS) and Distributed Real-Time Specification for Java (DRTSJ). Even though the authors of [31] do not provide a verdict or have not carried out a ranking, however, a classification can be made.

CORBA/RT-CORBA has the following advantages: it is based on a very mature technology, one involved in a wide range of applications [31], such as Software Defined Radios [32] and Industrial Robotics [33]; RT-CORBA entities validate

the development of critical real-time applications; from the point of view of scheduling, the RT-CORBA provides static scheduling based on Fixed Priority Scheduling (FPS), the use of threads as schedulable entities, control of the competition degree on the servers using thread pools, deterministic access to shared resources, the use of different scheduling policies, and the use of distributable threads as a schedulable entity; as far as network resource management is concerned, it provides mechanism for the fine-tuning of network properties, it uses private connections and definitions of priority-banded connections; it is the only standard which provides mechanisms for the specification of scheduling parameters which may be used during execution; facilitate interoperability between implementations (GIOP - General Inter-ORB Protocol); TAO implementation is the most popular and updated open-source implementation for RT-CORBA. As disadvantages of RT-CORBA, we can mention: unlike the CORBA specification updated in [34], RT-CORBA is not currently in the attention of the Object Management Group (OMG), the last update being performed in 2005 [35][36]; it does not take into consideration the network scheduling; it uses TCP/IP stack which means that even the use of Ethernet switches is unsuitable for implementation of hard real-time systems; TAO implementation does not provide synchronous protocols (it is based on the operating system); it does not implement the priority transforms model, the use of buffers to store remote requests in thread pools nor the borrowing of threads among thread pool lanes.

The DDS has the following advantages: it is considered a mature technology involved in several real-time applications [31] in the fields such as Defence [17], Automation [37], and Space [38]; supports anonymous and asynchronous dissemination of information; has specific requirements for distributed applications such as control systems, sensor networks, and industrial automation systems; it is a data-centric middleware [18] and, therefore, it is aware of the contents of the interchanged data which can be directly managed; it provides multiplatform and multi-language support; the types of shared data can be defined by using IDL language [34]; interoperability between different implementations is provided by DDS Interoperability Wire Protocol (DDSI) [39]; it is a recently updated specification, OMG provides specification for the Extensible and Dynamic Topic Types [40], which provides support in order to define and modify dynamic (on runtime) data for the extension and evolution of systems based on DDS; the DDS model defines a strongly typed Global Data Space where publishers (Data Writer (DW)) can write (provide) data and subscribers (Data Reader (DR)) can read (consume) data allowing the middleware to focus on obtaining data independent of their origin; the standard was explicitly designed for distributed real-time systems; specifications define a set of QoS parameters in order to configure non-functional properties for each entity and allow the change of some of them during an operation; a subset of QoS parameters allows the control of temporal behaviour and improves the application predictability; it defines different mechanisms meant to validate the communication between entities (polling, synchronous mode and asynchronous mode for the DR entity)

and provides the opportunity to notify the application by Polling, Listeners, Conditions, and Wait-sets; there are both commercial (CoreDX or RTI-DDS) and open source (OpenSplice or OpenDDS) implementations. Among the DDS disadvantages, we can mention: there are no evaluations in detail done on the DDS real-time performance (an attempt can be found in [41]); it does not explicitly address the scheduling of threads at processor level; it is oriented on IP networks and not on the real-time networks (still lists a set of requirements for network support); considers only network policies based on fixed priority scheduling and excludes any other type of predictable network used in industry; some internal middleware operations generate meta-traffic thus introducing an override that must be taken into account in the analysis of behaviour in time; DDSI has an indefinite number of sub-messages; there is still no profile for safety-critical applications.

The DSA and DRTSJ are not competitive for real time as CORBA/RT-CORBA and DDS. The DSA [31] was specifically designed to support predictable applications and several features, which ensure determinism, are left to application implementation; while the DRTSJ [31] specification is not complete, there are still problems which were not addressed and there is no formal DRTSJ specification (only a draft). On the other hand, all these protocols and their implementations for real-time communication use IP-based networks. Even if local networks that use switches are used, real time is not easily achieved.

For the application level, the design issues taken into account are: easy embedding and integration of several technologies (OPC DA, OPC .NET, OPC UA, and CORBA); default communication between application objects by defining a "software bus" so that the application objects communicate with each other and the implementation, at the current level, of the gateway function between different technologies; decoupling of the company's activities and specific production processes, which requires a high degree of security; the publication of some information on the Internet; platform-independent communication between the instances of several applications; establishment of a connection with the usual databases which benefit from a specialized middleware for data communication.

The IoT framework of the system proposed in this article, in order to integrate IoT in the monitoring and control of the industrial processes, is presented in Fig. 1. The proposed framework is based on the OPC specification, DDS and CORBA middleware (TAO implementation). Furthermore, the framework will be presented from the point of view of CORBA and DDS middlewares. These middlewares were used because they allow the development of applications distributable on the Internet. In industry, CORBA middleware is not widely used although there is the DAIS [42] standard which describes how to develop SCADA applications based on CORBA. In the proposed framework, new TAO servers and clients are considered supplementary uses, which, just as DAIS, are based on the OPC DA 2.05 specification. Our solution is easier to implement compared to DAIS.

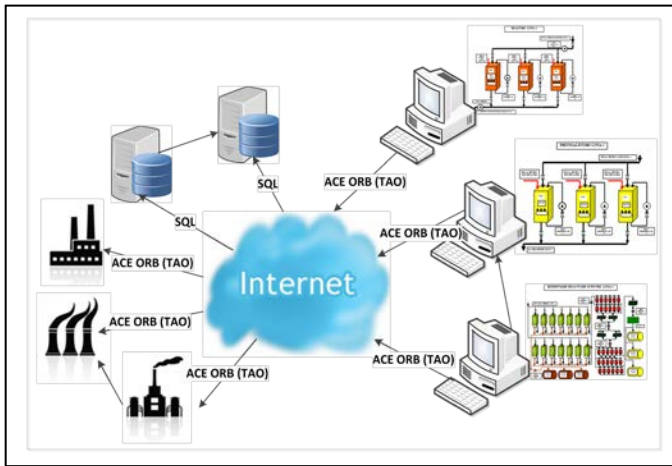


Fig. 1. Distribution on Internet of the proposed framework

From the point of view of implementation, the proposed framework consists of two main functional modules: the data-acquisition module (which will be referred to as the server module) and the Human-Machine Interface (HMI) module (for the information retrieved from the server modules) which will be referred to as the Human Machine Interface - Process Control and Monitoring (HMI-PCM) module, and it is mainly a client application for TAO and OPC servers. DDS is implemented as an object in the HMI-PCM. By using the TAO, the information acquired from the industrial process can be distributed on the Internet, in a client-server manner, as noticeable in Fig. 1. A functional (complex) system can be composed of multiple servers and multiple HMI-PCM clients. A HMI-PCM client can connect to multiple server modules and database servers, as described in the following sections. Clients can generate history based on the data read from the server, history stored in a database which can be consulted later by the client who generated this history or by other clients.

B. Server module

The architecture of the server module is shown in Fig. 2. This architecture is structured on three main levels. On the lower level, we have the drivers which acquire data from the fieldbuses and store it on the cache located on the upper level. This level is integrated in the device level of the IIoT framework. Its main role is the implementation of the acquisition cycle which is specific to the fieldbuses protocol used for communication. On this level there are more software modules, each module specific for one fieldbus. Furthermore, these modules receive data from the top level, which will be sent to the fieldbuses (e.g. commands for actuators). These modules receive all the data which must be updated continuously from the top level (data that is in at least one client's subscription list). This data is included in the acquisition cycle implemented in the drivers. Furthermore, these modules implement mechanisms for the data read on request (asynchronously read). They rely on a running platform (Linux or Windows) and are developed as independent modules (as libraries). This allows the development of new drivers without recompiling the other server modules. Between this level and the upper level, there is a well-defined interface that allows the integration of drivers for new fieldbus protocols (API 1 from Fig. 2).

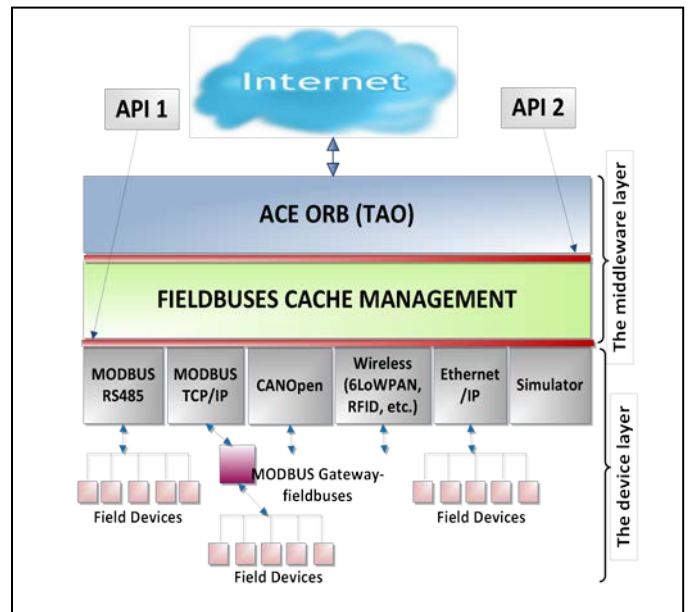


Fig. 2. The server module architecture

On the intermediate level, we have the Fieldbuses Cache Management (FCM) module which deals with the management of the cache which stores the data read from the fieldbuses, and which is also developed as an independent module (as a library). This memory cache is necessary to achieve a rapid response to the requests received from the upper level. The cache memory is a resource shared by several threads and has all the access control mechanisms implemented to ensure data consistency. Furthermore, this module stores a list of data on which clients are subscribed to ensure continuous updating of the cache (data update is provided only for that list of data). The data received from clients, which must be submitted to the devices connected on fieldbuses, are stored in the cache and are forwarded to the appropriate network driver. This level is part of the middleware level. Between this level and the upper level (the server itself), there is a well-defined interface which allows the adaptation of the FCM to any desired type of server, including TAO server (API 2 from Fig. 2).

On the top level, we have the server which provides support to access the cache with both read and in writing operations, in other words, the access to field devices connected to networks. Furthermore, the server integrates the TAO middleware which provides services for the transmission/reception of data to/from the HMI-PCM clients. To ensure these services, a CORBA IDL interface was defined, one which has been integrated into the server and the client modules. The interface is based on the OPC DA 2.05 classical specification. So, four interfaces were defined, namely: DataServer, an interface with a Register (server connection) and DeRegister (disconnected from server) methods; IServer interface with Addgroup, RemoveGroup, and SetState methods (edits the properties of the group); IGroup interface with AddItems and RemoveItems methods; IUpdate interface with OnDataChange (updates data to the client group) and Disconnect (server being offline) methods; IBrowse interface with BrowseAddressSpace (accesses the server address space), ChangeBrowsePosition (browses the address space server),

GetItemID (takes over the address space identifier of server), QueryAvailableProperties (reads the properties of an Item), SyncRead (synchronously reads the value and quality of a list of items, from the cache or device), and SyncWrite methods (synchronously writes the value and quality of a list of item, from the cache or device). We detail the implementation based on TAO because it is a less used a solution in industrial environments compared with servers based on OPC specifications.

C. HMI-PCM – Human Machine Interface -Process Control and Monitoring

The client application (HMI-PCM) is an environment that can instantiate many objects (controls). There are three types of objects: graphical objects, middleware objects and expression objects. They expose data members in the HMI-PCM environment. The data members can be interconnected in order to transfer data between objects, or can be used in different math expressions to which other objects can connect (subscribe) by using a standard interface (API 3 from Fig. 3). Middleware objects connect to data providers (servers) based on different middleware packages (OPC.NET objects to transport data from/to OPC.NET data servers, OPC DA objects to transport data from/to OPC DA servers; OPC UA objects to transport data from/to OPC UA servers; TAO objects to transport data from/to CORBA servers). The architecture of the HMI-PCM module is presented in Fig. 3.

OpenDDS is an open source implementation of the DDS specification based on TAO. The DDS objects from HMI-PCM environment ensure the interoperability between different HMI-PCM applications running anywhere (on the same computer, the computers interconnected throughout local network or computers interconnected throughout the Internet). The objects can expose the HMI-PCM address space, including middleware objects that partially or fully expose the server address space (see subsection D).

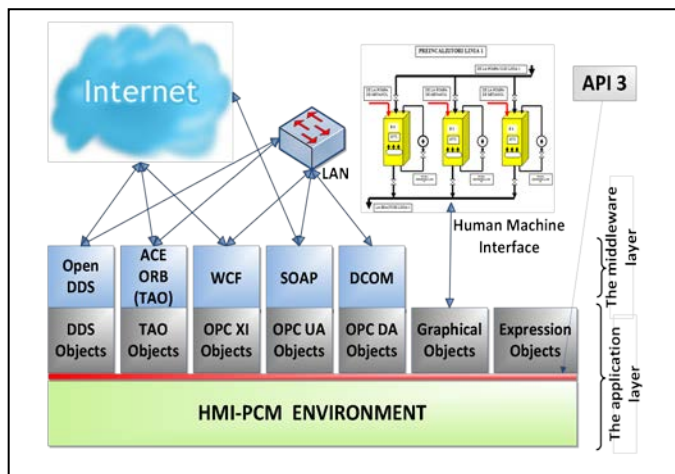


Fig. 3. The HMI-PCM module architecture

The most important feature of this application is that it allows the interconnection of objects in the HMI-PCM. Each object has data members that can be connected to each other or to the data members of other objects from the HMI-PCM. Thus, to display the data from the server, a graphical object is

used, one that connects to the TAO objects that are connected to these servers. With this feature, the HMI-PCM application can be easily configured according to the user’s requirements and preferences. Another important feature of the HMI-PCM is that new objects can be added as dynamic libraries. They must comply with HMI-PCM standard interface (API 3 from Fig. 3) that enables communication between the HMI-PCM objects (objects derived from a basic object). So, it is not necessary to compile the whole clients (only the object added).

D. Implementation considerations

The server is developed and implemented as an application in C++. For each fieldbus, there is a library which implements the function specific to the fieldbus. It was implemented a library for MODBUS RTU (with a RS485–RS232 interface), a library for MODBUS TCP/IP and a library for CANOpen (with a USB-CAN interface). The libraries for EtherCAT and Ethernet/IP are under development process. Since there are many Modbus TCP/IP gateways to other fieldbuses, these systems can be easily integrated into the proposed framework (should be considered the differences in terms of real time between fieldbuses and MODBUS TCP / IP because TCP/IP stack is best effort type and not real time). For the transport protocol between server and client, the following protocol was employed: IIOP (default) Internet Inter-ORB Protocol, SHMIOP - shared memory transport protocol, IIOP over Secure Sockets Layer (SSL), HTTP Tunnelling Inter-ORB Protocol, and ZIOP – IIOP with compression).

Due to the modular software architecture of the server (see Fig. 2), servers based on more middleware types were developed, while this paper deals with the server based on the TAO middleware (version 6.2.5). The server will expose data as a collection of industrial networks, each network having a collection of devices. Every device connected to an industrial process can be seen as a collection of objects. For this reason, a dictionary of objects was developed, managed by the FCM, exposing all the capabilities of the devices. Each object can have multiple data members and each data member can be characterized by properties such as value, data type, access rights, or other property that can be defined by the user based on the application. The content of the object dictionary (data provider) forms the address space of the server. Each middleware object will expose this address space to the client. A natural question is how to create this address space? FCM has on the upper level a defined standard interface for server connection (API 2 from Fig. 2), and another one at the bottom for connection to the fieldbus-specific drivers (API 1 from Fig. 2). Any driver that implements this interface is loaded without recompiling the entire application.

At this point, another question appears: how are the system devices described? Among the various solutions (EDDL - Electronic Device Description Language, FDT –Field Device Tool, FDI-Field Device Integration, EDS - Electronic Data Sheet), for simplicity reasons, a solution was adopted, based on the CiA DS 306 D3 v1.3 specification (EDS). This specification has been extended to support Modbus, M-Bus and ASCII-DCON protocols in addition to CANOpen. Modbus TCP/IP gateway connects to other protocol implementing devices, such as Profibus, Profinet, EtherCAT, EP PowerLink, Ethernet / IP, LonWork, etc. From a device, one cannot get

more than the information that is defined in the corresponding EDS. For example, for the Modbus protocol, a new section called [Communication] was added. This section of the EDS file describes the commands required to access objects like:

[IndexObject]: Request: FC: SFC-x: ADR: L-x:E:
ADR:L-x:E: ADR, L-x:/

Response: FC: SFC-x: ADR: L-x:/

Where: IndexObject – Process Data Object (PDO) or Service Data Object (SDO) that describes the data. Request: the format of request commands: FC – function code; SFC – sub-function code; ADR – address; L-x – length or count, x = number of bytes of this field; E – The extension of the commands. Response: the response format that is optional. For the functions of the MODBUS protocol, the answer can be built depending on request commands. If the PDO or SDO objects have a defined separate area for read and write operation, subsections [read] and [write] may be used. “:” - fields’ separator (if a field is missing from a MODBUS command/ response, only a separator, “/” – terminator is used).

In the (automatic or manual) configuration process, a file is created in order to attach a driver to each fieldbus (a specific dynamic library) and an ID and an EDS file to each device from the fieldbus. This file is used by the FCM, which sends the path to the EDS files of the active devices from the fieldbus to the driver. There may be several fieldbuses of the same type and more identical devices in one fieldbus. A configuration file associated with the server and built on EDS files contains the entire tree structure of the information that can be accessed and forms the address space of the server. This address space can be accessed by the TAO object from the HMI-PCM application through the IDL interfaces defined at the end of subsection III.B.

Once the server address space is defined, the server will expose this information to the clients, using the interfaces defined in IDL (see the end of subsection III.B). The main implementing objectives of the server refer mainly to the service name, client management, client-associated group management, group-associated items management, updating groups, reading and writing items, browsing in the address space, security information, and QoS.

The HMI-PCM is developed and implemented as an application in C#. Each object (see Fig. 3) is a library which exports a class derived from a base object. For the TAO object, a wrapper was used to marshal data from C++ to C# (TAO is developed for C++ application). The HMI-PCM application is developed in C#, as it offers the possibility of rapidly developing graphical applications and for productivity reasons. The HMI-PCM application is very interesting, allowing the communication between servers implemented with different technologies. Each server has one or more simulation drivers (a client can write or read to simulating some functionalities which can read or write by other clients). In addition to their role of simulation, these drivers allow the implementation of a relay function (gateway) between different types of servers. For example, suppose that the HMI-PCM has activated two middleware objects, one for OPC UA (data profile) and one of TAO type. A TAO user wishes to expose, to TAO clients, the

nodes of the OPC UA. Firstly, it must create an EDS file for the simulation driver for the TAO server with the desired objects that are visible from the OPC UA object (the compatibility of data types must be ensured). The objects exposed by the TAO object based on the EDS file will be found in the FCM dictionary of objects. In the HMI-PCM client, any item of the OPC UA object (from the ones chosen and described in the EDS file from TAO) can be connected to a corresponding item exposed by the TAO server based on the EDS file for the simulator (read or write - IN or OUT).

All the TAO clients can read or write properly from/in the items exposed by the simulator. There can be any number of simulators (depending on the host system resources). This type of relay can be attained between any of the middleware objects using a simulation driver and its attached EDS file. Connection can also be made directly, with the specification that an item should be output (or bidirectional) and the other input (or bidirectional), and the data types must be compatible. In addition, one can connect an intermediate expression object which can operate on source value using a mathematical expression.

For low power communication stack, there is the MICRO PROFILE and COMPACT PROFILE as part of CORBA/e (and it is implemented in TAO), while reliable communications and Internet-enabled communications are provided by TAO through transport protocols and naming service.

E. Security

Security features are presented at different levels of the proposed framework. In general, at the fieldbus protocols, security features are not provided, because they introduce an additional overhead and are non-deterministic components. In order to use the FCM component, the server must authenticate throughout a unique identification key. In the absence of authentication, the exported functions of the FCM module do not work correctly. The same thing happens with the fieldbus drivers. The current security level of the application is sent to the FCM in order to enable/disable the controls from the windows of the network manager, the connection manager, and from other configuration windows exposed by FCM and fieldbus drivers. The server application has an access panel that requires a user name and a password in order to view and change configuration parameters of the fieldbuses. Users are divided into groups, for users, manager, administrators and guests, each group having restricted access to the functionalities of the server, except for the administrator group. The server configuration is stored in an encrypted XML file (hidden somewhere in the system). The same vision is applied to the HMI-PCM application.

At the middleware level, in TAO there is the possibility to comprise messages (using pluggable ZIOP protocols) and to secure the communication (using SSLIOP pluggable protocol that is based on SSL). In the original DDS specification, related to the security, only the following is specified: “the application could attach security credentials via the USER_DATA policy that can be used by the remote application to authenticate the source”. The new DDS security specification [43] (request for proposal) proposes interesting solutions based on Domains Secure and Confidential Topics. RTI has a wide range of

security solutions such as: domain separation, access control and secure bridging; deep packet inspection; data filtering; secure operating system; secure transport; improved paradigm for secure distributed infrastructure [44]. OpenSplice ensure DDS Secure Networking Service and Access Control [45]. For OpenDDS, we integrated the SSLIOP (from TAO) through Extensible Transport Framework, in order to enable confidentiality and authentication.

OPC DA security for the communication is based on DCOM security, OPC.NET has different binding modes and types of authentication security modes depending on the type of binding (Named piped, TCP, HTTP Basic and HTTP WS) more types of authentication are being offered. OPC UA contains the philosophy related to the security in the specification, namely OPC UA part 2 - Security Model [46]. OPC UA is Secure-by-default, encryption enabled, and uses advanced certificate handling.

IV. EXPERIMENTAL RESULTS

This section presents the tests performed for the proposed solution based on TAO (with 3 transport protocols: IIOP, SSLIOP, ZIOP) when it is used in a local network. First, the bandwidth used by the server based on TAO was compared with the one used by the server based on OPC DA, OPC UA and OPC.NET. Tests were performed in a network composed of eight computers, identical in terms of hardware and software, and a switch with 100Mbps Ethernet ports. Each computer had an AMD Athlon (tm) 64 X2 Dual Core Processor 4200+ 2.21GHz, 1GB of RAM and a Windows operating system. On one computer (which will be referred to as the server), are executed in turn the data server based on OPC DA, OPC UA, OPC.NET, and the server based on TAO. All these servers use the same data provider (a simulator that generates random values for items and stores them in the cache memory of the server). For the experimental test, we used version 6.2.5 for TAO and the IIOP, SSLIOP, and ZIOP protocols. On the other computers, the HMI-PCM application is executed in turn with TAO, OPC.NET HTTP, OPC.NET TCP, OPC UA BIN (data profile), OPC.DA objects connected to TAO, OPC.NET HTTP, OPC.NET TCP, OPC UA BIN (data profile), and respectively, OPC DA servers. For the TAO objects, the IIOP, SSLIOP and ZIOP were used, as transport protocols. Clients will make a group/subscription/list (the names are specific to the used middleware) that contains 16 items/nodes whose data type is BYTE.

With Colasoft Capsa software package, the traffic speed on the server computer was measured. It should be noted that there is no network traffic generated by other applications (the LAN is not connected to the Internet). The software architecture of the tests performed is shown in Fig. 4.

The first test consisted in determining the transfer rate when data is updated at a rate of 100ms. The test results are shown in Fig. 5. In this figure, we can see that the bandwidth occupied when using TAO with IIOP and SSLIOP is higher than when using the OPC DA, OPC UA BIN and OPC.NET TCP, and smaller than when using the OPC.NET HTTP, but is lowest when ZIOP is used as transport protocol.

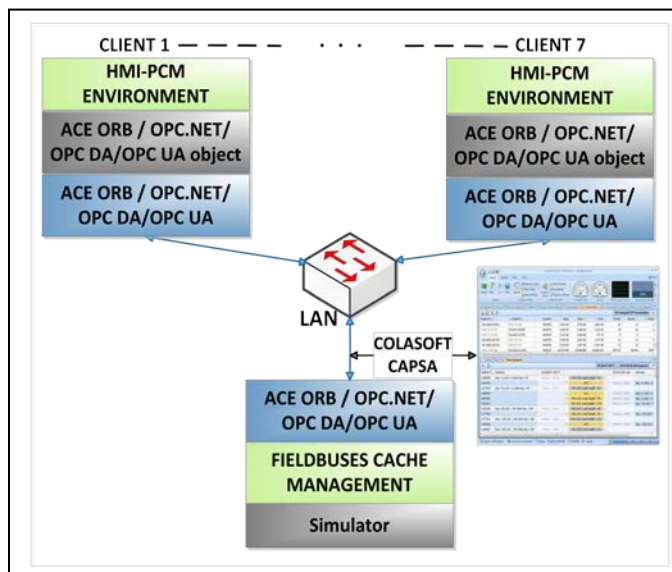


Fig. 4. The software architecture of the tests performed

The second test consisted in determining the transfer rate when data is updated at a rate of 500ms. The test results are shown in Fig. 5. From this figure, we can see that the occupied bandwidth when TAO is used is higher than when OPC DA and OPC.NET TCP are used, and lower than when OPC.NET HTTP or OPC UA BIN is used. Unlike the previous test, the bandwidth occupied by TAO is much closer to the bandwidth occupied by OPC.NET TCP and OPC DA.

The third test consisted in determining the transfer rate when data is updated at a rate of 1000ms. The test results are shown in Fig. 5. As in the previous tests, the same approximation trend of the band occupied by TAO with the band occupied by OPC DA and OPC.NET TCP can be noticed.

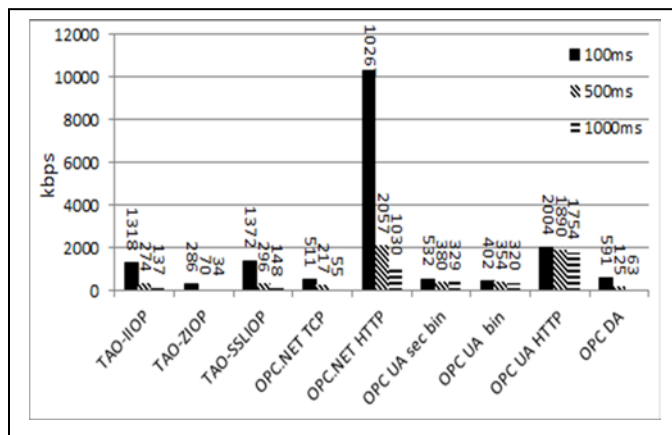


Fig. 5. Bandwidth occupied for a refresh rate of 100ms

Fig. 6 presents a synthesis of the 3 cases presented so far. An approximation trend of the bandwidth occupied by TAO with the bandwidth occupied by OPC.DA and OPC .NET TCP can be easily noticed. It should be noted that the tests were done in a local network, a framework widely used in the operation of industrial SCADA applications.

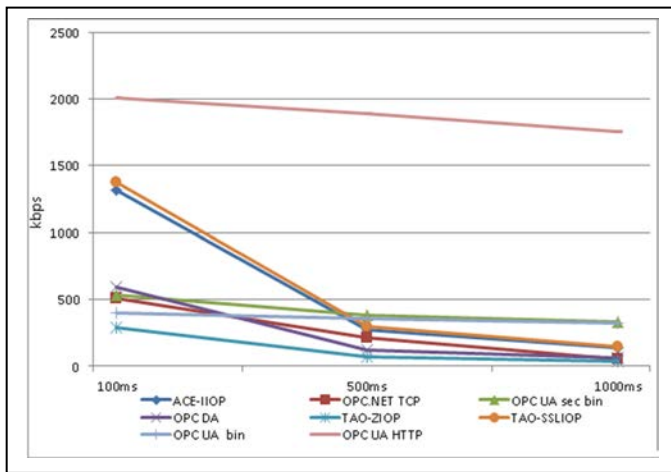


Fig. 6. Comparison for the bandwidth occupied

The proposed framework is designed to provide access to data via the Internet, where the response time cannot be guaranteed. It is unlikely to apply a refresh rate of 100ms for a client to connect to a server through the Internet, and very likely to use refresh rates of around 1000ms (in the Internet, this refresh rate cannot be guaranteed because the communication protocols are best-effort type depending on the network load).

The performances of the application based on TAO IIOP are very close to the performances of the applications based on OPC DA and OPC.NET TCP at an update rate of around 1000ms, but OPC DA is based on DCOM technology that works in a LAN network and OPC.NET TCP is dependent on .NET platform, based on Windows Communication Foundation. Furthermore, to use OPC.NET and to get the source code, you must be a member of the OPC Foundation. One advantage of using the TAO middleware is that it is an open source.

TAO with ZIOP transport protocol is the best because the messages are archived, but it does not provide any security mechanism. The use of security and encryption of the messages with SSLIOP transport protocol (based on SSL) introduces an additional overhead of the messages related to the IIOP transport protocol, which can be seen in the graphs, due to certificate exchanges and the increasing of the message size. The same difference can be seen for OPC.UA binary and OPC.UA binary with security and encryption of the messages. The use of the HTTP protocol leads to a significant increase of the messages that can be seen for OPC.NET HTTP and OPC.UA HTTP. In the case of the OPC.UA middleware, an important traffic generated by the keep alive mechanism (observed with Wireshark tool) can be observed. This traffic is much lower in the TAO implementation. From Wireshark tool, it can be seen that if the encryption or archiving mechanisms are not used, the data can easily be identified in the messages.

Table 1 presents the number of bytes of Ethernet frames and TAO for the three transport protocols (IIOP, ZIOP, and SSLIOP) sent by the server in order to update a group consisting of 1, 2, 4, 8, and 16 items. This information was obtained with Wireshark tool. As expected, the smallest frames are obtained by activating the ZIOP transport protocol. If the

messages are small, and the size of the archived message (plus archived message header) is higher than the size of the original message (with IIOP), then the message is no longer archived and it is sent using the IIOP transport protocol.

TABLE I. THE MESSAGE FOR TAO TRANSPORT PROTOCOLS

	TAO-IIOP	TAO-ZIOP	TAO-SSLIOP
1 items	296B/1 frame	314B/1 frame	351B/1 frame
2 items	402B/1 frame	324B/1 frame	475B/1 frame
4 items	674B/1 frame	335B/1 frame	810B/2 frames
8 items	1174B/1 frame	357B/1 frame	1310/2 frames
16 items	2236/2 frames	397B/1 frame	2401/3 frames

From the point of view of the memory footprint, the working set for the server with TAO is about 11MB for IIOP, increased to about 36MB for ZIOP and reaches about 38MB for SSLIOP, while the processor load depends on the refresh rate of the items, reaching 8% for a refresh rate of 100ms. On the other hand, OPC UA has a working set that varies from 48MB (without encryption and security) and reaches about 174MB with encryption and security. The processor load at a refresh rate of 100ms is about 50%. This may be due to the development mode of the server, which is developed in C# and the code is interpreted, while TAO is implemented in C++.

V. CONCLUSION

In the rather poor landscape of IIoT architectures, the proposed framework can be a starting point, especially since efforts are being made to implement and to perform a practical demonstration of the proposed functionalities.

This model was referred as framework and not as architecture because it is concerned with the IIoT device platform that transport the specific messages (little data) and which, through the DDS objects, can connect to the IoT services and applications (big data).

The framework enjoys several powerful points. First, it is based on mature and very mature standards and it can say that it is highly standardised. At device level, a unified method to describe the devices based on the EDS specification from CiA was defined. It was extended, among others, for the MODBUS protocol. Currently, there are many MODBUS TCP/IP - other protocol gateways which have their own mechanism of describing the devices; it can be depicted by the EDS modified for MODBUS. This solution resolved the challenge related on the large number of fieldbuses. The standardised interface from the lower level of FCM is scalable, allowing the integration of drivers specific for other fieldbus protocols without recompiling the FCM module and the server. The presence of the objects' dictionary, which creates the server address space, is the reason for the decoupling (virtualisation) between server and the complexity of fieldbuses, and a unified way of describing them. The configuration interfaces of the fieldbuses have a semi-automatic behaviour (drivers identify the field devices and the display on the server objects which can be exposed, and the server automatically restores the last saved configuration).

At the middleware level, several technologies were selected and implemented (OPC, TAO, and DDS) which enable a proper adaptation to the specific application. The PCM-HMI

application allows easy exchange of information between servers and, by implementing the DDS objects; it allows the publisher / subscriber a type of communication between PCM-HMI applications on the same computer, in the local networks or on the Internet. This solution resolved the challenge related on middleware choice and separation of the industrial activity.

Sensing the weaknesses of the framework, the authors intend: to clearly define the vertical planes such as security, timing and management; to improve support for automatic configuration of fieldbuses; to directly connect the DDS object to the FCM in order to retrieve data from fieldbuses through the objects' dictionary (there is a risk of creating a security breach, because the same object has direct access to process data and may publish the data acquired from the sensors and transducers on the Internet and can take commands from the Internet for the actuators); to be embedded, even partially, based on a new profile, in TAO and DDS, the address space concept and the information model from OPC UA; to develop tools for the easy configuration of DDS objects; to develop OPC UA security concepts in OpenDDS.

ACKNOWLEDGMENT

This work was partially supported from the project "Integrated Centre for research, development and innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for fabrication and control", Contract No. 671/09.04.2015, Sectoral Operational Program for Increase of the Economic Competitiveness co-funded from the European Regional Development Fund.

REFERENCES

- [1] K. Ashton, "Internet of Things," RFID Journal, June 22 2009.
- [2] Qazi Mamoon Ashraf, Mohamed Hadi Habaebi, „Autonomic schemes for threat mitigation in Internet of Things," Journal of Network and Computer Applications, Volume 49, March 2015, Pages 112-127, ISSN 1084-8045, <http://dx.doi.org/10.1016/j.jnca.2014.11.011>.
- [3] R. van Kranenburg, The Internet of Things: A Critique of Ambient Technology and the All-Seeing Network of RFID. Institute of Network Cultures, 2008
- [4] Jordán Pascual Espada, Ronald R. Yager, Bin Guo, Internet of things: Smart things network and communication, Journal of Network and Computer Applications, Volume 42, June 2014, Pages 118-119, ISSN 1084-8045, <http://dx.doi.org/10.1016/j.jnca.2014.03.003>.
- [5] Scott MacDonald, Whitney Rockley, McRock CAPITAL, The Industrial Internet of THINGS – IIoT Report, 2014.
- [6] Roselli, L.; Mariotti, C.; Mezzanotte, P.; Alimenti, F.; Orecchini, G.; Virili, M.; Carvalho, N.B., "Review of the present technologies concurrently contributing to the implementation of the Internet of Things (IoT) paradigm: RFID, Green Electronics, WPT and Energy Harvesting," Wireless Sensors and Sensor Networks (WiSNet), 2015 IEEE Topical Conference on , vol., no., pp.1,3, 25-28 Jan. 2015.
- [7] Bolic, M.; Rostamian, M.; Djuric, P.M., "Proximity Detection with RFID: A Step Toward the Internet of Things," Pervasive Computing, IEEE , vol.14, no.2, pp.70,76, Apr.-June 2015.
- [8] Eugster, P.; Sundaram, V.; Xiangyu Zhang, "Debugging the Internet of Things: The Case of Wireless Sensor Networks," Software, IEEE, vol.32, no.1, pp.38,49, Jan.-Feb. 2015.
- [9] Senouci, Mustapha Reda, et al. "WSNs deployment framework based on the theory of belief functions." Computer Networks 88 (2015): 12-26..
- [10] Palattella, M.R.; Accettura, N.; Vilajosana, X.; Watteyne, T.; Grieco, L.A.; Boggia, G.; Dohler, M., "Standardized Protocol Stack for the Internet of (Important) Things," Communications Surveys & Tutorials, IEEE , vol.15, no.3, pp.1389,1406, Third Quarter 2013, doi: 10.1109/SURV.2012.111412.00158.
- [11] Sanchez, Luis, et al. "SmartSantander: IoT experimentation over a smart city testbed." Computer Networks 61 (2014): 217-238.
- [12] Jian An, Xiaolin Gui, Wendong Zhang, Jinhua Jiang, Jianwei Yang, Research on social relations cognitive model of mobile nodes in Internet of Things, Journal of Network and Computer Applications, Volume 36, Issue 2, March 2013, Pages 799-810, ISSN 1084-8045, <http://dx.doi.org/10.1016/j.jnca.2012.12.004>.
- [13] Zheng Yan, Peng Zhang, Athanasios V. Vasilakos, A survey on trust management for Internet of Things, Journal of Network and Computer Applications, Volume 42, June 2014, Pages 120-134, ISSN 1084-8045, <http://dx.doi.org/10.1016/j.jnca.2014.01.014>.
- [14] Satyanarayanan, M.; Simoens, P.; Yu Xiao; Pillai, P.; Zhuo Chen; Kiryong Ha; Wenlu Hu; Amos, B., "Edge Analytics in the Internet of Things," Pervasive Computing, IEEE , vol.14, no.2, pp.24,31, Apr.-June 2015, doi: 10.1109/MPRV.2015.32
- [15] Therese Sullivan, The Cutting-Edge of IoT, How does the IoT really change the future of commercial building operations?, November 2014, AutomatedBuildings.com, November 2014, <http://www.automatedbuildings.com/news/nov14/articles/buildingcontext/141030095606bldgcntx.html>
- [16] Akram Hakiri, Pascal Berthoua, Aniruddha Gokhale, Douglas C. Schmidt, Gayraud Thierry, Supporting End-to-end Scalability and Real-time Event Dissemination in the OMG Data Distribution Service over Wide Area Networks , Elsevier Journal of Systems and Software, 2013.
- [17] D. C. Schmidt, A. Corsaro, and H. V. Hag. 2008. Addressing the challenges of tactical information management in net-centric systems with DDS. Journal of Defense Software Engineering, 24–29.
- [18] OMG. 2007. Data Distribution Service for Real-Time Systems. v1.2.
- [19] <http://www.omg.org/spec/DDS/1.2/>
- [20] Li Da Xu, Wu He, Shancang Li, Internet of Things in Industries: A Survey, DOI 10.1109/TII.2014.2300753, IEEE Transactions on Industrial Informatics, 2014.
- [21] Q. Wei, S. Zhu, C. Du, "Study on key technologies of Internet of Things perceiving mine," Procedia Engineering, vol.26, pp.2326-2333, 2011.
- [22] Bo Cheng, Xin Cheng, Junliang Chen, Lightweight monitoring and control system for coal mine safety using REST style, ISA Transactions, In Press, Corrected Proof, Available online 8 August 2014.
- [23] ACATECH – Recommendations for implementing the strategic initiative INDUSTRIE 4.0. April 2013. http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Webs_ite/Acatech/root/de/Material_fuer_Sonderseiten/Industrie_4.0/Final_report_Industrie_4.0_accessible.pdf
- [24] IoT@Work, <https://www.iot-at-work.eu/> (Accessed April 2016).
- [25] Herman Storey (co - chair ISA 100), Rick Bullota and Daniel Drolet. The Industrial Internet of Things, <http://www.csemag.com/single-article/the-industrial-internet-of-things/c98837a0efec387d9fc14c2de0a3b2f.html> (Accessed April 2016).
- [26] Ovidiu Vermesan, Peter Friess, Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems, pp158, ISBN: 978-87-92982-73-5, River Publishers, 2013.
- [27] International Telecommunications Union, ITU-T Y.2060, Overview of the Internet of things, 2012.
- [28] OMG, Data Distribution Service (DDS) <http://www.omg.org/hot-topics/dds.htm> (Accessed April 2016).
- [29] Vasile-Gheorghita Gaitan, Nicoleta-Cristina Gaitan, Ioan Ungurean, A flexible acquisition cycle for incompletely defined fieldbus protocols, ISA Transaction journal, Elsevier, Volume 53, Issue 3, pp. 776-786, May 2014.
- [30] Yucel Cetinceviz, Ramazan Bayindir, Design and implementation of an Internet based effective controlling and monitoring system with wireless fieldbus communications technologies for process automation—An experimental study, ISA Transactions journal, Elsevier, Volume 51, Issue 3Pages 461–470, May 2012.
- [31] H. Perez, J.J. Gutierrez, "A survey on Standards for real-time distribution middleware" Journal ACM Computing Surveys, vol. 46, issue 4, March 2014, article no.49.

- [32] J. Bard and V. J. Kovarik. 2007. Software Defined Radio: The Software Communications Architecture. Wiley-Blackwell. ISBN: 0-47086-518-0.
- [33] M. Amoretti, S. Caselli, and M. Reggiani. 2006. Designing distributed, component-based systems for industrial robotic applications. In *Industrial Robotics: Programming, Simulation and Applications*, Low Kin Huat (Ed.). ISBN: 3-86611-286-6, InTech, DOI:10.5772/4892.
- [34] OMG. 2012. Corba Core Specification. v3.3. <http://www.omg.org/spec/CORBA/3.3/>, or <http://www.omg.org/spec/ZIOP/> (Accessed April 2016).
- [35] OMG. 2005. Realtime Corba Specification. v1.2. <http://www.omg.org/spec/RT/1.2/> (Accessed April 2016).
- [36] D. C. Schmidt. 2005. TAO Developer's Guide: Building a Standard in Performance. Object Computing, Inc.
- [37] M. Ryll and S Ratchev. 2008. "Application of the data distribution service for flexible manufacturing automation." *International Journal of Aerospace and Mechanical Engineering* 2, 3, 193–200.
- [38] M. Gillen, J. Loyall, K. Z. Haigh, R. Walsh, C. Partridge, G. Lauer, and T. Strayer. 2012. Information dissemination in disadvantaged wireless communications using a data dissemination service and content data network. In *Proceedings of the SPIE Conference on Defense Transformation and Net-Centric Systems*, Vol. 8405.
- [39] OMG. 2009. The Real-Time Publish-Subscribe Wire Protocol. DDS interoperability wire protocol specification. v2.1. <http://www.omg.org/spec/ DDSI/2.1/>
- [40] OMG. 2012. Extensible and Dynamic Topic Types for DDS. v1.0. <http://www.omg.org/spec/ DDS-XTypes/1.0/> (Accessed April 2016).
- [41] H. P´erez, J. J. Guti´errez, and M. Harbour. 2012. Adapting the end-to-end flow model for distributed Ada to the Ravenscar profile. *Ada Letters* 33, 1, 53–63.
- [42] <http://www.omg.org/spec/DAIS/1.1/PDF> (Accessed April 2016).
- [43] <http://www.omg.org/cgi-bin/doc?omg/11-08-01.pdf> (Accessed April 2016).
- [44] https://www.rti.com/docs/RTI_Security_Solutions.pdf (Accessed April 2016).
- [45] [http://www.primstech.com/opensplice/resources/documentation, OpenSplice_SecurityConfiguration_Guide_A131. Pdf](http://www.primstech.com/opensplice/resources/documentation,OpenSplice_SecurityConfiguration_Guide_A131.Pdf) (Accessed April 2016).
- [46] <https://opcfoundation.org/developer-tools/specifications-unified-architecture/part-2-security-model/> (Accessed April 2016).

A Survey of IPv6 Deployment

Manal M. Alhassoun

Department of Information Technology, King Saud
University
Riyadh, Kingdom of Saudi Arabia

Sara R. Alghunaim

Department of Information Technology, King Saud
University
Riyadh, Kingdom of Saudi Arabia

Abstract—The next-generation Internet protocol (IPv6) was designed to overcome the limitation in IPv4 by using a 128-bit address instead of a 32-bit address. In addition to solving the address the limitations, IPv6 has many improved features. This research focused to survey IPv6 deployment all around the world. The objectives of this survey paper are to highlight the issues related to the IPv6 deployment and to look into the IPv4 to IPv6 transition mechanisms. Furthermore, provide insight on the global effort around the world to contribute in IPv6 deployment. In addition, identify the potential solutions or suggestions that could improve the IPv6 deployment rate. In order to achieve the said objectives we survey number of papers on IPv6 deployment from different countries and continents.

Keywords—IPv4; IPv6; deployment; Internet

I. INTRODUCTION

The use of the Internet is growing over the time. Many day to day activities are depending on the Internet and lot of services are provided through the Internet too such as: social networking websites, search engines, video calls and many more. In order to reach these services; people use devices connected to the Internet such as computers, mobile phones, Personal Digital Assistants (PDA). All these devices are communicating with each other through the network using Internet Protocol (IP) where each device is assigned a unique IP address.

Internet Protocol version 4 (IPv4), has been the standard protocol over the Internet for more than 20 years, it provides over 4 billion IP addresses [1]. However, with the rapid growth of devices that can connect to the Internet and upcoming technologies, the limited 32-bit address space of IPv4 will not be able to cope with the internet. Some studies expected that by 2020 there will be 50 billion devices online which are 10 times more devices than IPv4 can handle [2].

Besides the shortage of IP addresses, IPv4 has several major weaknesses that made it difficult to keep up with the rapid growth of the Internet, including the following:

- Security: IPv4 does not provide any security like authenticating or data encryption when transmitting packets
- Network Congestion: packets are sent to all addresses in the network at the same time, this broadcast feature may cause overload and congestion on the network
- Packet Loss: IPv4 Time to Live (TTL) feature set time of expiry for the datagram. So if the data was not able

to reach the destination on its time, it will be expired and the receiver will request it again from the sender. This delay and multiple resending of packets are not sufficient for real time data.

- Data Priority: the IPv4 cannot recognize the type of data being transmitted, so it cannot prioritize the transmission of high priority data like video streaming and others[3]

In general, the scarcity of IPv4 address is considered as a major limitation of IPv4 addressing system, thus various techniques used to bridge the gap and extend the life of the existing IPv4 infrastructure such as "Network Address Translation" (NAT) and "Classless Inter Domain Routing" (CIDR) (which are described later in this paper). However, these techniques have their own drawbacks.

For solving the problem, IETF (Internet Engineering Task Force) offered a new Internet protocol for the next generation called IPv6. IPv6 extends the address space from 32-bit to 128-bit. By doing this, it provides about four times larger address space than IPv4. This huge number of address will be sufficient to satisfy the need of IP addresses in the future.

This Internet protocol does not only solve the problem of the address space, but also includes many other features such as:

- Streamlined header format: some IPv4 header fields were removed or made optional in IPv6. The aim of this change is to lower the cost of packet processing and to reduce the bandwidth cost despite the increased size of the IPv6 addresses, as shown in Fig. 1 [7].
- Address auto-configuration: the main usage of the auto-configuration feature in IPv6 is to facilitate the large number of hosts. With this feature, any device connected to the network can easily discover it at any location and get a new globally unique IPv6 address.
- Improved Quality-of-Service (QoS): "Flow label" component in the IPv6 header insures fastest delivery and more efficient performance. This is done by specifying the route of the IPv6 packet till it reach its destination and preventing it from going through unnecessary/bad network route.
- Built-in security: IPv6 requires the support of IPsec in order to provide a standard-based solution to satisfy the security needs in the network and to provide more improbability between IPv6 implementations.

- Better support for mobility: mobility is one of the requirements of IPv6 which enable the roaming between different networks. This is done using a global notification when you leave a network to enter the other one. [4], [5], [6].

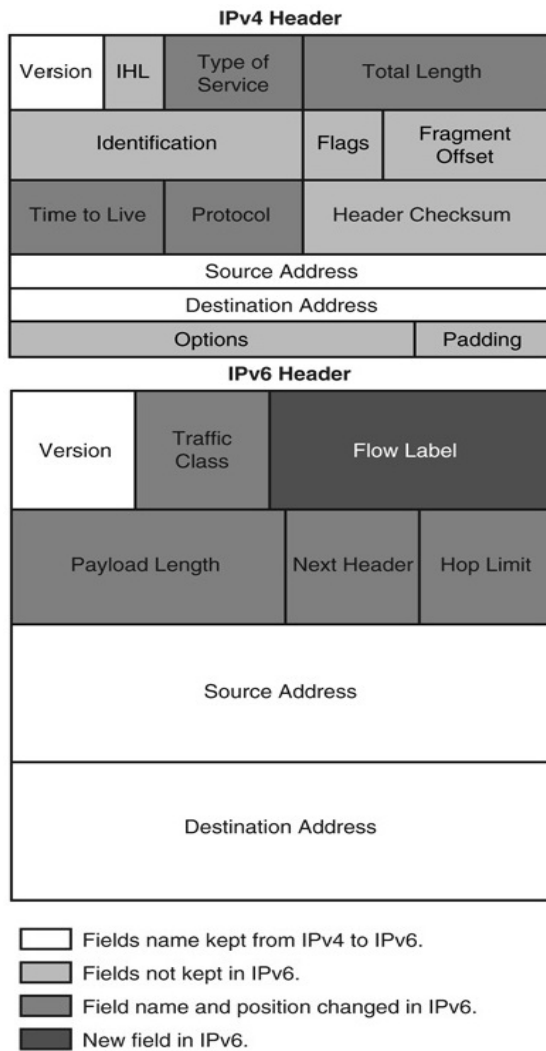


Fig. 1. IPv4 and IPv6 Headers

The rest of this document is structured as follows: Section II is dedicated to several challenges related to IPv6 deployment, Section III contains the critique of existing work and Section IV explains the future work direction. Finally, conclusions are drawn in Section V.

II. MAJOR RESEARCH CHALLENGES IN THE AREA

In this section, we outline several problems related to the IPv6 deployment. The techniques used to avoid or delay the complex process of migration to IPv6, the consequences of delay in IPv6 deployment and the hurdles encountered in IPv6 deployment.

A. Problems Related to Extend IPv4 Lifetime

The IPv4 is widely used and deployed in most of the internet architecture which make the transition to IPv6 very

risky and challenging. Consequently, various techniques have been used to prolong the lifespan of IPv4 and to avoid/delay the migration such as "Classless Inter-Domain Routing" (CIDR), "Network Address Translation" (NAT) and others.

Prior to the invention of IP addressing scheme "Classless Inter-Domain Routing" (CIDR), IP "Classful" Addressing was used, which divides the address space into different classes to determine the maximum potential size for a computer network. There were three major network classes (A,B,C), each Class A network can have over 16 million hosts, each class B network can have 65,535 hosts and each class C network can have 254 hosts. While the CIDR is based on variable-length subnet masking (VLSM), which allows a network to be divided into variously sized subnets, providing the opportunity to size a network more appropriately for local needs, hence, reduced the problem of wasted address space[8]. Unfortunately, this will fill the gap in the short-term only.

The idea of Network Address Translation (NAT) is to group many machines together and assign to them only one global unique address, while giving a hidden "private" address to each machine individually. The main benefit of this technique is to lower the number of IP addresses any organization may needs. However, this technique suffers from the filtering problem which reduces the network access performance [4].

As these techniques are inefficient in the long-term and the unallocated IPv4 addresses is expected to be exhausted soon or later, the ultimate solution is to move towards IPv6.

B. Problems Associated with Delaying IPv6 Deployment

Not adopting IPv6 may cause several issues. The future growth and global connectivity of the Internet will be negatively impacted. Individual users may not be able to reach IPv6-exclusive websites. Customers expecting or demanding IPv6-compatible or IPv6-enabled products and services may turn to the competitors for their needs, thus the companies will lose market share and revenues. Developers may not be able to introduce new services because they require an unusually high number of IP addresses (for instance, sensor and remote control systems being developed in many different industries including healthcare, automotive industry, disaster prevention, and many others). We may not be able to integrate applications and services because they may require IPv6 features which will not work in an IPv4 network [9].

C. IPv6 Deployment Challenges

Although the transition from IPv4 to IPv6 is necessary for the continuous running and growing of the internet, the IPv6 deployment growth rate is considerably slow. There are some challenges and factors that have contributed towards the slow rate of IPv6 deployment [4], [10], some of them are:

- IPv6 is not backwards-compatible with IPv4. The compatibility problem will create significant challenges for organizations as they move to IPv6.
- The benefits, strength points and necessity of IPv6 remain unknown for the end-users due to absence of campaigns or programs spreading the awareness about

IPv6. This will lower the end-users demand and need for moving towards IPv6.

- With the absence of IPv6 demands from end users, the service providers will not be able invest money in developing new hardware and software and charge their services for their customers.
- Many companies resist the migrating towards IPv6 since it will cost them money, time, resources and expertise.
- Many Internet Service Providers (ISP) and local operators view the IPv6 as a solution for providing more addresses to their clients. They still could not realize the real business value of IPv6.
- There is no enough participation and encourage from the Internet communities to move towards IPv6 which cause the limited number of IPv6 applications developed.
- Lack of practical experience

III. CRITIQUE OF EXISTING WORK

In this section, we survey some of the existing work related to IPv6 deployment in several countries/continents. In order to find the IPv4 to IPv6 transition mechanism used today. And to highlight the worldwide policies and initiatives used to promoting the IPv6 deployment. Additionally, identify the potential solutions and suggestions that could improve the IPv6 deployment rate around the world.

A. Transition to IPv6

The IPv6 and IPv4 are not compatible protocols, thus, the resources available over IPv6 cannot be reached by IPv4 node and vice versa. Fortunately, the network architecture allows the usage of these two protocols in parallel which make the transition from IPv4 to IPv6 done smoothly.

There are different strategies for transition from IPv4 to IPv6 such as [6]:

- Upgrade the whole network architecture along with the operating systems and applications to be IPv6 compatible. This option will guarantee the maximum benefit from all IPv6 features but it is very expensive.
- Wait for the last minute to deploy, which means nothing will be used from IPv6 features till IPv4 address exhaustion. This option is very risky and will lead to loss of market share.
- As a middle strategy, the deployment to IPv6 could be made at incremental levels, which guarantee the benefit from IPv6 features and at the same time it will lower the cost of deployment and allow the risk management.

From our research, we have found that many countries prefer to follow the incremental transition of IPv6 [6], [5],[11], and in order to continue working with their IPv4 infrastructure and to provide an final transition to an IPv6-only infrastructure, they have followed some mechanisms

1) *Using Both IPv4 and IPv6*: While the network infrastructure is being transmitted from IPv4 only, to IPv4 and IPv6 and at the end to IPv6 only; some services will be reachable over IPv6 only, while other services which still not updated to work with the two protocols will be reachable by IPv4 only. Therefore, this mechanism was implemented to allow network hosts to communicate by sending and receiving IPv4 and IPv6 packets at the same time. This requires the routers and applications to have the capability of Dual-Stack and the application layer needs to decide which protocol to follow.

To use the Internet layers for the two protocols on the same host, the host should be either a Dual IP layer architecture host or Dual stack architecture host.

- **Dual IP layer architecture**: This architecture has two separate Internet layers one for the IPv4 and the other is for the IPv6, while the Transport layer is implemented only once. The figure below illustrates the concept [6].

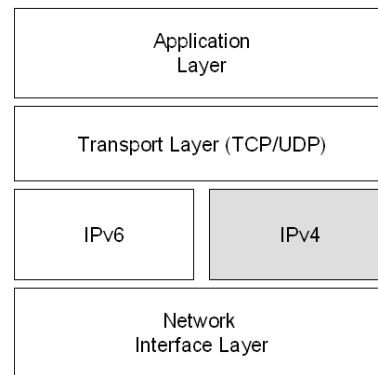


Fig. 2. Dual IP layer architecture

- **Dual stack architecture**: This architecture has two separate Internet layers one for the IPv4 and the other for the IPv6, and the Transport layer is implemented twice for each protocol. The figure below illustrates the concept [6].

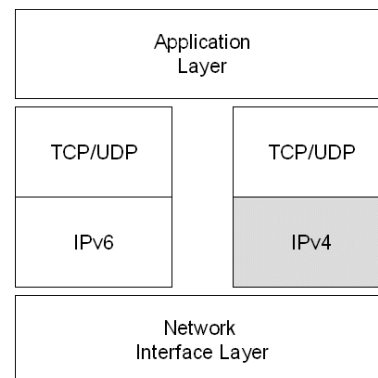


Fig. 3. Dual stack architecture

2) *Tunnelling*: The main usage of tunnelling is to enable a non-IPv6 device to communicate with other devices in IPv6

network. For example, a packet may be passed through an IPv6 network and suddenly reached one of the devices in its rout which was not upgraded to work with IPv6 packet. In this case the tunnelling is used. This is done by encapsulating the IPv6 packet into IPv4 capsule in order to be recognized and passed through IPv4 device normally. The tunnelling allows two different IPv6 networks to communicate even through IPv4 Networks. The changes on the IPv4 header are:

- The Protocol field value will be set to 41 to point to an encapsulated IPv6 packet.
- The Source and Destination fields are set to IPv4 addresses of the tunnel endpoints.

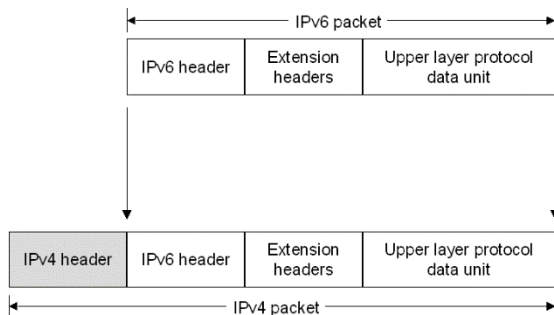


Fig. 4. IPv6 over IPv4 Tunnelling

3) *Protocol Translation*: This mechanism allows the communication between IPv6 only network and IPv4 only network. It uses translator which serves as an “interpreter” between the two networks. It just translates IPv4 packets to IPv6 Packets and vice versa. Some of the used protocol translation mechanisms are: Network Address Translation - Protocol Translation (NAT-PT6), Stateless IP/ICMP Translation (SIIT7), Bump-in-Stack (BIS8) and Bump-in-the-API (BIA9).

4) *DNS Infrastructure*: In this mechanism, the Domain Name System (DNS) infrastructure will be updated by populating the DNS servers with records to support IPv6 name-to-address and address-to-name resolutions.

B. Global Deployment Initiatives and Policies

Since the deployment of IPv6 has not been done at global scale in all countries, lot of deployment problems are still unknown. In order to uncover any problem related to the movement towards IPv6; government policies and global experiments and awareness campaigns have been set up such as World IPv6 day and Test-Bed.

1) *World IPv6 day*: the idea was started on the 6th of June 2012 when major Internet service providers (ISPs), home networking manufacturers and web companies around the world united to launch a new era for the Internet by collaborating in a 24-hour global experiment - World IPv6 Day. The goal of this day is to discover the problems and challenges regarding movement towards IPv6 and to find solutions for these problems. More than 300 organisations participating in the Day have enabled IPv6 for their products

and services and advertised both IPv4 and IPv6 addresses in the DNS [1], [2].

2) *Test-bed*: Test bed allows the examination of the IPv6 environment (development, testing, and deployment) without breaking the production network. There are various Test-bed experiments made around the world [4], [5], we will discuss some of them below.

- **Worldwide Test bed – The 6bone**: since many people and network manufactures and vendors started to implement and experiment the IPv6; this test bed was established to give more support for the evolution and development of IPv6. It was first started at 1996.
- **Indonesia Test Bed**: this started when the "Institute Teknologi Bandung" (ITB) in Indonesia was connected to the "Asian Internet Interconnection Initiatives" (AI3) in order to support the academic research on this field.

3) *International Policies*: a national IPv6 network was established to support the practical testing and usage of IPv6 technology by many developers, researchers and operators around the world. The network was established by many countries such as India, Korea and Japan. [6]. Some of the important examples are shown below:

- **India**: In 2004, the Minister of Communications and Information Technology declared the Ten Point Agenda to boost IT and communications, and includes the migration to IPv6.
- **Japan**: Japan believes that IPv6 is very helpful in leveraging the Internet to rejuvenate Japanese economy. Because of that, Japanese took a leadership to design a roadmap for IPv6 in 2000. Japanese government forced the incorporation of IPV6 and decided a deadline for upgrading all existing systems in both public and business sectors. On 2003, the Japanese government announce a tax credit program that eliminates the taxes from the purchase of any IPv6 routers.
- **South Korea**: In 2003, the South Korean Ministry of Information and Communication announced its funding to the IPv6 products and services as a promotion program.
- **China**: In 2003, the Chinese government started a plan to make their network fully operated on IPv6 by the end of 2005. The government issued licenses and assigned budget for the construction of the "China Next Generation Internet" (CGNI).

C. Solutions and Suggestions to Improve IPv6 Deployment Rate

In order to facilitate the rate of IPv6 deployment around the world, we have found many suggestions and solutions which based on the distribution of responsibilities [1], [12]. We have divided these responsibilities into three groups as follow:

1) *Government Organizations Responsibilities*:

- Establish a strategy programs to help in the essential deployment and development of IPv6 technologies and applications.
- Develop IPv6 awareness campaigns for the general populace in order to educate the people about IPv6 features and benefits, thus will create a motivation for switching to IPv6.
- Engage more in the global experiments of IPv6 keep updated on IPv6 activities around the world and understand the necessary requirements and needs to guarantee the readiness for transition at any time.

2) *Business Sectors Responsibilities:*

- Host content on IPv6 enabled websites
- Provide IPv6 enables hardware and software

3) *Civil Society and Academic Society Responsibilities:*

Build a demand and ensure competitive availability to IPv6 in both government and business sectors, build technical competency among universities and researchers by requesting more IPv6 services and products.

IV. CONCLUSION

Since 1978, IPv4 was deployed globally with the growth of the Internet, it has served as the Internet Protocol which is responsible of sending and receiving packets through the network. It uses 32-bit addresses, which limits the address space to about 4 billion addresses. Because of the demand of the growing Internet and the problems related to the limited address space in IPv4, IPv6 was developed. IPv6 provides much larger address space and it has lot of features. This protocol was designed to completely replace IPv4, but this will take several years before we completely migrate. The change is rather inevitable, therefore, lot of researches were made to find strategies and mechanisms for transmitting to IPv6 in an incremental level to maintain interconnectivity between the two protocols and allow these protocols to coexist without issues. The implementation and deployment of IPv6 is a challenge, risk and expensive job indeed, but it can be much easier and applicable with the good planning and the optimal choosing of implementation tools and methodologies.

V. DIRECTIONS FOR FUTURE WORK

Given the examination of IPv6 activities in various regions covered in this paper, the direction of future work would be to conduct similar analysis on the IPv6 deployment in Kingdom of Saudi Arabia (KSA). The aim of such work is to provide a comprehensive study of the current status of IPv6 in Saudi Arabia, the main reasons that make the transmission to IPv6 mandatory, the KSA migration plans and issues involved, the role of KSA in the global effort to deploy IPv6 and what

lessons Saudi Arabia can draw from deployment experiences acquired elsewhere.

Future studies also need to be carried out in perform a detailed analysis of a real life experiences of enterprises that had deployed IPv6. Identify what factors significantly contributed to adopt IPv6 in the enterprise. The challenges the enterprise faced during the deployment. Analyze the operational expenses and the risks associated with IPv6 transition efforts and identify the benefits it brings to the business. We hoped this would encourage other businesses to adopt IPv6.

Also, as future work it would be useful to test existing applications for IPv6 readiness. In addition, increase the awareness about IPv6 to the general public and why IPv6 is needed in practice and hopefully therefore increase the demand for IPv6 by customers. This most likely will raise the involvement of the Internet community in upgrading the IPv4 applications and services to IPv6.

REFERENCES

- [1] E. Agbaraji, F. Opara, and M. Airiguzo, "IPv6 Deployment Status, the Situation in Africa and Way Out," International Journal of Advances in Engineering & Technology (IJAET), Vol. 1, Issue 6, p. 315, Jan. 2012..
- [2] Worldipv6launch.org, 'World IPv6 Launch', 2015. [Online]. Available: <http://www.worldipv6launch.org/>. [Accessed: 12- Oct- 2015].
- [3] O. Babatunde, and O Al-Debagy, "A Comparative Review Of Internet Protocol Version 4 (IPv4) and Internet Protocol Version 6 (IPv6)", International Journal of Computer Trends and Technology (IJCTT), Vol. 13, No. 0, Jul. 2014.
- [4] R. Azlina, "IPv6 Deployment in Malaysia: The Issues and Challenges" SANS-GSEC White Paper, Apr. 2002.
- [5] A. Ferry, and T. Shin-ichi. "The Critical Needed of IPv6 Development in Indonesia." In Proc. of the IECEI, Japan Workshop. 2003.
- [6] S. Gurha, "IPv6 Deployment – Benefit & Opportunities in India with World-wide Experiences", International Journal of Advanced Technology in Engineering and Science (IJATES), Vol. 03, No. 02, Feb. 2015
- [7] D. Teare, and C. Paquet. Authorized Self-study Guide: Building Scalade Cisco Internetworks (BSCI). Cisco Press, 2007
- [8] R. Nagar, and A. Ali. "Survey and Study of Next Future Problems in IPV4 and IPV6 Created by Different Unreliable Network Issues," International Journal of Advance Research and Innovative Ideas in Education, Vol. 1, Issue 4, pp. 195-199, 2015.
- [9] Hagen, Silvia, Planning for IPv6, O'Reilly Media, Inc., 2011.
- [10] BlueCat. (2011). 6 Steps to IPv6 Readiness: A Practical Approach to Adopting IPv6 with IP Address Management (IPAM) [Online]. Available: http://www.mtechpro.com/2011/mconnect/december/eu/BlueCatNetworksWhitepaper-6_Steps_to_IPv6_Readiness.pdf.
- [11] Microsoft Corporation. (2008, February). Windows Server 2008, IPv6 Transition Technologies. Microsoft Corporation. [Online]. Available: <http://download.microsoft.com/download/1/2/4/124331bf-7970-4315-ad18-0c3948bdd2c4/IPv6Trans.doc>.
- [12] S. Limkar and R. Jha, "IPv6: Features, Current Deployment Scenario, Issues and Migration Status in India", in Proceeding of the International Conference on Software and Computing Technology, 2010, pp. 149–153.

Intelligent Image Watermarking based on Handwritten Signature

Saeid Shahmoradi

Department of computer, college of
Engineering technical
Bandar Abbas Branch, Islamic Azad
University
Bandar Abbas, Iran

Nasrollah Sahragard*

Department of Electrical and
Computer Engineering
University of Hormozgan,
Bandar Abbas, Iran

Ahmad Hatam

Department of Electrical and
Computer Engineering
University of Hormozgan,
Bandar Abbas, Iran

Abstract—With the growth of digital technology over the past decades, the issue of copyright protection has become especially important. Digital watermarking is a suitable way of addressing this issue. The main problem in the area of watermarking, is the balance between image transparency and resistance to attacks after watermarking, where an increase in either one of them will always cause a decrease in the other. Providing statistical and intelligent methods, is the most common way of optimizing resistance and transparency. In this paper, the intelligent method of genetic algorithm (GA) in watermarking will be examined and also the results of using this method will be compared with the results of a statistical SVD-based method. Also, by combining the issues of watermarking and authentication, a relatively higher security in these two issues can be achieved. In this scheme, the security of watermarking increases through the provision of a new method which is based on the combination of image watermarking with a person's handwritten signature. It must be mentioned that the section of signature recognition is implemented using neural networks. The results from implementing these two methods show that in this area, intelligent methods have a better performance compared to statistical methods. This method can also be used for tasks like passport or national identity card authentication.

Keywords—intelligent watermarking; genetic algorithm; neural networks; handwritten signature

I. INTRODUCTION

With the growth of digital technology over the past decades, sending and storing of electronic media have also increased, because copying data without any loss of quality and with a very small cost has become possible. Therefore, use of digital works without compliance with copyrights, document manipulation and use of forged documents, have found new dimensions. Use of traditional encryption systems made it possible that only a person possessing a key would be able to view an encrypted media text. But even in this condition, after data decryption, it will still be possible to use it illegally. Therefore, traditional encryption methods will not be efficient enough to prevent unauthorized use and malicious attacks. In such circumstances, intangible data embedding for prevention of unauthorized use, has a high commercial potential. Digital watermarking has been introduced to overcome this problem [1]. In watermarking methods, there are a series of requirements which need consideration. Of the most important requirements to mention are watermark transparency,

resistance, security and capacity. Here, watermarking transparency and resistance are more important than others. Transparency is the invisibility of the information hidden in an image, and resistance is the resistance of a watermark signal against various image processing techniques and intentional or inadvertent attacks. These two characteristics are in contrast with each other, that means, an increase in either one of them will always cause a decrease in the other. One of the factors effective at creating a balance between transparency and resistance, is the adaptive and optimized selection of watermark strength coefficient. Watermark strength coefficient, represents a watermark's injection into the host image. An increase in this coefficient, increases resistance and decreases transparency, and vice versa. The proposed algorithm in this scheme for intelligent watermarking is based on GA and the HVS. This scheme is an adaptive watermarking method in the area of the discrete cosine transform (DCT) for digital images [2].

II. GENETIC ALGORITHM (GA)

A genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. You can apply the genetic algorithm to solve problems that are not well suited for standard optimization algorithms, including problems in which the objective function is discontinuous, no differentiable, stochastic, or highly nonlinear [3].

Before a GA can be run, a suitable coding (or representation) for the problem must be devised. We also require a fitness function, which assigns a figure of merit to each coded solution. During the run, parents must be selected for reproduction, and recombined to generate offspring. The most common way to show chromosomes in genetic algorithms is binary strings. Decision variables are converted to binary form and then after these variables are joined together, a chromosome is created. This method is the most common coding method, but there are also other rapidly developing methods such as representing with real numbers.

Also a fitness function must be devised to assign a value to each coded solution. During the execution, parents are selected for reproduction, and are combined together via mating and mutation operators to produce new children. This process is repeated several times until the next population generation is produced. Next, this population will be investigated and if the convergence criteria are met, this process is ended [4].

III. INTELLIGENT IMAGE WATERMARKING IN THE PROPOSED SCHEME

The proposed algorithm in this scheme for intelligent watermarking is based on GA and HVS. This scheme is an adaptive watermarking method in the area of the discrete cosine transform (DCT) for digital images. In this method, the host image is classified into non-overlapping 8x8 blocks and watermark bits are embedded into the DCT coefficients of these blocks. To increase the security of this method, embedding locations are selected randomly. The selected blocks are classified into six different classes based on characteristics such as texture, brightness and proximity to edges. Also, a support vector machine (SVM) is used for the simulation of human visual system (HVS) for the classification of the blocks. One of the factors effective at creating a balance between transparency and resistance, is the adaptive and optimized selection of the watermark strength coefficient. Watermark strength coefficient represents the watermark injection into the host image. An increase in this coefficient, increases image resistance and decreases its transparency, and vice versa. Adaptive watermarking methods which are based on the HVS, determine the watermark strength coefficient in an adaptive way suited for the HVS [5].

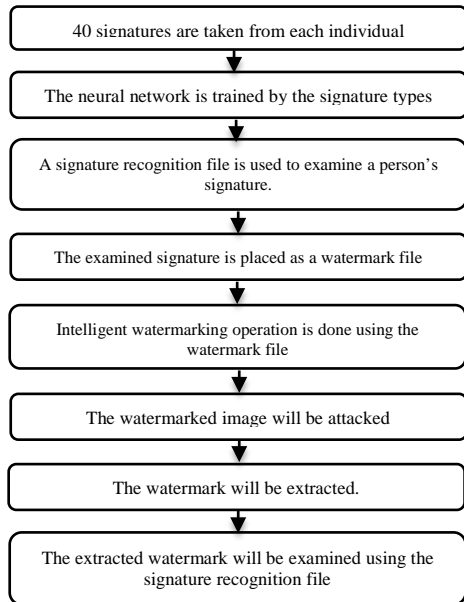


Fig. 1. A Flowchart Of The Overall Process Of The Proposed Scheme

A. Watermark embedding algorithm

The proposed watermark embedding method, is a blind watermarking approach in the area of the DCT. Therefore, watermark extraction does not require the original image. If the host image, is image 'I' with the dimensions of M*N and the

watermark image, is a binary image with the dimensions of * , to embed the watermark in the host image, first, image 'I' will be classified into non-overlapping 8x8 blocks, and a DCT is taken from each of them. Next, the DCT coefficients of each block, are adjusted in a zigzag form:

$$F_k = DCT(I_k) \quad 1 \leq k \leq k_t \quad (1)$$

F_k , includes the DCT coefficients of the k(th) block in host image 'I' which are adjusted in a zigzag form, and k_t shows the total number of 8x8 blocks $k_t = \left\lceil \frac{M}{8} \right\rceil \times \left\lceil \frac{N}{8} \right\rceil$. Each DCT block

$F_k(1)$	$F_k(2)$	$F_k(6)$	$F_k(7)$	$F_k(15)$	$F_k(16)$	$F_k(28)$	$F_k(29)$
$F_k(3)$	$F_k(5)$	$F_k(8)$	$F_k(14)$	$F_k(17)$	$F_k(27)$	$F_k(30)$	$F_k(43)$
$F_k(4)$	$F_k(9)$	$F_k(13)$	$F_k(18)$	$F_k(26)$	$F_k(31)$	$F_k(42)$	$F_k(44)$
$F_k(10)$	$F_k(12)$	$F_k(19)$	$F_k(25)$	$F_k(32)$	$F_k(41)$	$F_k(45)$	$F_k(54)$
$F_k(11)$	$F_k(20)$	$F_k(24)$	$F_k(33)$	$F_k(40)$	$F_k(46)$	$F_k(53)$	$F_k(55)$
$F_k(21)$	$F_k(23)$	$F_k(34)$	$F_k(39)$	$F_k(47)$	$F_k(52)$	$F_k(56)$	$F_k(61)$
$F_k(22)$	$F_k(35)$	$F_k(38)$	$F_k(48)$	$F_k(51)$	$F_k(57)$	$F_k(60)$	$F_k(62)$
$F_k(36)$	$F_k(37)$	$F_k(49)$	$F_k(50)$	$F_k(58)$	$F_k(59)$	$F_k(63)$	$F_k(64)$

is made of a combination of 64 coefficients. This is shown in figure 2-4.

Fig. 2. DCT coefficients adjustment method

In this proposed method, in order to increase the watermarking security, two keys of key1 and key2 have been used to determine the watermark embedding locations. It is assumed that the number of the watermark bits is lower than the 8x8 blocks in the host image $M_w * N_w \leq k_t$. Therefore, utmost one bit of watermark is embedded in each block. First, the DCT blocks are selected by the number of the watermark bits via key1, which are called B_k .

$$B_k \subseteq F_k \quad 1 \leq k \leq M_w * N_w \quad 1 \leq k \leq k_t \quad (2)$$

Next, via key2, in each selected block, one of the DCT coefficients in the intermediate frequency band will be selected for embedding the watermark bit. For the watermark strength coefficients to be minimized to the location of the DCT coefficients, only coefficients 11 to 15 will be used for this. In each DCT block, coefficients 11 to 15 will be used because we want to consider the minimum watermark strength coefficient for embedding so image transparency will be increased after watermarking. These coefficients are of the intermediate frequency band. The selected coefficient in block B_k will be called c_k . So, $B_k(c_k), 1 \leq k \leq M_w * N_w$, show the watermark embedding location. Now, in order to embed the watermark in the host image, first, we define $\tilde{B}_k(i)$ as follows:

$$\tilde{B}_k(i) = B_k(1) * R(i) \quad 1 \leq k \leq M_w * N_w \quad (3)$$

$\tilde{B}_k(i)$ is an approximate of $B_k(i)$ and is used as a reference value for watermark embedding and extraction. Ultimately, watermark embedding per bit is as follows:

$$\begin{aligned}
 & \text{if } w(k) = 0 \\
 & B_k(c_k) = \begin{cases} \text{Min}(B_k(c_k), \tilde{B}_k(c_k) - \alpha_k) \\ \text{Max}(B_k(c_k), \tilde{B}_k(c_k) + \alpha_k) \end{cases} \quad (4) \\
 & \text{if } w(k) = 1
 \end{aligned}$$

α_k is watermark strength coefficient in block B_k . After watermark embedding, a DCT reaction is taken from the resulting blocks so image I_w which is the watermark carrying image is obtained [6].

B. Watermark extraction algorithm

If image I' with the dimensions of $M * N$, is the image carrying the watermark, in order to extract the watermark, first, I' will be classified into non-overlapping 8×8 blocks and DCT is taken from each of them. In the next step, using key1, the blocks carrying the watermark (\tilde{B}_k), and then using key2, the coefficients in which the watermark is embedded (\tilde{c}_k), will be identified. Finally, watermark bits w will be extracted from each of the blocks carrying the watermark, which is shown below:

$$W(k) = \begin{cases} 1 & \text{if } B_k(c_k) \geq \tilde{B}_k(c_k) \\ 0 & \text{els} \end{cases} \quad (5)$$

C. Determining watermark strength coefficient

At this stage, the watermark strength coefficient will be adaptively determined. The adaptive determination of this coefficient, includes three stages:

- Extracting appropriate characteristics from the image blocks based on HVS characteristics.
- Categorizing the blocks into different classes based on the extracted characteristics.
- Determining the watermark strength coefficient for each class [5].

D. HVS-based characteristics extraction

Watermark embedding into a host image is actually the task of adding a weak noise to a strong signal, and as long as the noise power is below the just noticeable difference 1 (JND), the human eye cannot detect it. Studies have shown that the human eye is as follows:

- It has lower noise sensitivity in higher resolution groups.
- It has lower noise sensitivity in areas of an image with higher or lower brightness.
- It has lower noise sensitivity in areas with high texture, however, it has higher sensitivity to the proximity of edges.

Based on this, in a watermarking system, it is possible to consider different watermark strength coefficients for different areas of an image. For example, in noisy textures, a higher watermark strength coefficient can be used around the edges and in bright or dark areas. In this scheme, different image blocks are classified into six classes based on texture and

brightness, and different watermark strength coefficients are determined for each class. The six classes are as follows:

- T1: blocks with a smooth texture and high brightness
- T2: blocks with a smooth texture and average brightness
- T3: blocks with a smooth texture and low brightness
- T4: blocks with edges
- T5: blocks with a relatively noisy texture (coarse texture)
- T6: blocks with a very noisy texture (fine texture)

In this scheme, four characteristics are used to classify the blocks into the aforementioned classes: brightness level, entropy, variance and contrast. In other words, the four characteristics are extracted from each block, then based on them, the class corresponding to each block is determined. The brightness level of each block is obtained by averaging from its pixel values. When the brightness of a block is low or high, it is possible to use a higher watermark strength coefficient. Second moment or variance has a especial importance in texture description and is a criterion of intensity contrast which can be used to determine the relative smoothness of images. Entropy is a criterion which shows dispersion in pixels intensity. Thus, in images having a texture, it has higher values. Brightness level, variance and entropy are obtained from the equations below:

$$\text{Mean} = \sum_{i=0}^{l-1} (z_i P(z_i)) \quad (6)$$

$$\text{Var} = \sum_{i=0}^{l-1} ((z_i - \text{Mean})^2 P(z_i)) \quad (7)$$

$$\text{Ent} = - \sum_{i=0}^{l-1} (P(z_i) \log_2 P(z_i)) \quad (8)$$

In these equations, z_i represents brightness and l shows the number of gray levels. $P(z_i)$ represents the probability of brightness intensity z_i of an image, and equals the ratio of the number of pixels with brightness z_i to total image pixels. Texture criteria which are only calculated using a histogram, like the above criteria, do not carry any information about the relative location of pixels relative to each other. This point is important when describing a texture, and a method for embedding this type of information in a texture analysis process is to consider the relative location of each pixel. For this purpose, in this scheme, a second order elemental differential moment or in other words, a contrast criterion is used. For this purpose, first, a co-occurrence matrix is formed for the image block. A co-occurrence matrix is a matrix whose elements show the number of times when the pixel pairs of z_i and z_j are placed in a particular position relative to each other. Here, each of the elements of this matrix (g_{ij}), shows the number of times when the pixel pairs of z_i and z_j are placed next to each other in each other's eightfold neighborhood. The contrast criterion for each image block is calculated as follows:

$$\text{Cont} = \sum_{i=0}^{l-1} \sum_{j=0}^{l-1} ((z_i - z_j)^2) (P(z_i - z_j)) \quad (9)$$

z_i and z_j represent brightness intensity and $P(z_i, z_j)$ is the probability of them being placed next to each other which is calculated as follows:

$$P(z_i, z_j) = (g_{ij}/n) \quad (10)$$

n represents the total number of pixel pairs placed next to each other, and in other words, equals the sum of the co-occurrence matrix elements. As is clear from equation (10), if the difference in the brightness intensity of adjacent pixels is higher, the contrast will be higher. Variance, entropy and contrast are used to determine block texture. When entropy and variance have small values, the image is smooth. If entropy is high, if variance has a large value, it is edge or otherwise texture. In images with a very noisy texture (fine texture), contrast has large values [7].

E. Classification of blocks via SVM

In order to determine watermark strength coefficients α_k in embedding blocks B_k , the 4 characteristics of brightness level, variance, entropy and contrast are extracted, next, based on them, the blocks are classified into six classes. In this scheme, for block classification, a trained support vector machine (SVM) is used [8]. In previous papers where block classification has been used to make watermarking adaptive [6], mostly, classic classification methods based on thresholding for extracted characteristics are used. Such classifier will not provide the necessary accuracy, because the HVS is an entirely complicated and nonlinear system. Therefore, here, an SVM is used to find a relationship between the mentioned characteristics and their corresponding classes because of its high capability at simulating the HVS and its high capabilities at learning, nonlinear generalization and approximation. To use SVM in data classification, first, it must be trained. For this purpose, in this scheme, 1000 image blocks with different texture and brightness are used as training samples. At the end, the trained SVM will be used to determine the class of the B_k blocks.

$$\text{Class}(B_k) = \text{SVM}(\text{mean}_k, \text{var}_k, \text{ent}_k, \text{cont}_k) \quad (11)$$

$$\text{Class}(B_k) \in \{t1, t2, t3, t4, t5, t6\} \quad (12)$$

SVM, represents the process of data classification via the SVM. Since the number of the classes is more than 2, the one-against-one method is used for data classification ($\text{mean}_k, \text{var}_k, \text{ent}_k, \text{cont}_k$). The characteristic vector is extracted from block B_k which includes brightness level, variance, entropy, and contrast [9].

F. Determining watermark strength coefficients via a GA

After classifying the blocks into the six mentioned classes, the suitable watermark strength coefficient for each class is determined using a GA. We define vector S as below which shows the watermark strength coefficient in each class.

$$S = [s1, s2, s3, s4, s5, s6] \quad (13)$$

s_i represents the watermark strength coefficient in the blocks of t_i class $S_i = S(t_i)$. The objective is to find the optimal S via the GA. If S^* is the vector resulting from the GA, the values of α_k in the section of watermark embedding are obtained from the equation below:

$$\alpha_k = S^*(\text{Class}(B_k)) \quad (14)$$

The GA starts with a primary population of S vectors. For every S vector in the population, the resulting fit functions of a number of them are selected as the parents. After carrying out crossover and mutation operations on the parents and production of children, a new generation including a new population of S vectors will be formed. This process continues until the population is converted into an optimal vector. At the end, the vector with the greatest fit value in the last generation, will be the ultimate response of the algorithm (S^*). In order to define the fit function, the watermarking transparency and watermark's resistance will be calculated for every S vector. Therefore, for every S vector, a watermark embedding operation is carried out. Next, in order to evaluate the watermarking transparency, the similarity level of the watermarked image to the host image will be measured. Now the simplest and most useful criteria for measuring similarity are MSE and PSNR which are calculated from the equations below and the reason for it is the simple calculation and inclusion of an understandable physical meaning.

$$MSE = (1/M * N) \sum_{i=1}^M ((I(i, j) - \hat{I}(i, j))^2) \quad (15)$$

$$PSNR = 10 * \log_{10}((255^2)/MSE) \quad (16)$$

After evaluating the watermarking transparency level through the calculation of the structural similarity index between the host image and the watermarked image, for evaluating the watermark resistance level, a number of attacks and image processing operations are carried out on the watermarked image. Then the watermark is extracted from the images, and the BCR index for each of them, between the original watermark and the extracted watermark is calculated.

$$BCR(w, w^2) = (\sum_{i=1}^{M_w} \sum_{j=1}^{N_w} XOR(w(i, j), w^2(i, j))) \quad (17)$$

The maximum value for BCR is 1, and this value is obtained only if the original watermark and the extracted watermark are the same. The fit function for each S vector is defined as below:

$$F(s) = PSNR(I, \hat{I}) + 1/p \sum_{k=1}^p BCR_k(w, w^2) \quad (18)$$

I is the host image and \hat{I} is the watermarked image which was obtained via the separation of the watermark in image \hat{I} through the use of an S vector. P shows the number of attacks applied to image \hat{I} . W is the original watermark and \hat{W} is the watermark extracted from each of the attacked images. In this scheme, two attacks including a median filtering and a JPEG compression with a quality factor of 40 for evaluation of the watermark resistance level in the fit function [14] were used.

IV. SIGNATURE RECOGNITION IN THE PROPOSED SCHEME

The main difference between simple watermarking and handwritten signature-based watermarking is that in simple watermarking, an ordinary logo is placed behind the host image but in handwritten signature-based watermarking, a person's handwritten signature is used as the watermark. In the proposed scheme, signature recognition is done based on static signature characteristics. The purpose in this project, was network training for three people, which, 40 handwritten

signatures of every user have been taken and scanned. Next, the waste areas around the signature are removed. For this purpose, after acquiring the image size, the sum of the elements of its rows and columns is calculated and finally, a figure is shown from an area whose sum of row and column is lower than the row and column of the image, because the background of the signatures was white and they are considered as being monochrome and the signatures were black and zero. We place the cut image in that same primary matrix and then alter its size and change it into a 70×20 matrix. Now, we change the figure matrix into a column shaped one and change it into a 1400×1 matrix. Now using a Perceptron Neural Network with 3 nerves and with a target vector as [100,010,001] which represent the first, second and third person, respectively, these signatures are trained to the network. The interesting point is that the watermark acceptable for the host image in the watermarking in the proposed scheme, must be of a 70×20 size. Also, since the person's signature may be written in different colors, before the network training, the signatures are changed from the RGM mode into binary mode, so the watermark image finds a 0 and 1 mode. Therefore, during the network training, the signatures are primarily changed into the considered size and are given gray color and then will be trained to the network.

After the network is trained, one sample of a person's handwritten signature will be changed into the mentioned size and color and is hidden behind the host image as the watermark. After that, this watermark can be extracted before or after the attacks, and the signature's originality can be verified using the trained network, and it will be possible to verify whose signature it is. The flowchart of the process of network training for handwritten signatures is shown in figure (3). The interesting point about watermarks is that, given that a watermark is a person's handwritten signature, after it is altered to the size acceptable for watermarks that is 70×20 , only a part of the signature is selected and used as a watermark. An example of this is shown in figure 6.

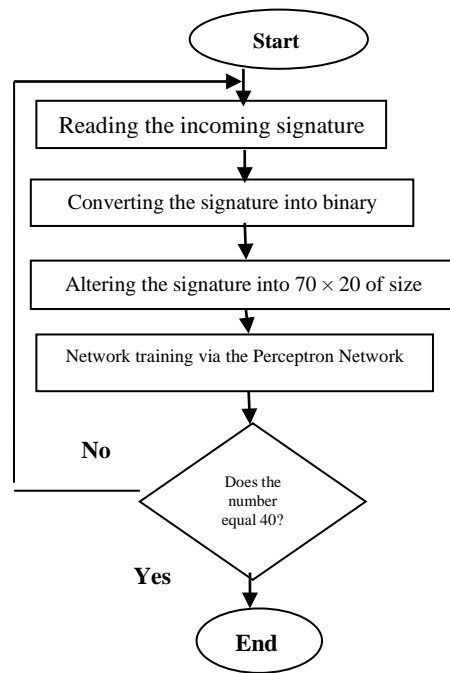


Fig. 3. flowchart of the signature recognition network training process

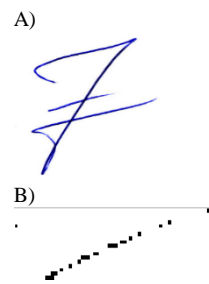


Fig. 4. A) original handwritten signature, B) signature prepared as the watermark

V. SVD STATISTICAL METHOD

Recently, Singular Values Decomposition (SVD) in watermarking has become very popular due to the matrix characteristics in its attractive mathematics. SVD is one of the useful tools in Linear Algebra with various uses in image compression, watermarking and other signal processing areas. The main idea of this method is that the SVD of the cover image is calculated and then special values are modified for

BCR watermark correctness in SVD	BCR watermark correctness in GA	Image	attack
0.50	0.96	Peppers	JPEG40
0.53	0.95	Mandril	
0.49	0.98	Cameraman	
0.61	0.98	Lena	Median filtering
0.65	0.97	Mandril	
0.62	0.98	Cameraman	

watermark embedding. If 'A' is an NXN matrix, after that, the SVD of this matrix can be defined as follows:

$$A=U*S*V^T \tag{19}$$

In this equation, U and V are orthogonal matrixes and S is a diagonal matrix. The diagonal elements of matrix S, are special values and follow the characteristic below:

$$S (1, 1) > S (2, 2) > S (3, 3) > > S (n, n) \tag{20}$$

The SVD-based watermarking scheme is provided by Kumar Gupta et al. (2010), this method is a combination of watermarking in the area of DWT and SVD. DWT, decomposes an image into four frequency groups: LL, HL, LH and HH. LL represents low frequency, HL and LH represent average frequency and HH represents high frequency. LL shows approximate details, HL shows horizontal details, LH provides vertical details and HH highlights the diagonal details of an image. In this suggestion, the HH group is selected for watermark embedding, because it includes more accurate details and provides a small contribution to image energy. Therefore, watermark embedding does not affect the image perceptual correctness [10]. The proposed scheme is based on an idea of replacing special values from the HH group with special values from the watermark. The special values of the HH group of various experimental images, have shown that these values are between 84 and 173. If a watermark is selected in a way whose special values are placed inside the given range, after that, the energy of the special values of the watermark will almost equal the special values of the HH group. For this purpose, the replacement of special values does not affect the perceptual image quality and the energy level of the HH group. In this method, the size of the host image, was considered as 512 × 512 and the watermark size was considered as 256 × 256, which during the process of watermarking, the size of the watermark changes and becomes equal with the size of the HH group.

VI. RESULTS OF IMPLEMENTATION OF THE PROPOSED INTELLIGENT SCHEME AND COMPARING IT WITH THE SVD METHOD

Based on the results of Gupta et al. (2010) which investigated the resistance and transparency of watermarked images via the SVD method, we will compare this method with the proposed method. In terms of image resistance and transparency, the GA method is better than the special values method. Experiments were performed on Bubble, Lena and Cameraman images. In table (1), the image transparency level in the two methods after watermarking is shown, and in table (2), image resistance to attacks is shown. The attacks tested in both methods include JPEG40 compression attacks and median filtering [10].

TABLE I. IMAGE TRANSPARENCY AFTER WATERMARK EMBEDDING

PSNR transparency in SVD	PSNR transparency in GA	Image
43.33	68.98	Peppers
50.67	64.74	Mandril
47.42	68.57	Cameraman

TABLE I. WATERMARK CORRECTNESS LEVEL AFTER VARIOUS ATTACKS

As is seen in table 1, image transparency in the GA method is more than the SVD statistical method. Also, based on the results in table 2, we will see that the correctness level of the watermark after various attacks, is better in the GA method than the SVD method. The reason for the superiority of the intelligent making GA scheme compared to the SVD statistical method at image transparency is that in the GA method, the watermark strength coefficients for embedding are considered as the smallest values possible, and smaller bits of the host image are placed for covering the watermark. For example, in the following, we will investigate the Cameraman image before and after watermarking and after the attacks in the intelligent method.

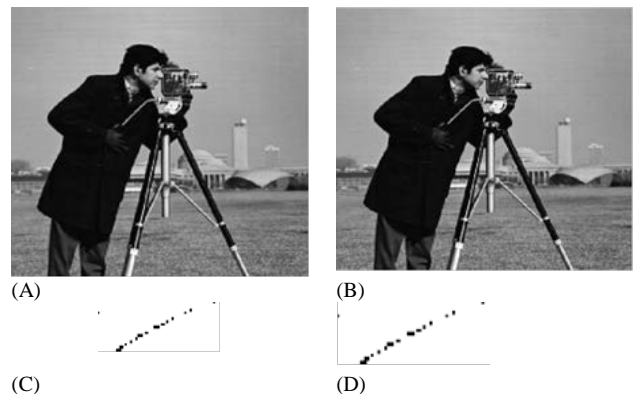


Fig. 5. A) main image, B) watermarked image, C) main watermark, D) extracted watermark

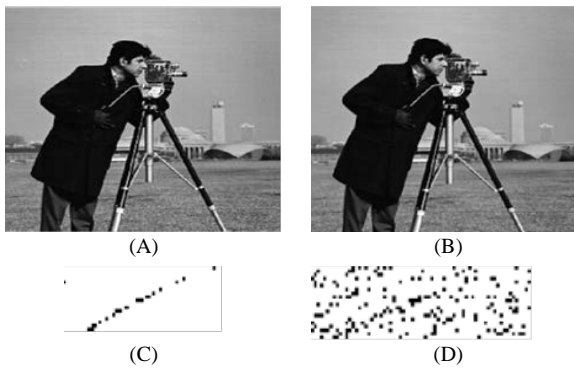


Fig. 6. A) watermarked image, B) image after median attack, C) watermark extracted before the attack, D) watermark extracted before median attack

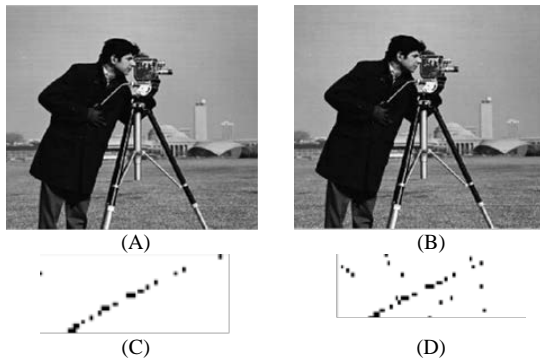


Fig. 7. A) watermarked image, B) image after a JPEG40 attack, C) watermark extracted before the attack, D) watermark extracted before a JPEG40 attack

VII. CONCLUSION

Biometrics are the most secure identity verification factors in the world of information and communication, which also provide improved accuracy, speed, ease and reduce costs. Here, authentication via a person's signature is one of the most widely used authentication methods and it is because of its importance in e-commerce security issues where the real identity of the person who signs documents is discovered. Watermarking is the act of hiding a data (watermark) in a covering data (cover) in order to exercise of the right of ownership on the cover. The difference between a normal watermarking and a handwritten signature-based watermarking is that in normal watermarking, one image with one sign is selected and used as the watermark, but in the method of handwritten signature-based watermarking, the person's signature is embedded in the image and prevents all attempts to copy it.

In images which are watermarked, the two topics of resistance to attacks (or any changes) and quality are important. The balance between resistance and quality can be properly achieved through the adjustment of the embedded parameters. In intelligent watermarking, evolutionary computing algorithms such as genetic algorithms and particle swarm optimization, automatically find the embedded parameters which are the results of optimization for each image. For intelligent watermarking, techniques such as fuzzy logic, genetic algorithms and artificial neural networks are used.

In the proposed scheme, a GA is used for intelligent watermarking. Watermark embedding in this scheme, was performed based on the DCT method and in an adaptive way and the results from implementing it compared to the results of the SVD statistical method, provided better transparency and resistance for the image. Signature recognition in the proposed scheme was performed via the perceptron neural network and given the static characteristics of the signature, the network was trained and it was tested. Since the person's signature was considered as the watermark, despite the addition of many noises after the attack to the watermark image, still, the signature recognition was properly done and the signatures of all the individuals were recognized without any mistakes. Based on what is mentioned above, handwritten signature-based intelligent watermarking will be a method that provides more security in authentication systems and copyright issues and is done with optimal quality and resistance.

VIII. RECOMMENDATIONS FOR FUTURE RESEARCH

In the future, the three items below can be applied to the proposed scheme to improve it:

1) In signature recognition, for more accurate recognition, instead of the static characteristics of a signature, its dynamic characteristics can be used.

2) The proposed method for handwritten signature-based intelligent watermarking can be tested with more attacks and its results can be investigated.

3) The present scheme was implemented on black and white images, and to further develop it, it can be used on black and white images as well.

REFERENCES

- [1] Z. Toni, providing an optimal method for image hiding using a learning algorithm, (Master's thesis), Sharif University of Technology, Iran.2010.
- [2] E. Vellasques, R. Sabourin, E. Granger, A high throughput system for intelligent watermarking of bi-tonal images. *Applied Soft Computing*, vol. 11, no. 8, pp. 5215-5229, 2011.
- [3] D.E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, 1989.
- [4] J.H. Holland, "Adaptation in Natural and Artificial Systems", The University of Michigan Press, Ann Arbor, MI, 1975.
- [5] P. Nafisifard, Darhami, A. M. Latif, Adaptive watermarking of digital images based on machine learning. *Intelligent Systems in Electrical Engineering*, vol. 2, no. 4, pp. 47-64, 2011.
- [6] Z. Lu, S. Jiang and H. Dong, "Adaptive watermarking algorithm based on human visual system", *Journal of Harbin Institute of Technology*, vol. 35, no. 2, pp. 138-141, 2003.
- [7] R. C. Gonzalez and R. E. Woods, "Digital image processing, 3rd Ed", Prentice-Hall, 2008.
- [8] J. Huang, Y. Q. Shi and R. Yao, "Adaptive image watermarking based on block classification", *Journal of Image and Graphics*, vol. 4, no. 8, pp. 640-643, 1999.
- [9] F. Meng, H. Peng, Z. Pei, and J. Wang, "A Novel Blind Image Watermarking Scheme Based on Support Vector Machine in DCT Domain", *IEEE International Conference on Computational Intelligence and Security*, 2008.
- [10] A. Kumar Gupta, S.M. Raval, A robust and secure watermarking scheme based on singular values replacement. *Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar 382 007, India*, vol. 37, no. 4, pp. 425-440, 2010.

Fuzzy Risk-based Decision Method for Vehicular Ad Hoc Networks

Riaz Ahmed Shaikh
Computer Science Department,
King Abdulaziz University,
Jeddah, Saudi Arabia

Abstract—A vehicular ad hoc network (VANET) is an emerging technology that has the potential to improve road safety and traveler comfort. In VANETs, mobile vehicles communicate with each other for the purpose of sharing various kinds of information. This information is very useful for preventing road accidents and traffic jams. On Contrary, bogus and inaccurate information may cause undesirable things, such as automobile fatalities and traffic congestion. Therefore, it is highly beneficial to consider risk before vehicle takes any decision based on the received information from the surrounding vehicles. To overcome these issues, we propose a new risk-based decision method for vehicular ad hoc networks. It determines a risk-based the three key elements: 1) application type and sensitivity level, 2) vehicle context and 3) driver's attitude. This paper also provides theoretical analysis and evaluation of the proposed method, and it also discusses the applications of the proposed model.

Keywords—Ad hoc networks; Decision methods; Risk management; Trust management; Vehicular Networks

I. INTRODUCTION

In vehicular ad hoc networks (VANET), vehicles and roadside units (RSUs) communicate with each other by sharing various types of information, such as safety-related warning messages. This collaborative communication is very helpful in avoiding road accidents and traffic congestion. However, accuracy and reliability of the received information need to be evaluated first before vehicle takes any decision, such as change of lane or road, etc. For this purpose, various researchers [1-7] recommend the use of trust management schemes in VANETs.

Many trust management methodologies [1-7] exist in the literature that mainly focuses on evaluating the trustworthiness of the received data in VANETs. However, less attention is given to the decision-making. From VANET's perspective, decision making is very critical, because consequences of any wrong decision could be disastrous, such as fatal road accidents. In most of the trust management approaches, the decision logic is straight forward. For example, a message having the maximum trust value is accepted [4], or any message that is coming from the vehicle whose reputation is good is accepted [5]. Some researchers used simple majority or weighted majority voting techniques [6] for the decision making. However, most of the existing works do not consider risk in the decision making in VANETs.

A risk can be defined as a level of uncertainty resulting from the potential for a negative outcome. In the VANETs, it can be viewed from an application sensitivity level. In general, VANET applications can broadly be categorized into three types: 1) safety applications, 2) traffic efficiency applications, and 3) infotainment applications. A message that is generated by a *safety application* may have higher sensitivity level than a message which is generated by a *traffic efficiency application*. Similarly, messages that are generated by a *traffic efficiency application* have higher sensitivity level than an *infotainment application* messages. So, the application which has high sensitivity level imposes high risk. Furthermore, two or more conflicting messages belonging to the same application may impose various risks depending upon its consequences. Other than the application type, risk can also be evaluated by considering other contextual attributes that can affect the physical and mental condition of the driver, etc. [8].

Consider two messages $m1$ ("Road X is icy at intersection X-Y") and $m2$ ("Road X is dry at intersection X-Y") are received by the vehicle from the surrounding vehicles. With the help of any existing trust management scheme, assume that the trust value of $m1$ is 0.4 that is higher than the trust value of the message $m2$ (e.g., 0.3). So, $m1$ is considered more trustworthy than $m2$. However, from the sensitivity level perspective of the application, trust value of $m1$ is less than the acceptable threshold value (e.g., 0.5). Now, the problem is: How much risk we are willing to take to accept some relatively trustworthy message? Or in other words: "If the trust level of the message is low with respect to the sensitivity level of the application, should we accept that message?" This is the critical and challenging problem, which is commonly overlooked during the decision-making process in vehicular networks.

In this work, we have proposed a new fuzzy risk-based decision method for vehicular ad hoc networks. In this method, the risk is estimated based on three key factors: 1) application type and its sensitivity level, 2) vehicle context and 3) driver's attitude. The use of these three dimensions in the risk estimation makes our proposal unique. Furthermore, it also provides a better degree of completeness. We have provided solutions to measure risk in both qualitative and quantitative manner, which will increase its applicability in various scenarios. Additionally, we have also provided a data set that can be used to create benchmarks for the comparison purposes.

The rest of the paper is organized as follows. Section 2

discusses the related work. Section 3 describes the proposed fuzzy risk-based decision method. Section 4 contains analysis and evaluation of the proposed method. Section 5 depicts the applications of the proposed method. Finally, section 6 concludes the paper and outline future work directions.

II. RELATED WORK

Decision making is usually described as a mental process of selecting the best one from judging multiple options or alternatives [9]. The decision-making process commonly involves the following five activities [9][10][11]:

- decision problem identification;
- relevant information collection and verification;
- identification of the decision substitutes;
- foresee the consequences of decisions;
- decision making;

In this work, we are mainly focusing on the decision-making process of trust models that are primarily used for making reliable decisions. However, existing trust models do not consider the consequence of a wrong decision, which we are referring here to a *risk*. Description about the decision-making process of existing state-of-the-art trust models is given below.

Shaikh and Alzahrani [1] have proposed a trust management scheme for VANETs. The unique thing about their method is that it evaluates trust in an environment where identities of nodes are not known. It works in three phases. Firstly, each receiver node will measure a confidence value for each received message that is based on four parameters: 1) Time closeness, 2) Location closeness, 3) Time verification and 4) Location verification. Secondly, it calculates the trust value for each unique message. Finally, it takes the decision. The decision process is comprised of two steps: First, it selects the message which has the highest trust value. Second, if the selected message's trust value is exceeding the least acceptable threshold value, then the message will be accepted. Otherwise, it will be discarded.

Cohen's *et al.* [2] proposed trust method first measured the confidence value for each received report that is based on various factors, such as, history, time, location and role. After that, the method will take the decision. For this purpose, they adopted majority-based trust model. If the majority confidence is greater than the acceptable threshold and the number of reports is greater than the pre-defined variable '*n*,' then the method will follow the advice in the report. Else, it will follow the advice of the report with the highest confidence value.

Mármol and Pérez [5] have proposed a trust and reputation method for vehicular networks. In their model, whenever a node receives a message, it first assesses the reputation of the sender. It is measured based on three factors: 1) history, 2) recommendation from the neighbor vehicle, and 3) recommendation from the central authority via roadside units (RSUs). Based on the reputation score, a sender is classified into one of the three trust levels: 1) untrusted, 2) +/- trust, and 3) trusted. These levels are represented as fuzzy sets. After

that, receiver node takes the decision. If the sender node belongs to first set (not trusted), then the message is rejected. If it belongs to a second set (+/- trust), then the message is considered reliable with tunable probability. However, it will not be broadcasted or forwarded to any other node. If the node belongs to a third set (trusted), then the message is accepted.

Wei and Chen [7] have proposed an adaptive decision-making method for improving the efficiency of the trust management system of VANETs. The objective of their work is to make the quick accurate decision. The decision-making process will trigger in two cases: 1) when the number of messages received is more than the specific threshold (M_{\max}^{rsu}), or 2) when the time delay is exceeding the specific threshold (t_{wait}^{rsu}). Once the decision-making process triggered, it will take the decision by looking at the overall trust value of the event. If the trust value is greater than the given threshold value (T_{thld}), then the positive decision will be taken. Else, the method will consider that message as untrustworthy one.

U. F. Minhas *et al.* [12] have proposed a multi-faceted trust modeling framework for VANETs. In that framework, authors have incorporated the concepts of role-based trust, experience based trust, and majority-based trust. Among these three aspects, first two are used to select the advisors. Based on the recommendation of the selected advisors, the system takes the decision by adopting majority voting technique.

Tajeddine *et al.* [13] have proposed a framework which focuses on trust, reputation, and privacy-preserving trust system. In their paper, authors have shown that group decision and trust calculation of the received message are useful to increase the security in the VANET. Through this framework, the privacy of the users is respected using group decision organized by group management. Also, it provides security through trust and reputation. From this model, many attributes of security affecting trust calculation can be established using the group decision.

Huang *et al.* [14] have shown that simple voting for decision-making, leads to oversampling and gives wrong results using their proposed research. From this, they claimed that decision making based on trust management in mobile ad hoc networks is not suitable to VANET.

According to [15], authors have mentioned that intelligent transport system is maintainable through the appropriate signal quality which improves the decision-making ability in the VANET. Through the appropriate channel, decision-making, and solutions based on signal interference, strength and quality can be improved. The *TrafficInfo* algorithm is used to minimize the risk based on collisions and adjusts the number of dissemination reports included in the transmission. Risks are reduced through the adaptive control of transmission size [16]. The quality of the route which influences with VANET depended on the multi-metric routing decisions is proposed in [17].

Yang *et al.* [18] have proposed a dynamic three-layer reputation evidence decision and management mechanism, which combine with Dempster-Shafer evidence integration mechanism to distinguish selfish nodes and the risks which are suspect collusion vehicle nodes. A hierarchical Reputation

Evidence Decision System (REDS) based on the Dempster-Shafer evidence theory is defined to establish the reputation management which increases the reputation accuracy in the VANET. Through this theory, the degree of trust can be calculated.

Fernandes *et al.* [19] have proposed a decentralized reputation system for vehicular networks (RS4VANET). The objective of this system is to guarantee the proper operation of a data dissemination application in the presence of malicious nodes. The proposed system follows an optimistic approach and uses various techniques to assess the trustworthiness of vehicles. For the decision-making, authors have adopted the concept of voting schemes.

III. PROPOSED SOLUTION

As discussed earlier, risk can be derived from the application sensitivity level. Other than the application, different other factors could also be used in risk estimation as shown in Fig. 1. Our proposed risk assessment framework is composed of three factors: 1) Application type, 2) Vehicle context and 3) Driver's attitude.

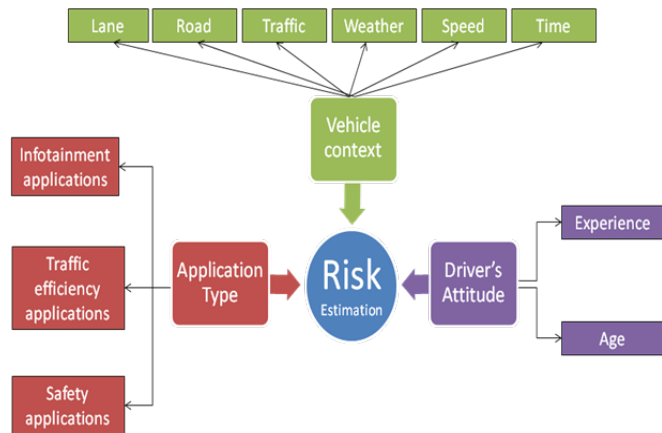


Fig. 1. Risk Estimation factors for VANET

The USA National Institute of Standards and Technology (NIST) [20] has defined risk as follows:

$$\text{Risk} = \text{Threat likelihood} \times \text{Impact} \quad (1)$$

At a high-level, we adopt the same definition. In our context, we determine the threat likelihood based on the vehicle context as well as driver's attitude and the impact based on the application sensitivity level.

A. Impact

The impact can be derived from a sensitivity value of the application. The higher the sensitivity value of the application, the higher the impact will be. There are two possible cases in which the impact can be derived: when the sensitivity value of an application is available or when it is unavailable.

Case 1: when the sensitivity value of an application is known

Let us assume that the sensitivity value [0, 10] is assigned to each application by the vendor. Assigning sensitivity value (S_v) from a specific range to an application is a subjective

matter. Therefore, we decided to use fuzzy logic to derive an impact level. This process is composed of two steps. In the first step, we determine the level of impact based on the sensitivity value, and in the second step, we determine the impact value in a quantitative manner by applying a mapping function.

Let us assume that there are three levels of impact: Low, Medium, and High, which can be considered as three sets. As compared to the classical set theory, the operations on fuzzy sets are based on the membership functions, which are typically linear and often take the shape of a triangle, trapezoid, or L [21]. The objective of the membership function is to determine the degree of truth that the element (i.e. sensitivity value) belongs to the particular set (i.e. low, medium or high).

In this work, we used Trapezoidal-shaped membership function. The reason for using this function is that it increases the flexibility, and it also allows an 'interval of values' that maximized the individual membership functions [22]. For example, Fig. 2 shows the trapezoid shape for low, medium and high fuzzy sets for sensitivity values. Mathematically, the membership functions that are shown in this figure are specified as below.

$$f^{\text{Low}}(S_v) = \max(\min(S_v + 1, 1, 3 - S_v), 0) \quad (2)$$

$$f^{\text{Medium}}(S_v) = \max(\min(S_v - 2, 1, 7 - S_v), 0) \quad (3)$$

$$f^{\text{High}}(S_v) = \max(\min(S_v - 6, 1, 11 - S_v), 0) \quad (4)$$

Note that the key feature of the fuzzy set is that there is no hard rule how boundaries of membership functions are defined. These can be set by taking input from the experts.

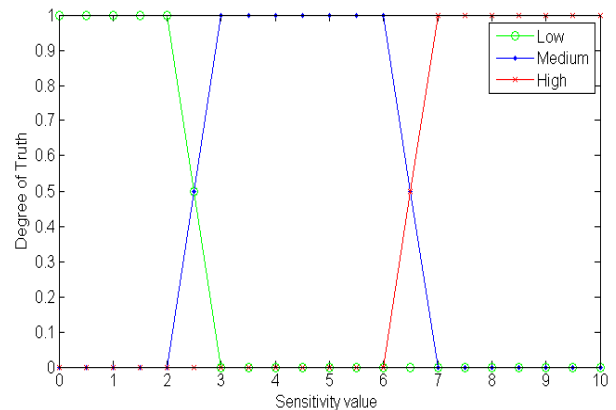


Fig. 2. Impact determination example using fuzzy set

After determining the impact level, we applied the following function to derive an impact value in a quantitative manner.

$$\text{Impact}(S_v) = \begin{cases} 0 & f^{\text{Low}}(S_v) == 1 & //\text{low impact} \\ 1 & f^{\text{Medium}}(S_v) == 1 & //\text{medium impact} \\ 2 & f^{\text{High}}(S_v) == 1 & //\text{high impact} \\ 2 & \text{else} & //\text{default value} \end{cases} \quad (5)$$

In this function, we mapped the low, medium and high 'impact levels' with 'impact values' zero, one, and two respectively. However, other values could also be used by keeping the following condition intact.

$$v(\text{Low}) < v(\text{Medium}) < v(\text{High}) \quad (6)$$

Where v represent the value. Note that in this function (Eq. 5), the default value is same as the highest value. The reason for adopting this pessimistic approach is to avoid any undesirable things that may occur in the vehicular networks.

Case 2: when the sensitivity value of an application is unknown

In the absence of sensitivity value, the impact can be derived from the application type. As discussed earlier, there are three types of applications: infotainment applications, traffic efficiency applications, and safety applications. Based on the application type, we can determine the sensitivity level, and that will be used to determine impact.

The sensitivity level of infotainment applications, traffic efficiency applications, and safety applications can be assigned as low, medium and high respectively as shown in Fig. 3.

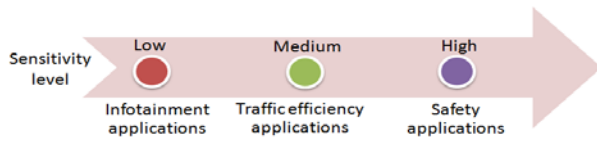


Fig. 3. Sensitivity level of VANET's applications

To estimate impact in a quantitative manner, we first assign numbers to the sensitivity levels of the application in the following manner.

$$S_1 = \{\text{Low} = 0, \text{Medium} = 1, \text{High} = 2\} \quad (7)$$

Note that different labels and numbers could also be used. Formally, we can define the relationship between the impact and S_1 in the following manner.

$$\text{Impact} \propto S_1 \quad (8)$$

The above relationship shows that the impact is directly proportional to the sensitivity level of the application (S_1).

B. Threat Likelihood

Threat Likelihood can be measured based on the vehicle context (V_{context}) and driver's attitude (D_{attitude}).

$$\text{Threat likelihood} = \beta_1 V_{\text{context}} + \beta_2 D_{\text{attitude}} ; \quad (9)$$

$$\beta_1 + \beta_2 = 1$$

Where β_1 and β_2 represent the weight values for V_{context} and D_{attitude} parameters respectively.

1) Vehicle context

Vehicle context can be determined based on the following factors:

- a) Lane (L_e), e.g., straight, curve, winding, Uphill, Downhill, Intersection, Corner.
- b) Road (R_d), e.g., dry, wet, icy.
- c) Traffic (T_c), e.g., a car in front, car on left, car on right, a car in rear.
- d) Weather (W_r), e.g., clear, raining, snowing, foggy, windy.

e) Speed (S_d), e.g., accelerating, decelerating.

f) Time (T_e), e.g., day, night, dusk, dawn.

Table 1 shows the fuzzy risk levels for the values of the above-mentioned factors.

TABLE I. VEHICLE CONTEXT PARAMETERS

Lane (L_e)	Weather (W_r)	Time (T_e)	Traffic (T_c)	Road (R_d)	Speed (S_d)
Straight (L)	Clear (L)	Day (L)	Car in front (H)	Dry (L)	Accelerating (H)
Curve (H)	Raining (L)	Night (H)	Car on left (M)	Wet (M)	Decelerating (L)
Winding (M)	Snowing (H)	Dusk (M)	Car on right (M)	Icy (H)	
Uphill (M)	Foggy (H)	Dawn (M)	Car in rear (H)		
Downhill (M)	Windy (M)				
Intersection (H)					L = Low Risk M = Medium Risk H = High Risk
Corner (H)					

Based on these factors of vehicle context, we can determine the risk in the following manner:

$$V_{\text{context}} = \max(w_1 L_e, w_2 R_d, w_3 T_c, w_4 W_r, w_5 S_d, w_6 T_e) ; \quad (10)$$

$$\sum_{i=1}^6 w_i = 1$$

Where w_i represents the weight value for the specific parameter.

2) Driver's attitude

To derive driver's attitude, we use two factors in our model: 1) Age (A) and 3) Experience (E).

$$D_{\text{attitude}} = \max(\alpha_1 A, \alpha_2 E) \quad (11)$$

Age and experience play a vital role in the driving performance. Younger drivers and old drivers are mostly involved in an accident more than middle age drivers. There is no consensus regarding higher threshold age for young drivers and lower threshold age for old drivers. As stated in [23], 65 is a most commonly accepted age for defining the older driver and it is an age where accident rate begins to increase. Upper limit defined for the younger drivers is 25 [24]. We can determine the risk based on the age factor in the following way.

$$A = \begin{cases} \text{Low} & \text{if } 35 < \text{Age} \leq 75 \\ \text{Medium} & \text{if } 25 < \text{Age} \leq 35 \\ \text{High} & \text{if } \text{Age} \leq 25 \text{ OR } \text{Age} \geq 75 \end{cases} \quad (12)$$

We can determine the risk based on the experience factor in the following way.

- Low, when the driver has experience of more than 30 years.
- Medium, when the driver has experience of more than ten years and less than 30 years.
- High, when the driver has experience of fewer than ten years.

Formally, we can define these rules in the following way.

$$E = \begin{cases} \text{Low} & \text{if } Experience \geq 30 \\ \text{Medium} & \text{if } 10 < Experience < 30 \\ \text{High} & \text{if } Experience \leq 10 \end{cases} \quad (13)$$

C. Qualitative Risk estimation

As mentioned in Equation 1, the risk is the product of threat likelihood and impact. This formulation gives us the quantitative value of the risk. In some scenarios, we may want to determine risk in a qualitative manner. In such situations, we need some mapping function that can be used to convert quantitative values in a qualitative manner.

Let us assume that the maximum value of a risk is n and the number of risk levels is k . Then we can use the following mapping function.

$$\text{Risk level} = \begin{cases} \text{level} - k & (k-1)n/k < Risk \leq n \\ \vdots & \vdots \\ \text{level} - 2 & n/k < Risk \leq 2n/k \\ \text{level} - 1 & 0 \leq Risk \leq n/k \end{cases} \quad (14)$$

Example: Assume that the maximum risk value is two, and there are three risk levels (low, medium and high). In this scenario, the above-mentioned mapping function will be simplified in the following way.

$$\text{Risk level} = \begin{cases} \text{High} & 2 \times 2/3 < Risk \leq 2 \\ \text{Medium} & 2/3 < Risk \leq 2 \times 2/3 \\ \text{Low} & 0 \leq Risk \leq 2/3 \end{cases} \quad (15)$$

IV. THEORETICAL ANALYSIS AND EVALUATION

As mentioned earlier, we have determined risk based on the product of impact and threat likelihood. We derived the impact based on the sensitivity level of the application and threat likelihood based derived from vehicle context and driver's attitude. To find out the impact of various parameters on the cumulative risk, we developed a small program which generates 100 different scenarios as shown in Table 2 (See Appendix).

For each scenario, first, we derived the impact from the sensitivity level [1-10] by implementing Equations 2 to 5. The Fig. 4 shows the impact for each scenario. One can see that the impact value fluctuates between 0 and 2, where 0 means low and 2 means high impact.

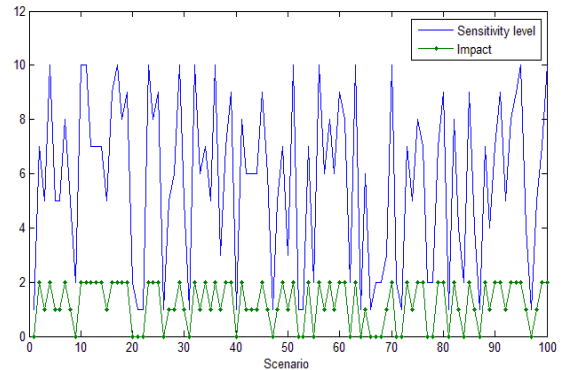


Fig. 4. Impact Analysis

To derive the threat likelihood, first, we calculate the vehicle context by implementing Equation 10 and driver's attitude by implementing Equation 11. Vehicle context is shown in the form of area chart (as shown in Fig. 5) which illustrate cumulative totals using number over each scenario. Note that, in this analysis; as indicated in the caption of Fig. 5, the weight value assign to each parameter are not same.

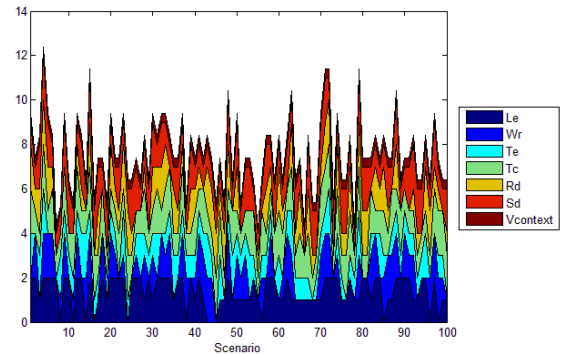


Fig. 5. Vehicle Context Analysis ($w_1 = w_4 = w_5 = w_6 = 0.2$; $w_2 = w_3 = 0.1$)

The impact of Age and Experience on driver's attitude is shown in Fig. 6 that is derived by applying equations 11 to 13.

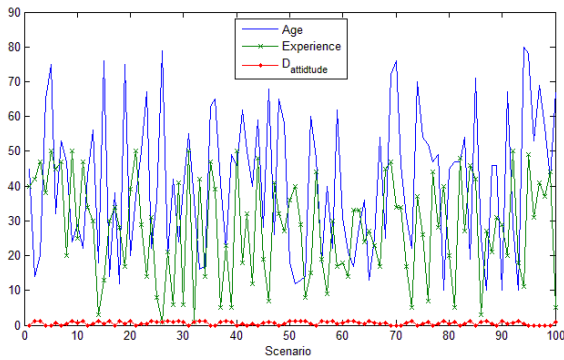


Fig. 6. Driver's attitude Analysis ($\alpha_1 = 0.6, \alpha_2 = 0.4$)

Threat likelihood is derived by implementing Equation 9. Fig. 7 shows the impact of vehicle context and driver's attitude on the threat likelihood. For the demonstration purposes, we have assigned a higher weight to the vehicle context ($\beta_1 = 0.6$) as compared to the driver's attribute ($\beta_2 = 0.4$).

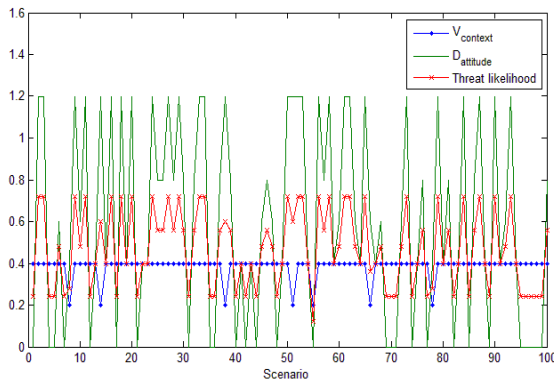


Fig. 7. Threat Likelihood Analysis ($\beta_1 = 0.6, \beta_2 = 0.4$)

Risk determination for each scenario based on threat likelihood and impact is shown in Fig. 8.

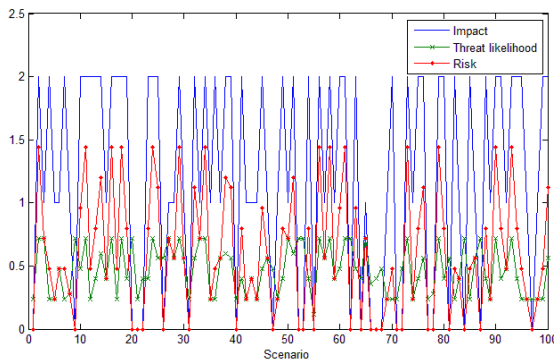
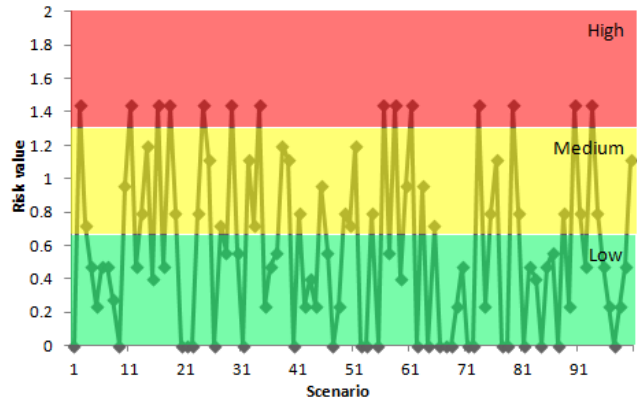
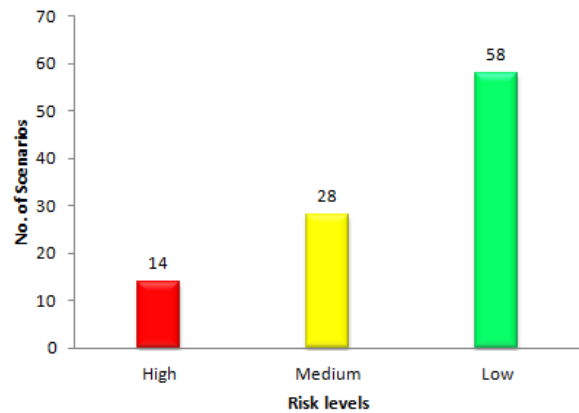


Fig. 8. Risk estimation based on Impact & Threat likelihood

To determine the risk in a qualitative manner, we applied the Equation 15. Results are shown in Fig. 9a. According to the results, risk values of 14, 28 and 58 scenarios are classified as high, medium and low respectively as indicated in Fig. 9b.



(a)



(b)

Fig. 9. Qualitative Risk estimation, (a) The risk value for each scenario, (b) Scenario classification w.r.t. risk level

V. APPLICATION OF THE PROPOSED METHOD

Our proposed method can assist existing trust management schemes of VANETS to make reliable decisions. For example, most of the trust management schemes [1, 2, 7], first measure the trust value based on various factors and then compare it with some pre-defined threshold value. If the trust value is greater than the threshold value, then the message is accepted. Otherwise, it will be discarded. Most of the researchers do not define how to calculate this pre-defined threshold value. The risk value that we are calculating in this work could be used as a threshold value. First order diagram of this concept is shown in Fig 10.

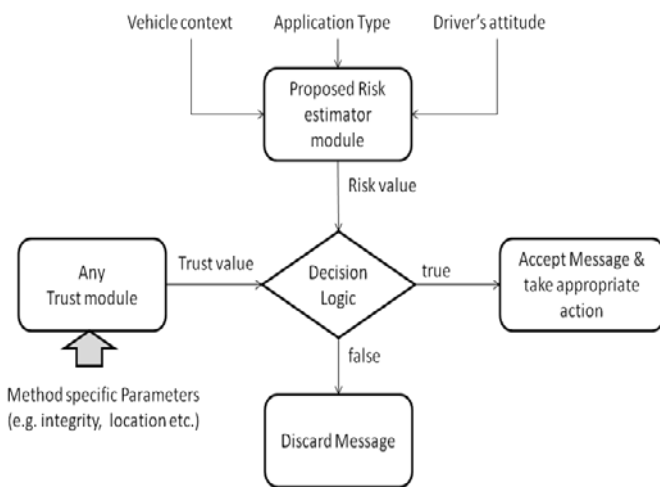


Fig. 10. Application of the Risk method

Another interesting application of our proposed model is its use in autonomous vehicles. Recently, many organizations like Google have developed prototypes of driverless cars. Google Inc. said their self-driving car will be ready by 2020 [25]. Driverless vehicles need similar techniques as standard programs which provide the decision dynamically. In this case, vehicle context and application type will be useful for direct applications because weather conditions and road situation are continuously changing with many parameters considered in our model. Although driver's attitude may not involve directly in the driverless vehicles, we can derive threat likelihood directly from the vehicle context. In the near future, we will see both regular and driverless vehicles on the roads. So, risk-based decision method will play a vital role. For example, when different drivers are driving their vehicles close to the driverless vehicles, this feature may be useful for providing safety and comfort to the passengers.

VI. CONCLUSION AND FUTURE WORK

Traditional decision methods are not suitable for vehicular ad hoc networks. Due to its sensitivity, a risk must be considered in the decision-making process. Therefore, in this work, we proposed a new fuzzy risk-based decision method for vehicular ad hoc networks. In this method, we derived the risk based on the three key factors: 1) application type and sensitivity level, 2) vehicle context, and 3) driver's attitude. We measured risk both in qualitative and quantitative manner. Since benchmarks are not available for this domain, therefore, we have provided detailed data set (See Appendix). Based on this data set, we have conducted a theoretical analysis and evaluation of the proposed method. We hope that this will be helpful for the researchers in this field to come up with better solutions and make a comparison. In future, we would like to extend our model by adding more parameters like gender, traffic density, speed limits etc.

REFERENCES

[1] Shaikh, R. A., & Alzahrani, A. S., "Intrusion-aware trust model for vehicular ad hoc networks", *Security and Communication Networks*, vol. 7 (11), pp. 1652-1669, 2014

[2] Cohen, R., Zhang, J., Finsson, J., Tran, T., & Minhas, U. F., "A Trust-Based Framework for Vehicular Travel with Non-Binary Reports and Its

Validation via an Extensive Simulation Testbed", *Journal of Trust Management*, vol. 1(1), 2014.

[3] Huang, D., Hong, X., & Gerla, M., "Situation-aware trust architecture for vehicular networks", *Communications Magazine, IEEE*, vol. 48(11), pp. 128-135, 2010.

[4] Zhang, J., "Trust management for vanets: Challenges, desired properties and future directions", *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 3(1), pp. 48-62, 2012.

[5] Gómez Mármol, F., & Martínez Pérez, G., "TRIP, a trust and reputation infrastructure-based proposal for vehicular ad hoc networks", *Journal of Network and Computer Applications*, vol. 35(3), pp. 934-941, 2012.

[6] Ostermaier, B., Dotzer, F., & Strassberger, M., "Enhancing the security of local danger warnings in vanets-a simulative analysis of voting schemes", In the 2nd International Conference on Availability, Reliability and Security (ARES 2007), Vienna, Austria, April 2007, pp. 422-431.

[7] Wei, Y. C., & Chen, Y. M., "Adaptive Decision Making for Improving Trust Establishment in VANET", 16th Asia-Pacific Network Operation and Management Symposium (APNOMS), Taiwan, Sep 2014, pp. 1-4

[8] Vijey T., Alzahrani, A., & Qureshi, M. S., "Risk prediction system based on MIMO system for vehicle users", *Life science Journal*, vol. 10(4), pp. 3055-3061, 2013.

[9] Bohanec, Marko. "Decision making: A computer-science and information-technology viewpoint." *Interdisciplinary Description of Complex Systems*, vol. 7(2), pp. 22-37, 2009.

[10] Skinner, D.C., "Introduction to Decision Analysis", Probabilistic Publishing, Gainesville, 3rd Edition, 2009

[11] Clemen, R.T., "Making Hard Decisions: An Introduction to Decision Analysis", Duxbury Press, Pacific Grove, 1996

[12] Minhas U, Zhang J, Tran T, Cohen R., "Towards expanded trust management for agents in vehicular ad-hoc networks", *International Journal of Computational Intelligence: Theory and Practice (IJCTP)*, vol. 5(1), pp. 3-15, 2010.

[13] Tajeddine, A., Kayssi, A., & Chehab, A., "A Privacy-Preserving Trust Model for VANETs", 10th IEEE International Conference on Computer and Information Technology (CIT 2010), 2010, pp. 832-837.

[14] Huang, Z., Ruj, S., Cavenaghi, M., & Nayak, A., "Limitations of Trust Management Schemes in VANET and Countermeasures" 2011 IEEE 22nd International Symposium on personal, indoor and mobile radio communications, 2011, pp. 1228-1232

[15] el mouna Zhioua, G., Tabbane, N., Labiod, H., & Tabbane, S., "A Fuzzy Multi-Metric QoS-Balancing Gateway Selection Algorithm in a Clustered VANET to LTE Advanced Hybrid Cellular Network" *IEEE Trans. on Vehicular Technology*, vol. 64(2), pp. 804-817, 2015.

[16] Zhong, T., Xu, B., & Wolfson, O., "Disseminating Real-Time Traffic Information in Vehicular Ad-Hoc Networks", 2008 IEEE Intelligent Vehicles Symposium, Jun 2008, pp. 1056-1061.

[17] Wang, X. B., Yang, Y. L., & An, J. W., "Multi-Metric Routing Decisions in VANET", 8th IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, 2009, pp. 551-556.

[18] Yang, Y., Gao, Z., Qiu, X., Liu, Q., Hao, Y., & Zheng, J., "A Hierarchical Reputation Evidence Decision System (REDS) in VANET", *International Journal of Distributed Sensor Networks*, 2015

[19] Fernandes, C. P., de Simas, I., de Mello, E. R., & Wingham, M. S., "RS4VANETs - a decentralized reputation system for assessing the trustworthiness of nodes in vehicular networks," in 2015 International Wireless Communications and Mobile Computing Conference (IWCMC), pp.268-273, Aug. 2015

[20] Stoneburner G, Goguen A, Feringa A., "Risk management guide for information technology systems", Tech. Rep. SP 800e30. NIST; July 2002.

[21] Shang, K., and Hossen, Z., "Applying Fuzzy Logic to Risk Assessment and Decision-Making." Research Report sponsored by Casualty Actuarial Society, Canadian Institute of Actuaries, 2013, Available at: <http://www.soa.org/Files/Research/Projects/research-2013-fuzzy-logic.pdf>

[22] Miao, S., Hammell II, R. J., Hanratty, T., & Tang, Z., "Comparison of Fuzzy Membership Functions for Value of Information Determination",

Proceedings of the 25th Modern Artificial Intelligence and Cognitive Science Conference, Spokane, Washington, USA, April 2014, pp. 53-60

[23] Young, K., Lee, J. D., & Regan, M. A. (Eds.), "Driver distraction: Theory, effects, and mitigation" CRC Press, 2008
[24] McKnight, A. J., & McKnight, A. S., "The effect of cellular phone use

upon driver attention." Accident Analysis & Prevention, vol. 25(3), pp. 259-265, 1993.

[25] Halleck, T., "Google Inc. Says Self-Driving Car Will Be Ready By 2020". International Business Times, 15 Jan 2015, Available at: <http://www.ibtimes.com/google-inc-says-self-driving-car-will-be-ready-2020-1784150>

APPENDIX

TABLE I. DATA SET FOR RISK ESTIMATION

#	Sensitivity level	Vehicle context						Driver's attitude	
		Lane	Weather	Time	Traffic	Road	Speed	Age	Experience
1	1	Corner	Clear	Night	Car in rear	Icy	Decelerating	45	40
2	7	Corner	Foggy	Day	Car on right	Wet	Decelerating	14	42
3	5	Downhill	Raining	Night	Car on left	Icy	Accelerating	20	47
4	10	Curve	Snowing	Night	Car in front	Icy	Accelerating	65	38
5	5	Intersection	Snowing	Day	Car on right	Icy	Accelerating	75	50
6	5	Corner	Snowing	Day	Car in front	Wet	Decelerating	32	45
7	8	Curve	Clear	Day	Car on left	Dry	Decelerating	53	47
8	5	Straight	Windy	Night	Car on left	Dry	Decelerating	47	20
9	2	Corner	Foggy	Dusk	Car in rear	Dry	Accelerating	24	50
10	10	Corner	Raining	Dawn	Car on left	Dry	Accelerating	29	25
11	10	Straight	Windy	Dawn	Car in front	Dry	Decelerating	22	47
12	7	Curve	Foggy	Dusk	Car in rear	Wet	Decelerating	43	34
13	7	Intersection	Clear	Night	Car on left	Wet	Accelerating	56	30
14	7	Straight	Snowing	Night	Car on right	Dry	Decelerating	18	3
15	5	Intersection	Foggy	Dawn	Car in front	Icy	Accelerating	76	13
16	9	Straight	Clear	Dawn	Car in rear	Dry	Accelerating	14	30
17	10	Uphill	Clear	Dawn	Car on right	Icy	Accelerating	38	34
18	8	Curve	Foggy	Day	Car in rear	Dry	Decelerating	12	28
19	9	Downhill	Clear	Dawn	Car on left	Dry	Accelerating	75	17
20	2	Corner	Foggy	Dawn	Car on right	Icy	Decelerating	20	39
21	1	Corner	Windy	Dusk	Car on left	Wet	Decelerating	37	50
22	1	Corner	Raining	Day	Car in front	Wet	Accelerating	50	29
23	10	Corner	Windy	Night	Car on right	Icy	Decelerating	67	14
24	8	Straight	Raining	Dawn	Car in front	Icy	Decelerating	22	31
25	9	Intersection	Raining	Day	Car on right	Wet	Accelerating	36	8
26	1	Corner	Windy	Dawn	Car on left	Dry	Accelerating	79	1
27	5	Intersection	Clear	Night	Car on right	Dry	Decelerating	20	21
28	6	Downhill	Snowing	Dusk	Car in front	Wet	Decelerating	42	6
29	10	Corner	Raining	Dawn	Car on left	Dry	Accelerating	24	41
30	5	Downhill	Snowing	Dusk	Car on left	Icy	Accelerating	39	6
31	1	Corner	Clear	Night	Car in front	Wet	Decelerating	55	50
32	10	Corner	Foggy	Day	Car on right	Icy	Accelerating	36	1
33	6	Corner	Windy	Dawn	Car in rear	Icy	Decelerating	16	42
34	7	Curve	Foggy	Dawn	Car in front	Dry	Decelerating	17	14
35	5	Winding	Raining	Night	Car on left	Icy	Decelerating	63	47
36	10	Curve	Clear	Dawn	Car on left	Wet	Accelerating	65	39
37	3	Curve	Snowing	Dawn	Car in rear	Wet	Decelerating	41	5
38	7	Straight	Foggy	Dawn	Car on right	Dry	Decelerating	22	23
39	9	Corner	Clear	Dawn	Car on left	Icy	Accelerating	49	5
40	1	Uphill	Windy	Dawn	Car in rear	Wet	Decelerating	46	50
41	8	Curve	Snowing	Dawn	Car on right	Wet	Decelerating	62	18
42	6	Downhill	Foggy	Dawn	Car in front	Dry	Decelerating	50	32
43	6	Straight	Foggy	Night	Car on right	Icy	Decelerating	40	12
44	6	Straight	Foggy	Dusk	Car on right	Icy	Decelerating	59	48
45	9	Straight	Clear	Day	Car in rear	Icy	Decelerating	28	19
46	6	Uphill	Clear	Night	Car in front	Wet	Decelerating	68	7
47	1	Downhill	Clear	Day	Car on left	Icy	Decelerating	26	41
48	5	Intersection	Foggy	Dawn	Car on right	Icy	Accelerating	65	32
49	7	Uphill	Clear	Night	Car in front	Dry	Decelerating	58	27
50	3	Uphill	Snowing	Dusk	Car on right	Icy	Accelerating	18	36
51	10	Downhill	Windy	Dawn	Car on right	Dry	Decelerating	12	40
52	1	Winding	Foggy	Dusk	Car on right	Dry	Accelerating	13	29

53	1	Winding	Clear	Night	Car in rear	Dry	Accelerating	14	8
54	7	Curve	Clear	Dawn	Car in rear	Dry	Decelerating	60	15
55	2	Winding	Raining	Day	Car on right	Wet	Decelerating	48	44
56	10	Straight	Foggy	Dusk	Car in front	Dry	Decelerating	18	19
57	6	Corner	Raining	Dawn	Car on left	Icy	Accelerating	40	9
58	8	Intersection	Foggy	Dawn	Car on right	Wet	Decelerating	22	30
59	6	Corner	Raining	Day	Car in rear	Dry	Accelerating	62	17
60	9	Uphill	Raining	Night	Car in rear	Icy	Decelerating	31	18
61	8	Uphill	Windy	Day	Car in rear	Wet	Decelerating	21	14
62	2	Intersection	Foggy	Dusk	Car on right	Dry	Accelerating	17	33
63	10	Winding	Foggy	Night	Car in rear	Icy	Decelerating	28	33
64	1	Downhill	Clear	Dawn	Car on left	Wet	Accelerating	36	24
65	6	Winding	Clear	Dawn	Car in front	Icy	Decelerating	13	27
66	1	Winding	Raining	Dawn	Car on left	Dry	Decelerating	25	23
67	2	Downhill	Raining	Dusk	Car in rear	Icy	Accelerating	54	17
68	2	Downhill	Clear	Day	Car on right	Wet	Accelerating	25	45
69	3	Winding	Raining	Dawn	Car on left	Dry	Accelerating	72	47
70	10	Winding	Snowing	Dusk	Car in front	Wet	Accelerating	76	34
71	2	Intersection	Foggy	Dusk	Car in rear	Icy	Accelerating	46	34
72	1	Corner	Snowing	Night	Car in front	Icy	Decelerating	31	17
73	7	Curve	Clear	Day	Car on right	Dry	Accelerating	22	5
74	5	Intersection	Windy	Dusk	Car in rear	Icy	Decelerating	70	37
75	8	Uphill	Clear	Dusk	Car on left	Wet	Accelerating	54	26
76	7	Downhill	Raining	Day	Car in front	Wet	Accelerating	52	7
77	2	Intersection	Raining	Dusk	Car in rear	Icy	Decelerating	47	44
78	2	Downhill	Clear	Dusk	Car on left	Dry	Decelerating	49	28
79	7	Uphill	Snowing	Night	Car in front	Icy	Accelerating	10	40
80	9	Intersection	Raining	Day	Car on left	Icy	Accelerating	45	20
81	1	Corner	Raining	Day	Car on left	Icy	Accelerating	47	5
82	8	Winding	Snowing	Dawn	Car in front	Dry	Decelerating	47	48
83	4	Intersection	Foggy	Day	Car on right	Icy	Decelerating	54	27
84	2	Corner	Clear	Night	Car on right	Wet	Decelerating	19	46
85	9	Straight	Snowing	Dawn	Car in front	Icy	Decelerating	71	42
86	4	Uphill	Snowing	Day	Car on right	Wet	Accelerating	25	3
87	1	Downhill	Foggy	Dusk	Car in front	Dry	Decelerating	10	27
88	7	Curve	Foggy	Dusk	Car on left	Icy	Accelerating	46	21
89	4	Curve	Clear	Dawn	Car in front	Dry	Decelerating	46	31
90	7	Corner	Snowing	Dusk	Car on right	Dry	Decelerating	10	29
91	9	Curve	Windy	Dawn	Car on right	Dry	Accelerating	67	20
92	5	Downhill	Foggy	Dawn	Car on right	Wet	Accelerating	29	50
93	8	Winding	Raining	Dawn	Car on right	Wet	Accelerating	10	18
94	9	Curve	Clear	Dawn	Car on right	Dry	Accelerating	80	11
95	10	Intersection	Clear	Night	Car on left	Icy	Decelerating	78	49
96	4	Straight	Snowing	Day	Car on left	Icy	Decelerating	53	31
97	1	Curve	Foggy	Day	Car in rear	Wet	Accelerating	69	41
98	5	Straight	Snowing	Dawn	Car in front	Wet	Decelerating	57	37
99	7	Winding	Windy	Dawn	Car in front	Dry	Decelerating	42	44
100	10	Downhill	Raining	Day	Car in rear	Wet	Accelerating	67	5

Good Quasi-Cyclic Codes from Circulant Matrices Concatenation using a Heuristic Method

Bouchaib AYLAIJ
LIMA Lab, Faculty of Sciences
Chouaib Doukkali University
El jadida, Morocco

Mostafa BELKASMI
SIME Labo, ENSIAS
Mohammed V University
Rabat, Morocco

Said NOUH
TIM Lab, Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

Hamid ZOUAKI
LIMA Lab, Faculty of Sciences
Chouaib Doukkali University
El jadida, Morocco

Abstract—In this paper we present a method to search q circulant matrices; the concatenation of these circulant matrices with circulant identity matrix generates quasi-cyclic codes with high various code rate $q/(q+1)$ (q an integer).

This method searches circulant matrices in order to find the good quasi-cyclic code (QCC) having the largest minimum distance. A modified simulated annealing algorithm is used as an evaluator tool of the minimum distance of the obtained QCC codes. Based on this method we found 16 good quasi-cyclic codes with rates (1/2, 2/3 and 3/4), their estimated minimum distance reaches the lower bounds of codes considered to be the better linear block codes in Brouwer's database.

Keywords—Circulant matrix; quasi-cyclic Codes; Minimum Distance; Simulated Annealing; Linear Error Correcting codes

I. INTRODUCTION

In coding theory, a large side of research has been interested in design and construction of error correcting codes families which are the basis of the channel coding element in the digital communication system. This research is not an easy problem. Moreover, the sphere packing problem is equivalent to finding a linear code with largest minimum Hamming weight in a given space [1]. The term good codes in this work, refers to maximizing the minimum distance for a binary linear code of a given parameters: length and dimension or various-code rate and/or high-code rate.

The author in [2] used the canonical form based in circulant matrices to found many good codes: quadratic residue codes and high quality group codes, and the author in [3] found the best quadratic residues with the same circulant property over the field $GF(3)$

More generally, the author in [4] proposes a quadratic double circulant codes schemes which are a generalization over any field $GF(q)$ and for any length code, on the contrary, of the construction methods cited in [1] [2].

The design of good error correcting codes is a difficult problem, which remains open in coding theory. Recently this

problem is attacked with meta-heuristic methods. Some of these works, A. El Gamal et al. [5] used simulated annealing to build good source codes, good channel codes and spherical codes. In [6] Chatonnay et al. introduced genetic algorithms for finding good linear codes. In [7] [8] the authors found good double and triple circulant codes, using the multiple pulse method and genetic algorithms. Comellas et al. [9] used genetic algorithms to design constant weight codes. Walice et al. [10] have presented a comparative study of meta-heuristic techniques applied to estimate the minimum distance of BCH Codes.

The determination of the minimum distance of linear block codes (minimum Hamming weight) by classical methods is hardly feasible; in general, this is an NP-hard problem [11]. The combinatorial nature of the problem requires an enumeration of the codewords for a linear code in order to find one with the minimum weight. Unfortunately, exhaustive exploration of the search space, is not possible, especially when the length n increases [12][13], which means that the size of the search space that is 2^k codewords, becomes prohibitively high, where k is the dimension of code. Hence, the need of a met-heuristic technique to estimate the minimum Hamming weight value or in some cases, to find its true value.

We present in this paper, a method to search a good quasi-cyclic codes with rate $q/(q+1)$ (where q is an integer) based in extensive random search for circulant matrices, and we chose the heuristic simulated annealing method (SA) to find the value of the minimum distance of quasi-cyclic codes that we have constructed.

The remainder of this paper is presented in six sections. On the next section, we give an introduction on quasi-circulant codes, the minimum distance of linear block codes, encoding operations and simulated annealing method. In section III we present the method for searching the good quasi-cyclic codes. The modified Simulated Annealing method is presented in section IV. The obtained codes and experiment results are presented in section V. Finally, concluding remarks and perspectives of this work are given in section VI.

If ($Task == Task_1$) **then** determine a neighbor information vector (D_{i+1}) from task_1;
Else determine a neighbor information vector (D_{i+1}) from task_2;
End if
 Evaluate $\Delta F = F(D_{i+1}) - F(D_i)$;
If $\Delta F \leq 0$ **then** $D_i \leftarrow D_{i+1}$;
 Generate $q = random [0, I]$;
Else if ($q \leq Exp(-\Delta F/T)$)
Then $D_i \leftarrow D_{i+1}$;
End if
End if
Until (iterations number < N)

If (Transition criterion is satisfied == yes) **Then**
 switch between $Task_1$ and $Task_2$;
End if
 $T \leftarrow a.T$;

Until ($T > T_p$)

Task_1: Let $D_i = (D_{i1}, \dots, D_{ik})$ be the current information vector over $GF(2)^k$ and $S_i = (s_{i1}, \dots, s_{ik})$ a switch vector over $GF(2)^k$, randomly generated, where $1 \leq W_H(S_i) \leq k$. The neighborhood information vector D_{i+1} is defined as follows:

Step1. $D_{i+1} = D_i \oplus S_i \pmod{2}$ (6)

Step2. $W_H(D_{i+1})$ must be between 1 and minimum distance upper bound of the QC code

Step3. $D_{i+1} \in GF(2)^k - \{0\}$

Task_2: Let Γ_p be the cyclic shift of p places of elements

The neighborhood information vector D_{i+1} is produced by generate a random integer number p over $[1, k-1]$, and we apply the cyclic shift Γ_p on D_i .

$$D_{i+1} = \Gamma_p(D_i) \tag{7}$$

Criterion of transition between Task_1 and Task_2

The transition between Task-1 and Task-2 is made randomly from a uniform distribution.

V. COMPUTATIONAL EXPERIMENT RESULTS

We performed the computational experiments with:

- Software: program developed in language C
- Hardware: CPU CORE 2Duo 2GHz and 2GB of RAM

We used the parameters in algorithm 2 for simulated annealing algorithm.

All good quasi-cyclic codes that we found by this method, using the modified simulated annealing method, have been verified and validated independently using the well known computer algebra package, MAGMA [18].

Here, the term good quasi-cyclic code refers to a binary quasi-cyclic code with the largest d_{min} for a given length n and dimension k . In cases where there is more than one good code, only one is chosen.

The Tables I, II and III as following summarize the obtained good quasi-cyclic codes with code rate $q/(q+1)$ where q is an integer between 1 and 3.

Note that LB and UP , respectively, denote Lower Bound and Upper Bound on the minimum distance of a linear code for a given parameters, these limits are taken from the Brouser's data base [19]. d_{magma} is the minimum distance calculated by the calculator algebraic Magma [18] and d_{found} is the minimum distance of QCC obtained by the modified simulated annealing algorithm. The obtained QCC codes seem to be good codes because their estimated minimum distance is equal to their lower bounds.

TABLE I. GOOD QUASI-CYCLIC CODES FOUND USING ALGORITHM 1, WITH Q=1, CODE RATE T=1/2

Rate	QCC	Binary Total Header TH	d_{found}	d_{magma}	LB	UB
1/2	C(60,30)	000010111001111001000000110000	12	12	12	14
	C(52,26)	00010111000000010010111110	10	10	10	12
	C(58,29)	01110111110010100111010101010	12	12	12	14
	C(76,38)	1111101001111001101111011011011000011	14	14	14	18
	C(94,47)	10001101010110011011010000000111110110001110010	16	16	16	22

TABLE II. GOOD QUASI-CYCLIC CODES FOUND USING ALGORITHM 1, WITH Q=2, CODE RATE T=2/3

Rate	Codes	Binary Total Header TH	d_{found}	d_{magma}	LB	UB
2/3	QCC(93,62)	011111011110001101011101110110101111011100101101001000000100	10	10	10	14
	QCC(99,66)	011011111101101111011110000001110010100111100110111111011110010101	10	10	10	14
	QCC(105,70)	010110010000010001110000100001100001110001100100010000001110001001100	10	10	10	15
	QCC(123,82)	11000110001000011111010001101010010100101100001111111111010110110011011001111	12	12	12	17
	QCC(150,100)	01111100111011001100011000010110010100010000000010111110001100101010000111111100100000001111011111	14	14	14	20
	QCC(156,104)	0000010110100010010111011000011011011101110011000011001000011100010010101000011011000110001100010001	14	14	14	22

TABLE III. GOOD QUASI-CYCLIC CODES FOUND USING ALGORITHM 1, WITH $Q=3$, CODE RATE $T=3/4$

Rate	Codes	Binary Total Header TH	d_{found}	d_{magma}	LB	UB
3/4	QCC(68,51)	000110111111101011101101001100100101110100011011001	6	6	6	8
	QCC(72,54)	101000111000000110000010011001100001111110010110011111	6	6	6	8
	QCC(92,69)	100101011111100011111011101111001000010010101001010011001010111100111	8	8	8	10
	QCC(96,72)	010011110110100010110000110110000101100000110000001000000100101110001101	8	8	8	10
	QCC(108,81)	11110011111101011001111010111001101110111100101111110001011101111101111001101	8	8	8	11

VI. CONCLUSION

We gave a method to search good quasi-cyclic codes with different rate $q/(q+1)$ (where q is an integer) and we presented 16 new quasi-cyclic codes with minimum distances equal to lower bounds of the good linear codes in Brouwer’s database. The fact that we have integrated a modified simulated annealing in the search algorithm speeded up the extensive random search process. In the future work, we will try to search with this efficient technique others better linear block codes, and to test the obtained codes in Encoder/Decoder systems for computational complexity and BER performance.

REFERENCES

[1] Conway J.H. and Sloane N.J.A. Sphere Packings, Lattices and Groups, 3rd ed., Springer, New York, 1999.
 [2] Karlin M. IEEE Trans. Inform. Theory, 15, 81-92, 1969.
 [3] Pless V. J. Combinatorial Theory, 12, 119-142, 1972.
 [4] Gaborit P. J. Combin. Theory, A97, 85-107, 2002.
 [5] El Gamal A.A., Hemachandra L.A., Shperling I. & Wei V.K.W. IEEE Transactions on Information Theory, 33(1),pp 116-123, 1987
 [6] Jérôme Lacan, Pascal Chatonnay, “Search of Optimal Error Correcting Codes with Genetic Algorithms,” Proceedings of the 6th International Conference on Computational Intelligence, Theory and Applications: Fuzzy Days, Springer-Verlag, 1999.
 [7] A. Azouaoui, M.Askali and M. Belkasmi , “A genetic algorithm to search of good double-circulant codes”, IEEE International Conference on Multimedia Computing and Systems (ICMCS’11) proceeding ,pp 829- 833, Ouarzazate, Morocco, April 07-09,2011.
 [8] Askali M., et al. Discovery of Good Double and Triple Circulant Codes using Multiple Impulse Method. Advances in Computational Research,

ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 5, Issue 1, pp.-141-148, 2013.
 [9] Comellas F., Roca R. Using genetic algorithms design constant weight codes. In applications of Neural Networks to Telecommunications, 119-124, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1993.
 [10] J. Wallis and K. Houghten, “A Comparative Study of Search Techniques Applied to the Minimum Distance of BCH Codes,” Conference on Artificial Intelligence and Soft Computing, Banff, 17-19 July 2002.
 [11] A. Vardy, “The Intractability of Computing the Minimum Distance of a Code”, IEEE Transaction on Information Theory, vol. 43, N.6, 1997.
 [12] Pless V.S. and Huffman W.C. Handbook of Coding Theory, Elsevier, Amsterdam, 1998.
 [13] Coley D. An Introduction to Genetic Algorithms for Scientists and Engineers, World Scientific, 1999.
 [14] McWilliams F.J. and Sloane N.J.A. The Theory of Error-Correcting Codes, Amsterdam, The Netherlands: North-Holland Mathematical Library, 1977.
 [15] Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. Journal of Chemical Physics, 21 :1987–1091, 1953.
 [16] Kirkpatrick, S , Gelatt, C.D., Vecchi, M.P. Optimization by Simulated Annealing. Science, vol220, No. 4598, pp671-680, 1983.
 [17] Aylaj B. and Belkasmi M. New Simulated Annealing Algorithm for Computing the Minimum Distance of Linear Block Codes. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 6, Issue 1, pp.-153-158, 2014.
 [18] Bosma, W., Cannon, J.J., and Playoust, C.: ‘The Magma algebra system I: the user language’, J. Symb. Comput., 24, pp. 235–266, 1997.
 [19] Brouwer, A.E.: ‘Bounds on the size of linear codes’, in Pless, V.S., and Huffman, W.C. (Eds.): ‘Handbook of coding theory’ (Elsevier, North Holland), http://www.codetables.de/, 1998.

Balanced Distribution of Load on Grid Resources using Cellular Automata

Amir Akbarian Sadeghi

Department of Computer
Engineering
South Tehran Branch, Islamic Azad
University, Tehran, Iran

Ahmad Khademzadeh

Research Institute for Information &
Communication Technology
Tehran, Iran

Mohammad Reza Salehnamadi

Department of Computer
Engineering
South Tehran Branch, Islamic Azad
University, Tehran, Iran

Abstract—Load balancing is a technique for equal and fair distribution of workloads on resources and maximizing their performance as well as reducing the overall execution time. However, meeting all of these goals in a single algorithm is not possible due to their inherent conflict, so some of the features must be given priority based on requirements and objectives of the system and the desired algorithm must be designed with their orientation. In this article, a decentralized load balancing algorithm based on Cellular Automata and Fuzzy Logic has been presented which has capabilities needed for fair distribution of resources in Grid level.

Each computing node in this algorithm has been modeled as a Cellular Automata's cell and has been provided with the help of Fuzzy Logic in which each node can be an expert system and have a decisive role which is the best choice in a dynamic environment and uncertain data.

Each node is mapped of one of the VL, L, VN, H and VH state based on information exchange on certain time periods with its neighboring nodes and based on fuzzy logic tries to decrease the communication overhead and estimate the state of the other nodes in subsequent. The decision to send or receive the workload is made based on each node state. Thus, an appropriate structure for the system can greatly improve the efficiency of the algorithm. Fuzzy control does not search and optimize, just makes decisions based on inputs which are effective internal parameters of the system and are mostly based on incomplete and nonspecific information.

Each node based on information exchange at specific time periods with its neighboring nodes, and according to Fuzzy Logic rules is mapped of one of the VL, L, N, H and VH states. To reduce communication overhead, with the help of Fuzzy Logic tries to estimate the state of the other nodes in subsequent periods, and based on the status of each node, makes a decision to send or receive workloads. Thus an appropriate structure for the system can improve the efficiency of the algorithm. In fact, Fuzzy Logic does not search and optimize, just makes decisions based on the input parameters which are often incomplete and imprecise.

Keywords—Computing Grid; Load balancing; Cellular Automata; Fuzzy Logic

I. INTRODUCTION

The need for high computational power and organizational limitations have created a new type of shared computing environment, which is called grid computing. Grid computing is a computing infrastructure Which provides access to high-performance computing resources. End users and applications

see this environment as a large virtual computing system. Systems that are connected to Grid may be distributed globally and be running on different hardware platforms and operating systems and belong to various organizations. In a short definition, Grid can be considered as a system for distributed resource sharing on a large scale and indeed without borders. To improve the global throughput of the Grid computing, requests must be divided evenly among the available resources. Resource management is a major and infrastructure issues in this environment. The overall objective of resource management is the effective scheduling to run jobs that need to use resources in Grid environment. In a general definition, the purpose of load balancing algorithms is improving the distribution of workloads across resources, maximizing throughput, and minimizing response time, which means the difference between the overloaded and under-loaded resources should be minimal. The desirable characteristics of a load balancing solution include: scalability, adaptability, stability, application transparency, fault tolerant and minimal overhead. The load balancing methods are generally classified as centralized or decentralized, static or dynamic, periodic or non-periodic, and with threshold or without threshold.

Cellular Automata answer this question that How complex systems can be studied. There is the ability to predict the next state of cells in this system based on the status of each cell and its neighboring cells which can help in proper distribution of load among nodes. Fuzzy Logic can make effective decision by imprecise and incomplete Information. Cellular Automata can evaluate the current and future status of each resource by Fuzzy Logic rules to improve load distribution among the heterogeneous resources [14-19].

This article introduce a Grid load balancing algorithms based on Cellular Automata and fuzzy rules. The remainder of the paper is organized as follows: In section 2, the definition of concepts such as Grid, Load balancing, Cellular Automata and Fuzzy Logic. Section 3 describes our proposed algorithm in detail. Section 4 discusses our simulation and results of evaluation. Finally, section 5 concludes this paper

II. DEFINITION OF CONCEPTS

Advances in areas technical constantly need to have faster computing, but computer hardware manufacturers have reached fundamental limitations in the physical speed [1]. Electronics and hardware advances in technology alone cannot meet the demand for increased computing speed. Parallel

processing is the emerging response to this problem in which large problems can often be divided into smaller ones, which can then be solved at the same time on several processors [2, 3].

Although writing code that is flexible enough to be split among several processors is more complicated, but the tendency toward parallel processing hardware and software has increased [2]. Instead of limiting the execution of a task in a processor, parallel processing divide many calculations among several processors and solve this issue through team work [3]. Reduce the cost of computers and advances in communication networks have increased the tendency for the use of large-scale parallel systems and distributed computing systems. In fact, recent studies in the field of computing architecture have led to the emergence of a new computing paradigm which is Grid computing[11]. A computational Grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities [12]. This technology is a type of distributed system that collects computer resources from multiple locations to reach a common goal to solve a single task. Nowadays, a variety of Grid systems is manufactured with various definitions and facilities which have different objectives. Thus, providing a single definition that covers all aspects of Grid computing technology is not easy nor true. Various experts have provided different definitions according to different pursued goals.

Ian Foster who was the main inventor of Grid and founder of Globus defines Grid as follows [12]:

“Grid technology is seeking to create the possibility of large-scale and controlled resource sharing which is flexible and is after creating protocols, services and software packages.”

Grid is defined as follows in IBM Company which is among pioneers of Grid:

“Grid is a set of distributed computing resources in a local area network or a wide area network which seems like a computer and virtual computing system for end-user or applications. Its main goal is creating dynamic virtual organizations through sharing resources using coordinated and safe methods among users, universities and organizations.”

Grid is a distributed system which contains following items [8]:

- Resources (software and hardware) are heterogeneous
- Resources are coordinated but are not under a centralized management
- The use of all-purpose standard protocols and interfaces
- Grid may have Multiple administrative domains or in other words be made of several Virtual Organizations (VO)
- Ensuring the quality of the services provided

Resource management is one of the important and infrastructure issues in such environment. The overall objective of resource management is the effective scheduling to run jobs

that need to use resources in Grid environment. In a general definition, the purpose of load balancing algorithms is improving the distribution of workloads across resources, and maximizing their performance as well as reducing the overall execution time [9]. In another definition, the load balancing algorithm is an algorithm which ultimately allows all nodes to task at once [10-11]. The desirable characteristics of a load balancing solution include: Scalability, Adaptability, Stability, Application Transparency, Fault Tolerant and minimum overhead imposed on the system. The mentioned specifications are interdependent. For example, delays such as Computation Delay and Communication Delay have abnormal effects on the stability and thus comparability of the algorithm. Due to the many parameters involved in the problem of load balancing as well as contradictory of some mentioned features, meeting all the them in the form of a single algorithm is practically difficult or even impossible. Most of the existing methods try to satisfy one or more of the above objectives [12-14].

For better efficiency and more use of dynamic algorithms and considering that the main focus of this article is on the same set of algorithms, in general, the process of dynamic load balancing algorithms has four main procedures:

A. *Load Measuring procedure*

B. *Information Exchange procedure*

C. *Initiation procedure*

D. *The final load balancing procedure*

Load measuring procedure is an expression of CPU load in a way that heavier load on processors will increase it, and its reduction will reduce it. Because this routine repeatedly execute in this algorithms, the calculation of it should be as simple and efficient as possible [17]. Information Exchange procedure determine the method of collecting necessary task load for load balancing decisions. Initiation procedure decides about the time of starting load balancing. This decision-making is along with determining the ratio of efficiency to imposing overhead. Load balancing methods attempt to achieve goals such as minimizing the average response time for processing or maximizing resource efficiency by running processes on distributed resources. This target may initially be a demand or take place after the start of its execution. Of course, in any case, a good and efficient algorithm must consider the cost of the communications as well [18]. Cellular Automata (CA) is an answer to this question that how to study complex systems. Cellular Automata can be a complex system in itself and yet provide appropriate methods to study complex systems like these - Complex systems – [19-20].

III. THE PROPOSED ALGORITHM OF FUZZY LOAD DISTRIBUTION USING CELLULAR AUTOMATA (FUZZY LOAD BALANCING CELLULAR AUTOMATA)

The main idea of this project is using a cell of Cellular Automata to show a computational node in which the state of the cell shows the status of that node. In this method, a global load balancing solution can only be produced just using local load balancing. This method of load distribution is in the form of a wave motion.

All parameters that each processor considers during the proposed load balancing algorithm are described below:

M: Number of heterogeneous computing nodes in the system (P_1, P_2, \dots, P_M)

x: Number of jobs executed in the system (J_1, J_2, \dots, J_x).

T_s : Information exchange time.

T_e : The estimated time period.

N_i : Buddy set of node P_i .

$S_i(T_n)$: State of node 'I' at time T_n .

m_j : Number of migration of a job.

$Q_i(t)$: The number of jobs waiting in the execution queue at the node P_i at time t.

W_i : Processing power at P_i .

$Z(J_x)$: Size of job(x).

$TET_{i,t}$: Total waiting time for execution of waiting job at P_i queue.

$RET_{i,t}$: The remaining execution time of the job being processed at the P_i .

$LD_{i,t}$: Load of P_i at time t, comes from (1).

$$LD_{i,t} = TET_{i,t} + RET_{i,t} \quad (1)$$

$NLD_{i,t}$: Normalized average load in the buddy set of node P_i at time t.

BW_{ij} : Bandwidth communication between processors i and j

$ArrTime(J_x)$: Arrival time of J_x job.

$endTime(J_x)$: End time of J_x job.

$ETC(J_x, P_i)$: Estimated execution time of J_x at P_i , comes from (2).

$$ETC(J_x, P_i) = \frac{ETC(J_x, P_{std})}{W_i} \quad (2)$$

$T_{com}(J_x, P_i, P_j, t)$: The time required for transfer J_x job from P_i to P_j at time t.

$EFC(J_x, P_i, P_j, t)$: Estimation of finish time of J_x job when transfer from P_i to P_j at time t.

if $T_{com}(J_x, S_i, S_j, t) \geq LD_{j,t}$

$$EFC(J_x, S_i, S_j, t) = T_{com}(J_x, S_i, S_j, t) + ETC(J_x, S_j)$$

ELSE

$$EFC(J_x, S_i, S_j, t) = LD_{j,t} + ETC(J_x, S_j) \quad (3)$$

$B_x(P_i, P_j)$: Benefit of execution of the J_x job at P_j compared to execution at P_i .

$$B_x = EFC(J_x, P_i, P_i, t) - EFC(J_x, P_i, P_j, t) \quad (4)$$

The general procedure of the proposed Load balancing algorithm is in the way when a new task enters the computational node, that node will decide based on cell's conditions should carry out this task itself or migrate it to another node.

This algorithm consists of several main procedure:

- Determining the state of nodes
- Making decision to migrate the task
- Selecting the best node to carry out the task

A. Determining the state of nodes

The overall basis for all decisions is the state of each node. In fact, the essential criterion in deciding to send a job is the state of the node, and the main criterion for selecting a node to perform the job is also the state of the node. Thus, determining the state of each node is crucial in load balance in the whole system.

To determine the state of each node and its neighbors for each execution and migration, the information is needed to determine the status of nodes. There will be a huge communication overhead in the system if all nodes exchange their status information. Thus, regular intervals are used to determine the state of nodes which are called information exchange periods (T_s). T_s which is greater than the period of time for running and migration of jobs is performed between nodes which estimate the state of nodes between these time periods. T_e tries to reduce communication overhead and more accurate decisions (Fig. 1).

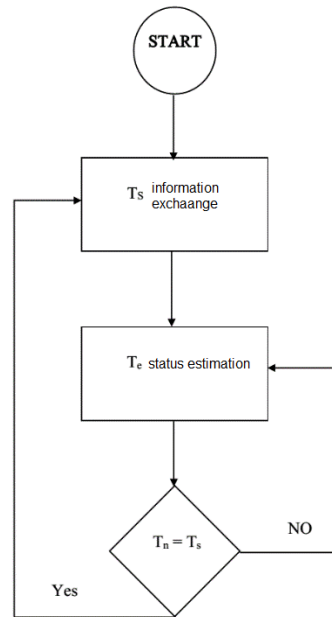


Fig. 1. Determining the state of nodes in T_n

Determining the state of the node is done in two methods:

- Information Exchange (T_s)
- State estimation (T_e)

1) Determining the state of nodes through the information exchange

It contains three parts of sending load information, calculating the average load and determining the status of nodes using Fuzzy Logic (Fig. 2).

a) Sending load information

The nodes send their status information to their buddy set in regular time periods of T_s which are called information exchange periods and receive their information. Then each node records the received information from each neighboring node in the neighboring table.

b) Calculate the average load

Each node calculates its average load and the load related to neighboring collection using (5) and considers the obtained index as normalized load average.

$$NLD_{i,t} = \frac{\sum_{j \in N_i} LD_{j,t}}{N \times \sum_{i \in N_i} w_i} \quad (5)$$

c) Determining the state of nodes using Fuzzy Logic

The normalized load average is considered as the point of balance and map the state of each node to one of Very Light, Light, Normal, Heavy, Very Heavy forms using Fuzzy Logic. This step is called mapping input values to the fuzzy mode and the state of each node will be recorded in the neighboring table.

2) Determining the state of node using estimation

If this data transfer is done in intervals with a short distance, there will be a high communication overhead imposed on the system. Thus, these intervals must be increased and estimated T_e state in information exchange periods using fuzzy rules.

To discover these laws, the algorithm runs without estimation periods and records the data on each node. The related fuzzy rules are extracted by using Matlab software. To reduce communication overhead caused by data exchange, exchange periods will be increased, and the status of each node using obtained rules to reduce errors in decision making will be estimated.

B. The decision to send task

When job J_x enters node, and that node is in one of VL, L, VN states, it will be queued for processing and will be waiting to run according to respective priority, and the higher rate of migration (m) leads to increased running priority.

Otherwise if J_x enters node, and that node is in one of H, VH states, and it migrated from another node, it will not be accepted, on the other hand if the job belongs to that node, if the node is in VH state then it calculates its own load and the normal load, and if it has a neighbor with VL state, it sends $\frac{1}{2}$ of its extra load and if there is a neighbor with L state, it sends $\frac{1}{4}$ of its extra load. If its state is H, it will calculate the difference between its own load and the normal load, and if there is a neighbor with VL state, it will select $\frac{1}{4}$ of its extra load for migration, and selects $\frac{1}{8}$ if there is a neighbor with L state.

C. Selecting the most appropriate node to execute the job

At the beginning, nodes in the buddy set are marked with the weight of 1 for L nodes, and 2 for VL nodes. Then the execution time of the job J_x is estimated on specified nodes and based on $B_x(P_i, P_j)$ benefit which is the difference between execution time in node P_j compared to P_i node, if this benefit is positive and bigger than the threshold, a weight is given to each of them. In this way, the higher execution benefit leads to higher score. Finally, considering the weight of each node and the weight of execution benefit in P_j node, the decision to transfer job to P_j is taken. the higher weight will lead to a higher possibility of sending the task to that node (Fig. 3).

If the execution time of a job is equal in the source and destination nodes, which means running the job in the source node or migrating it to another node may result in similar end time, and the benefit will be nil or smaller than threshold. In this case, the job is not allowed to be migrated to that node, because it imposes communication overhead on the system and the bandwidth between resources is engaged even for small time.

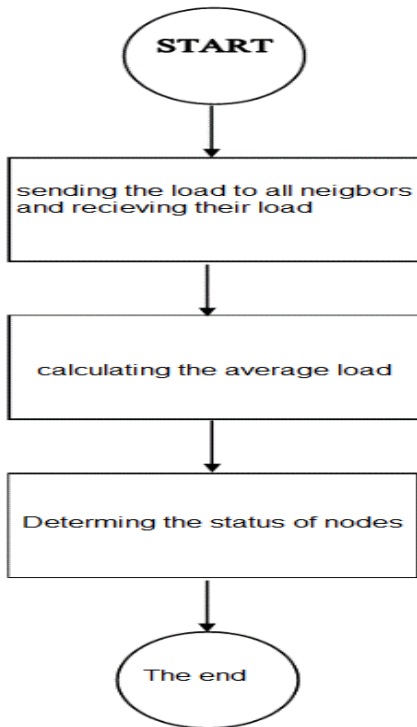


Fig. 2. Information exchange

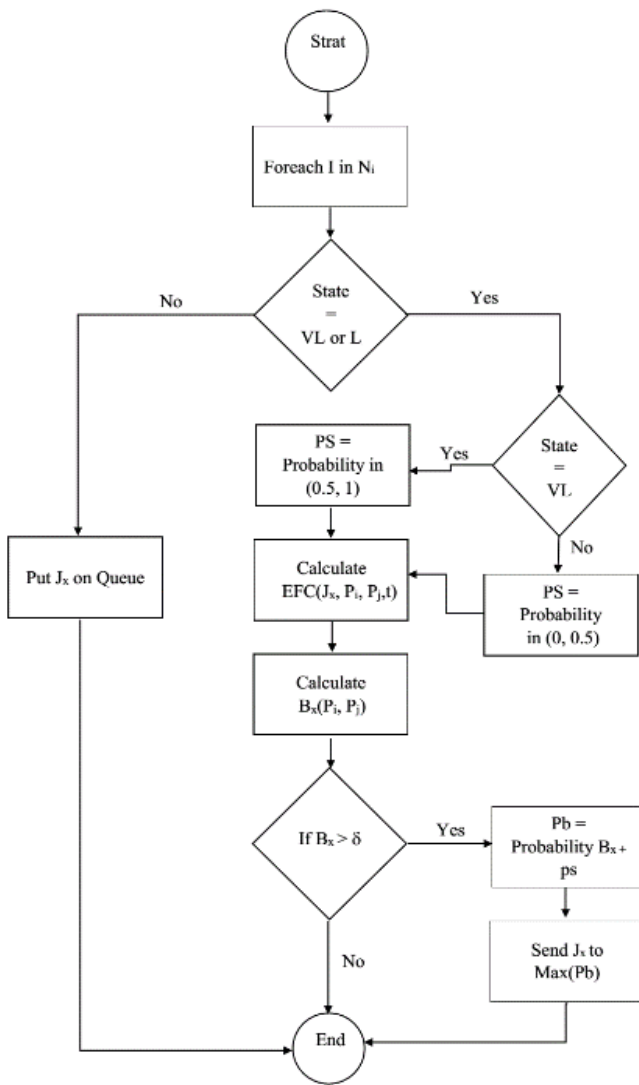


Fig. 3. Selecting the most appropriate node to execute the job

IV. SIMULATION AND RESULTS

Simulation is an imitation of the real world. Simulators enable programmers, developers and Grid developers who need to test their programs, tools and services to ensure the proper function of their program before final production and using them in a real world.

The proposed FLBCA algorithm has been simulated using C# programming language in Visual Studio 2010 environment, and its algorithm has been compared with MELISA and ELISA algorithms.

A. Performance metrics

Following two parameters have been used to evaluate the performance of the proposed algorithm [21-23]:

1) Average Response Time (ART):

The average response time is the average amount of time a job must wait before execution.

2) Average time of Resources Used (ARU):

The ratio of working time to the total time of a system:

$$U_i = \frac{Busy_i}{Busy_i + Idle_i} \quad (6)$$

$$ARU = \frac{\sum_{i=1}^M U_i}{M} \quad (7)$$

B. Simulation Model

This simulation is formed by 30 heterogeneous computing nodes which their processing power follows random distribution in the range of [1, 10] and the relation between two nodes has been formed by a heterogeneous communication network in a way that their communication bandwidth is variable from 1 Mbps to 10 Mbps.

10,000 independent tasks have been used in this simulation in a way that running time of each task has been generated randomly in the range of [1, 100]. These tasks enter the system based on Poisson distribution with the rate of [1, 4], and the volume of each task follows a normal distribution with the mean of 5 MB and standard deviation of 1 MB.

The time for information exchange (T_s) is assumed to be 20 units and the estimation time of status is considered to be five units.

C. Simulation results

This algorithm has been evaluated regarding performance metric and under the effect of parameters such as the number of tasks, time and period of service transition and estimation interval.

1) The effect of the jobs entered in the homogeneous environment

In a homogeneous environment where processing power of each CPU is 1, and the communication bandwidth between any two nodes is constant and equal to 10 Mbps, the number of tasks has been added from 0,000 to 50,000 to measure these factors. In these conditions where the number of tasks is 10000. The average response time is about the same among all three algorithms although after increasing jobs, the efficiency of ELISA algorithm is better than MELISA and the proposed algorithm (FLBCA), and the efficiency of FLBCA algorithm is slightly better than MELISA algorithm (Fig. 4). The total run time is about the same in all three algorithms (Fig. 5).

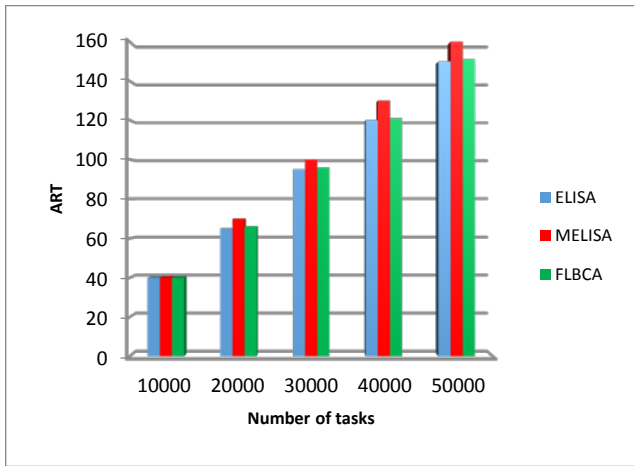


Fig. 4. The average response time in case of homogeneous

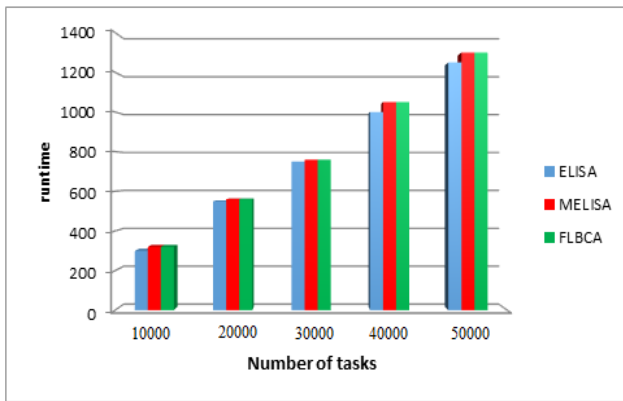


Fig. 5. The total runtime in the homogeneous environment

2) *The effect of jobs entered into the heterogeneous environment*

Since this algorithm has been designed for a heterogeneous environments, it is compared in a heterogeneous environment with different processing power in the range of [1, 10] and various communication bandwidths between any two nodes in the range of 1 Mbps to 10 Mbps with ELISA and MELISA algorithms. To measure the effectiveness of the jobs, the number of jobs has been increased from 10,000 to 50,000. The average response time is much better in FLBCA and MELISA algorithms than the ELISA algorithm. The average response time is initially about the same in FLBCA and MELISA algorithms, but is gets better with increasing number of tasks in FLBCA algorithm, and it shows better and faster decision making in Fuzzy Logic (Fig. 6).

The total runtime is similar to all three algorithms. The runtime is a function of the rate of entering jobs into the system, and the runtime is about the same due to using same data (Fig. 7).

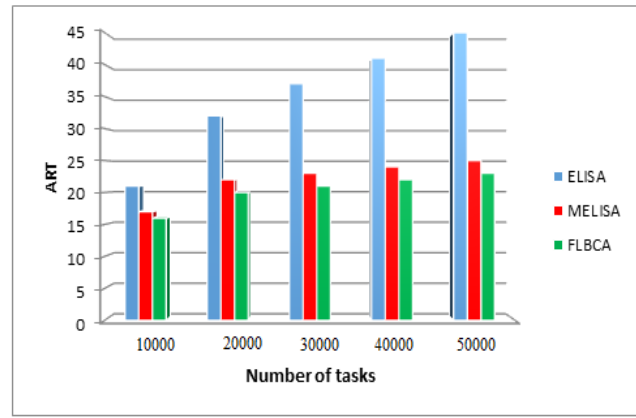


Fig. 6. Comparing the average response time in the heterogeneous environment

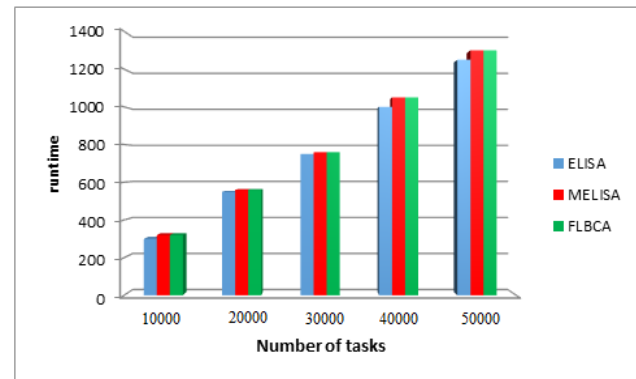


Fig. 7. Comparing the total runtime in the heterogeneous environment

3) *The effect of job size*

This section tries to evaluate the effect of changing tasks volume from 5 MB to 50 MB on the average response time in the proposed algorithm. The number of migration reduces and the average response time increases (Fig. 8).

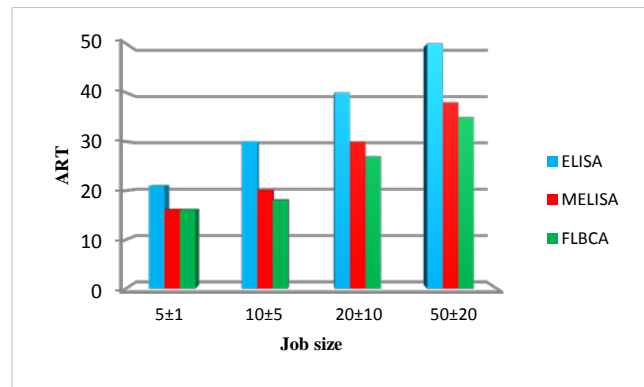


Fig. 8. Average response time for different job sizes

4) The effect of job runtime

This section tries to evaluate the effect of changing the average runtime of tasks from 10 to 150 units on the efficiency of the algorithm. The average response time increases with large rate by increasing the runtime (Fig. 9).

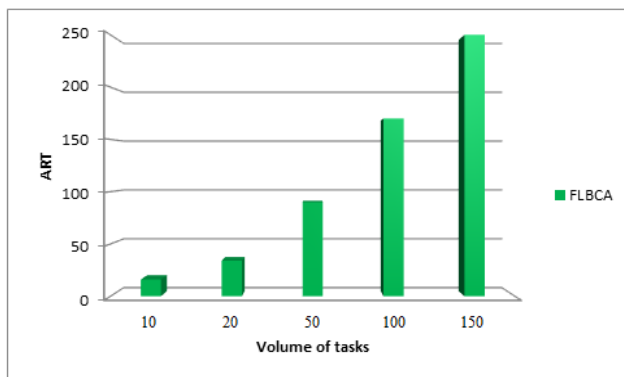


Fig. 9. Average response time for various run times

5) The effect of information exchanges time

This algorithm was tested with different times in range of 2 units to 40 units to find a best time for information exchanges which is suitable in perspective of efficiency criteria (Fig. 10).

The accuracy of information reduces with increasing time of the information exchange. The distribute the load between nodes is done with less precision and average response time increases for this purpose.

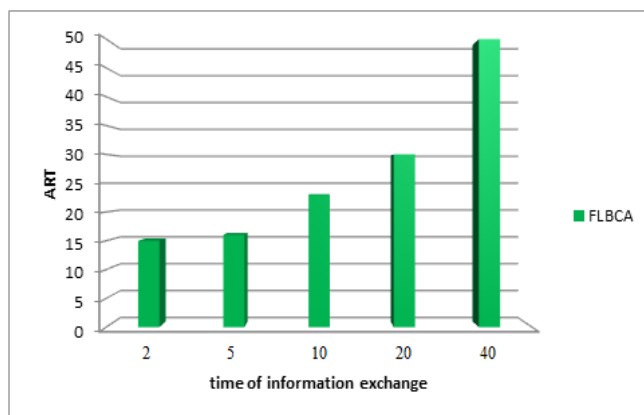


Fig. 10. Average response time for different times of exchanging information

6) The effect of state estimation time

To access a proper state estimation time from the point of efficiency criteria, the information exchange time must be considered to be 20 units by default. Then by changing the estimation time in different intervals from 2 units to 10 units, the most effective time will be found (Fig. 11).

The average response time increases by reducing estimation time due to increased computational overhead and the most optimal time is reached at times of 4 and 5 and the average response time increases again by increasing this time due to reduced accuracy of data.

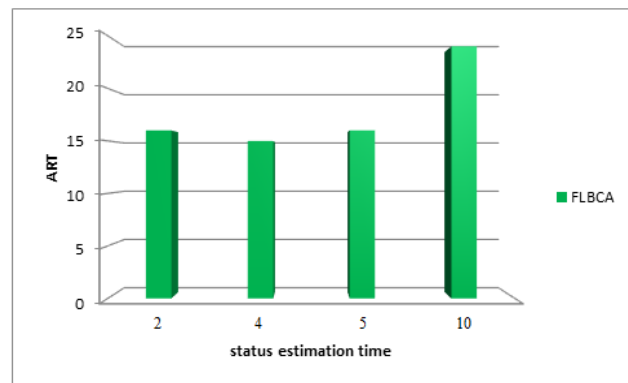


Fig. 11. Comparing the average response time for different times

V. CONCLUSION

A load balancing model has been provided in this research for Grid computing environment which is called FLBCA. This algorithm tries to meet needs and characteristic required for a load balancing algorithm such as scalability, adaptability, stability, application transparency, and minimal communication overhead as much as possible. Because the Fuzzy Logic can deal with imprecision and uncertainty information, and increase accuracy in decision making, the algorithm based on Fuzzy Logic has been proposed for dynamic load balancing in the computing Grid. The major purpose of using this algorithm is increasing efficiency and productivity of Grid system which will lead to reducing organization's costs, and increasing productivity and saving energy, and using energy efficiently saves it which is the top priority in our life today and helps to protect the environment and nature.

In the following, it seems hierarchical Cellular Automata is an appropriate structure for designing these types of algorithms, which can provide a better view of the whole conditions of the Grid System. In addition, the usage of Fuzzy Logic which leads to accuracy of decision-making in uncertain environments can be used to improve the efficiency of parallel algorithms. The combination of Fuzzy Logic and Cellular Automata can be a good technique for a lot of parallel algorithms.

REFERENCES

- [1] M. Fathi, F. Mehryari., "the principles and concepts of grid computing technology and its applications in various fields", Noorpardazan, .10, 2009 - Tehran, pp.
- [2] M. Amini Salehi, H. Deldari, load balancing in the grid resources using agent-based resource management, Master's thesis, Department of Computer Engineering, Ferdowsi University of Mashhad, 2005.
- [3] S. Ghanbari, balance and self-organization in the grid computing using learning automata, Master's thesis, Department of Computer Engineering, Amirkabir University of Technology, 2004.
- [4] R. Tlili, Y. Slimani, A Hierarchical Dynamic Load Balancing Strategy for Distributed Data Mining, International Journal of Advanced Science and Technology, Vol. 39, February, 2012
- [5] S. Adabi, reducing power consumption in wireless sensor networks based on cellular automata, Master's thesis, Department of Computer Engineering, Islamic Azad University of Science and Research Branch, 2010.

- [6] L. Anand, D. Ghose, V. Mani, ELISA: An Estimated Load Information Scheduling Algorithm for Distributed Computing Systems, An International computers & mathematics with applications, 1999.
- [7] L. Rostami, A. Rahmani, An adaptive Load Balancing Algorithm with use of cellular Automata for Computational Grid Systems, Euro-Par 2011 Parallel Processing, Lecture Notes in Computer Science Volume 6852, 2011, pp 419-430.
- [8] A. Karimi, F. Zarafshan, A. Jantan, Anew Fuzzy Approach for Dynamic Load Balancing Algorithm, International Journal of Computer Science and Information Security, Vol. 6, No. 1, 2009.
- [9] M. Marinov, Intuitinistic Fuzzy Load balancing in cloud computing, 8th Int. Workshop on IFSs, Ocy 2012.
- [10] S. Mousavi Nejad, S. Mortazavi, B. Vosoughi Vahdat, Design and set optimal control and intelligent load balancing based on fuzzy logic in distributed systems, first regional conference on new approaches in computer engineering, 2011.
- [11] I. Foster, and C. Kesselman, "The Grid: Blueprint for a New Computing Infrastructure" Morgan Kaufmann and Elsevier, Second Edition, USA, ISBN: 1-55860-933-4, 2004.
- [12] I. FOSTER, C. KESSELMAN, M. NICK J, S. TUECKE, "Grid services for distributed system integration," vol. 35, 6, 2002.
- [13] C. J., K. E., L. M. Anderson D P., "SETI @ home: an experiment in public-resource computing," vol. 45 (11).
- [14] S. Graupner, J. Pruyne, S. Singhal, "Making the Utility Data Center a Power Station for the Enterprise Grid," 2003.
- [15] J. Liu, X. Jin, and Y. Wang, Agent-Based Load Balancing on Homogeneous Minigrids: Macroscopic Modeling and Characterization, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 16 (7) 2005.
- [16] L.SUN CHEUNG, A fuzzy approach to load balancing in a distributed object computing network, In Proc. Of 6th Int IEEE Conf. HPDC, 2000.
- [17] T. L. Casavants and J. G. Kuhl, A taxonomy of scheduling in general-purpose distributed computing systems, IEEE Trans. Software Eng., Vol. SE-14 (2), pp. 141-154, 1988.
- [18] H. Kameda, J. Li, C. Kim, and Y. Zhang, Optimal Load Balancing in Distributed Computer Systems. London, U.K.: Springer-Verlag, 1997.
- [19] Z. Zeng and B. Veeravalli, Rate-Based and Queue-Based Dynamic Load Balancing Algorithms in Distributed Systems, 10th Int. Conference on Parallel and Distributed Systems, IEEE 2000.
- [20] Abubakar, Haroon Rashid and Usman Aftab, Evaluation of Load Balancing Strategies, National Conference on Emerging Technologies 2004.
- [21] J.Cao, Daniel P. Spooner, Agent-Based Grid Load Balancing Using Performance-Driven Task Scheduling, In Proc. of 17th IEEE Int. Parallel & Distributed Processing Symposium (IPDPS 2003), Nice, France, April 2003.
- [22] A. Shaout and P. McAuliffe, Job scheduling using fuzzy load balancing in distributed system, in Proc. of 6st conf ICPAD, 1998.
- [23] J. Liu, X. Jin, and Y. Wang, Agent-Based Load Balancing on Homogeneous Minigrids: Macroscopic Modeling and Characterization, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 16 (7), 2005.

Camera Self-Calibration with Varying Intrinsic Parameters by an Unknown Three-Dimensional Scene

B. SATOURI

LIIAN, Department of Mathematics and informatics
Faculty of Sciences Dhar-Mahraz P.O.Box 1796 Atlas- Fes,
Morocco

A. EL ABDERRAHMANI

LIIAN, Department of Mathematics and informatics
Larache Poly disciplinary School,
LARACHE, Morocco

H. TAIRI

LIIAN, Department of Mathematics and informatics
Faculty of Sciences Dhar-Mahraz P.O.Box 1796 Atlas- Fes, Morocco

K. SATORI

LIIAN, Department of Mathematics and informatics
Faculty of Sciences Dhar-Mahraz P.O.Box 1796 Atlas- Fes,
Morocco

Abstract—In the present paper, we will propose a new and robust method of camera self-calibration having varying intrinsic parameters from a sequence of images of an unknown 3D object. The projection of two points of the 3D scene in the image planes is used to determine the projection matrices. The present method is based on the formulation of a non linear cost function from the determination of a relationship between two points of the scene with their opposite relative to the axis of abscise and their projections in the image planes. The resolution of this function with genetic algorithm enables us to estimate the intrinsic parameters of different cameras. The important of our approach reside in the use of a single pair of images which provides fewer equations, simplifies the mathematical complexities and minimizes the execution time of the application, the use of the data of the first image only without the use of the data of the second image, the use of any camera which makes the intrinsic parameters variable not constant and the use of a 3D scene reduces the planarity constraints. The experimental results on synthetic and real data prove the performance and robustness of our approach.

Keywords—Self-calibration; varying intrinsic parameters; non linear optimization; Interests points; Matching; Fundamental matrix

I. INTRODUCTION

Computer vision is the science of vision machines. It is a scientist discipline who is interested in building artificial systems that obtain information from images. The input data can take many forms: photographs, video footage, multiple camera images or multidimensional data medical scanner. Subdomains of computer vision are for example the Reconstruction of scenes, detection of events, object recognition, learning and image restoration.

The Reconstruction of 3D scenes is a research path which became very important and active with the advent of visualization by computer. As a matter of fact this technique will be found in various fields almost all of them situated on the crossroads of IT(data processing), mathematics and some of robotics related disciplines. The major objectif is always to extract information on the three-dimensional scene from a set of images gathered by numerical cameras with or without a priori knowledge of the scene. Therefore it will become clear

and necessary to begin by modeling the camera. The parameters of the cameras can be estimated by two major methods: calibration [1, 2, 3, 4] and self-calibration. In this paper, we are interested in the self-calibration methods that can calibrate the cameras without any prior knowledge about the scene. The standard process of most of these methods is to search for equations according to intrinsic parameters and the invariants in the images, whose aim generally is to solve a nonlinear equation system. The algorithm used to solve this system requires two steps, initialization and optimization of a cost function. Self-calibration of the cameras is the main step to obtain three-dimensional coordinates of points from matches between pairs of images. Several methods of camera self-calibration with constant intrinsic parameters [5–14] and those with varying intrinsic parameters [15–25] are treated in this area.

Our approach is a new and robust method for camera self-calibration having the varying intrinsic parameters by the use of an unknown three-dimensional scene. After the detection of interests points in the images by the Harris method [26] and the matching of these points in each pair of images by the correlation measure *ZNCC* [27], the fundamental matrix can be estimated from eight matches by the RANSAC algorithm [28]. This matrix is used with the projection of four points of the 3D scene in images taken by different views in order to formulate linear equations. Solving these equations allows the estimation of the projection matrices. The determination of a relationship between the four points of the 3D scene and their projections in the planes of the images *g* and *d* and the relationships between the images of the absolute conic allow the formulation of a nonlinear cost function. The minimization of this function by the genetic algorithms [29] allows the estimation of the intrinsic parameters of the cameras used.

Our method presents a novelty: two images only are sufficient to estimate the cameras' intrinsic parameters, the use of the data of the first image only, the use of any camera (with varying intrinsic parameters) and the use of an unknown 3D scene. These advantages allow us, on the one hand, to solve some problems related to the self-calibration system and, on the other hand, to work freely in the domain of self-calibration with fewer constraints.

Our work is organized as follows: In the third part, we present the camera model and matching containing three subparts: The first subpart comprises the camera model, the second is the Interest points' detection. It is a preliminary step in many computer vision processes; many methods have been advanced to extract points of interest. In this paper, we used Harris interest point detector. The third is the *Matching*: Finding in two images of the same scene, taken at different positions, pairs of pixels which are the projections of the same point of the scene. In this phase, the detected interest points are matched by ZNCC (Zero mean Normalized Cross Correlation) correlation measure. The most important section is related to the estimation of the projection matrices and self-calibration equations in section four and five. The experiment results are discussed in the Seven part, and finally, in section ten we will proceed to make a general conclusion.

II. SURVEY OF THE PREVIOUS WORKS

In order to make the self-calibration with the intrinsic constant parameters some of the concepts are based on simplified models which are designed to make the equations less complex which often allows them to converge into good results [30,31] Others are based on particular movements of the camera [32,33]. In there is also the category of the concepts which exploit general movements of constant intrinsic parameters [7] that takes into account cyclical points within a key view, in a flat scene and homographies. Other authors have shown various studies based on the Kruppa equations [34, 35] which on the one hand simplify the self-calibration process by a direct estimation without making a projective reconstruction and on the other one by eliminating the infinite plane (the projection matrices have disappeared and only the fundamental matrices and the epipoles are present). In the same context other practical methods are proposed: [36] a method which presents an analytical reduction of the Kruppa equations. [37] This article presents a framework for random sampling nonlinear optimization for the self-calibration with modeling intrinsic parameters space of the camera, the focal length is modeled by using a Gaussian distribution originated by the Kruppa equations while the optical center is close to the center of the picture, this model allows the cost of the calculations.

[38] This article deals with the problem of self-calibrating a moving camera with constant parameters. This method proposes a new set of quartic trivariate polynomial equations within the unknown coordinates of the indefinite plane derived from the hypothesis of no-skew, these new equations allow to better respect the constant of the principal point in all the images when recuperating the infinite plan. [39] In the present article a new method that combines the parallelism plane and the self-calibrating constraints of Mendonça/Cipolla. In this technique each pair of images is used independently and therefore presents a pair of different parallel planes not necessarily visible in the other images.

In order to solve the problem of the self-calibration of the cameras our interest nowadays goes to the category of concepts that exploit the varying intrinsic parameters. Amongst them we recognize various approaches. One of them [40] consists of the determination of the dual absolute quadric

of which the image is the dual absolute conic. The idea is then to transfer these constraints to the dual absolute quadric; once this matrices is known it suffices to determine the transformation that will replace the dual absolute in its canonical position. [41] A recent analogue method consists in the use of the dual absolute line (ALQ) instead of the absolute conic. [42, 43] They have solved the problem of the self-calibration through the use of Kruppa equations with the study of the varying intrinsic parameters case.

[44] A new method of self-calibration and stratified metric reconstruction for zooming/refocusing cameras is proposed sticking to the circular movement and its constraints: the ambiguity is then solved with the hypothesis of the square pixel of the camera and this flexibility allows the focal length and the principal point to vary. [45] This article details a method of self-calibration of the varying internal parameters of the camera that is based on a dual absolute quadric transformation of the image. This method may lead to a considerable improvement of the stability and robustness of the results. [46] This article describes a new method for the self-calibration of a sequence of images with the varying intrinsic parameters of the camera. This article is based on the Kruppa equations. This technique is based on the Kruppa equations with two upper triangular matrixes with which a relational matrix should be in place. Utilizing this way the epipolar geometry relationship of absolute conic to obtain the intrinsic parameters of the camera. [47] This article proposes a new method of initialization using a minimum of two images. The basic idea is that the minimum deviation of the intrinsic parameter will result in a more stable result. [48] This article presents an algorithm of sequential filtering to reach a simulated estimation of the 3D scenes. The auto calibration in this article uses the standard projective parameters of the focal distance and the principal point with two coefficients of the radial distortion. [49] A solid linear method for the self-calibration of a moving camera starting from a sequence of images is presented. The proposed approach uses known linear equations which are weighted by variable factors. The experiments show that this modification reduces the problems with the critical motion sequences.

III. CAMERA MODEL AND MATCHING

A. Pinhole Camera Model

The pinhole model (Figure 1) projects the scene in the image planes, for the camera g , it is defined by $K_g(R_g t_g)$ with a matrix $(R_g t_g)$ containing extrinsic parameters R_g the rotation matrix, and t_g the translation vector of camera in space, K_g is a matrix containing the intrinsic parameters and is expressed as follows:

$$K_g = \begin{pmatrix} f_g & \tau = 0 & u_{0g} \\ 0 & \varepsilon_g f_g & v_{0g} \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

f_g is the focal length

ε_g is the scale factor

τ is the skew factor

$(u_{0g} \ v_{0g})$ represent the coordinates of the principal point in the images.

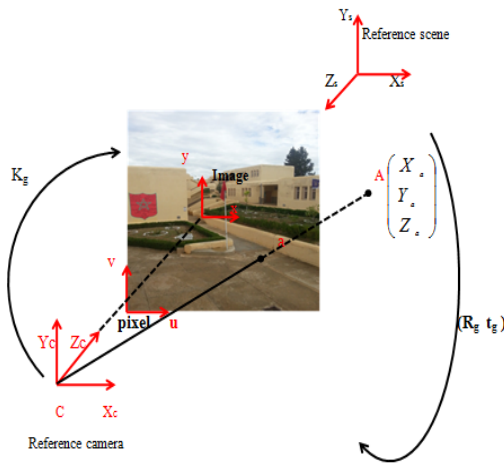


Fig. 1. Pinhole camera model

B. Interests points

Interest points are characteristic points of the image that are particularly holders of information. The detection of these points in stereoscopic images is an essential step in the field of Computer Vision and especially in the Three-dimensional Reconstruction. It is to match the projections of the same entity in the scene. The pairing of the corners points is based on the calculation and evaluation of degree of similarity of two pixel areas. For a pixel and its vicinity in the left image, we look at the other image pixel and its neighborhood that suits them the most. This step requires robust algorithms vis-a-vis different geometrical transformations and changes in illuminance disturbances.

We begin by extracting the corners points with the Harris detector that exists in the literature:

1) Detection Of Interest Points

The detectors of interests points found in the literature are: Moravec detector [27], Harris detector, Susan detector [28],... Harris [23],[26] developed the method Moravec for calculating local maxima in the images used by a matrix N related to the autocorrelation function which takes into account the first derivatives of the values of the signal I on a window in image space. N matrix is calculated by the following formula:

$$N = \begin{pmatrix} \left(\frac{\partial I}{\partial u} \right)^2 & \left(\frac{\partial I}{\partial u} \right) \left(\frac{\partial I}{\partial v} \right) \\ \left(\frac{\partial I}{\partial u} \right) \left(\frac{\partial I}{\partial v} \right) & \left(\frac{\partial I}{\partial v} \right)^2 \end{pmatrix} \quad (2)$$

In order not to extract the values of this matrix, Harris uses a variable r which is greater than zero in the case of a corner(interest point), its value is given by:

$$r = \det(N) - \gamma [\text{trace}(N)]^2 \quad (3)$$

with

$$\gamma = 0.04 \text{ (Value fixed by Harris)}$$

The detected primitive type is given by the values of r, three cases:

$r < 0$: in the vicinity of an edge

$r = 0$: in a homogeneous region

$r > \text{threshold}$: near a point of corners

2) Correlation Measure

Corresponding points between two images of the sequence are the points of Harris previously detected in each image and matched by the correlation measure ZNCC(Zero mean Normalized Cross Correlation) [24], [25] which is invariant to local linear change luminance .m and m' are two interest points detected in the left and right image respectively. Measurement correlation ZNCC(m, m') is given by the formula follows:

$$\text{ZNCC}(m, m') = \frac{\sum_i ((I(m+i) - \bar{I}(m))(I'(m'+i) - \bar{I}'(m'))))}{\sqrt{\sum_i (I(m+i) - \bar{I}(m))^2 \sum_i (I'(m'+i) - \bar{I}'(m'))^2}} \quad (4)$$

With I (m) and I'(m') the average luminance of pixels in a window of size 11x11 respectively in m and m' and i varies from Ito n.

IV. ESTIMATION OF THE PROJECTION MATRICES

Considering two points A_1, A_2 and their opposite A_3, A_4 relative to the axis of abscise on the 3D scene. Let π a plane that contains these four points, we consider R an euclidean reference (O X Y Z) such as O is the midpoint of the chord $[A_1A_2]$, and Z is perpendicular on the plan of the scene ($Z \perp \pi$). The homogeneous coordinates of the four points A_1, A_2, A_3 and A_4 (Figure 2) in the reference R (O X Y) are given as follows:

$$A_1 = (d \cos \varphi, d \sin \varphi, 1)^T$$

$$A_2 = (-d \cos \varphi, -d \sin \varphi, 1)^T$$

$$A_3 = (d \cos \varphi, -d \sin \varphi, 1)^T$$

$$A_4 = (-d \cos \varphi, d \sin \varphi, 1)^T$$

Where $d=A_1A_2/2$ and φ is the angle between the chord $[A_1A_2]$ and the X -axis. Considering two homographies H_g and H_d that can project the plane π in images g and d , therefore, the projection of the four points A_1, A_2, A_3 and A_4 can be given by the following expressions:

$$a_{gk} = H_g A_k \quad (5)$$

Where $k = 1, 2, 3, 4$ and a_{gk} represent the points in the images g that are the projections of the four vertices A_1, A_2, A_3 and A_4 of the 3D scene, and H_g represent the homography matrices that are expressed as follows:

$$H_g \sim K_g R_g \begin{pmatrix} 1 & 0 & & \\ 0 & 1 & R_g^T t_g & \\ 0 & 0 & & \end{pmatrix} \quad (6)$$

Expressions (5) can be written as follows:

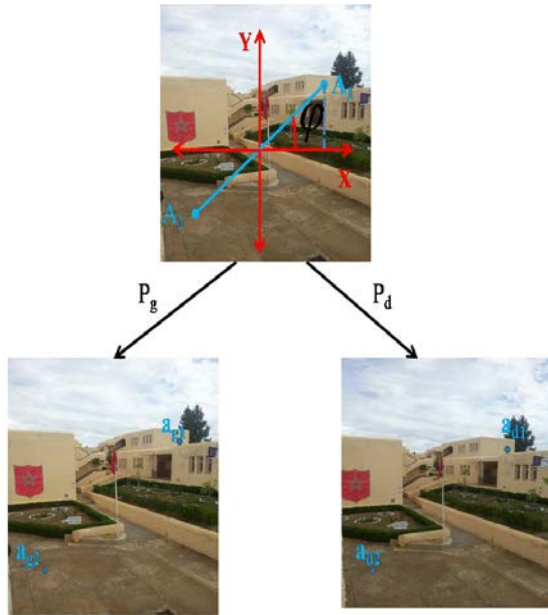


Fig. 2. The projection of the two points A_1, A_2 in the planes of images g and d

Expression (9) gives:

$$\begin{aligned} a_{g1} &\sim P_g A_1 \\ a_{g2} &\sim P_g A_2 \\ a_{g3} &\sim P_g A_3 \\ a_{g4} &\sim P_g A_4 \end{aligned} \quad (10)$$

$$a_{gk} \sim H_g M A_k' \quad (7)$$

Where $M = \begin{pmatrix} d \cos \varphi & 0 & 0 \\ 0 & d \sin \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and

$$A_k' = \begin{pmatrix} m_1 \\ m_2 \\ 1 \end{pmatrix} \begin{cases} k=1 & m_1=1 & m_2=1 \\ k=2 & \Leftrightarrow & m_1=-1 & m_2=-1 \\ k=3 & & m_1=1 & m_2=-1 \\ k=4 & & m_1=-1 & m_2=1 \end{cases}$$

We can represent the projection matrices by:

$$P_g \sim H_g M \quad (8)$$

Where P_g represent the projection matrices of the four points A_1, A_2, A_3, A_4 in images (Figure 2).

Formula (8) give:

$$a_{gk} \sim P_g A_k' \quad (9)$$

The latter four relations give eight equations with eight unknowns, which are the P_g elements.

So, the P_g parameters can be estimated from these eight equations with eight unknowns.

V. SELF-CALIBRATION EQUATIONS

A nonlinear cost function will be defined in the main part of this work from the determination of the relationships

between the images of the absolute conic (ω_g) and from the relationships between two points (A_1, A_2) and his oppositely relative to the X axis(A_3, A_4) of the 3D scene and their projections ($A_{g1}, A_{g2}, A_{g3}, A_{g4}$) and in the planes of the images g, respectively. The different relationships are based on some techniques of projective geometry. The defined cost function will be minimized by the genetic algorithms to estimate the ω_g elements and, finally, by the intrinsic parameters of the cameras used.

Expression (9) gives

$$\alpha_{gk} a_{gk} = P_g A_k' \quad (11)$$

Where

$$P_g = \begin{pmatrix} P_{g11} & P_{g12} & P_{g13} \\ P_{g21} & P_{g22} & P_{g23} \\ P_{g31} & P_{g32} & P_{g33} \end{pmatrix}, \quad a_{gk} = \begin{pmatrix} x_{gk} \\ y_{gk} \\ 1 \end{pmatrix} \quad \text{and}$$

$$\alpha_{gk} = m_1 P_{g31} + m_2 P_{g32} + P_{g33}$$

α_{gk} is a nonzero scale factor that is used to realized the transition between equality with a scale factor \sim to precise equality $=$. The value of α_{gk} is determined from expression (11).

Therefore, formula (11) leads to:

$$\alpha_{gk} a_{gk} = P_g A_k'' \quad (12)$$

Where

$$a_{gk}' = \begin{pmatrix} x_{gk} & \frac{P_{g12}}{\alpha_{gk}} & \frac{P_{g13}}{\alpha_{gk}} \\ y_{gk} & \frac{P_{g22}}{\alpha_{gk}} & \frac{P_{g23}}{\alpha_{gk}} \\ 1 & \frac{P_{g32}}{\alpha_{gk}} & \frac{P_{g33}}{\alpha_{gk}} \end{pmatrix} \quad \text{and} \quad A_k'' = \begin{pmatrix} m_1 & 0 & 0 \\ m_2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Expression (12) gives:

$$P_g \sim a_{gk}' A_k''^{-1} \quad (13)$$

Expressions (6) and (8) give:

$$P_g \sim K_g R_g \begin{pmatrix} m_1 & 0 \\ m_2 & 1 \\ 1 & 0 \end{pmatrix} R_g^T t_g M \quad (14)$$

The previous formula leads to:

$$K_g^{-1} P_g \sim R_g \begin{pmatrix} m_1 & 0 \\ m_2 & 1 \\ 1 & 0 \end{pmatrix} R_g^T t_g M \quad (15)$$

According to the formula (15) we have:

$$P_g^T \omega_g P_g \sim \begin{pmatrix} M^T M' & M^T R_g^T t_g \\ t_g^T R_g M' & t_g^T t_g \end{pmatrix} \quad (16)$$

Where $\omega_g = (K_g K_g^T)^{-1}$ is the image of the absolute conic,

$$\text{and } M' = \begin{pmatrix} d \cos \varphi & 0 \\ 0 & d \sin \varphi \\ 0 & 0 \end{pmatrix}$$

Expressions (13) and (16) give:

$$(a_{gk}' A_k''^{-1})^T \omega_g (a_{gk}' A_k''^{-1}) \sim \begin{pmatrix} M^T M' & M^T R_g^T t_g \\ t_g^T R_g M' & t_g^T t_g \end{pmatrix} \quad (17)$$

The previous expression gives:

$$(a_{g1}' A_1''^{-1})^T \omega_g (a_{g1}' A_1''^{-1}) \sim \begin{pmatrix} M^T M' & M^T R_g^T t_g \\ t_g^T R_g M' & t_g^T t_g \end{pmatrix} \quad (18)$$

$$(a_{g2}' A_2''^{-1})^T \omega_g (a_{g2}' A_2''^{-1}) \sim \begin{pmatrix} M^T M' & M^T R_g^T t_g \\ t_g^T R_g M' & t_g^T t_g \end{pmatrix} \quad (19)$$

$$(a_{g3}' A_3''^{-1})^T \omega_g (a_{g3}' A_3''^{-1}) \sim \begin{pmatrix} M^T M' & M^T R_g^T t_g \\ t_g^T R_g M' & t_g^T t_g \end{pmatrix} \quad (20)$$

$$(a_{g4}' A_4''^{-1})^T \omega_g (a_{g4}' A_4''^{-1}) \sim \begin{pmatrix} M^T M' & M^T R_g^T t_g \\ t_g^T R_g M' & t_g^T t_g \end{pmatrix} \quad (21)$$

The expressions (18), (19), (20) and (21) give:

$$(a_{g1}' A_1''^{-1})^T \omega_g (a_{g1}' A_1''^{-1}) \sim (a_{g2}' A_2''^{-1})^T \omega_g (a_{g2}' A_2''^{-1}) \sim (a_{g3}' A_3''^{-1})^T \omega_g (a_{g3}' A_3''^{-1}) \sim (a_{g4}' A_4''^{-1})^T \omega_g (a_{g4}' A_4''^{-1}) \quad (22)$$

Let

$$B = \begin{pmatrix} b_{11g} & b_{12g} & b_{13g} \\ b_{21g} & b_{22g} & b_{23g} \\ b_{31g} & b_{32g} & b_{33g} \end{pmatrix} \quad \text{denotes the matrix corresponding to}$$

$$(a_{g1}' A_1''^{-1})^T \omega_g (a_{g1}' A_1''^{-1})$$

$$C = \begin{pmatrix} c_{11g} & c_{12g} & c_{13g} \\ c_{21g} & c_{22g} & c_{23g} \\ c_{31g} & c_{32g} & c_{33g} \end{pmatrix} \quad \text{denotes the matrix corresponding to}$$

$$(a_{g2}' A_2''^{-1})^T \omega_g (a_{g2}' A_2''^{-1})$$

$$D = \begin{pmatrix} d_{11g} & d_{12g} & d_{13g} \\ d_{21g} & d_{22g} & d_{23g} \\ d_{31g} & d_{32g} & d_{33g} \end{pmatrix} \quad \text{denotes the matrix corresponding to}$$

$$(a_{g3}' A_3''^{-1})^T \omega_g (a_{g3}' A_3''^{-1})$$

$$E = \begin{pmatrix} e_{11g} & e_{12g} & e_{13g} \\ e_{21g} & e_{22g} & e_{23g} \\ e_{31g} & e_{32g} & e_{33g} \end{pmatrix} \text{ denotes the matrix corresponding to}$$

$$(a'_{g4}A_4''^{-1})^T \omega_g (a'_{g4}A_4''^{-1})$$

$$\text{and } M^T M' = \begin{pmatrix} d^2 \cos^2 \varphi & 0 \\ 0 & d^2 \sin^2 \varphi \end{pmatrix}$$

Therefore, the formula (22) with the previous gives:

$$\begin{cases} b_{12g} = 0, \\ c_{12g} = 0, \\ d_{12g} = 0, \\ e_{12g} = 0, \\ \frac{b_{11g}}{b_{13g}} = \frac{c_{11g}}{c_{13g}} = \frac{d_{11g}}{d_{13g}} = \frac{e_{11g}}{e_{13g}} \\ \frac{b_{13g}}{b_{22g}} = \frac{c_{13g}}{c_{22g}} = \frac{d_{13g}}{d_{22g}} = \frac{e_{13g}}{e_{22g}} \\ \frac{b_{22g}}{b_{23g}} = \frac{c_{22g}}{c_{23g}} = \frac{d_{22g}}{d_{23g}} = \frac{e_{22g}}{e_{23g}} \\ \frac{b_{23g}}{b_{33g}} = \frac{c_{23g}}{c_{33g}} = \frac{d_{23g}}{d_{33g}} = \frac{e_{23g}}{e_{33g}} \end{cases} \quad (23)$$

The previous equations contain twenty eight equations with five unknowns that are the elements of ω_g . This system is non-linear. So to solve it, we try to minimize the objective function with the genetic algorithms:

$$\begin{cases} b_{12g} = 0, & c_{12g} = 0, & d_{12g} = 0, & e_{12g} = 0, \\ b_{11g}c_{13g} - c_{11g}b_{13g} = 0, & b_{11g}d_{13g} - d_{11g}b_{13g} = 0, & b_{11g}e_{13g} - e_{11g}b_{13g} = 0, \\ c_{11g}d_{13g} - d_{11g}c_{13g} = 0, & c_{11g}e_{13g} - e_{11g}c_{13g} = 0, & d_{11g}e_{13g} - e_{11g}d_{13g} = 0, \\ b_{13g}c_{22g} - c_{13g}b_{22g} = 0, & b_{13g}d_{22g} - d_{13g}b_{22g} = 0, & b_{13g}e_{22g} - e_{13g}b_{22g} = 0, \\ c_{13g}d_{22g} - d_{13g}c_{22g} = 0, & c_{13g}e_{22g} - e_{13g}c_{22g} = 0, & d_{13g}e_{22g} - e_{13g}d_{22g} = 0, \\ b_{13g}c_{22g} - c_{13g}b_{22g} = 0, & b_{13g}d_{22g} - d_{13g}b_{22g} = 0, & b_{13g}e_{22g} - e_{13g}b_{22g} = 0, \\ c_{13g}d_{22g} - d_{13g}c_{22g} = 0, & c_{13g}e_{22g} - e_{13g}c_{22g} = 0, & d_{13g}e_{22g} - e_{13g}d_{22g} = 0, \\ b_{22g}c_{23g} - c_{22g}b_{23g} = 0, & b_{22g}d_{23g} - d_{22g}b_{23g} = 0, & b_{22g}e_{23g} - e_{22g}b_{23g} = 0, \\ c_{22g}d_{23g} - d_{22g}c_{23g} = 0, & c_{22g}e_{23g} - e_{22g}c_{23g} = 0, & d_{22g}e_{23g} - e_{22g}d_{23g} = 0, \\ b_{23g}c_{33g} - c_{23g}b_{33g} = 0, & b_{23g}d_{33g} - d_{23g}b_{33g} = 0, & b_{23g}e_{33g} - e_{23g}b_{33g} = 0, \\ c_{23g}d_{33g} - d_{23g}c_{33g} = 0, & c_{23g}e_{33g} - e_{23g}c_{33g} = 0, & d_{23g}e_{33g} - e_{23g}d_{33g} = 0. \end{cases} \quad (24)$$

VI. MINIMIZATION AND INITIALIZATION OF THE NON-LINEAR OBJECTIVE FUNCTION

To solve the equations (24), in practice there isn't a direct method to solve them. So to solve this problem, we minimized the following non-linear objective function:

$$F(p_l) = \sum_{g=1}^n (\beta_g^2 + \gamma_g^2 + \delta_g^2 + \eta_g^2 + i_g^2 + \kappa_g^2 + \mu_g^2 + \nu_g^2 + o_g^2 + \pi_g^2 + \varpi_g^2 + \theta_g^2 + \varrho_g^2 + \rho_g^2 + \sigma_g^2 + \zeta_g^2 + \tau_g^2 + \upsilon_g^2 + \omega_g^2 + \xi_g^2 + \psi_g^2 + \zeta_g^2 + L_g^2 + M_g^2 + N_g^2 + Q_g^2 + R_g^2 + S_g^2 + T_g^2 + G_g^2 + H_g^2 + J_g^2 + \delta_g^2 + \lambda_g^2) \quad (25)$$

With n is the number of images,

$$\left\{ \begin{array}{l} \beta_i = b_{12g}, \quad \gamma_i = c_{12g}, \quad \delta_i = d_{12g}, \quad \eta_i = e_{12g}, \\ t_i = b_{11g}c_{13g} - c_{11g}b_{13g}, \quad \kappa_i = b_{11g}d_{13g} - d_{11g}b_{13g}, \quad \mu_i = b_{11g}e_{13g} - e_{11g}b_{13g}, \\ v_i = c_{11g}d_{13g} - d_{11g}c_{13g}, \quad o_i = c_{11g}e_{13g} - e_{11g}c_{13g}, \quad \pi_i = d_{11g}e_{13g} - e_{11g}d_{13g}, \\ \varpi_i = b_{13g}c_{22g} - c_{13g}b_{22g}, \quad \theta_i = b_{13g}d_{22g} - d_{13g}b_{22g}, \quad \varrho_i = b_{13g}e_{22g} - e_{13g}b_{22g}, \\ \rho_i = c_{13g}d_{22g} - d_{13g}c_{22g}, \quad \sigma_i = c_{13g}e_{22g} - e_{13g}c_{22g}, \quad \varsigma_i = d_{13g}e_{22g} - e_{13g}d_{22g}, \\ \tau_i = b_{13g}c_{22g} - c_{13g}b_{22g}, \quad \upsilon_i = b_{13g}d_{22g} - d_{13g}b_{22g}, \quad \omega_i = b_{13g}e_{22g} - e_{13g}b_{22g}, \\ \xi_i = c_{13g}d_{22g} - d_{13g}c_{22g}, \quad \psi_i = c_{13g}e_{22g} - e_{13g}c_{22g}, \quad \zeta_i = d_{13g}e_{22g} - e_{13g}d_{22g}, \\ L_i = b_{22g}c_{23g} - c_{22g}b_{23g}, \quad M_i = b_{22g}d_{23g} - d_{22g}b_{23g}, \quad N_i = b_{22g}e_{23g} - e_{22g}b_{23g}, \\ Q_i = c_{22g}d_{23g} - d_{22g}c_{23g}, \quad R_i = c_{22g}e_{23g} - e_{22g}c_{23g}, \quad S_i = d_{22g}e_{23g} - e_{22g}d_{23g}, \\ T_i = b_{23g}c_{33g} - c_{23g}b_{33g}, \quad G_i = b_{23g}d_{33g} - d_{23g}b_{33g}, \quad H_i = b_{23g}e_{33g} - e_{23g}b_{33g}, \\ J_i = c_{23g}d_{33g} - d_{23g}c_{33g}, \quad \hat{o}_i = c_{23g}e_{33g} - e_{23g}c_{33g}, \quad \hat{\lambda}_i = d_{23g}e_{33g} - e_{23g}d_{33g}. \end{array} \right.$$

To solve the non-linear objective function (25), we used genetic algorithms [29] which require an important initialization step which is to calculate the unknowns assuming certain conditions were verified. Replacing at the end these parameters in the system of equations (24) allows the estimation of the intrinsic camera parameters.

The initialization values are selected such that each parameter of the camera belongs to a specific interval:

TABLE I. VARIATION INTERVAL OF THE CAMERA SETTING

	Variation interval
f_g	[800 2000]
\mathcal{E}_g	[0 1]
τ	[0 1]
u_{0g}	[200 300]
v_{0g}	[200 300]

VII. EXPERIMENTATION

1) Real Data

In this section, a sequence of two images of a checkerboard pattern is simulated to test the performance and robustness of the present approach. After the detection of interests points by the Harris algorithm [26], the matches between each pair of images are determined by the correlation function *ZNCC* [27]. The pattern is projected in images taken from different views with Gaussian noise of standard

deviation σ , which is added to all image pixels. The projection of the four points in the image planes allows formulating the linear equations, and the solution of these equations gives the projection matrices. The determination of a relationship between the four points and their projections in the image g and the relationships between images of the absolute conic can define a non linear cost function. The minimization of this function by the genetic algorithms [29] allows estimating the intrinsic parameters of the cameras used.

To achieve our theoretical idea and have a practical result, we based on open source tools to offer an application that implements the algorithms used in our article.

The tools used are:

- A robust programming language in the field, object-oriented open source such as Java.
- Sophisticated and open source APIs to solve mathematical complexities such as JAMA, Jscientific and JAI.
- The Swing API for loading images and the realization of different graphical interfaces.

A pair of images is loaded into our application (Figure 3), after the interests points are computed in each image, using the implementation of Harris algorithm (Figure 4), and then the matching of interest points is calculated in the phase matching by implementing the correlation algorithm *ZNCC* (Figure 5).

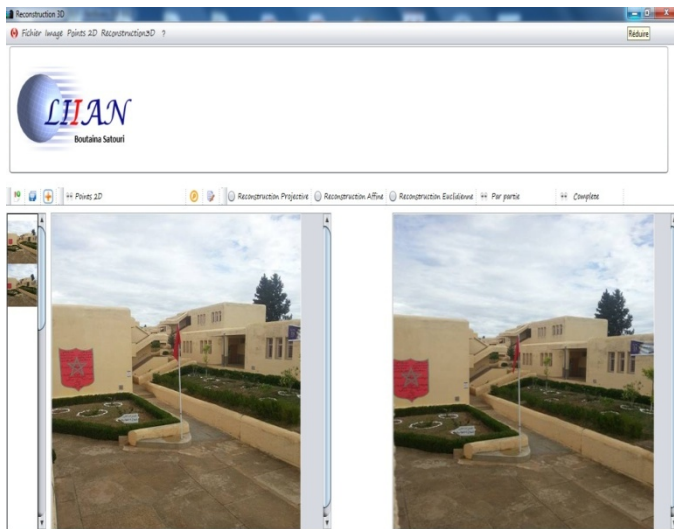


Fig. 3. loading apair of images



Fig. 4. detection of Interest points for each image

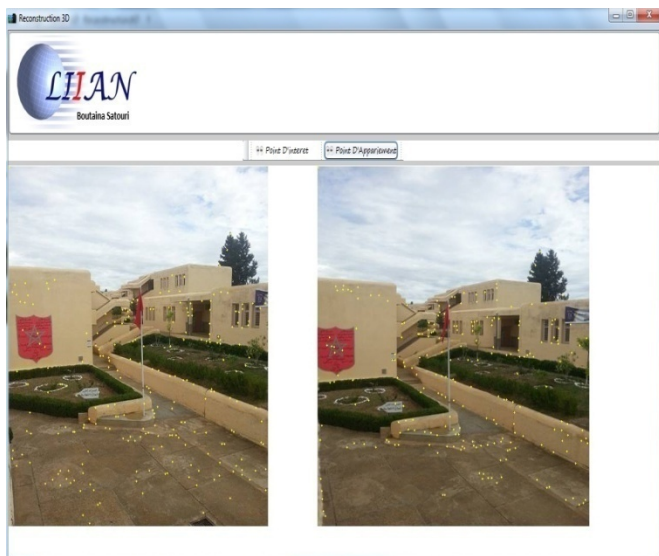


Fig. 5. Matching Interest points by the Correlation algorithm ZNCC

To estimate the intrinsic parameters of each camera by our approach two phases must be implemented initialization to provide a initial solution and the minimization of cost function (24) to find an optimal solution (Table 2).

VIII. RESULTS AND COMPARAISON OF OUR METHOD WITH OTHER

TABLE II. THE RESULTS OF THE INTRINSIC CAMERA PARAMETERS ESTIMATED BY THE THREE METHODS

		Optimal solution				
		cameras	f	ε	u_0	v_0
Our approach	Camera1	1487	0,98	257	263	0,03
	Camera2	1480	0,97	261	258	0,02
	Camera3	1483	0,93	258	261	0,05
	Camera4	1484	0,95	260	255	0,07
Jiang	Camera1	1477	1	263	258	0
	Camera2	1472	0,98	260	263	0,12
	Camera3	1474	0,92	254	261	0,15
	Camera4	1471	0,94	257	255	0,13
El akkad	Camera1	1492	0,95	240	260	0,05
	Camera2	1490	0,92	248	255	0,06
	Camera3	1491	0,97	253	258	0,04
	Camera4	1489	0,93	252	262	0,03

In order to show the performance and robustness of our method presented in this paper, the simulation results are compared to those obtained by several efficient methods of Jiang [14], and El akkad [50].

The loading of images is shown in Figure 3, the corner points and the matches between these two images are shown in Figure 4, 5 and the intrinsic parameters estimated by three methods(the present method, El akkad [50], and Jiang’s [14]) are shown in Table 2below.

After comparing the results on the synthetic data, the results of the present approach on real data are compared to those obtained by El akkad [50] and Jiang [14] on the same data. The reading and the analysis of the intrinsic camera parameters presented in Table 2show that the results of the present approach are a little different from those obtained by Jiang and El akkad. Therefore, this method provides a robust performance, and it is very close to the other well-established methods. In addition, this method has several advantages: it is based only on the data of the first image without the use the data of the second image for the estimation of the intrinsic camera parameters, the use of any camera and the use of an unknown3D scene.

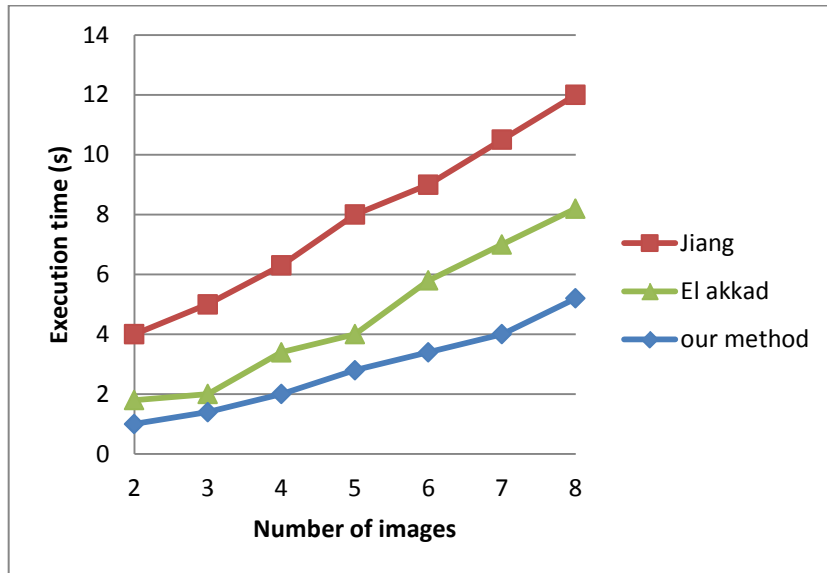


Fig. 6. The execution time according to the number of images

Figure 6 shows that the execution time of the different methods increases, and finally it shows the effect of the use of a large image number.

IX. CONCLUSION

In this paper, a robust method of camera self-calibration by an unknown three-dimensional scene is presented. The lack of information about the scene requires to be based on mathematical complexities to find the camera settings. This new method is based on the determination of a relationship between two points and oppositely relative to the X axis in the 3D scene and their projections in the planes of the images g and d and between the relationships between the images of the absolute conic. These relationships give a nonlinear cost function, and the minimization of this function provides the

intrinsic parameters of the cameras used. Our technology is used to provide more information about the scene which makes the calculation of the parameters in images very simple.

Our method allows easily and without the use or of the data of the second image or the geometrical entities defined between the pair of images estimating the intrinsic parameters of each camera independently from each other. The robustness, the power and the rapidity ofthis method is shown by the results of the experiments andthe simulations conducted.

REFERENCES

- [1] Xiaoqiao Meng, Hua Li and Zhanyi Hu. A New Easy Camera Calibration Technique Based on Circular Points, BMVC2000.
- [2] Liangfu Li, Zuren Feng, Yuanjing Feng. Accurate Calibration of Stereo Cameras for Machine Vision. JCS&T Vol. 4 No. 3, October 2004.

- [3] Jin Sun, Hongbin Gu. Research of Linear Camera Calibration Based on Planar Pattern. World Academy of Science, Engineering and Technology 60, 2009.
- [4] Tarek A. Al-Saeed¹, Nahed H. Solouma², and Said M. El-Sherbiny³. Retinal Motion Detection and 3D Structure Recovery From Two Perspective Views. GVIP Special Issue on Medical Image Processing, March, 2006.
- [5] A.Elabderrahmani, A.Saaidi and K.Satori. Planar Self-Calibration with Less Constraint. IJCST Vol. 2, Issue 2, June 2011.
- [6] A.El abderrahmani, A.Saaidi and K. Satori. Robust Technique for Self-Calibration of Cameras based on a Circle. ICGST-GVIP, Volume 10, Issue 5, December 2010
- [7] Triggs, B.: Autocalibration from planar scenes. In: Proceedings of the 5th European Conference on Computer Vision, pp. 89–105 (1998)
- [8] Sturm, P.: A case against Kruppa's equations for camera selfcalibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1199–1204 (2000)
- [9] Saaidi, A., Halli, A., Tairi, H., Satori, K.: Self-calibration using a particular motion of camera. WSEAS Tran. Comput. Res. **3**(5), 295–299 (2008)
- [10] Zhao, Y., Lv, X.D.: An approach for camera self-calibration usingvanishing-line. Inf. Technol. J. **112**, 276–282 (2012)
- [11] Zhang, W.: A simple method for 3D reconstruction from twoviews. In: GVIP 05 Conference (2005)
- [12] Saaidi, A., Halli, A., Tairi, H., Satori, K.: Self-calibration usinga planar scene and parallelogram. In: ICGST-GVIP, pp. 41–47(2009)
- [13] Liu, P., Shi, J., Zhou, J., Jiang, L.: Camera self-calibration usingthe geometric structure in real scenes. In: Proceedings of the ComputerGraphics International, pp. 262–265 (2003)
- [14] Baataoui, A., El batteoui, I., Saaidi, A., Satori, K.: Camera selfcalibrationby an equilateral triangle. Int. J. Comput. Appl., 29–34(2012)
- [15] P.Gurdjos and P.Sturm. Methods and Geometry for Plane-Based Self-Calibration. CVPR, pp. 491-496, 2003.
- [16] Manolis I.A. Lourakis and R.Deriche. Camera self-calibration using the kruppa equations and the SVD of the fundamental matrix: the case of varying intrinsic parameters. Technical Report 3911, INRIA, 2000.
- [17] Sturm, P.: Critical motion sequences for the self-calibration ofcameras and stereo systems with variable focal length. Image Vis.Comput. **20**, 415–426 (2002)
- [18] Cao, X., Xiao, J., Foroosh, H., Shah, M.: Self-calibration from turntable sequences in presence of zoom and focus. Comput. Vis.Image Underst. **103**(2), 227–237 (2006)
- [19] Jiang, Z., Liu, S.: The self-calibration of varying internal camera parameters based on image of dual absolute quadric transformation. In: Information and Automation, Communications in Computer and Information Science, vol. 86, pp. 452–461. Springer, Berlin (2011)
- [20] Shang, Y., Yue, Z., Chen, M., Song, Q.: A new method of camera self-calibration based on relative lengths. Inf. Technol. J. **11**(3), 376–379 (2012)
- [21] Jiang, Z., Liu, S.: Self-calibration of varying internal camera parameters algorithm based on quasi-affine reconstructio J. Comput. **7**(3), 774–778 (2012)
- [22] Zhao, Y., Hu, X., Lv, X., Wang, H.: Solving the camera intrinsic parameters with the positive tri-prism based on the circular points. Inf. Technol. J. **11**(7), 926–930 (2012)
- [23] Gao, Y., Radha, H.: A multistage camera self-calibration algorithm. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 537–540 (2004)
- [24] El akkad, N., Saaidi, A., Satori, K.: Self-calibration based on a circle of the cameras having the varying intrinsic parameters. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems, pp. 161–166 (2012)
- [25] Hemayed, E.E.: A survey of camera self-calibration. In: Proceedings of the IEEE Conference on Advanced Video and Signal BasedSurveillance, pp. 351–357 (2003)
- [26] C. Harris and M. Stephens, "A combined corner and edge detector," In Alvey Vision Conference, 1988.
- [27] Chambon, S., Crouzil, A.: Similarity measures for image matching despite occlusions in stereo vision. Pattern Recognit. **44**(9), 2063–2075 (2011)
- [28] Torr, P.H.S., Murray, D.W.: The development and comparison of robust methods for estimating the fundamental matrix. Int. J.Comput. Vis. **24**, 271–300 (1997)
- [29] J. H., Holland. Adaptation in Natural and Artificial Systems. *University Bibliographieof Michigan Press: Ann Arbor, 1975.*
- [30] M.J. Brooks, W. Chojnacki, and L. Baumela. Determining the ego-motion of an uncalibrated camera from instantaneous optical flow. Journal of the Optical Society of America, 14(10), October 1997.
- [31] M.J. Brooks, L. de Agapito, D.Q. Huynh, and L. Baumela. Direct methods for self-calibration of a moving stereo head. In B. Buxton and R. Cipolla, editors, Proceedings of the 4th European Conference on Computer Vision, Cambridge, England, volume 1065 of Lecture Notes in Computer Science, pages 415.426. Springer-Verlag, April 1996.
- [32] M. Armstrong, A. Zisserman, and R. Hartley. Self-calibration from image triplets. In B. Buxton and R. Cipolla, editors, Proceedings of the 4th European Conference on Computer Vision, Cambridge, England, volume 1064 of Lecture Notes in Computer Science, pages 3.16. Springer-Verlag, April 1996.
- [33] C. Wiles and M. Brady. Ground plane motion camera models. In B. Buxton and R. Cipolla, editors, Proceedings of the 4th European Conference on Computer Vision, Cambridge, England, volume 1065 of Lecture Notes in Computer Science, pages 238.247. Springer-Verlag, April 1996.
- [34] P.Sturm. A case against Kruppa's equations for camera self-calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, Issue 10, pp. 1199-1204, October 2000.
- [35] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, seconde édition, Cambridge University Press, ISBN 0-521-54051-8, 2004.
- [36] Houman Rastgar, Eric Dubois, Liang Zhang. Random Sampling Nonlinear Optimization for Camera Self-calibration with Modeling of Intrinsic Parameter Space. Advances in Visual Computing, 6th International Symposium, ISVC 2010, Las Vegas, NV, USA, November 29 - December 1, 2010, Proceedings, Part III, pp 189-198.
- [37] B. Boudine, A. El abderrahmani, A. Saaidi et K. Satori. Détection des points d'intérêt entre les détecteurs Harris, Sift et Surf. Communication au Workshop en Imagerie, Systèmes et Applications. Mai 23-24, 2013, FP-Taza Maroc.
- [38] Adlane Habed, Kassem Al Ismaeil, David Fofi. A New Set of Quartic Trivariate Polynomial Equations for Stratified Camera Self-calibration under Zero-Skew and Constant Parameters Assumptions. Computer Vision– ECCV 2012 Lecture Notes in Computer Science Volume 7577, 2012, pp 710-723.
- [39] Adlane Habed, Tarik Elamsy, Boubakeur Boufama. Combining Mendonça-Cipolla Self-calibration and Scene Constraints. Advances in Image and Video Technology Lecture Notes in Computer Science Volume 7088, 2012, pp 168-179.
- [40] B. Triggs, Autocalibration and the Absolute Quadric, Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, Porto Rico, USA, p. 609-614, 17-19 juin 1997.
- [41] A. Vladés, J. Ronda and G. Gallego, The Absolute Line Quadric And Camera Autocalibration, International Journal of Computer Vision, vol. 66, n° 3, p. 283-303, mars 2006.
- [42] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, seconde édition, Cambridge University Press, ISBN 0-521-54051-8, 2004.
- [43] P. Sturm, On Focal Length Calibration from Two Views, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, Hawaï, vol. 2, p. 145-150, 8-14 décembre 2001.
- [44] Y. Li, W. K. Tang , Y. S. Hung Stratified Self Calibration and Metric Reconstruction for Zooming/Refocusing Circular Motion Sequences. Journal of Mathematical Imaging and Vision. May 2008, Volume 31, Issue 1 , pp 105-118.
- [45] Ze-tao Jiang, Shan-chao Liu. The Self-calibration of Varying Internal Camera Parameters Based on Image of Dual Absolute Quadric

- Transformation. Information and Automation Communications in Computer and Information Science Volume 86, 2011, pp 452-461.
- [46] Zetao Jiang, Shutao Guo, Lianggang Jia. Sequences Images Based Camera Self-calibration Method. Intelligent Science and Intelligent Data Engineering Lecture Notes in Computer Science Volume 7202, 2012, pp 538-545.
- [47] Jong-Eun Ha, Dong-Joong Kang. Initialization Method for the Self-Calibration Using Minimal Two Images. Computational Science and Its Applications – ICCSA 2004 Lecture Notes in Computer Science Volume 3046, 2004, pp 915-923.
- [48] Javier Civera, Andrew J. Davison, José María Martínez Montiel. Self-calibration. Structure from Motion using the Extended Kalman Filter Springer Tracts in Advanced Robotics Volume 75, 2012, pp 111-122.
- [49] Thorsten Thormählen, Hellward Broszio, Patrick Mikulastik. Robust Linear Auto-calibration of a Moving Camera from Image Sequences. Computer Vision – ACCV 2006. Lecture Notes in Computer Science Volume 3852, 2006, pp 71-80.
- [50] N. El akkad, M. Merras, A. Saaidi and K. Satori. Camera Self-Calibration with Varying Intrinsic Parameters by an Unknown Three-Dimensional Scene. *The Visual Computer (Springer)*. Vol. 30, No. 5, pp. 519-530, 2014.

On the Internal Multi-Model Control of Uncertain Discrete-Time Systems

Chakra Othman

Laboratory L.A.R.A, National
Engineering School of Tunis
Tunis Elmanar University
Tunis, Tunisia

Ikbel Ben Cheikh

Laboratory L.A.R.A, National
Engineering School of Tunis
Tunis Elmanar University
Tunis, Tunisia

Dhaou Soudani

Laboratory, L.A.R.A, National
Engineering School of Tunis
Tunis Elmanar University
Tunis, Tunisia

Abstract—In this paper, new approaches of internal multi-model control are proposed to be applied for the case of the discrete-time systems with parametric uncertainty. In this sense, two implantation structures of the internal multi-model control are adopted; the first is based on the principle of switching and the second on the residues techniques. The stability's study of these control structures is based on the Kharitonov theorem, thus two extensions of this theorem have been applied to define the internal models. To illustrate these approaches, simulation results are presented at the end of this article.

Keywords—Internal model control IMC; Internal multi-model control IMMC; Kharitonov theorem; Switching method; Residues techniques; discrete-time systems; uncertain systems

I. INTRODUCTION

The robustness problem in the case of parametric uncertainties aroused great interest among researchers. Robustness means the preservation of system characteristics such as stability or performance in the presence of unknown disturbances and noise.

Different control methods to solve this problem have been proposed. The internal model control has always been considered as an efficient approach in control systems, due to its high accuracy and robustness against internal and external disturbances. This method is generally used because of its robustness; it includes an inspired internal model of the process and a controller. It's preferable that this controller is the inverse of the internal model to ensure a perfect tracking of the reference. In this article, the internal multi-model control approach for the case of discrete-time uncertain systems is proposed to be applied.

Our contribution consists in implanting a control structure based on a multi-model controller in the discrete area by adopting two synthesis techniques namely switching technology and residues technique in order to minimize errors due to the modeling imperfections. These two techniques will be developed in this paper.

This structure contains instead of one internal model a set of models representing the process in different operating points by using multi-model approach and by consequent, a set of controllers that based on two specific inversion methods.

The system under consideration in this paper is a class of complex systems which is the discrete-time uncertain system with parametric uncertainty.

The multi-model approach is used to obtain the internal models inspired from the process of this control structure. It's a mathematical approach designed to represent the best possible the dynamic operation of a complex process, using linear time-invariant models. The multi-model approach allows representing complex systems in the form of interpolation between linear models. Each local model is a dynamic linear time invariant system valid around an operating point. [1]

Kharitonov method is used with these two theorems [15] for the discrete-time uncertain systems to determine the internal models of the multi-model control structure.

In these control structures, synthesis of the controller is reduced to a problem of internal models inverse construction. In addition, the direct inversion of the models is often impossible. Thus, the proposed controller synthesis approach is based on a specific inversion method. This approach has been modified to improve the accuracy of the controlled system. [3,4]

II. DISCRETE-TIME UNCERTAIN SYSTEMS

In practice there are many uncertainties that affect the physical system and therefore its model. In general, two uncertainty classes are distinguished, the structured uncertainties that affect the physical parameters value of the process model and the unstructured uncertainties defined by an upper bound of the model difference in the frequency domain. [2,13]

This article focuses on a class of uncertain systems where uncertainty is parametric.

III. STABILITY STUDY OF THE DISCRETE-TIME UNCERTAIN SYSTEMS USING KHARITONOV METHOD

The Kharitonov theorem is an important combination, generalizing the Routh-Hurwitz criterion. [10] The application of the Kharitonov theorem in the continuous case leads to false results for the uncertain discrete-time systems. This has required the development of this theorem in the discrete-time case. [14, 15, 16]

A. First extension of the Kharitonov method

Let $I(z)$ be the polynomials family of the following form:

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0 \quad (1)$$

where the coefficients belong to a box A :

$$A := \{a = (a_0, \dots, a_n) / a_i \in [a_i^-, a_i^+], i = 0, \dots, n\} \quad (2)$$

By introducing the vertices V and edges E of the box A:

$$V := \{(a_0, \dots, a_n), a_i = a_i^- \text{ or } a_i = a_i^+, i = 0, \dots, n\} \quad (3)$$

$$E_k := \left\{ (a_0, \dots, a_n) / a_i = a_i^- \text{ or } a_i^+, i = 0, \dots, n, \right. \\ \left. i \neq k, a_k \in [a_k^-, a_k^+] \right\} \quad (4)$$

$$\text{And } E = \bigcup_{k=0}^n E_k \quad (5)$$

The corresponding families of vertices and edges polynomials are defined by:

$$I_v(z) := \{P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0, (a_n, \dots, a_0) \in V\} \quad (6)$$

$$I_E(z) := \{P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0, (a_n, \dots, a_0) \in E\} \quad (7)$$

Theorem 1: Let $n > 1$ and assume that in the family $I(z)$ we have fixed upper order coefficients such that $a_i^- = a_i^+$ for $i = n/2 + 1, \dots, n$ if n is even and $i = (n+1)/2, \dots, n$ if n is odd. Then the entire family $I(z)$ is stable if and only if the family of vertex polynomials $I_v(z)$ is stable.

B. Second extension of the Kharitonov theorem

In the above theorem, the upper order coefficients are fixed, now let us consider that all the coefficients are allowed to vary, so let 'nu' be defined as follow:

$$nu = \left\{ \frac{n}{2} + 1, \frac{n}{2} + 2, \dots, n \right\} \quad \text{if } n \text{ is even} \quad (8)$$

$$nu = \left\{ \frac{(n+1)}{2} + 1, \frac{(n+1)}{2} + 2, \dots, n \right\} \quad \text{if } n \text{ is odd} \quad (9)$$

Let consider the upper edges E^* which can be defined by:

$$E_k^* := \left\{ (a_0, \dots, a_n) / a_i = a_i^- \text{ ou } a_i^+, i = 0, \dots, n, i \neq k, a_k \in [a_k^-, a_k^+], \right. \\ \left. k \in nu \right\} \quad (10)$$

The upper edge polynomials are obtained by varying a single higher order parameter and fixing others at their minimum or maximum values.

The family of higher edge polynomials can be defined by:

$$I_E^*(z) := \{P(z) = a_n z^n + a_{n-1} z^{n-1} \dots + a_0, (a_n, \dots, a_0) \in E^*\} \quad (11)$$

A typical upper edge in $I_E^*(z)$ is defined by:

$$a_n z^n + \dots + (\lambda a_k^- + (1 - \lambda) a_k^+) z^k \dots + a_0$$

$$k \in nu, a_i = a_i^- \text{ or } a_i = a_i^+, i = 0, \dots, n, i \neq k$$

There are: $\left(\frac{n}{2}\right) 2^n$ upper edges if n is even

$$\left(\frac{n+1}{2}\right) 2^n \text{ upper edges if } n \text{ is odd}$$

Theorem 2: The family of polynomials $I(z)$ is stable if and only if the family of edge polynomials $I_E^*(z)$ is stable.

IV. INTERNAL MODEL CONTROL STRUCTURE

This internal model control strategy acquired interest due to its robustness. The main advantage of this structure is the simplicity of its construction, and the easy interpretation of the roles of its blocks.[3,8]

The internal model control structure includes an internal model 'M' which is an explicit model of the process to be controlled and a regulator 'C' which can be chosen the inverse of the model and if necessary a robustness filter 'F' as indicated in figure (1). 'R', 'd', 'Y', are respectively the reference to reach, the modeling error and the system output. 'P' is a disturbance added at the output of the process

The internal model control structure used as a control signal the difference between the output of the process and its internal model.

In the basic structure of the IMC, the command signal "U" outcome from the corrector 'C' is applied simultaneously to the process 'G' and its model 'M'. The IMC exploits the behavior gap to correct the error on the reference. The error signal includes the influence of external disturbances and modeling errors.

Generally the internal model control structure includes a robustness filter usually introduced in the feedback loop. Its role is to introduce certain robustness against the modeling errors.

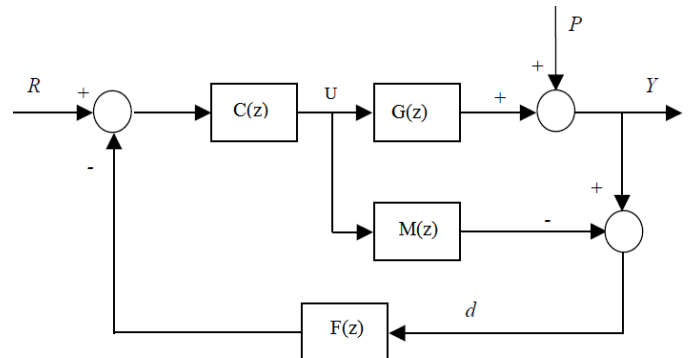


Fig. 1. Basic structure of the internal model control structure

In this article, the presence of the filter is not taken into account.

In this control structure, the controller is chosen equal to the model inverse to ensure the equality between the process output and the reference despite the added disturbance at the output. [3]

In addition, the direct inversion of the model is often impossible, especially when the model is with no-minimum phase or presents a delay, thus, inversion methods are used. The implementation method of the approximated inverse is used for systems with a transfer function whose order of the numerator is less than the order of the denominator, non-minimum phase systems and delay systems.

The following diagram is considered with $M(z)$ is the model transfer function and A_1 is a gain to choose. [5,6,7,8,9]

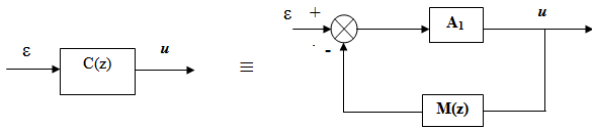


Fig. 2. Basic idea to obtain the approximated inverse

The global transfer function of the scheme (2) is:

$$C(z) = \frac{A_1}{1 + A_1 M(z)} \quad (12)$$

For sufficiently high values of the gain A_1 , the controller $C(z)$ approaches the inverse of internal model $M(z)$:

$$C(z) \approx \frac{1}{M(z)} \quad (13)$$

Thus, the global transfer function $C(z)$ is the approximated inverse of the model transfer function $M(z)$.

For some classes of systems, the gain A_1 that ensures stability of the loop that realize the controller $C(z)$, may not be very high, which does not allow us to obtain the approximated inverse, therefore, a gain A_2 is added to ensure a null static error. Thus, a second structure of the corrector is proposed: [5, 6, 8]

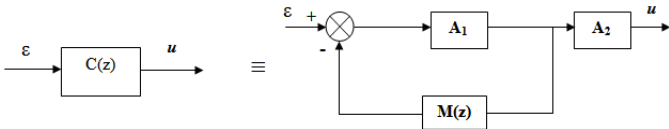


Fig. 3. Structure of the second corrector used

The gain A_2 is used to ensure the desired accuracy, it is described by the following expression:

$$A_2 = \frac{1 + A_1 M(1)}{A_1 M(1)} \quad (14)$$

V. INTERNAL MULTI-MODEL CONTROL APPROACH OF UNCERTAIN DISCRETE-TIME SYSTEMS

In order to reduce the complexity of dynamic process, the tendency has been to use linear time invariants models (LTI).

The multi-model represents complex system as an interpolation between in general linear or affine local models. Each local model is a dynamic system LTI (Linear Time Invariant) valid around an operating point.[1]

Uncertain systems can be represented by a library of linear models. These linear models are at the origin of the elaboration of a new control structure called internal multi-model control structure denoted IMMC. By combining the internal model control structure and the multi-model approach, the internal multi-model control approach is obtained.

The internal multi-model control structure for uncertain discrete-time systems was developed from the structure

described in the latest paragraph. It uses instead of a single internal model a library of models after the application of the Kharitonov's theorem for this class of discrete-time uncertain systems.[12]

This IMMC structure exploits the difference between the output of the process and the library of models outputs.

Let's consider the following diagram of the internal multi-model control structure: [6, 8, 10, 11]

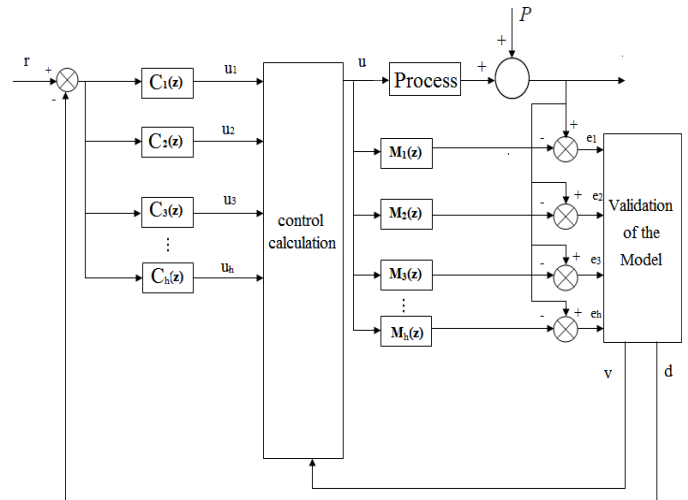


Fig. 4. Internal multi-model control structure of uncertain discrete-time systems

In this structure, the process is the uncertain discrete-time system to be controlled, $M_1(z), M_2(z), M_3(z), \dots, M_h(z)$ for $i=1, \dots, h$, represent the transfer functions of the internal models and $C_1(z), C_2(z), C_3(z), \dots, C_h(z)$ for $i=1, \dots, h$, are the transfer functions of the controllers.

In this control structure $M_i(z)$ for $i = 1, \dots, h$, are the linear models library inspired from the uncertain process, 'd' is the modeling error and 'v' is the validation index of the nearest model.

The proposed regulators for this control structure are the M_i^{-1} inverse models library that represents the inverse of the internal models M_i for $i=1, \dots, h$.

Several fusion methods were employed in the literature.

The choice of the control signal to be applied in this article is based firstly, on the switching method and secondly on the fusion method known as the residues techniques.

A. First IMMC structure based on the switching principle

This first method consists of determining the closest model to the process that allows to have the least modeling error. The control signal to be applied is therefore the signal that corresponds to the model that leads to the slightest error.

Using the first method to realize the approximated inverse [5,6], this diagram of the internal multi-model control structure based on the switching technique is obtained:

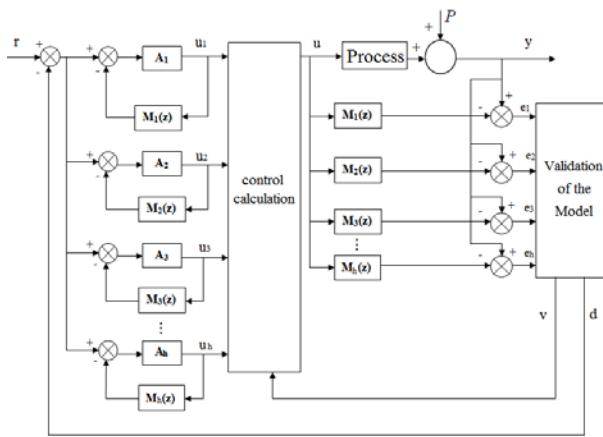


Fig. 5. Diagram of the first internal multi-model control structure of uncertain discrete-time systems based on switching method

In the basic diagram of this structure $A_1, A_2, A_3, \dots, A_h$ are the gains used for the inverse models.

Among the different Kharitonov models, the model that has the slightest error is chosen. The selected controller is then obtained from the model, whose output is nearest to the process, the validation block ensures the choice of this model.

The figures (6) and (7) represent the block diagram of the model validation method and the diagram describing the principle of control calculation.

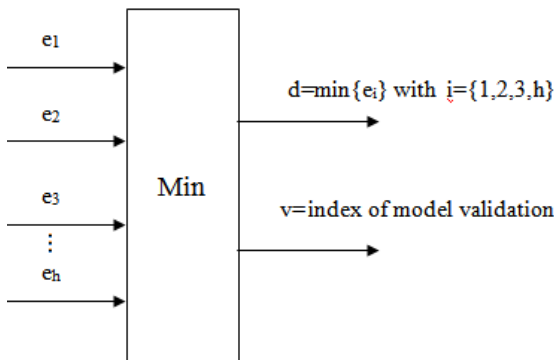


Fig. 6. Basic diagram of the model validation method

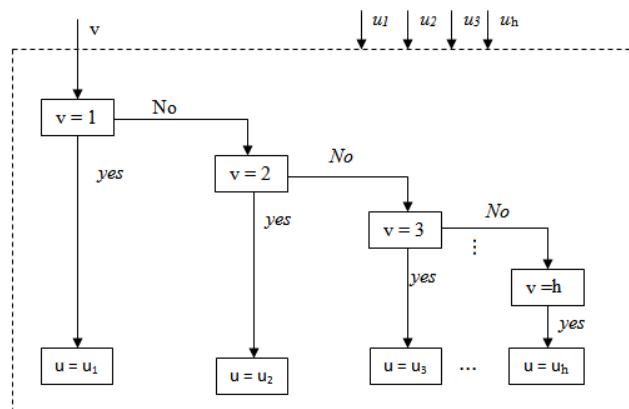


Fig. 7. Basic diagram for computing the control signal

For some classes of systems, the gain A_i with $i \in [1, \dots, h]$ that ensures stability of the loop that realize the controllers

$C_i(z)$ for $i=1, \dots, h$, may not be very high, which does not allow us to obtain the approximated inverses, therefore, a gain A_{2i} for $i=1, \dots, h$ is added to ensure a null static error. Thus, using the second structure of the proposed corrector described in figure (3), a second internal multi-model control structure which is based on the switching principle is obtained:

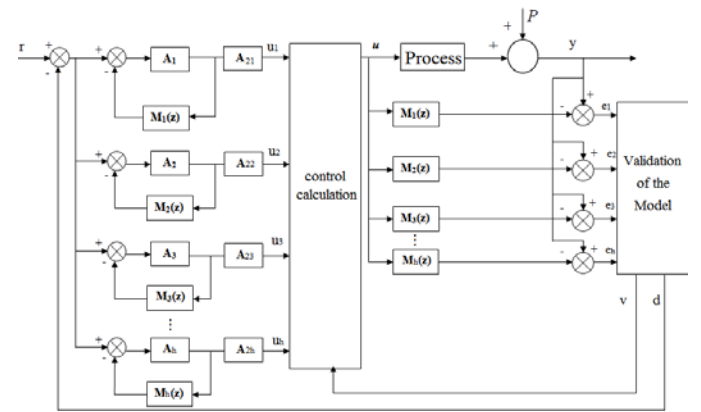


Fig. 8. Diagram of the second internal multi-model control structure of uncertain discrete-time systems based on switching method

B. Second IMMC structure based on the residues techniques of the uncertain discrete-time systems

The second internal multi-model control structure assumes the same internal models but the choice of the control signal to be applied is based on the principle of fusion known as residues techniques.

Using the inversion methods described previously for the realization of the inverse models, the first diagram of the internal multi-model control structure based on the residues techniques is obtained:

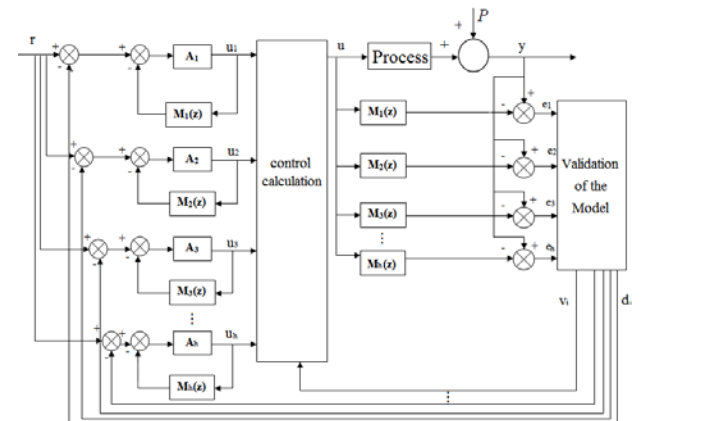


Fig. 9. First diagram of the internal multi-model control structure for discrete-time uncertain systems based on residues techniques

The calculation of the global command to apply to the system depends on partial control signals related to the models M_i and on the validities of these models.

Validity indexes are inversely proportional to the difference between the system output and the outputs of the internal models that can be defined by:

$$d_i(t) = y(t) - y_i(t) \quad \text{for } i=1, \dots, h \quad (15)$$

The validity can be expressed by the expression (16):

$$v_i = \frac{\left\| \frac{1}{d_i} \right\|}{\sum_{j=1}^h \left\| \frac{1}{d_j} \right\|} \quad (16)$$

Thus, the global control signal can be defined by the expression (17):

$$u(t) = \sum_{i=1}^h v_i(t)u_i(t) \quad (17)$$

Using the second structure of the proposed corrector, a second internal multi-model control structure which is based on the residues technique is obtained:

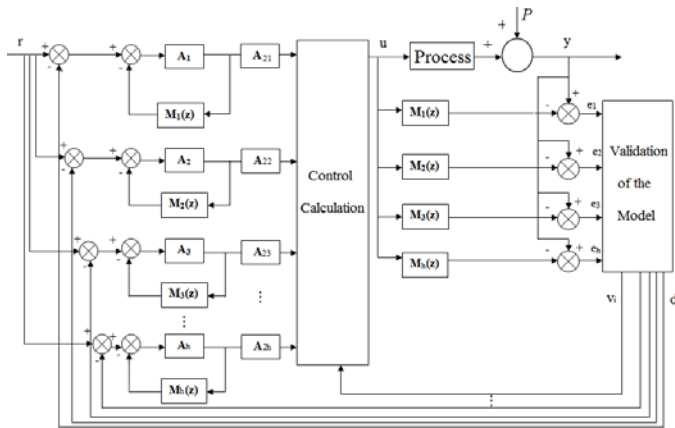


Fig. 10. Second diagram of the internal multi-model control structure for discrete-time uncertain systems based on residues techniques

VI. APPLICATION

In this paragraph, there are two parts. In the first part, the first extension of the Kharitonov theorem is used and in the second part, the second Kharitonov theorem extension is applied. In each part, there are two sections. In the first section the internal multi-model control structure based on the switching method is applied, however, in the second section, the second internal multi-model control structure based on residues techniques is considered.

A. Example 1: Using the first Kharitonov theorem

1) Application of the multi-model control structure based on the switching method:

Let's consider the following transfer function:

$$G(z) = \frac{z - 0.5}{z^2 + a_1 z + a_0} \quad (18)$$

where:

$$a_0 \in [0.1, 0.15]$$

$$a_1 \in [-0.9, -0.6]$$

Applying the first Kharitonov theorem previously defined, the following four Kharitonov models are obtained:

$$M_1(z) = \frac{z - 0.5}{z^2 - 0.9z + 0.1} \quad (19)$$

$$M_2(z) = \frac{z - 0.5}{z^2 - 0.9z + 0.15} \quad (20)$$

$$M_3(z) = \frac{z - 0.5}{z^2 - 0.6z + 0.1} \quad (21)$$

$$M_4(z) = \frac{z - 0.5}{z^2 - 0.6z + 0.15} \quad (22)$$

The system is stable for values of gains $A_1 < 1.33$, $A_2 < 1.36$, $A_3 < 1.13$ and $A_4 < 1.16$.

During the whole application:

The sampling period is considered equal to $T=0.1s$

The input signal takes the form of a unit step reference

The disturbance takes the form of step with amplitude equal to 0.5 applied at $k=20$

The output signal by applying the first internal multi-model control structure based on the switching method for $A_i=1$ for $i=1, \dots, 4$ is presented in the figure 11. The output signal oscillate at startup, this is due to the switching of the control signal. Also, the system presents a non-null error on the steady-state, this is because the gains A_i for $i=1, \dots, 4$ that ensure stability of the loop that realize the controller $C(z)$ are not very high, which does not allow us to obtain the approximated inverses, thus, a gains A_{2i} for $i=1, \dots, 4$ are added to ensure a null static error. It is therefore preferable to apply the second internal multi-model control structure based on the switching technique.

The output signal for $A_i=1$ for $i=1, \dots, 4$ and $A_{21}=1.4$, $A_{22}=1.5$, $A_{23}=2$, $A_{24}=2.11$ is displayed in the figure 12. The output of this uncertain process presents oscillation in transient regime and quickly reaches the reference in steady state.

By adding a disturbance at the output, the output signal for the same gains values A_i and A_{2i} for $i=1, \dots, 4$ is displayed in the figure 13. This control structure has rejected the external disturbance.

The figure 14 displays the validity signal, this signal lets us show the chosen model and therefore the selected controller applied.

For a sampling period $T=1s$ and for the same gains values A_i and A_{2i} , the figure 15 shows the output signal. The system takes longer time to stabilize that with a period ten times smaller.

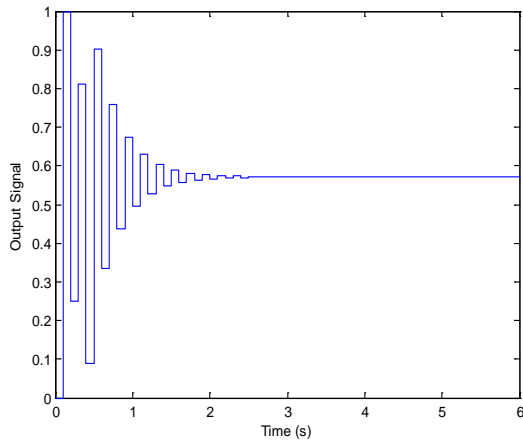


Fig. 11. Output signal for $A_i=1$ for $i=1,\dots,4$ and for a sampling period $T=0.1s$

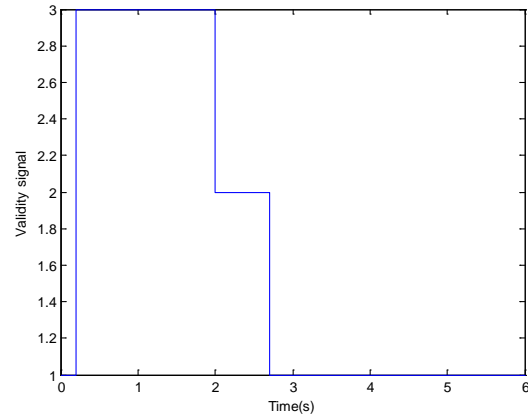


Fig. 14. Validity signal for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ and for a sampling period $T=0.1s$

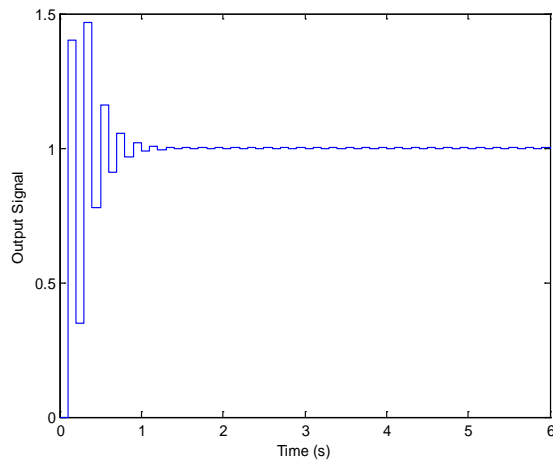


Fig. 12. Output signal for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ and for a sampling period $T=0.1s$

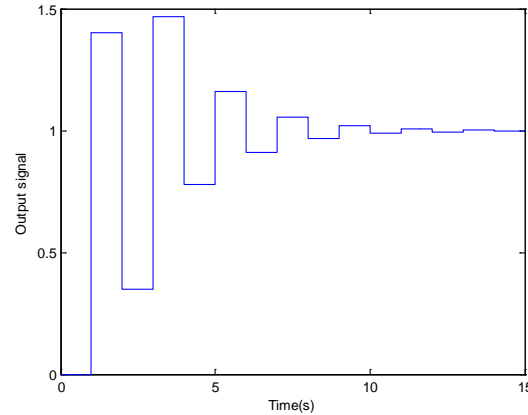


Fig. 15. Output signal for a sampling period $T=1s$, for $A_i=1$ for $i=1,\dots,4$ and for $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$

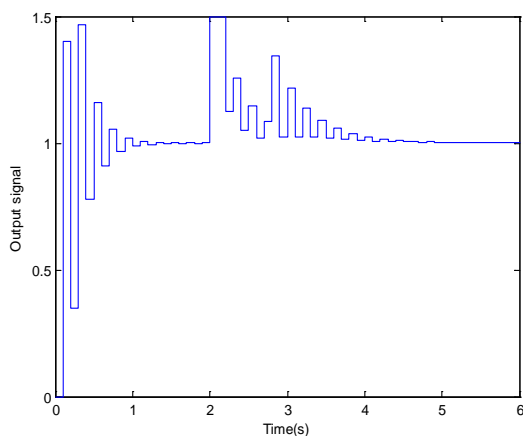


Fig. 13. Output signal by adding a disturbance for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ and for a sampling period $T=0.1s$

2) *Application of the multi-model control structure based on the residues techniques:*

The same example previously studied in the last paragraph and defined by the transfer function (18) and the four internal models (19, 20, 21, 22) is considered in this paragraph.

The output signal by applying the first internal multi-model control structure based on residues techniques for $A_i=1$ for $i=1,\dots,4$ is presented in the figure 16. As previously, this first control structure with internal multi-model based on residues techniques does not allow us to obtain perfect results, the system output presents static errors. It is preferable to apply the second internal multi-model control structure based on residues techniques.

By applying the second internal multi-model control structure based on residues techniques, the output signal for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ is displayed in the figure 17. It's noted that the transient state presents oscillation then the process output converges well to the reference.

By adding a disturbance at the output, the output signal for the same gains values is presented in the figure 18. This second internal multi-model control structure allows to reject the external disturbance. The figure 19 displayed the validity signals of the different models. These signals are used for the computing of the control signal.

For a period $T=1s$, the output signal for the same gains values is displayed in the figure 20. By increasing the value of the sampling period T , the system takes more time to stabilise.

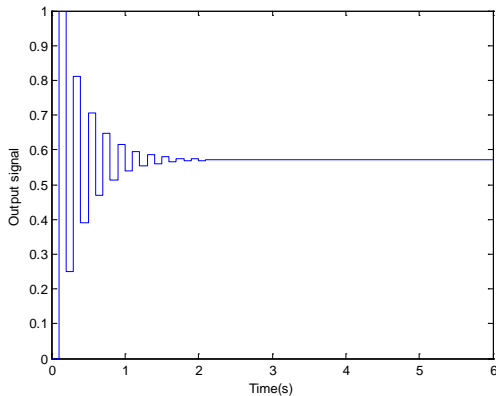


Fig. 16. Output signal for $A_i=1$ for $i=1,\dots,4$

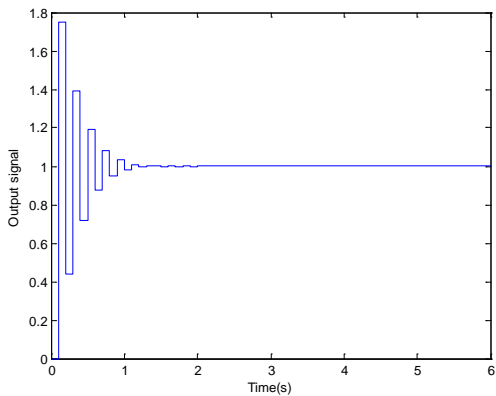


Fig. 17. Output signal for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ and for a sampling period $T=0.1s$

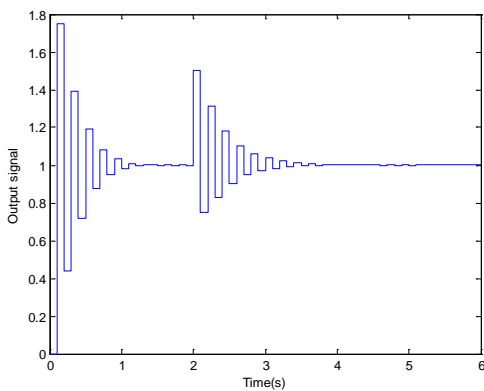


Fig. 18. Output signal by adding a disturbance for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ and for a sampling period $T=0.1s$

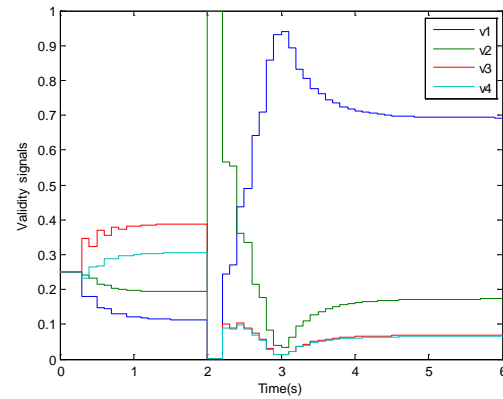


Fig. 19. Validities signals for $A_i=1$ for $i=1,\dots,4$ and $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$ and for a sampling period $T=0.1s$

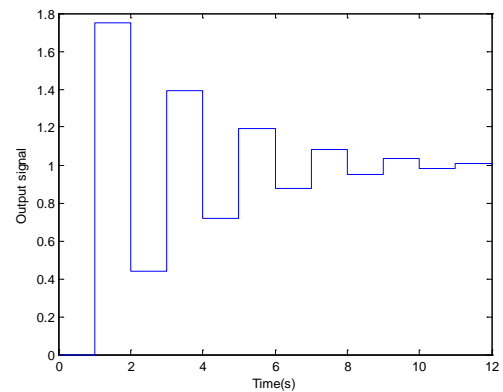


Fig. 20. Output signal for a sampling period $T=1s, A_i=1$ for $i=1,\dots,4$ and for $A_{21}=1.4, A_{22}=1.5, A_{23}=2, A_{24}=2.11$

Satisfactory results have been obtained with this two internal multi-model control structure. The amplitude of the oscillations in the transient state for the case of the second structure based on the fusion method is somewhat higher compared to the first structure based on the switching method. However, by using this second structure, the system has reached the reference and has rejected the disturbance more rapidly.

B. Example 2: Using the second Kharitonov theorem

1) Application of the multi-model control structure based on the switching method:

Let's consider the transfer function of the uncertain system defined by the expression (23):

$$G(z) = \frac{(z - 0.5)}{a_2 z^2 + a_1 z + a_0} \quad (23)$$

where:

$$a_0 \in [0.2, 0.8]$$

$$a_1 \in [-0.1, 0.25]$$

$$a_2 \in [2.5, 4]$$

It is clear that the system can be written in the poly-topical form, therefore by applying the second theorem, the following twelve edges polynomials can be determined by varying one parameter and setting the others.

$$E_1(\lambda, z) = 2.5z^2 - 0.1z + (0.8 - 0.6\lambda) \quad (24)$$

$$E_2(\lambda, z) = 2.5z^2 + 0.25z + (0.8 - 0.6\lambda) \quad (25)$$

$$E_3(\lambda, z) = 4z^2 - 0.1z + (0.8 - 0.6\lambda) \quad (26)$$

$$E_4(\lambda, z) = 4z^2 + 0.25z + (0.8 - 0.6\lambda) \quad (27)$$

$$E_5(\lambda, z) = 2.5z^2 + (0.25 - 0.35\lambda)z + 0.2 \quad (28)$$

$$E_6(\lambda, z) = 2.5z^2 + (0.25 - 0.35\lambda)z + 0.8 \quad (29)$$

$$E_7(\lambda, z) = 4z^2 + (0.25 - 0.35\lambda)z + 0.2 \quad (30)$$

$$E_8(\lambda, z) = 4z^2 + (0.25 - 0.35\lambda)z + 0.8 \quad (31)$$

$$E_9(\lambda, z) = (4 - 1.5\lambda)z^2 - 0.1z + 0.2 \quad (32)$$

$$E_{10}(\lambda, z) = (4 - 1.5\lambda)z^2 - 0.1z + 0.8 \quad (33)$$

$$E_{11}(\lambda, z) = (4 - 1.5\lambda)z^2 + 0.25z + 0.2 \quad (34)$$

$$E_{12}(\lambda, z) = (4 - 1.5\lambda)z^2 + 0.25z + 0.8 \quad (35)$$

According to the second theorem defined above, the stability of the uncertain system can be checked by studying the four higher models. This system is stable if and only if these four higher polynomials are stable.

By studying these four models, we find that the uncertain system is stable thereafter, the four higher Kharitonov models for $\lambda=0.5$ are proposed to be considered as the four internal models of the multi-model control structure.

$$M_1(z) = \frac{z - 0.5}{3.25z^2 - 0.1z + 0.2} \quad (36)$$

$$M_2(z) = \frac{z - 0.5}{3.25z^2 - 0.1z + 0.8} \quad (37)$$

$$M_3(z) = \frac{z - 0.5}{3.25z^2 + 0.25z + 0.2} \quad (38)$$

$$M_4(z) = \frac{z - 0.5}{3.25z^2 + 0.25z + 0.8} \quad (39)$$

The process is stable for gains $A_1 < 2.36$, $A_2 < 2.76$, $A_3 < 2.13$, $A_4 < 2.86$.

The output signal by applying the first internal multi-model control structure based on the switching method for $A_i=1$, for $i=1, \dots, 4$ is displayed in the figure 21. The static error is different to zero in the steady state that's because the gains A_i for $i=1, \dots, 4$ that ensure stability of the loop that realize the controller $C(z)$ are not very high. Thus the second internal multi-model control structure based on the switching method is proposed to be applied.

By applying the second internal multi-model control structure based on the switching method, the output signal for $A_i=1$ for $i=1, \dots, 4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ is shown in the figure 22.

The figure 23 represents the validity signal for the same gains values A_i and A_{2i} for $i=1, \dots, 4$. This figure shows the switching between the models and consequently between the control signals in the transient state. After, one second the validity signal has stabilized and the second controller has commanded the system.

By adding a disturbance at the output, the output signal is shown in the figure 24. The use of this control structure allowed us to reject the external disturbance.

The addition of external disturbance leads to more switching between models. This is shown in the figure 25 that displayed the validity signal for this case.

For a sampling period $T=1s$, the output signal for the same gains values A_i and A_{2i} for $i=1, \dots, 4$ is displayed in the figure 26. From this figure, it's clear that for the same value of the gains A_i and A_{2i} for $i=1, \dots, 4$, the system takes more times to stabilize.

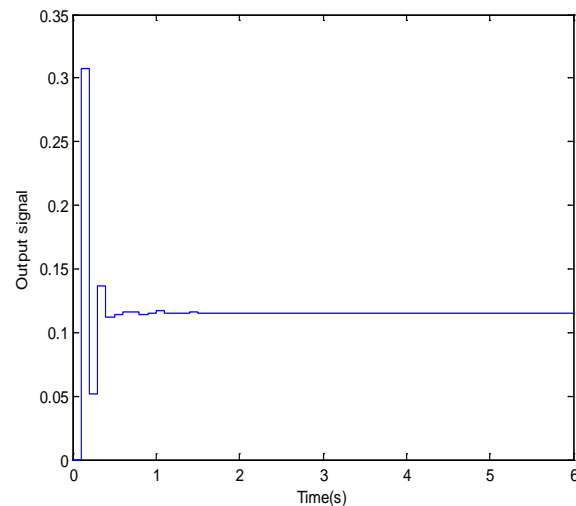
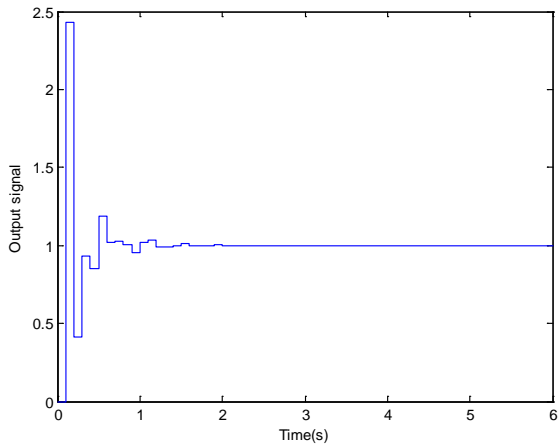


Fig. 21. Output signal for $A_i=1$, for $i=1, \dots, 4$ and $T=0.1s$



(b) : Output signal

Fig. 22. Output signal for $A_i=1$, for $i=1, \dots, 4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for a sampling period $T=0.1s$

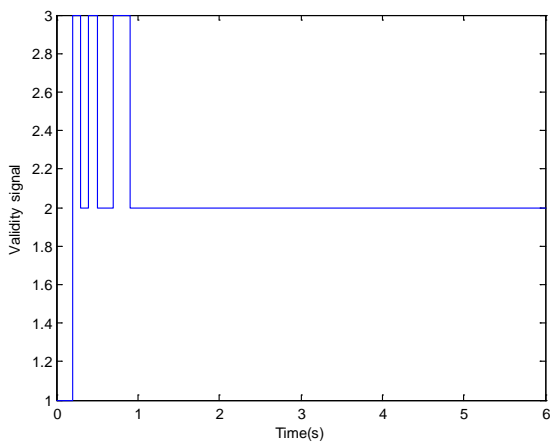


Fig. 23. Validity signal for $A_i=1$, for $i=1, \dots, 4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for a sampling period $T=0.1s$

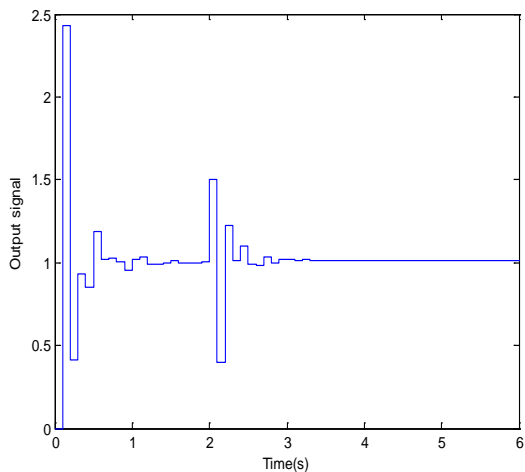


Fig. 24. Output signal by adding external disturbance for $A_i=1$, for $i=1, \dots, 4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for $T=0.1s$

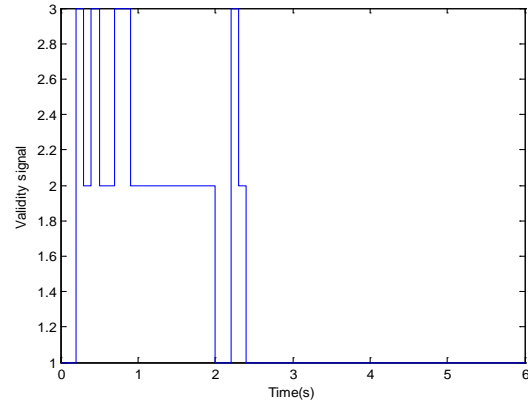


Fig. 25. Validity signal for the case of the disturbance presence and for $A_i=1$, for $i=1, \dots, 4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for $T=0.1s$

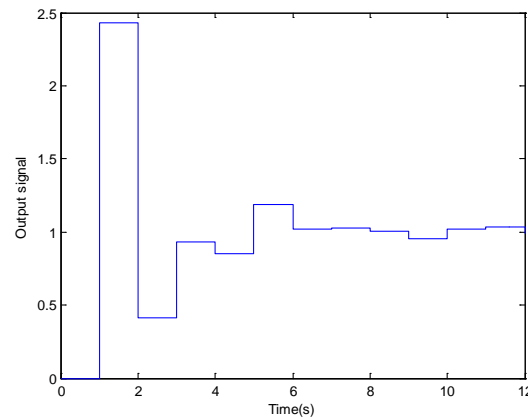


Fig. 26. Output signal for a sampling period $T=1s$ and for $A_i=1$, for $i=1, \dots, 4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$

2) *Application of the multi-model control structure based on the residues techniques:*

Let's consider the same example previously studied in the preceding paragraph where the transfer function of the system is defined by (23) and the four internal models are given by (36, 37, 38 and 39).

The output signal by applying the first internal multi-model control structure based on fusion method for $A_i=1$ for $i=1, \dots, 4$ is presented in the figure 27. The system output does not follow properly the reference and it present low oscillations in transient state. Thus, it's better to apply the second internal multi-model control structure based on the residues techniques.

By applying the second internal multi-model control structure based on the residues techniques, the output signal for $A_i=1$, $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$ and $A_{24}=9.6$ for $i=1, \dots, 4$ is shown in the figure 28. Satisfactory results have been obtained by the application of this second internal multi-model control structure based on residues techniques, it has enabled us to have null static errors.

By adding disturbance at the output, the figure 29 displays the output signal. This second internal multi-model control

structure based on residues techniques leads to reject the added disturbance.

The figure 30 displays the validity signals of the different models. This figure shows the validity of the different internal models used for the calculation of the control signal.

For a sampling period $T=1s$, the output signal for the same gains values A_i and A_{2i} for $i=1,\dots,4$ are presented in figures 31. From this figure, it's clear that the system takes longer time to reach the reference.

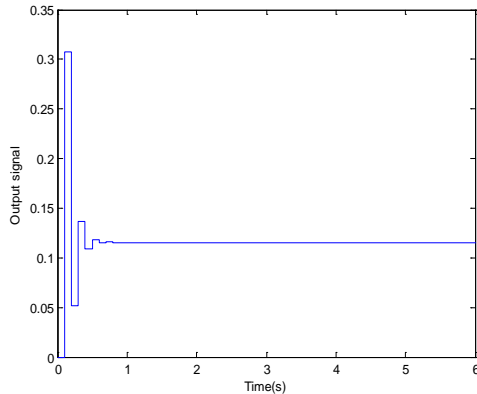


Fig. 27. Output signal for $A_i=1$, for $i=1,\dots,4$ and for a sampling period $T=0.1s$

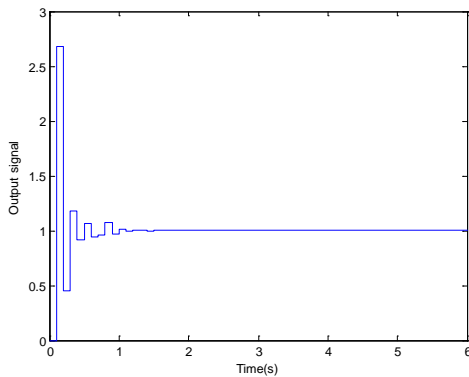


Fig. 28. Output signal for $A_i=1$, for $i=1,\dots,4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for a sampling period $T=0.1s$

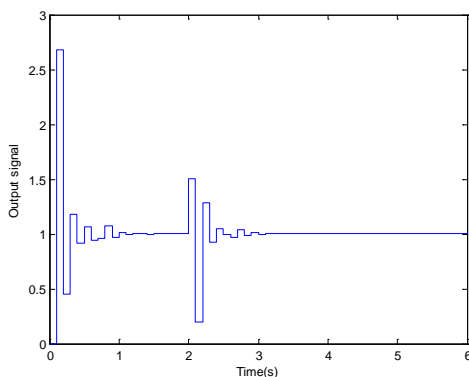


Fig. 29. Output signal by adding external disturbance for $A_i=1$, for $i=1,\dots,4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for $T=0.1s$

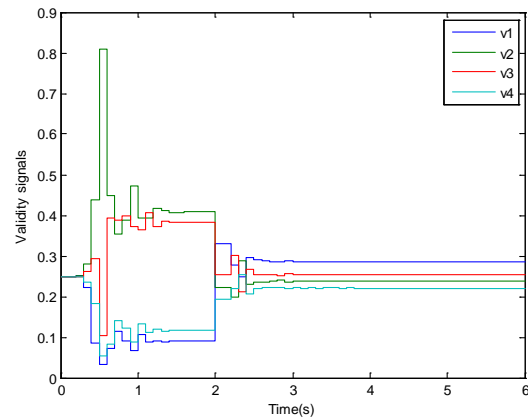


Fig. 30. Validity signals for $A_i=1$, for $i=1,\dots,4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for $T=0.1s$

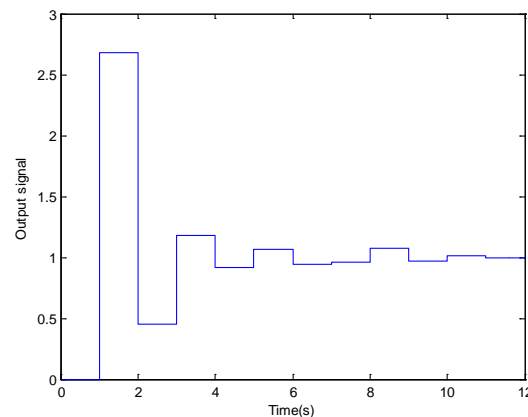


Fig. 31. Output signal for $A_i=1$, for $i=1,\dots,4$ and $A_{21}=7.9$, $A_{22}=8.9$, $A_{23}=8.4$, $A_{24}=9.6$ and for $T=1s$

Satisfactory results have been obtained with this two internal multi-model control structure.

The second Kharitonov method leads to best results compared to the first method. The oscillations in the transient state decreased and the output signals have reached the reference more rapidly relatively.

VII. CONCLUSION

In this article, the internal multi-model control structures IMMC have been applied for the case of the discrete-time uncertain systems. The first internal multi-model control structure is based on the approach of switching between the different models. On the other hand, the second internal multi-model control structure is based on the residues techniques.

The Kharitonov method has been used for the study of the uncertain discrete-time systems. Linear models were obtained by applying this method with their two extensions for the case of the discrete-time uncertain systems.

These two control approaches have been applied for this class of systems, the discrete-time uncertain systems which can be represented by a linear model library. These linear models

obtained by using the Kharitonov method are considered as internal models of these control structures.

The controller synthesis approach proposed is based on a specific inversion method. This approach has been modified to improve the accuracy of the controlled system.

These different control structures have been successfully applied for the case of uncertain discrete-time systems.

By applying the first extension of the Kharitonov method, the second internal multi-model control structure based on residues techniques led to good results in terms of speed compared to the internal multi-model control structure based on the switching method.

The second extension of the Kharitonov method leads to better results compared to the first method. The oscillations in the transient state have decreased clearly.

In this article, the robustness of the proposed internal multi-model control approaches overlooked the modeling errors and the external disturbances has been shown.

The choice of the sampling frequency strongly influences on the response of the system, a high value of the sampling frequency leads to more time taken to stabilize.

However, these satisfactory results invite us to improve the structure of our control approach whatsoever at the level of our controller or at the control loop level to improve the system performance in particular rejecting the effect of uncertainty whatever the synthesis approach.

REFERENCES

- [1] M. Chadli, "Stabilité et commande de systèmes décrits par des structures multimodèles," Thèse de doctorat, Institut National Polytechnique de Lorraine, december 2002.
- [2] M. Morari and E. Zafiriou, "Robust process control", Prentice Hall, Englewood Cliffs, 1989.
- [3] C.E. Garcia and M. Morari, "Internal Model Control 1- A unifying review and some results", *Ind. Eng. Chem. Process Des. Dev.*, vol. 21, pp. 403-411, 1982.
- [4] L. Saidi, "Commande a modèle interne: Inversion et équivalence structurelle", Thèse de doctorat, INSA de Lyon, France, 1990.
- [5] M. Benrejeb, M. Naceur and D. Soudani., "On an internal model controller based on the use of a specific inverse model", *International Conference on Machine Intelligence, ACIDCA'2005, Tozeur*, pp. 623-626, 2005.
- [6] M. Naceur, "Sur la commande par modèle interne des systèmes dynamiques continus et échantillonnés", Thèse de doctorat, Ecole Nationale d'Ingénieurs de Tunis, february 2008.
- [7] M. Naceur, D. Soudani, M. Benrejeb. "Sur la commande par modèle interne des systèmes échantillonnés basée sur une inversion spécifique", *JTEA' 2006, Hammamet 2006*.
- [8] C.Othman, I. Ben Cheikh, D. Soudani, "Application of the internal model control method for the stability study of uncertain sampled systems", *CISTEM 2014 IEEE Tunis*.
- [9] M. Naceur, I. Ben Cheikh, D. Soudani, M. Benrejeb, "On the Internal Model Control of Uncertain Systems", 978-1-4244-7534-6/10/\$26.00 ©IEEE, Juin 2010.
- [10] D. Soudani, M. Naceur, K. Ben Saad. and M. Benrejeb, "On an internal multimodel control for nonlinear systems – A comparative study", *Int. J. Modelling, Identification and Control*, Vol. 5, No. 4, pp. 320-326, 2008.
- [11] M. Naceur, D. Soudani, M. Benrejeb, P. Borne "On internal multimodel control for nonlinear system," *IMACS, CESA 2006, Beijing*, pp 306-310,2006.
- [12] Z. Sun, J. Chen, X. Zhu, "Multi-model internal model control applied in temperature reduction system", *Proceeding of the 11th World Congress on Intelligent Control and Automation Shenyang, China, June 29 - July 4, 2014*.
- [13] N.K.Sinha and Whou Qi-Jie, " Discrete_time approximation of multivariable continuous_time systems", *IEE PROC.*, Vol 130, Pt. D, No. 3. May 1983.
- [14] K. Yeung, S. Nang, "A simple proof of Kharitonov's Theorem", *IEEE Trans. On Automatic Control*, Vol. AC-32, No. 9, Sept 1987.
- [15] P. P. Vaidyanathan, "A New Breakthrough In Linear-System Theory: Kharitonov's Result", *California Institute of Technology, Pasadena, CA 91125*, 1988.
- [16] S.P. Bhattacharyya, H. Chapellat, L.H. Keel, "Robust Control - The Parametric Approach", *Prentice Hall PTR*, 1995.

High Performance Computing Over Parallel Mobile Systems

Doha Ehab Attia

Department of Computer Science
Faculty of Computers and
Information,
Cairo University
Cairo, Egypt

Abeer Mohamed ElKorany

Department of Computer Science
Faculty of Computers and
Information,
Cairo University
Cairo, Egypt

Ahmed Shawky Moussa

Department of Computer Science
Faculty of Computers and
Information,
Cairo University
Cairo, Egypt

Abstract—There are currently more mobile devices than people on the planet. This number is likely to multiply many folds with the Internet of Things revolution in the next few years. This may treasure an unprecedented computational power especially with the wide spread of multicore processors on mobile phones. This paper investigates and proposes a new methodology for mobile cluster computing, where multiple mobile devices including their multicore processors can be combined to perform possibly massively parallel applications. The paper presents in details the steps for building and testing the mobile cluster using the proposed methodology and proving the successful implementation.

Keywords—Parallel computing; High-performance computing; Mobile computing; Cluster computing; Android OS

I. INTRODUCTION

The International Telecommunication Union (ITU) estimates that by the end of 2016 there will be almost 7.3 billion mobile subscriptions [1], which is equivalent to 95 percent of the world population. The number of mobile broadband subscriptions is also rising significantly due to the ever increasing popularity of mobile devices (e.g., smartphones, tablets, notebooks, iPads, etc.) with an estimate that the number of subscribers will reach 2.3 billions globally by the end of 2016.

Another important prediction is that the number of mobile devices per capita, already increasing significantly, is expected to reach 1.5 by 2020 [2]. These numbers and predictions understandably made the mobile communications one of the fastest growing fields of technology. Many people are starting to depend on their smartphones as a primary and, sometimes, only computing device. The constant development in the capabilities of mobile devices enabled those devices to be in some cases an adequate replacement for traditional computers where it can compute with comparable performance.

With both the rapidly growing capabilities and wide spread of mobile devices, the authors expect further growth of using mobile devices in general and smartphones in particular for more performance-intensive computing tasks that were previously handled by traditional computers. To deal with this tidal wave of high-performance applications the quality of experience on devices based on single core CPUs rapidly degrades when users run several applications concurrently, or run performance-intensive applications. Therefore, smartphone

industry transitioned to multicore CPUs to cope with the performance challenges. The availability of multicore mobile devices makes them a prime candidate for parallel computing applications. Similar to what happened with traditional computers, the authors expect another extension of mobile parallel computing to include multi-node High Performance Computing (HPC) clusters.

Another critical factor to consider is the underutilization of mobile devices compared with the energy consumed. Most mobile computing and communication devices recharge their batteries because of the power leakage and the power consumed just to keep the devices on, without actual use. This observation applies more to the corporate mobile systems. This was another motivation for both the research community and industry to develop solutions to use the wasted energy in distributed mobile systems for larger combined computational power. For example, HTC and Samsung recently launched mobile applications (HTC Power to Give [3] and Samsung Power Sleep [4]) to enable smartphone owners to contribute unused computational power in mobile grids.

Using smartphones as a portable high performance computing infrastructure can be beneficial in many ways:

- Situations where access to HPC machines is non-existent and the only means for accomplishing resource-intensive computations are mobile devices. Those situations turned out to be widely spread in many military as well as civilian applications.
- Performance and energy-efficiency gains. By splitting the processing among several devices. This will, not only speed up the processing, but will also distribute the battery drain across all devices. Several solutions have been proposed to enhance the CPU performance [5], [6] and to manage the disk and screen in an intelligent manner to reduce power consumption [7], [8]. However, these solutions require changes in the structure of mobile devices, or they require a new hardware that results in cost increase and may not be feasible for all mobile devices. But distributing the workload over a large number of processors may reduce the power consumption of each device separately.
- Using the mobile cluster for educational purposes by creating an on spot parallel computing cluster using the

available smartphones in the classroom, especially when obtaining HPC machinery for education purposes is not feasible or possible.

- Another usage for the smartphone-based computing infrastructure is that an enterprise could benefit from significant energy savings and better operation by offloading tasks to clusters of smartphones and converging the clusters to the larger cloud.

Based on the above introduction and motivation, the research team investigated a solution based on developing a portable parallel computing cluster made of available smartphones. A key point is that the cluster uses standard parallel computing languages and techniques to facilitate the migration between the developed solution and standard HPC systems and clouds. Therefore, the solution is entirely based on standard C programming with Message Passing Interface (MPI).

The rest of the paper is organized as follows: Section 2 reviews the previous approaches identifying the shortcomings of each, which lead to the current research. Section 3, describes the details of building the proposed solution. Testing and evaluation of the developed mobile cluster are introduced in Section 4. Finally, Section 5 provides the conclusions and potential future applications.

II. PREVIOUS APPROACHES

There have been several parallel processing attempts on mobile devices, specifically on smartphones. In 2008 Daniel Doolan, Sabin Tabirca, and Laurence Yang implemented an MPI library on a Bluetooth network in a Java based environment [9]. The authors of the current paper attempted reproducing the reported Bluetooth cluster as described in the paper with no success. Despite the failure to replicate, we have three reservation on the published work today: (1) Doolan et al. reportedly redefined the MPI library to fit with the Bluetooth network. Hence, the MPI implementation is no longer standard to fit, and integrate, with other cross-compiled MPI systems, (2) The Java-based environment is on the decline [10] which makes the resulting cluster based on outdated technology, and (3) Typical HPC software systems are not developed with Java due to the virtual machine nature preventing full performance control.

Another attempt was carried out by Hinojos et al. [11]. The project is named BlueHoc, a system that enables distributed computation on Android smartphones via Bluetooth networks. It was designed for military fields where there might be no access to a fixed computing infrastructure. MPI was not used for the parallel programming on the BlueHoc system. An alternate approach was developed using the radio frequency communication (RFCOMM) protocol. This is clearly a nonstandard HPC programming library which, again, impedes the integration of parallel mobile systems into clouds and fixed HPC infrastructures.

Felix Büsching, Sebastian Schildt, and Lars Wolf developed another approach titled DroidCluster, based on installing Debian ARM on the SD card of a smartphone rather than direct Android programming. Their published paper lists a

number of potential applications of portable clusters and uses the LINPACK benchmark to test the cluster performance [12].

In 2013 Chien-Chung Wu and Juyn-Jie Huang published a study implementing Android parallel programming based on the dual-core Cortex A9 [13]. This study focused on single node multicore programming with OpenMP rather than cluster computing with MPI.

Finally, Iulian Vîrtejanu and Costică Nitu developed a mobile cluster based on cross compiling the MPICH2 library for Android phones. The paper does not mention the communication protocol used to form the mobile cluster (e.g. ssh). Also, the detailed steps for cluster building were not mentioned [14].

III. CLUSTER BUILDING

A. Software

In July 2016, Android's market share was 66.01%, exceeding all other smartphone's operating systems [10]. The availability of Android phones was not the only reason for choosing it for building our mobile cluster. The fact that Android is open source, Linux-based operating system allowed us to assume that all Linux applications can be easily ported to Android mobile devices. But that was proven incorrect. While Android is Linux-based it does not fully utilize the standard Linux kernel. Because Linux is open-source, the Android developers modified the Linux kernel to fit their needs [11][16]. There are some major differences between Android and Linux. For example, some of the standard GNU libraries are not included in Android. It does not rely on the GNU libc, for instance; it uses bionic instead [11].

Bionic is a C library that is not only much smaller in size than the GNU libc but also has less memory requirements. This means one cannot simply run Linux applications and libraries on Android [11]. For example, one of the libraries that cannot be directly ported to Android is the Message Passing Interface (MPI) library. MPI is an open source, portable library used in high-performance computing for message passing between parallel computing nodes. In order to build an MPI-based mobile cluster, we had to cross compile an MPI version with a library other than Bionic because Bionic does not support the full C/C++ standard [11]. In the current research project, the authors have chosen to cross compile and statically link MPICH which is a freely available implementation of the MPI standard to the Uclibc library.

Uclibc is a C library that is smaller in size than glibc; it was intended for Linux-based embedded systems [17]. A Uclibc based GNU toolchain was generated for the ARM architecture, the processor's architecture of the devices used in the current research. The authors used Buildroot, an open source tool to generate embedded Linux systems [18], to generate the required toolchain for the ARM architecture. In addition, The research team used the Android Debug Bridge (ADB) tool to access the used smartphone's shell. Finally, The researchers had to unlock the bootloader of the Android devices used in the experiment to enable access to the devices with root-level permission as a super user.

B. Hardware

Table I shows the hardware specifications for each of the two devices used in building the mobile cluster.

TABLE I. HARDWARE SPECIFICATIONS

specifications	Devices used	
	Device 1	Device 2
Number of cores	4 (1.2GHz)	2(1.5GHz)
Operating system	Android™ 4.4.2	Android™ 4.1.1
RAM	1GB	1GB
ROM	8 GB	8 GB

C. The steps for building a cluster of Android devices

1) Generate the toolchain. After downloading Buildroot software, go to the Buildroot directory and use `make menuconfig` command to open Buildroot configuration tool. After choosing the suitable configuration options, start the building process using `make` command. The Output will be found in `'/PathToBuildroot/output/host/usr/bin'` directory.

2) Compile MPICH for the ARM architecture. After downloading MPICH, create a new folder with the name "build". Then use the following command to statically build MPICH (from the MPICH directory):

```
./configure --prefix=/mpich/build/  
CC=/PathToBuildroot/output/host/usr/bin/arm-buildroot-  
linux-uclibcgnueabi-gcc  
CFLAGS="-I/PathToBuildroot/output/host/usr/include/"  
LDFLAGS="-L/PathToBuildroot/output/host/usr/lib/ "--  
host=arm-linux --disable-shared --enable-static --disable-  
fortran."
```

3) Use `/mpich/build/bin/mpicc` command to compile your C code. e.g.: `/mpich/build/bin/mpicc -o myCode mycode.c` Then copy the generated "myCode" file to each device `/data` directory.

4) Change `/system/etc/hosts` file on each device to contain all IP addresses. Any file explorer application that requires root permission can be used. The hosts file should look like this:

```
10.0.0.1 localhost  
10.1.1.1 hostname1  
10.2.2.2 hostname2
```

NOTE: The default hostname "localhost" was changed on all devices and each device was give a distinct hostname.

5) Copy the files in the generated `/mpich/build/bin` directory to each device in `/system/xbin` directory. The files can be copied from your computer to the mobile phones using ADB. Connect your phone to the computer, open the terminal and run:

```
adb push -p /mpich/build/bin /system/xbin
```

In case that 'Read-only file system' message appears you can mount your system to read/write by running the following commands from your terminal: type `adb shell`, the phone's terminal will open. Then type `su` command to have super user

permissions. By using the following command `mount -o rw, remount /system` you should be able to copy the files. In case that 'permission denied' message appears: Run `chmod 777 /system` on the phone's terminal. Now MPICH should be running on the device to test it you can run the following command from the phone terminal: `mpixec -n 1 /data/myCode`.

6) Setup ssh for communication between devices. First you have to download ssh server for your Android phones. In our case we used Android ports "Unix command line packages built for Android". Use the mobile device's shell to run this command:

`opkg install dropbear openssh`. Then generate public and private key pairs in `/.ssh/id_rsa` (in `system(root)`) using this command `/data/local/bin/ssh-keygen -t rsa`. After copying the public key to the other phone `/.ssh/authorized_keys` (if the file does not exist, create it). Start dropbear on all devices by running: `/data/local/bin/dropbear`. Then Check that ssh is working by trying this command `/data/local/bin/ssh root@ipaddress`.

NOTE: The hotspot was enabled on one of the devices and the other devices connected to it without being connected to the Internet just to create a local network.

IV. TESTING AND EVALUATION

A matrix multiplication program was developed to test the mobile cluster. The cluster was tested by executing a sequential version of the program on a single device. Then multiple parallel versions with a different number of processors were executed on the mobile cluster. Finally, we compared the execution time of all the runs.

The following is the pseudocode of the matrix multiplication program that was executed on the developed mobile cluster

Set the matrix size

If rank equal zero

Initialize the two matrices

Distribute the matrices based on the matrix size and the number of processes

If rank greater than zero

Multiply the assigned part of the matrix based on the process number

Send the process result

If rank equal zero

Receive/collect results from other processes

Table II shows the results of running the matrix multiplication code using square matrices of increasing sizes.

The table contains the time taken in seconds to execute a matrix multiplication C code using one processor on device1 then using four processors on device one then using six processors on device one and device two.

TABLE II. THE RESULTS OF RUNNING MATRIX MULTIPLICATION CODE ON MOBILE CLUSTER

Matrix Dimensions	Number of processes		
	1	4	6
200*200	1.136	0.552	2.082
400*400	7.452	3.296	5.802
600*600	27.03	9.686	11.962
800*800	68.194	25.652	34.436
1000*1000	136.152	49.192	56.168
2000*2000	1120.506	479.924	385.81
3000*3000	3994.886667	2231.788	1706.038
4000*4000	13107.08667	6638.176667	4844.5725
5000*5000	23148.62333	13920.55667	9351.23

Fig 1. illustrates the results. It shows that when the matrix size was small, the run time for the program was almost the same on single core and multiple cores. As the matrix size increase, the time taken increased dramatically on the single core, a clear indication of the need for extra computational power to cope with the growing problem size. As the number of processors increases the execution time decreases.

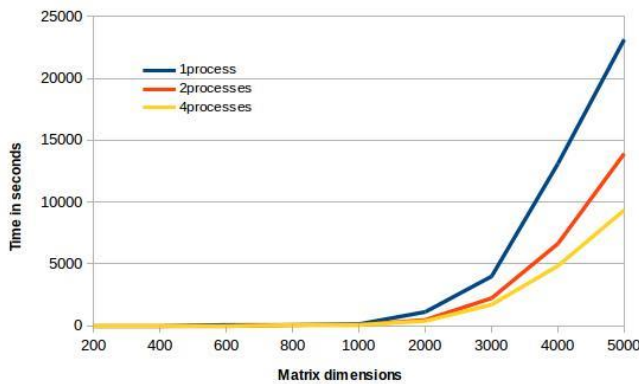


Fig. 1. The result of running matrix multiplication on mobile cluster

V. CONCLUSIONS AND POTENTIAL APPLICATIONS

A. Conclusions

The results show that using a cluster of smartphones as a high performance computing infrastructure is possible and can be in some cases an adequate alternative to the traditional cluster. It can also open the way to a plethora of new applications where the computational power of the smartphones are used to their fullest potential.

We also found some limitations in the mobile cluster, one of which is that the nodes of the cluster are constantly moving which means that some nodes may leave the cluster before finishing the job. Another limitation is that nodes of the cluster have to be predefined, the IP address of each node has to be mentioned in the hosts file before executing any job on the cluster.

B. Potential applications

In addition to the previously identified applications in the literature [12], the current research authors suggest and plant to work on the following potential applications:

- Personal identification by parallel biometrics computing using mobile devices. To overcome the limitations of the existing password-based authentication services on the Internet, we integrate personal features (ex: fingerprints, palmprints, hand geometry and face) into a hierarchical structure for fast and reliable personal identification and verification. To increase the speed and flexibility of the process, mobile devices can be used as a tool for parallel implementation in a distributed environment. The benefit of using a corporate or cloud mobile cluster for this application is two-fold. On one hand, it is increased security with distributed verification and manipulation of biomarkers. On the other hand, the use of parallel mobile clusters saves energy, computational resources, and consequently operation cost due to the saving of no longer needed dedicated infrastructure.
- Distributed key agreement. Securing the access to certain files/ places within the same institution by distributing the access key over different mobile nodes. The dynamic re-allocation of the access keys serves like the multiple keys safes to increase the security and enables the tracking of intrusion. This highly optimizes cloud security.
- Field data collection and processing. Capture, process and share data in places (such as military fields, agriculture fields, desert and geological excavations and navigations, etc.) where access to HPC (high performance computing) machines is non-existent. The utilization of the processing power of available devices, dynamic, mobile, ad hoc clusters can be used for the initial data collection and preprocessing. This may reduce or eliminate the need for data and program transmission.
- Video and image understanding applications.

ACKNOWLEDGMENT

The research team acknowledges the use and contribution of the HiPer-FC cluster and HPC group at the faculty of computers and information, Cairo Uuniversity.

REFERENCES

- [1] Statistics. (2016). ITU. Retrieved 30 August 2016, from <http://www.itu.int/en/ITU/Statistics/Pages/stat/default.aspx>
- [2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper. (2016). Cisco. Retrieved 31 August 2016, from <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html>
- [3] HTC Power To Give | HTC United States. (2016). HTC. Retrieved 30 August 2016, from <http://www.htc.com/us/go/power-to-give/>
- [4] Power Sleep - DO GOOD WHILE YOU SLEEP. (2016). Samsung.com. Retrieved 30 August 2016, from <http://www.samsung.com/at/microsite/powersleep/en/>

- [5] Kakerow R. "Low power design methodologies for mobile communication", In Proceedings of IEEE International Conference on Computer Design: VLSI in Computers and Processors, 2003; 8.
- [6] Paulson LD. "Low-power chips for high-powered handhelds". IEEE Computer Society Magazine 2003; 36(1): 21.
- [7] Davis JW. "Power benchmark strategy for systems employing power management", In Proceedings of the IEEE International Symposium on Electronics and the Environment, 2002; 117.
- [8] Mayo RN, Ranganathan P. "Energy consumption in mobile devices: why future systems need requirements aware energy scale-down", In Proceedings of the Workshop on Power-Aware Computing Systems, 2003.
- [9] D. Doolan, S. Tabirca and L. Yang, "MMPI a message passing interface for the mobile environment", Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia - MoMM '08, 2008.
- [10] "Operating system market share", Netmarketshare.com, 2016. [Online]. Available: <https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=9&qpcustomb=1>.
- [11] G. Hinojos, C. Tade, S. Park, D. Shires, and D. Bruno, "Bluehoc: Bluetooth ad-hoc network android distributed computing", Int. Conf. on Parallel and Distrib. Process. Tech. and Appl.(PDPTA), pp.468-473, 2013.
- [12] F. Busching, S. Schildt, and L. Wolf, "DroidCluster: Towards Smartphone Cluster Computing The Streets are Paved with Potential Computer Clusters", In Distributed Computing Systems Workshops (ICDCSW), pp.114-117, 2012.
- [13] C. Wu and J. Huang, "The Study of Android Parallel Programming Based on the Dual-Core Cortex-A9", 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2013.
- [14] I. VÎRTEJANU and C. NIȚU, "Programming distributed applications for mobile platforms using MPI", U.P.B. Sci. Bull., vol. 75, no. 4, 2013.
- [15] "Bionic (software)", Wikipedia, 2016. [Online]. Available: [https://en.wikipedia.org/wiki/Bionic_\(software\)](https://en.wikipedia.org/wiki/Bionic_(software)).
- [16] W. Project!, "Android Open Source Project", Source.android.com, 2016. [Online]. Available: <https://source.android.com/>.
- [17] "uClibc", Uclibc.org, 2016. [Online]. Available: <https://www.uclibc.org/>.
- [18] "Buildroot - Making Embedded Linux Easy", Buildroot.org, 2016. [Online]. Available: <https://buildroot.org/>.

ROHDIP: Resource Oriented Heterogeneous Data Integration Platform

Wael Shehab

Computers & Control Dept.
Faculty of Engineering Tanta Univ.
Tanta, Egypt

Sherin M. ElGokhy

Computers & Control Dept.
Faculty of Engineering Tanta Univ.
Tanta, Egypt

ElSayed Sallam

Computers & Control Dept.
Faculty of Engineering, Tanta Univ.
Tanta, Egypt

Abstract—During the last few years, the revolution of social networks such as Facebook, Twitter, and Instagram led to a daily increasing of data that are heterogeneous in their sources, data models, and platforms. Heterogeneous data sources have many forms such as the www, deep web, relational databases systems, No-SQL database systems, hierarchal data systems, semi-structured files, in which data are usually allocated on different machines (distributed) and have different data models (heterogeneous).

Large-scale data integration efforts demonstrate that their most valuable contribution is implementing a data integration platform that provides a uniform access to the heterogeneous data sources, as well as the different versions of data reported by the same data source over time. Furthermore, the platform must be able to integrate data from a broad range of data authoring devices and database management systems. It also should be accessible by almost types of data querying devices to ensure globally querying the integration platform from any place on earth anytime and receiving the query result in any data format.

In this paper, we create a resource oriented heterogeneous data integration platform (ROHDIP) that facilitates the data integration process and implements the objectives discussed above. We use the resource oriented architecture ROA to support the uniform access by most types of data querying devices from anywhere and to improve the query response time.

Keywords—Data Integration; Data heterogeneity; SOA; ROA; Restful; ROHDIP

I. INTRODUCTION

The enduring utilization of information technology raises data sharing as a challenging problem for many enterprises. Most enterprise information management systems adopted the opinion of establishing isolated database management systems in departments that may be geographically dispersed or have different business type to improve production, management, and efficiency. However, these systems are developed using different software companies at various times, on different platforms as well, which inevitably will lead to the coexistence of heterogeneous databases [1]. Heterogeneous data are often collected from an unknown or an unlimited number of sources in different formats. Two types of data heterogeneity; namely, structural heterogeneity and semantic heterogeneity are counted. In structural heterogeneity, the information systems store data in several structures. While semantic heterogeneity concerns with both the data item content and its intended meaning. The rapidly increasing number of structured, semi-structured data sources results in a crucial need for uniform and

flexible query interfaces to access data that are distributed on heterogeneous and autonomous sources [2] [3] [4] [5].

Data integration system allows users to specify what information is needed without providing detailed instructions of the methodology followed to obtain that information or even specifying its location. In order to have the capacity to do so, data integration system must be able to do the following process: communication and interaction with data sources, unifying different queries in requester specifying vocabulary (ontology) across multiple autonomous, distributed and heterogeneous data sources, mapping techniques between requester ontology and the data source ontology, extracting information from the query with respect to the target data sources, and finally translating the query results to the requester vocabulary [4].

Recently, many approaches to data integration were developed including manual integration, application-based integration, middleware data integration, physical data integration and virtual integration. In manual integration (common user interface), users manage all relevant information, accessing all the source systems and there is no unified view exists for the data. Application-based integration requires the particular applications to achieve all the integration efforts; therefore, this approach is manageable only in case of a limited number of applications. The approach that transfers the integration logic from particular applications to a new middleware layer is called middleware data integration. However, this approach does not ensure achieving the practical requirements. Physical data integration usually creates a new system that copies the data from different source systems to be stored and managed independently of the original system. The most well-known implementation of this approach is called data warehouse (DW) [6][7] which combines data from different sources (such as mainframes, databases, flat files). However, the need for a separate system to handle the vast volumes of data constitutes demerit of this approach. The final approach is virtual integration which leaves data in the source systems and defines a set of views to provide the customer with a unified view of the whole enterprise. For example, when we need to query specific data, it will be retrieved only from its data source [8] [9].

Virtual integration approach has several advantages that make it one of the most successful data integration approaches. It succeeds to propagate the data update from the source system to the integration system with almost zero latency.

Also, virtual integration has no need to copy any data from the data source to the integration system. Also, it does not need to unify the distributed data sources. Based on that, this paper is concerned with proposing an algorithm that adopts the virtual integration approach.

The rest of this paper is organized as follows: the related work is discussed in section II. Section III illustrates the problem statement and presents the proposed framework. The experimental results are exhibited and analyzed in section IV. Finally, Section V concludes the paper ideas and suggests future research ideas.

II. RELATED WORK

Several data integration platforms have been developed to provide a uniform query interface that has the capability to query the heterogeneous data sources [10] [11].

Specifically, an extensive wide variety of methods is proposed to achieve virtual integration approach, each of which was targeting the same goal but with its autonomous way. These methods are categorized into two approaches: Global-As-View (GAV) and Local-As-View (LAV). GAV produces a top abstract level that constitutes a single mediated schema described as views (mappings) of all local data sources [12] [13], while LAV describes the local data sources as views over a global schema [2] [14]. After that, many approaches have been derived from LAV and GAV [11] [15] making the best use of these two modern technologies, such as SOA "Service Oriented Architecture" that is used in implementing dynamic and flexible integration systems; namely, Service Oriented Data Integration systems. Then, various data integration frameworks based on SOA have been developed in the last few years such as SODIA architecture [11]. SODIA merges the data at various, distributed, heterogeneous and autonomous data sources into a single dynamic view. Service providers publish their data sources as data access services, which may be detected instantly at the time they are needed and released after use. Hence, variations of organization structures, backend data sources, data structures, or semantics could be managed and potentially the maintenance cost is reduced.

In 2009, an architecture for "Internet of Things" has been proposed to connect millions of different devices together based on service-oriented approach [16]. The architecture hides the heterogeneity of hardware, software, data formats and communication protocols. The specifications of the architecture support open and standardized communication via web services at all layers. Services abstract all functionality offered by networked devices. A runtime for the execution of the composed services was provided.

After that, Sanz et al. proposed an approach to integrate several technologies, such as the JSF, Spring and Hibernate frameworks in a multilayer architecture. SOA architecture provides services to allow collaborative work, using the independent development of components in different layers. The approach relies on developing a global software system where the presentation layers for different end devices are

separated from the business logic layer, whose services are reused for three types of user interfaces without changing the code [17].

The challenges of interconnection and communication of different protocols between heterogeneous systems have been investigated in 2010 [18]. An integration platform is constructed to achieve the synchronization and transformation of data between heterogeneous systems through registering, mapping the various service components and constructing the SOA framework of enterprise based on the service component. The platform uses XML as a middleware for mapping several data sources into a unified model depending on a set of mapping transformations. The element of each mapping model has one service component which indicates the source or destination of elements. So, a path from a data source to another data source must exist to define and achieve data synchronization.

A web service middleware framework that provides an interface for external clients to enable them to access different local data sources with a transparent manner has been developed [19]. It has a module for configuring the middleware with the information of the heterogeneous data sources. As a new query is submitted to the middleware, it is routed based on the registered information at the middleware then the query is locally wrapped into different forms. In addition, the result of the query is combined into large XML dataset that is returned to the client who initiates the query.

Kester et al. succeeded to develop a system that integrates several drug stores, which are incorporated based on SOA concepts with web services [20]. The database systems of the drug stores have been incorporated via a service bus such that drugs can be queried from all registered geographically distributed data stores. The nearest geographical location of drug result can be monitored and tracked.

Each of the previously mentioned systems has its own desirable features, but all of them suffer from some limitations. Thus, we propose a new data integration platform called Resource Oriented Heterogeneous Data Integration Platform (ROHDIP) to overcome these limitations.

III. METHODOLOGY

The pre-integrated system design is shown in Fig. 1. which consists of: applications, network connections, and local databases. Applications are allocated on different machines that utilize different operating systems (Windows, UNIX, Linux...etc.). Each application is written in any programming language (e.g. C#, Java, JS, Ruby ...etc.) A network is required for direct connection between each application and its corresponding local data source. This network can be LAN, MAN or WAN. There are several types of local databases each may have a specific data model with different database management systems (such as Relational, Object, Tree, Hierarchy, Flat file). Each application is able only to query its corresponding DBMS(Database Management System) that is installed on it.

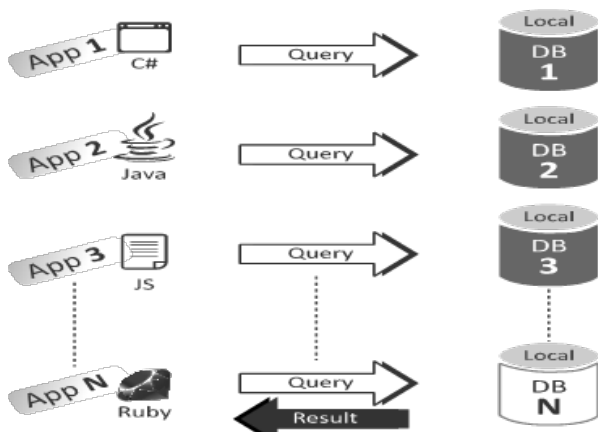


Fig. 1. Pre-Integration App(s)/Data source diagram

The previous architecture suffers from an obvious shortcoming that each application is restricted to query its local database only, to defeat this limitation the research attitude turned to focus on virtual integration approach.

The most challenging issue in the virtual data integration architecture is the network communication between the mediated schema and the data sources; see Fig. 2. This issue was solved by using Service-oriented architecture (SOA), as it exchanges data across the platform in the standard way through web services [21]. However, SOA data integration platforms still have some disadvantages [22] [23]. For example, size multiplication of the transmitted data leads to a negative impact on the network traffic and the system performance, especially when treating a large amount of data. Also, SOA platform suffers from higher latency and processing delay. Moreover, not all machines support the SOAP protocol (e.g. mobiles and embedded systems) as a native protocol. To overcome these limitations the Restful architecture is used.

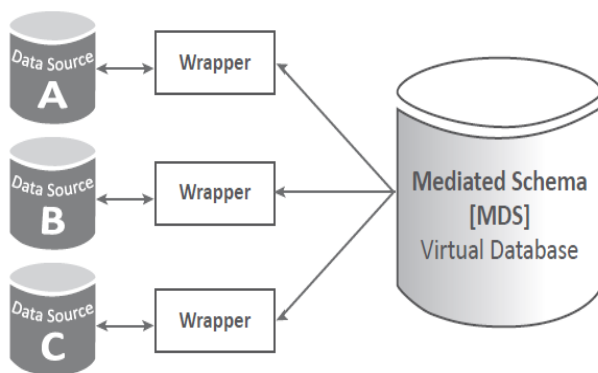


Fig. 2. Virtual data integration architecture

REST is a lightweight, easy and better alternative for the SOAP. Implementing the data exchanges across the platform using the Restful architecture of web service is more efficient in terms of both the network bandwidth utilization of the service requests transmitting over the Internet, as well as the latency incurred during these requests [22][24].

A. The Proposed Platform: ROHDIP

We propose a platform; namely, Resource Oriented Data Integration Platform (ROHDIP) that depends on the resource-oriented architecture (ROA) instead of the SOA architecture. ROA uses Representational State Transfer (RESTful) service based on HTTP protocol for communicating local data sources with mediated schema. ROHDIP is designed as a collection of collaborative RESTful resources allocated on distributed machines with different operating systems and constructed according to the ROA principles as shown in Fig. 3.

B. ROHDIP Architecture

The proposed platform architecture consists of three major steps: mediated schema creation, Data Source subscription, and mediated schema querying.

1) Mediated Schema Creation

TABLE I. MEDIATED SCHEMAS METADATA

Mediated Schema ID	Mediated Schema Name	Schema Definition (JSON)	Subscribed Data Sources (JSON)
mdsStudents	StudentsVDB	{"StudentID":"","Name":"","Mobile":""}	[{"DataSourceId":"ds2","DataSourceName":"Engineering"}]
mdsStaff	StaffVDB	{"StaffID":"","Name":"","Mobile":""}	[{"DataSourceId":"ds3","DataSourceName":"Commerce"}]
.....	{.....}
mdSchN	mdEmployee	{"EmployeeID":"","Name":"","Mobile":""}	[{"DataSourceId":"ds2","DataSourceName":"Medicine"}, {"DataSourceId":"ds2","DataSourceName":"Engineering"}]

Every mediated schema “Virtual Database” has its metadata see TABLE I. The metadata contain: ID, name, corresponding schema definition in JSON (JavaScript Object Notation) format and a list of the subscribed data sources of the mediated schema in JSON format. Fig. 4. illustrates the mediated schemas metadata in JSON format.

2) Subscribed Data Sources

When a new data source needs to join the ROHDIP, it must be added to the subscribed data sources metadata; see TABLE II. The metadata contain ID, URI, name, data model/DBMS, connection information between the data source and its wrapper service in JSON format, schema definition in JSON format, wrapper schema transformation rules in JSON format and the result data format. The subscribed data sources metadata in JSON format are also illustrated in Fig. 4.

The wrapper schema transformation rules from the data source schema to the mediated schema (e.g. mdsStudents.StudentID = StdentNO, mdsStudents.StudentName = StdName, StdTel = null) are required as they enable the mediated schema to map the requested query to the data source schema semantics. Furthermore, the data source result format (e.g. JSON, XML, delimited text) is provided to enable the mediated schema to read the result and convert it to the requester desired format.

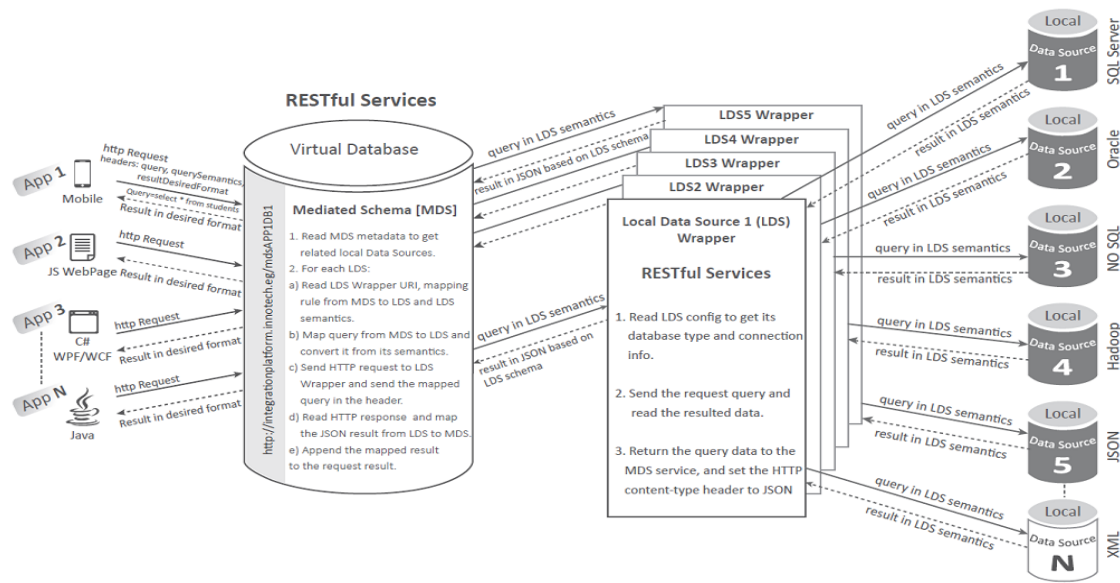


Fig. 3. Proposed ROHDIP System Design



Fig. 4. Mediated schemas, Subscribed data sources metadata in JSON format

TABLE II. SUBSCRIBED DATA SOURCES METADATA

DataSourceID	URI	DataSource Name	DataModel/DBMS	Connection-Information	DataDefinition	Wrapper	ResultDataFormat
DN1	http://78.110.9.246:9010	Medicine	Relational/SQL Server	{ "ConnectionType": "OLEDB", "ConnectionDetails": [{"Server": "192.168.200.2", "DataSource": "studentsDB", "UserName": "studentsAdmin", "Password": "123xx321"}] }	{ "StdentNO": "", "StdName": "", "StdTel": "" }	[{ "MediatedSchemaID": "mdsStudents", "MappingRules": { "StdentNO": "mdsStudents.StudentID", "StdName": "mdsStudents.Name", "StdTel": "" } }]	JSON
DN2	http://78.110.9.246:9050	Commerce	XML	{ }	{ }	[{ }]	XML
DNm	http://78.110.9.246:9090	Engineering	JSON	{ }	{ }	[{ }]	CSV

3) Mediated Schema “Virtual Database” Query

In order to query the mediated schema from any location and from any querying device, using the HTTP verb “GET”, we need to send HTTP request to the mediated schema URI i.e. <http://IntegrationPlatform.Innotech.com>, in which we have

to fill the requested mediated schema ID HTTP header i.e. “mdsStudents” and feed the “query” HTTP header with the desired query string i.e. “select * from mdsStudents”. Upon receiving the HTTP request; the mediated schema RESTful

service will check if the mediated schema ID is valid by examining the mediated schemas metadata.

If the mediated schema is valid, the mediated schema RESTful service will iterate through all its corresponding subscribed data sources by getting them from the mediated schemas metadata. Then, it will read its data model/DBMS, connection information, data definition, and wrapper details “mapping rules”. After that, HTTP request will be sent to all the subscribed data sources wrapper services URI(s) to retrieve the HTTP request result and append it to the mediated schema query result JSON data.

Finally, convert the JSON result consolidated from all the data sources to the requester data format then return it back to the requester in the body section of the requester query HTTP response.

IV. RESULTS AND DISCUSSIONS

The proposed platform is evaluated using the KDD’99 dataset, which includes 41 features extracted from DARPA (Defense Advanced Research Projects Agency) TCP dump in 1998 [25] [7]. The kdd’99 dataset consists of 494,021 connection records. We divide the KDD’99 dataset into six groups of smaller n-record datasets where n equals 5, 50, 500, 1000, 5000 and 10,000 records. The generated datasets are distributed over three servers each of which are of type PowerEdge R220 Rack Server, processor Xeon CPU e3-1220 v3 3.1GHZ, and Ram 24 GB.

We compare the performance of our proposed platform with SOA data integration framework, in terms of the end-to-end response time each query takes to retrieve a different number of rows. The response time is estimated for 5000, 25000, 50000, 75000, 100000 and 125000 rows. The proposed platform outperforms SOA, considering all the mentioned retrieved data sizes.

Fig. 5 through 10 illustrate the significant performance progress of the proposed ROHDIP platform comparing to the SOA framework regarding different sizes of retrieved query result. ROHDIP achieves the required integration of data retrieved from a query with minimum response time compared to SOA among a different number of data sets. The results clarify that the gap between ROA and SOA increases as the query result size increases. The results demonstrate that ROA is better than SOA in the data integration field.

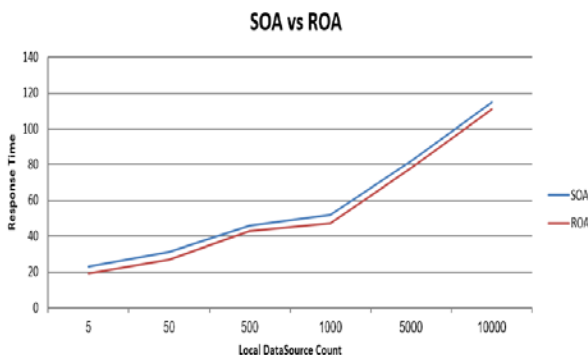


Fig. 5. Response time of ROHDIP vs. SOA for 5000 rows as a query result

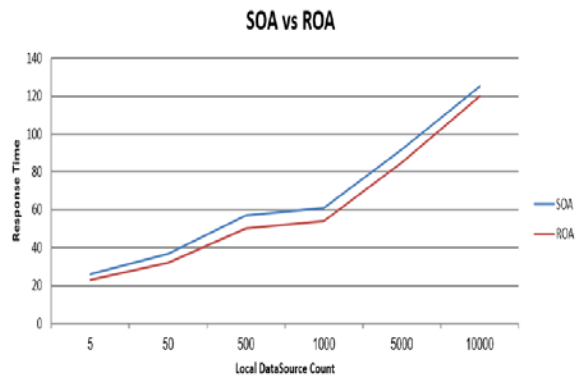


Fig. 6. Response time of ROHDIP vs. SOA for 25000 rows as a query result

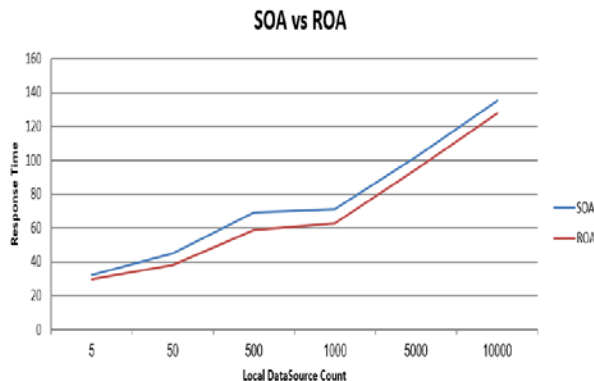


Fig. 7. Response time of ROHDIP vs. SOA for 50000 rows as query a result

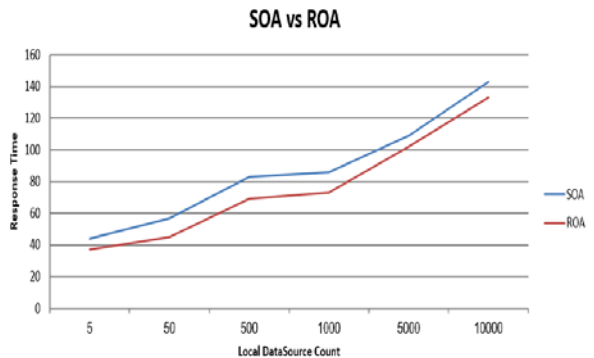


Fig. 8. Response time of ROHDIP vs. SOA for 75000 rows as a query result

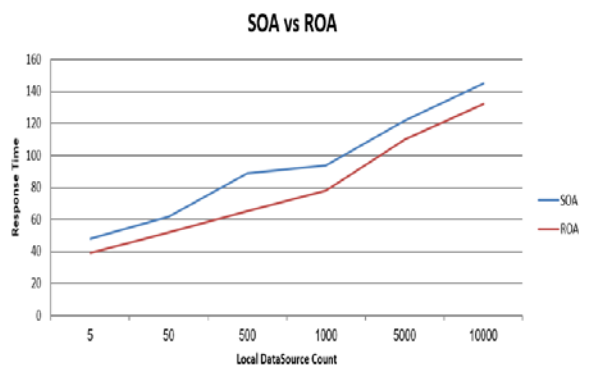


Fig. 9. Response time of ROHDIP vs. SOA for 100000 rows as a query result

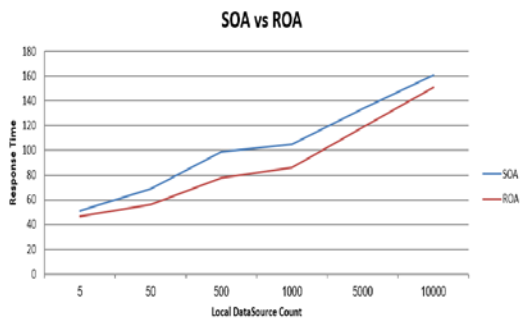


Fig. 10. Response time of ROHDIP vs. SOA for 125000 rows as a query result

V. CONCLUSION AND FUTURE WORK

Data integration is considered as the most urgent data task due to the daily increasing of data in heterogeneous data sources. In this paper, Resource Oriented Heterogeneous Data Integration Platform (ROHDIP) is proposed in order to integrate data from multiple heterogeneous data sources providing a unified query interface. The results evidence that ROA outperforms SOA for any query result size on a variety of distributed data sources achieving the minimum response time.

We believe that the vision and research contribution described in this paper will serve large-scale data gathering and integration studies in the near future.

As mentioned in the paper, the heterogeneous data sources are distributed and allocated on different machines, so, our future vision is to apply parallel processing or parallel querying between the mediated schema RESTful service and the wrappers RESTful services. Further investigations are needed concerning the security issues.

ACKNOWLEDGMENT

This paper is supported by iNOTECH development corporate which provided us with the suitable devices for this research.

REFERENCES

- [1] Y. Liu and M. Xia, "Research of heterogeneous database integration based on XML," ICMET 2010 - 2010 Int. Conf. Mech. Electr. Technol. Proc., pp. 793–796, 2010.
- [2] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Information Integration: Conceptual modeling and reasoning support," 3rd IFCIS Int. Conf. Coop. Inf. Syst., pp. 280–289, 1998.
- [3] A. Y. Levy, "Logic-based techniques in data integration," Logic-based Artif. Intell., pp. 575–595, 2000.
- [4] J. A. R. Castillo, A. Silvescu, D. Caragea, J. Pathak, and V. G. Honavar, "Information extraction and integration from heterogeneous, distributed, autonomous information sources - A federated ontology-driven query-centric approach," Proc. 2003 IEEE Int. Conf. Inf. Reuse Integr. IRI 2003, pp. 183–191, 2003.
- [5] J. A. Reinoso-castillo, "Ontology-driven information extraction and integration from heterogeneous distributed autonomous data sources: A federated query centric approach," Architecture, 2002.

- [6] P. Ziegler and K. R. Dittrich, "Three Decades of Data Integration — All Problems Solved?" 18th IFIP World Comput. Congr. (WCC 2004), vol. 12, pp. 3–12, 2004.
- [7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," Commun. ACM, vol. 39, no. 11, pp. 27–34, 1996.
- [8] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, "Ontology-based integration of information—a survey of existing approaches," IJCAI Work. Ontol. Inf. Shar., pp. 108–117, 2001.
- [9] S. Abiteboul, O. Benjelloun, and T. Milo, "Web services and data integration," Proc. Third Int. Conf. Web Inf. Syst. Eng. 2002. WISE 2002., pp. 3–6, 2002.
- [10] G. Elsheikh, M. Y. Elnainay, S. Elshehaby, and M. S. Abougabal, "SODIM: Service Oriented Data Integration based on MapReduce," Alexandria Eng. J., vol. 52, no. 3, pp. 313–318, 2013.
- [11] F. Zhu, M. Turner, I. Kotsiopoulos, K. Bennett, M. Russell, D. Budgen, P. Brereton, J. Keane, P. Layzell, M. Rigby, and J. Xu, "Dynamic Data Integration Using Web Services," Int. Conf. Web Serv., 2004.
- [12] H. Garcia-Molina and others, "The {TSIMMIS} Approach to Mediation: Data Models and Languages," J. Intell. Inf. Syst., vol. 8, no. 2, pp. 117–132, 1997.
- [13] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," ACM Comput. Surv., vol. 18, no. 4, pp. 323–364, 1986.
- [14] A. Y. Levy, A. Rajaraman, and J. J. Ordille, "Querying Heterogeneous Information Sources Using Source Descriptions," Proc. 22th Int. Conf. Very Large Data Bases, vol. 1, pp. 1–26, 1996.
- [15] S. Sathya and M. Victor Jose, "Application of Hadoop MapReduce technique to Virtual Database system design," 2011 Int. Conf. Emerg. Trends Electr. Comput. Technol. ICETECT 2011, pp. 892–896, 2011.
- [16] P. Spiess, S. Karnouskos, D. Guinard, D. Savio, O. Baecker, L. M. S. de Souza, and V. Trifa, "SOA-Based Integration of the Internet of Things in Enterprise Services," pp. 968–975, 2009.
- [17] A. L. Sanz, M. N. García, and V. F. Batista, "XML based integration of web, mobile and desktop components in a service oriented architecture," Adv. Soft Comput., vol. 50, p. 565, 2009.
- [18] S. B. Li, Y. Hu, and Q. S. Xie, "Heterogeneous System Integration Based on Service Component," Appl. Mech. Mater., vol. 20–23, pp. 1305–1310, 2010.
- [19] X. Wei, "Heterogeneous Database Integration Middleware Based on Web Services," Phys. Procedia, vol. 24, pp. 877–882, 2012.
- [20] Q. Kester and A. I. Kayode, "Using SOA with Web Services for effective data integration of Enterprise Pharmaceutical Information Systems," pp. 1–8.
- [21] P. Version, "Mumbaikar, S., & Padiya, P. (2013). Web services based on soap and rest principles. International Journal of Scientific and Research Publications, 3(5). Chicago," Int. J. Sci. Res. Publ. 3(5). Chicago, vol. 3, no. 5, 2013.
- [22] G. Mulligan and D. Gračanin, "A comparison of soap and rest implementations of a service based interaction independence middleware framework," Proc. - Winter Simul. Conf., pp. 1423–1432, 2009.
- [23] K. P. Pavan, A. Sanjay, and P. Zornitza, "Comparing Performance of Web Service Interaction Styles: SOAP vs. REST," 2012 Proc. Conf. Inf. Syst. Appl. Res., pp. 1–24, 2012.
- [24] H. Hamad, M. Saad, and R. Abed, "Performance evaluation of restful web services for mobile devices," Int. Arab J. e-Technology, vol. 1, no. 3, pp. 72–78, 2010.
- [25] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA online intrusion detection evaluation," Comput. Networks, vol. 34, no. 4, pp. 579–595, 2000.

Improving the Emergency Services for Accident Care in Saudi Arabia

Dr. Amr Jadi

Department of Computer Science and Software Engineering
College of Computer Science and Engineering, University of Hail
Hail, Saudi Arabia

Abstract—The road safety is one of the serious challenges faced by most of the governments due to the involvement of various issues. Being perfect in driving is not enough on the roads but tackling the mistakes of other persons is also an important aspect of the present day driving. Dealing with the accidents, injured personals, communicating the emergency services and dealing with other legal formalities is a serious challenge in present conditions. Providing emergency services is a real challenge due to increased population, heavy traffic and communication problems.

In this paper, a novel technique is being introduced to avoid delays and major setbacks by emergency services at the time of accidents. The proposed technique works along with traffic control system of Kingdom of Saudi Arabia (KSA). By introducing such system in the healthcare, the serious drawbacks of communication can be avoided to a maximum extent. The proposed system can prove to be very effective at a place like Saudi Arabia, where millions of Hajj pilgrims visit for socio-religious gatherings.

Keywords—Accidents; Communication; Emergency Services; Hajj Pilgrims; Healthcare; Saudi Arabia

I. INTRODUCTION

Saudi Arabia (SA) is a country with a population of about 28.83 million giving shelter to many (2.5 million) visitors to holy cities like Mecca and Medina every year to perform Hajj [1]. Apart from these more than 6 million visitors go throughout the sites for the whole year. These numbers are increasing year by year and managing people and facilitating the emergency services always been a challenging task for the officials to deal with. Many technological aids are being provided to the pilgrims to ensure their safe and peaceful stay during the Hajj. But some mishaps takes place due to various reasons keep the authorities on their heels. Most of the accidents and demand for the emergency services needed by these pilgrims are reported due to lack of awareness, ignorance, over enthusiasm, lack of understanding, communication problems and rush to deal with formalities. Apart from these unauthorized pilgrims from nearest cities/ countries will overcrowd or burden the facilities provided by the local authorities. Due to these reasons, many people face serious problems of hygiene, health, and demand for emergency services increases.

A. Background

At least hundreds of people will suffer or lose their life due to accidents in KSA at the time of Hajj. More than 270

pilgrims were stampede killed during devil ritual in May 1994 [1]. A similar incident was seen in Mina stampede where more than two thousand pilgrims suffocated and many crushed to death at the time of 2015 Hajj in Mecca. Various accidents and number of deaths due to increasing pilgrims reported by Ministry of Healthcare in KSA are shown in Table 1.

TABLE I. DIFFERENT ACCIDENTS AND NUMBER OF DEATHS REGISTERED DURING HAJJ

Accident Type	Stampede		Airplane Crashes		Fire		Protests and Violence	
	Year	Deaths	Year	Deaths	Year	Deaths	Year	Deaths
1	July 1990	1,426	Jan 1973	176	Dec 1975	200	July 1987	400
2	May 1994	270	Dec 1974	191	April 1997	343	July 1989	17
3	April 1998	118	Nov 1978	170	Nov 2011	2		
4	March 2001	35	Nov 1979	156				
5	Feb 2003	14	Aug 1980	287				
6	Feb 2004	251	July 1991	247				
7	Jan 2006	334						
8	Sept 2015	4,173						

Note: Data accumulated from various sources.

Only deaths were reported in the above Table 1 and injuries are plenty to be considered at the time of such accidents. Apart from these many road accidents due various reasons as listed below are increasing the number of death [11] and injury cases in KSA.

- Due to bad behavior of drivers
- Failure to follow the regulations due to lack of awareness and inability to read Arabic on the sign boards by most of the foreigners
- Over speeding, jumping the signals, and wrong route driving
- Using mobile while driving
- Driving under the influence of drugs and alcohol
- Tampering with vehicles on the road
- Laxity in dealing with vehicles
- Disregard for roads and patrons or people

The emergency services in Saudi Arabia needs to cover long distances due to a nation with larger in size and population as compared to any other middle east countries.

Strong winds in diserts also comes out to be one of the important reasons for serious accidents in Saudi Arabia. Recent accident in Mecca due to crane collapse is due to strong winds.

B. Motivation

The ultimate goal of this work is to provide a usable approach for people living in Saudi in the emergency situations as the services provided by the authorities being absolute at the time of rush hours. In emergency situations even a common man must be able to get in contact with the services like police, hospitals and insurance agencies when an accident is reported. The proposed technique in this paper will help the authorities to establish a free corridor for the people suffering due to accidents or any kind of health related issues in emergency times.

II. RELATED WORK

Earlier many authors worked in this area to establish the best services at the time of disasters and to manage the emergency situations. Alrajeh and Bounabat [2] proposed decisional reactive agent (DRA) based approach for formal modeling and checking the disasters at the early stages. Rapid assessment and intervention team (RAIT) was established to respond quickly to different events which perform the initial assessment of the accident so as to provide needful assistance. However the communication between people visiting from different countries faced the language problem during Hajj times. To solve the issue Mohandes proposed a near field communication (NFC) technology [3], which helped to improve the services in an efficient manner. NFC helps to identify the nearest checkpoints, medical camps and to maintain the medical records of all pilgrims during emergency situations. This technology tracks the pilgrim status (alive, dead or injured), guides them in emergencies, things to do, hotels, camps, and works as an information platform.

An automatic and intelligent system was introduced by Ullah et al. [4] to observe and report the patients on time. These systems are using sensory networks to communicate the signs of heart rate, respiratory rate and mental status of patients. Such records and continuous monitoring will help the emergency services to deal with such patients in crowded gathering easily as they are traced using GPRS systems installed in the devices.

The importance of using information technology (IT) services and their applications during emergency situations were discussed by Hijji et al. [5]. The role of disaster management cycle to mitigate, preparedness, response and recovery were explained in detail. They conducted a study on the emergency situation at Jeddah in 2009 due to the flash flood which killed 163 people and affected more than 10,000 citizens. Such high intensity flash flood is expected in the city for about next 10 to 15 years [5] needs a preparedness of keeping emergency services available not only for its citizens but also for the pilgrims visiting every year. Yang et al. [6] suggested an intelligent shelter allotment (ISA) for such emergency situations, by which it assigns the route and information of destination to reduce the evacuation time.

To tackle with asthma patients in emergency situations, a runtime monitoring system was proposed by Dowaihi et al. [7].

The technique proposed by these authors will help many pilgrims in the country, who are within the range of emergency service providers. Necessary care from the hospital sources will be provided for such patients with the help of android based mobile applications and wireless web-based applications. Alerts will be delivered to the patients' mobiles and emails at the emergency times by using this method. Harrou et al. [8] proposed an early detection method for overcrowding issues in most of the emergency departments (ED). The key challenge of ED is to handle the emergency situation and early detection of abnormal patients. Harrou et al. discussed a statistical technique to detect the indications of abnormal situations observed with patients using ED [9]. Similar research is being carried out in India by Sangle and Kadam to establish a real time tracking system for pilgrims at Kumbh Mela [10] by using the embedded devices included with global positioning system (GPS) modules with different sensors.

In the above discussions, most of the authors are providing a solution based on the accidents or upon receiving an emergency situation. However, the present system proposed in this paper will help the healthcare industry and public to communicate and get aid easily at the time of emergencies.

III. EXISTING EMERGENCY SERVICE SYSTEM IN SAUDI ARABIA

Present emergency system will have to depend on the traditional method of communication systems when an accident takes place. During this time people around the patient will call the healthcare emergency numbers and the ambulance services will be alerted to reach the accident location [5, 6]. In the meantime, they have to go through all the hurdles of traffic and need to maintain a continuous track of the patient location until they reach the destination.

In case of pilgrims from other countries the problem of language will be a huge challenge for the emergency service providers and doctors; hence they need to depend on alternate mode of communication. Possible failures of the present emergency system to reach the accident locations are due to (a) communications failures [10] (b) delayed facilities (c) ignorance of people and (d) crowded roads.

There should be a serious alerting system functioning in KSA to deal with accidents and emergency services due to unending rush to the kingdom with pilgrims throughout the year. The alert system must be able to establish a communication with hospitals, ambulance, police for traffic controlling, insurance agencies and rest of the people passing through the accident routes. Such alert system will help authorities to clear the premises and restore the services to a normal position.

IV. PROPOSED TECHNIQUE

There are various parameters to be considered before proposing a new technique for smooth emergency services during different emergency conditions. A careful attention was taken to draw some of the reasons for emergency conditions are explained below. Most of the emergencies listed below are very commonly seen in Saudi Arabia at normal and rush times

[12]. During Hajj times the health problems include variety of challenges to emergency service providers.

- Due to rash driving, signal jumping and driving against rules
- Due to sudden health problems (i.e., heart attacks, blood pressure, glucose levels, etc.)
- Due to protests, wars, terrorist activities, etc.
- Due to ignorance of others
- Due to bad light or weather conditions (heavy rains, floods, etc.)

- Due to fire accidents and short circuits, etc.

Hence the arrangement of emergency services is unavoidable in any kind of accident scenario to save the life of victims/ patients. The entities which perform key role at the time of emergency situation are shown in Fig. 1. In such scenarios establishing a good communication seems to be an important criterion which needs to be of less effort and easy to use. In this paper, the author proposed an efficient communication system based on Google applications based priority checks for road traffic detection, GSM for continuous updates and wireless networks for continuous communication.

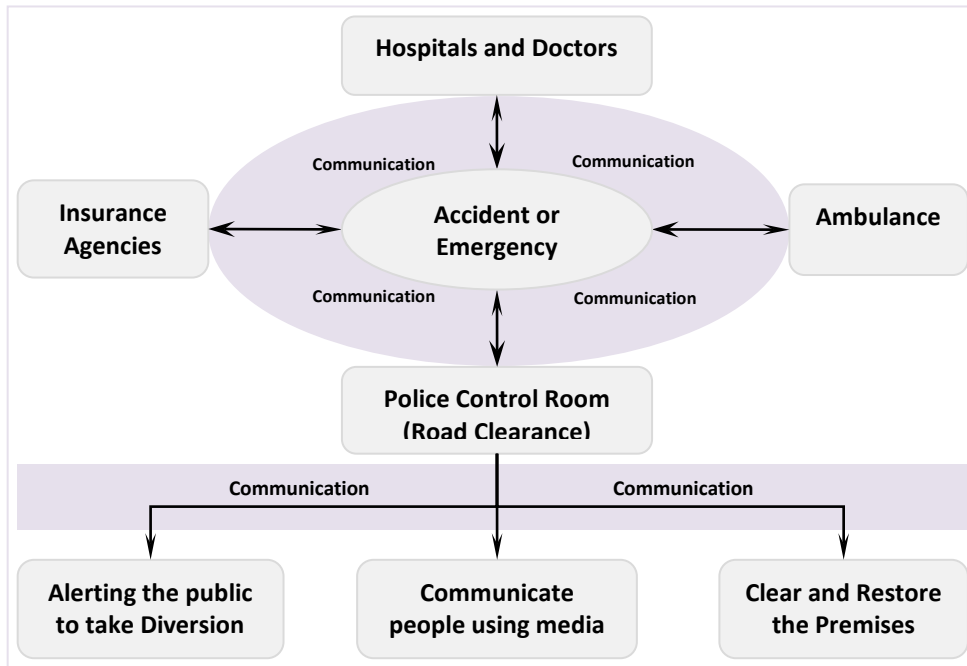


Fig. 1. Key entities performing important roles during emergency

In the case of an accident or an emergency the patient or any nearest common person will pass the information to the emergency service providers. The location of the person calling will be traced by GSM (Global System for Mobile Communication) technology [13] and the same will be communicated (raising alerts) to hospitals, ambulance, police and insurance agencies. Hospitals will decide the doctors and ambulances based on the requirement and distance to reduce the time in reaching the location.

The Google provides traffic information in one of its applications, i.e., Google Maps. Very clearly it shows the less traffic to heavy traffic with a color code of green to red respectively. The decisions taken by the emergency services in the proposed method are mostly based on the scientific approach to identifying the best route to be followed to reach the accident location. It explains the emergency services best possible route with less traffic and based on a road which is free from traffic as shown in Fig. 2. In the Fig. 2, there are three routes defined to reach a hospital with nearest possible distances. However, heavy traffic is indicated by red lines [15]. So based on the distance and less number of red lines in the

possible routes the decision will be taken. Now-a-day most of the vehicles are using GPRS (General Packet Radio Service) system [14] to locate the vehicle movement and the corresponding vehicle information carrying or about to carry the patients will be provided to the police control room. Based on the decisions by the proposed runtime monitoring system the drivers will be advised to follow the road.

On the other hand the police control room establishes a free corridor for the emergency services to ensure easy supply of facilities and emergency activities. The traffic police at all check posts for a distance of 3 – 5 km will get alerts from control room to ensure that the ambulance is moving freely till the hospital. This is possible only when the people are aware of such events and for which the proposed technique establish a free communication between police control room and local media (like radios, TVs, etc.) to cover such events as an alert for the common public during their journey. An awareness always helps the emergency service providers, police to control the situation and also to the authorities to clear and restore the regular services at the earliest possible. Otherwise a lot of

delay due to overcrowded people is observed in most of the accident locations or emergency scenarios.

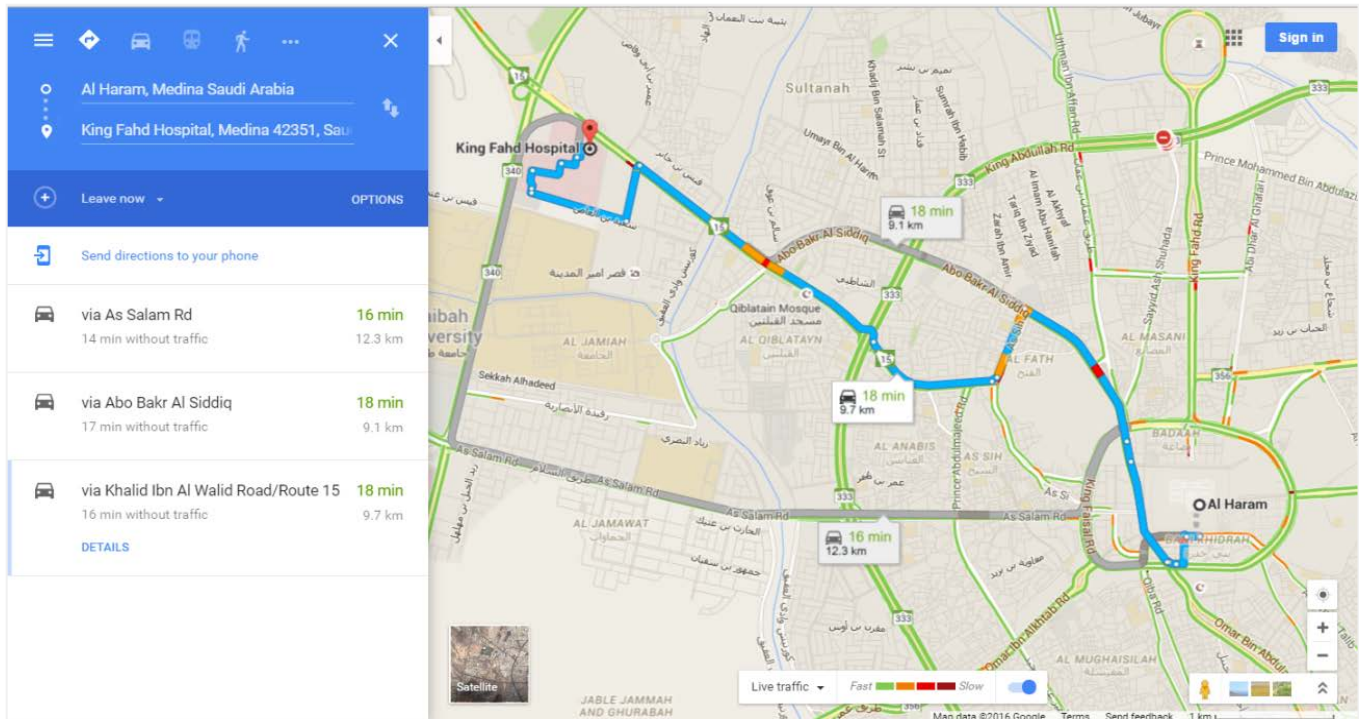


Fig. 2. Decision making during Emergency Hours

At the same time it is observed that almost all the roads near the mosques in KSA are observed to be very busy as shown in Fig. 3. Being an Islamic State, the number of mosques in SA is more and majority of its people do visit these

holy places for their prayers. At such places the authorities needs to ensure a special corridor for the smooth passage at the time of emergency situations.

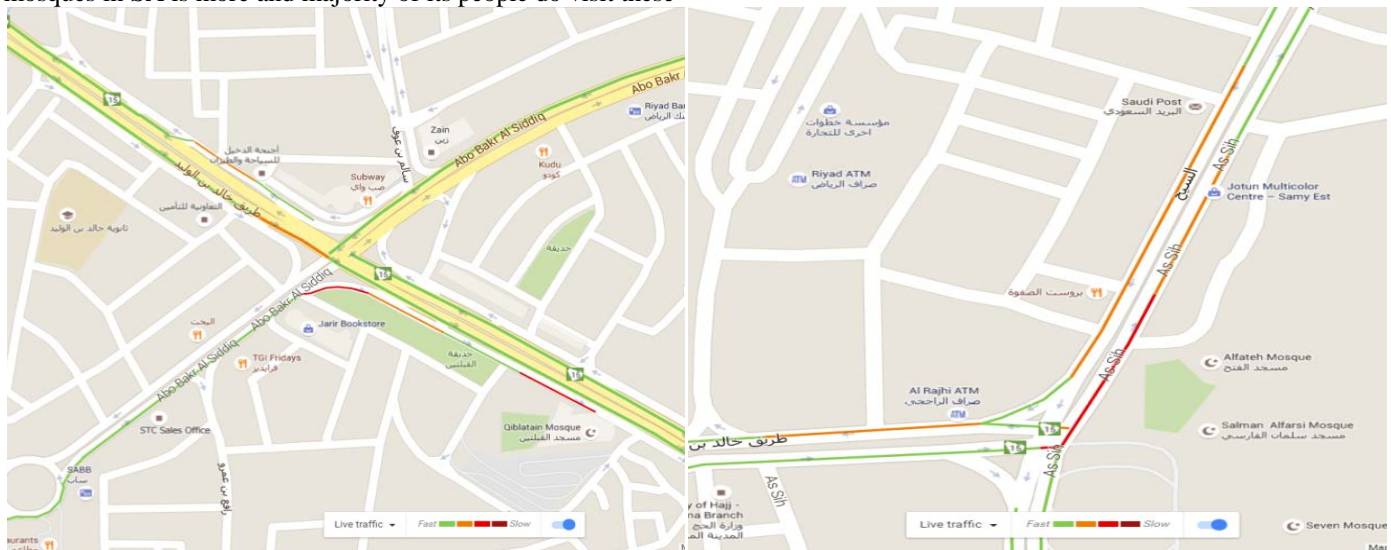


Fig. 3. Indication of Heavy Traffic Density near the Mosques

The need of passing information to the insurance agencies (as shown in Fig. 1) is essential to maintain a clean record of the emergency events to avoid delays in sanctioning the funds and to avoid wrong people to claim the policies. At the same time the insurance companies can assess the damage and help the patients on the spot and the financial aid or support without the knowledge of the diseased patients is possible by this

method/ approach. Otherwise, in general the patients family members need to claim the bills after damage is over and patients may need to suffer extra tensions at the time of emergency to arrange hospital bills at a sudden note. So this technique will not only help the patients and insurance agencies but also to the police in terms of reducing extra time to verify the details after a considerable gap.

V. CASE STUDY: ASSESSMENT OF EVENTS DURING EMERGENCIES

A simple flow diagram of events according to the proposed technique during emergencies is given below (Fig. 4). Some of the highlights of the proposed technique are listed below:

- Tracing of emergency location will be done automatically based on the communicating mobile or telephone service provider.
- Based on the situation and demand the nearest hospital / ambulance will be selected automatically.
- The police control room will generate an alert to corresponding police stations in the selected route. Once the ambulance and doctors finish their task at the emergency locations, local authorities will be allowed to restore the operations immediately.

Insurance agencies will updates their database and check the eligibility of patients for policies and immediate funding process will be sanctioned if they are eligible.

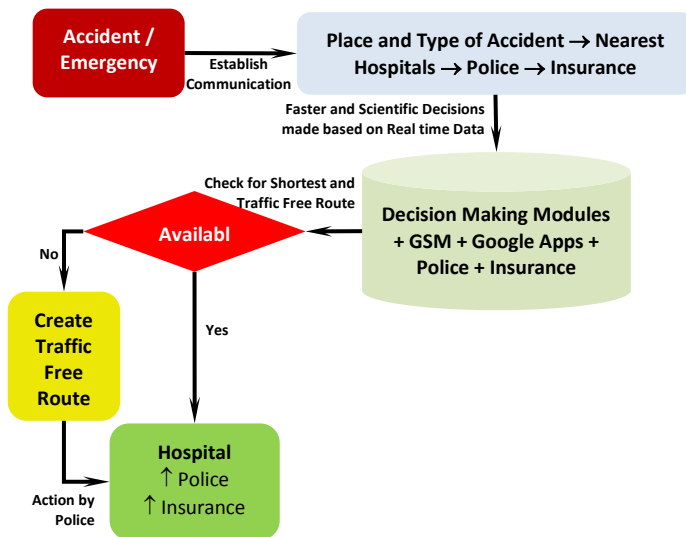


Fig. 4. Events based actions during Emergency Conditions

A system boundary for emergency services to the proposed method is shown in Fig. 5. The actions performed by the emergency services during emergency situations are shown in Fig. 6 using a sequence diagram.

- 1) The details of accident area and patient conditions are being noted carefully.
- 2) The patient or a random person will communicate the emergency services for the help.
- 3) A decision will be delivered to hospitals, ambulance, police and insurance agencies and municipal authorities with respective actions.
- 4) The ambulance with appropriate doctors and staff will reach the patients locations in the traffic free corridor.
- 5) Patient will be shifted to hospital and other support from insurance agencies and municipal authorities will be into

action for providing the insurance funds to the patient and to restore the services on accident area respectively.

6) Finally the patient will be treated with dignity and maximum care will be provided and with reduced pain for his family members and authorities.



Fig. 5. The designed System boundary for Emergency Services

The system boundary shown in Fig. 5 explains the rules and procedures to follow at the time of emergency situations. It will try to identify the people calling to the emergency services along with location and try to identify the kind of aid needed by the patients, so that, a strategic plan or decision is made to react immediately. A detailed description of actions and interactions between components at the time of emergency are shown in Fig. 6. Establishing a faster communication, decision-making and quick reactions to the problem are the key aspects of the entire proposed work.

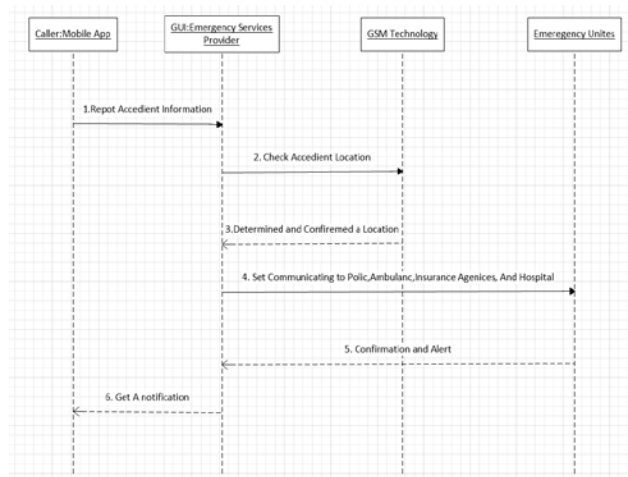


Fig. 6. Interaction components of events during emergencies in Saudi Arabia

The prototype class diagram representing different attributes and their operations in the proposed architecture are shown in Fig. 7. It explains the actions and the departments to be responded at the time of emergencies.

The emergency service provider will ensure to get the complete details of the caller and further procures the relevant location information from the GSM. Different units will be communicated and decision making process to direct each unit will be carried out automatically based on different scenarios and priorities based or risk levels.

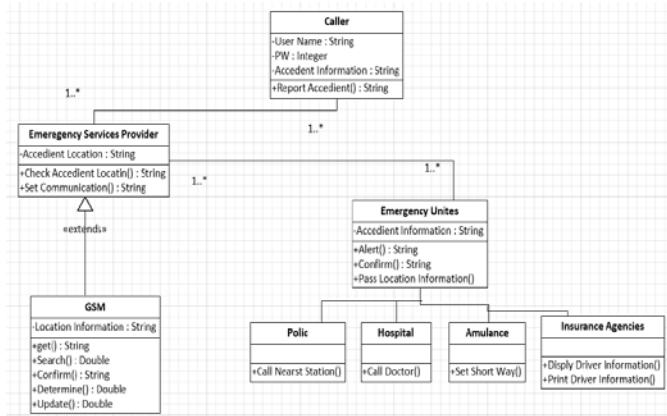


Fig. 7. The Proposed Prototype Class Diagram

The corresponding user interface of the architecture is shown in Fig. 8. This interface is user friendly and even a person with minimum knowledge of English can comprehend and operate it easily.

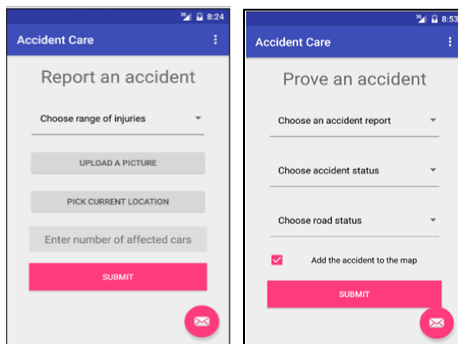


Fig. 8. The User Interface of the Proposed

Such a system will help the patients, hospitals, police, insurance agencies and local authorities to serve the people effectively in emergency situations.

VI. CONCLUSIONS

The proposed technique to enhance the emergency services in Saudi Arabia can be implemented with ease due to its usability and flexibility to adopt with existing system. Few modifications and certain modifications with low budget also can review the proposed model in Saudi Emergency services. So far the usage of traffic density system was not included in any kind of applications in SA. Hence by adopting the proposed technique the KSA will be able to create a healthy environment to tackle with emergency situations. Number of patients and possible casualties can be treated in less time effectively and properly during the accidents and emergency conditions so that a reduction in number of deaths can be seen in the nation. In the future work, testing of the modules in real-time environment can be extended for uncertain situations.

ACKNOWLEDGEMENT

The author would like to thank first the Almighty, Allah for his grace and blessings. The author also thanks his parents, family and teachers for their unconditional support and encouragement throughout the career. Lastly but not least author would like to thank Mr. Kamalakar Pallela for his technical support at different levels to make this successful.

REFERENCES

- [1] M.A. Mohandes. "Mobile Technology for Socio-Religious Events: A Case Study of NFC Technology." *IEEE Technology and Society Magazine* 34, no. 1 (2015): 73-79.
- [2] N.A. Alrajeh, and B. Bounabat. "Formal specification of humanitarian disaster management processes." In *2012 6th International Symposium on Medical Information and Communication Technology (ISMICT)*, pp. 1-4. IEEE, 2012.
- [3] M.A. Mohandes. "Near field communication for pilgrim services." In *Computing Technology and Information Management (ICCM), 2012 8th International Conference on*, vol. 2, pp. 771-774. IEEE, 2012.
- [4] F. Ullah, A. Khelil, A. A. Sheikh, E. Felemban, and H. M. Bojan. "Towards automated self-tagging in emergency health cases." In *Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on*, pp. 658-663. IEEE, 2013.
- [5] M. Hijji, S. Amin, R. Iqbal, and W. Harrop. "A Critical Evaluation of the Rational Need for an IT Management System for Flash Flood Events in Jeddah, Saudi Arabia." In *Developments in eSystems Engineering (DeSE), 2013 Sixth International Conference on*, pp. 209-214. IEEE, 2013.
- [6] K. Yang, A.H. Shekhar, F.U. Rehman, H. Lahza, S. Basalamah, S. Shekhar, I. Ahmed, and A. Ghafoor. "Intelligent shelter allotment for emergency evacuation planning: A case study of makkah." *IEEE Intelligent Systems* 30, no. 5 (2015): 66-76.
- [7] D. Al-Dowaihi, M. Al-Ajlan, N. Al-Zahrani, N. Al-Quwayfili, N. Al-Jwiser, and E. Kanjo. "Mbreath: Asthma monitoring system on the go." In *Proceedings of international conference on computer medical applications (ICCM)*, pp. 1-4. 2013.
- [8] F. Harrou, Y. Sun, F. Kadri, S. Chaabane, and C. Tahon. "Early detection of abnormal patient arrivals at hospital emergency department." In *Industrial Engineering and Systems Management (IESM), 2015 International Conference on*, pp. 221-227. IEEE, 2015.
- [9] F. Harrou, Y. Sun, and F. Kadri. "Enhanced monitoring of abnormal emergency department demands." In *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 692-696. IEEE, 2015.
- [10] S. Sangle, and S. Kadam. "Real time tracking and EHR for pilgrim." In *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pp. 116-120. IEEE, 2015.
- [11] K. Hojjati-Emami, B. S. Dhillon, and K. Jenab. "The integrative time-dependent modeling of the reliability and failure of the causes of drivers' error leading to road accidents." *International Journal of Strategic Decision Sciences (IJSDS)* 4.1 (2013): 25-39.
- [12] M. Al-Atawi, R. Kumar, and W. Saleh. "A framework for accident reduction and risk identification and assessment in Saudi Arabia." *World journal of science, technology and sustainable development* 11.3 (2014): 214-223.
- [13] S. Sonika, K. Sathiyasekar, and S. Jaishree. "Intelligent accident identification system using GPS, GSM modem." *International Journal of Advanced Research in Computer and Communication Engineering* 3.2 (2014).
- [14] M.S. Amin, J. Jalil, and M. B. I. Reaz. "Accident detection and reporting system using GPS, GPRS and GSM technology." *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*. IEEE, 2012.
- [15] K. Boriboonsomsin, M. J. Barth, W. Zhu, and A. Vu. "Eco-routing navigation system based on multisource historical and real-time traffic information." *IEEE Transactions on Intelligent Transportation Systems* 13, no. 4 (2012): 1694-1704.

Analysis of Purchasing Tendency using ID-POS Data of Social Login User

Case study of golf portal site

Kohei Otake

Faculty of Science and Engineering,
Chuo University
Tokyo, Japan

Takashi Namatame

Faculty of Science and Engineering,
Chuo University
Tokyo, Japan

Abstract—This study targets social login registrants on an EC site and aims to clarify the difference between the purchasing tendency of social login registrants and general members by analyzing product purchasing history. The authors focused on the golf portal site that is the subject of this research. The authors analyzed the purchasing data comparing social login registrants with general members. It became clear that the social login registrants and general members have different distribution regarding the number of purchases and purchase type. Moreover, the social login registrants have a larger range of purchase types per purchase and they are purchasing from a variety of genres. In addition, the authors analyzed them with a focus on the relationship between products purchased. As the results of network analysis, it became clear that the existence of specific product combinations (concentrated sets on the network) more readily purchased simultaneously by Facebook users than by general members. Moreover, the authors compared each network tendency using a network index (degree, closeness and betweenness centrality). As the results, it became clear that social login registrants have less resistance to purchasing expensive products on an EC site compared with general members and golf gears act as a bridge for purchasing.

Keywords—Social Networking Service; Consumer Behavior; Network Analysis; ID-POS Data

I. INTRODUCTION

There has been a general upward trend in the scale of consumer orientated electronic trading (Internet shopping) around the world in recent years. The Internet has ceased to serve simply as a media for the transmission of information, but rather continues to evolve as a platform (EC sites) for trading. As such, the importance of Internet marketing is on the rise [1].

There is an increasing interest in Consumer Generated Media (CGM) as represented by Social Networking Services (SNS) as a usable source of information for Internet marketing[2][3]. CGM refers to media created by consumers using the Internet. The content of CGM ranges from information exchange regarding various products and services, to everyday events. The onset of CGM has allowed companies operating EC sites easy access to conventionally unavailable consumer feedback. Combined analysis of information accumulated by EC sites (purchasing history and access log data) and information broadcast by consumers is the subject of

tireless research and development to realize more accurate behavioral analysis of consumers[4].

It is in such a context that more and more companies are adding social login features to their EC sites. A social login resembles a single sign on service offered by companies. Social login registrants can login to the company's EC site using their Facebook (SNS), Twitter or other such account. Well known companies offering a social login service include American technology company GIGYA¹. Utilizing social login enables users to streamline the registration process, while also alleviating the need for separate accounts on each EC site and reducing the risk of losing the password. On the other hand, social login can significantly reduce registration barriers and increase their ability to create new customers for companies. Moreover, social login facilitates the connection of registered consumers with a unique ID and social media account. In other words, companies can simplify the acquisition and analysis of information of their customers.

Social login is being introduced on the EC sites of various companies. Many studies looking at the affects of social login services on EC sites focus on access history such as pages viewed and page views per visit. Page view and pages viewed per visit are an important index when evaluating the effect of social login. However, the authors think that a more detailed analysis about purchase behavior focused on actual products is necessary. It is thought that close observation of customer tendency and purchase behavior will boost the affect of marketing. Consequently, this paper reports the analytical findings of research conducted to clarify the difference between the purchasing tendency of social login registrants and general members based on product purchasing history.

This paper is organized as follows. In chapter II, the authors describe the prior research focusing on behavior on social media. In chapter III, the authors describe the research objective. In chapter IV, the authors describe the result of the analysis using purchasing data compared social login registrants and general members. Based on the results, in chapter V, the authors describe the result of network analysis that focus on the relationship between products purchased. In chapter VI, the authors summarize all aforementioned analyses, review results and discuss. In chapter VII, concludes the paper.

¹GIGYA, <http://www.gigya.com/>, 2016/9/16 author checked.

II. PRIOR RESEACHES

This chapter is a short summary about research focusing on behavior on social media (actions on social media including access behavior and posts etc.) and purchase behavior. Many representative studies addressing behavior on social media and purchase behavior concern movie box-office records. Some such examples include a study [5] that divided the content of comments posted to YAHOO! Movies about a certain movie into positive and negative and analyzed the connection with that movie's box-office record, a study [6] analyzing the effect of negative posts on social blogs on a movie's box-office record, and a study [7] that modeled the combined effect on movie box-office records of both social blog posts and the volume of television advertisement. As Tsurumi [8] et al. have pointed out, it is thought that the reason these studies focus on movie box-office records is the ease of access to such box-office records and the related text data of reviews and comments posted by consumers.

Each of these relatively successful studies is extremely important as works focusing on behavior on social media and purchase behavior. On the other hand, many of them focus on movie box-office records. As Tsurumi [8] et al. have pointed out, movies exhibit a special characteristic in that many consumers comment after having watched the movie. It is thought that a detailed analysis of a more general product is required when applying social media to marketing activities [9]. This study differs from conventional research by selecting a golf EC site as the subject of research.

III. RESEARCH OBJECTIVES

This study targets social login registrants on an EC site and aims to clarify the difference between the purchasing tendency of social login registrants and general members by analyzing purchasing tendency based on product purchasing history. In addition, this study borrows data from Golf Digest Online Inc.² (herein referred to as GDO), the operator of the golf portal site that is the subject of this research. GDO is one of the largest golf portal sites in Japan boasting some 2 million members. Users can make golf course reservations, shop, manage and analyze scores online, and gain access to the latest golf news and product information. GDO has introduced social plus, a social login service operated by Feedforce Inc.³ that tenures a connection between the social login registrants' Facebook and Twitter accounts and the unique ID used on GDO's EC site. This study utilized the purchasing data from GDO's EC site.

IV. ANALYSYS OF PURCHASING DATA

A. Data Set

The authors will begin with an explanation of the data set utilized in this study. Purchasing data from the 24 month interval between January 2012 and December 2013 was used. The purchasing data listed the product name, date, unique ID, product purchase price and other relevant data for each product purchase. The authors extracted purchase data about social

login registrants from this purchasing data. Additionally, users who had registered in June 2013 were extracted to collect their purchasing information for the relevant period (Facebook: data set 1; Twitter: data set 2). This study utilizes these users as a social login registrant data set.

Next, a sampling was made of users not included in data set 1, 2 and who are not social login registrants. The sampling randomly selected the same number of users (1653 users) as Facebook social login registrants to collect their purchasing information for the relevant period (general members: data set 3). An overview of each data set is shown in TABLE I.

B. Classification Based on Purchase Price

This section presents the results of classification based on the purchasing tendency and purchase price of each data set extracted in the previous section. Firstly, the purchase price for users during the period was derived to confirm the purchasing tendency of each data set. User results were plotted in Fig. 1, 2, and 3 with the primary axis (left) as the purchase price per user (blue lines) and the secondary axis (right) as the cumulative purchase price ratio (red lines).

Secondly, a decile analysis of each data set was performed. Decile Analysis is a method of customer analysis that groups all customers into 10% categories (deciles) from highest to lowest purchase price [10]. The total purchase price of each decile is analyzed to derive the percent distribution in relation to overall sales and determine which customer segment contributes to sales. The results of decile analysis clearly showed that purchases by the top 30% of users accounted for some 80% of sales. Moreover, while each TABLE indicates that very few users make extremely expensive purchases, it shows that the purchase price of many users is low, confirming that purchasing tendency commonly follow the power law.

Thirdly, users were classified in each data set based on the cumulative purchase price ratio. More specifically, users with a cumulative purchase price ratio up to 70% were placed in the high purchase price group (High Group), users between 70% and 95% in the medium purchase price group (Middle group), and users with 95% and more in the low purchase price group (Low group). The numbers and ratio of each group are shown in TABLE II.

C. Comparison of Purchasing Tendency

This section presents the results of a comparison between the overall tendency and the purchasing tendency of each group classification using the classification based on purchasing price conducted in the previous section. For this comparison of purchasing tendency, this study focused on the number of purchases and purchase type per user during the relevant period. Compared with users purchase specific products, it is known that users purchase multiple types of products are higher purchase ratio when presented with product recommendations. Therefore, the authors think that product type is a necessary element when considering potential demand.

² Golf Digest Online, <http://www.golfdigest.co.jp/>, 2016/9/16 author checked.

³ FeedForce, <http://www.feedforce.jp/>, 2016/9/16 author checked.

TABLE I. OVERVIEW OF DATA SET

	Data set 1 (Facebook)	Data set 2 (Twitter)	Data set 3 (General members)
Users	1653	174	1653
Purchasing times	13999	997	28281
Average purchasing times	8.47	5.70	17.11

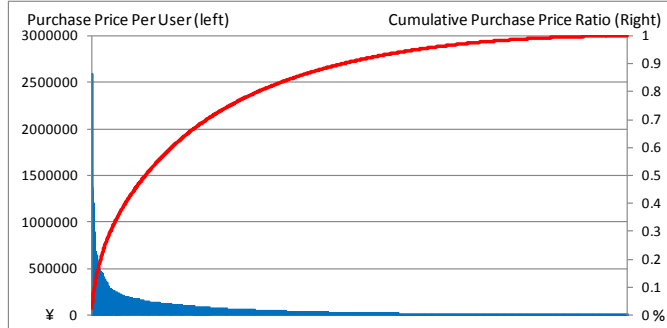


Fig. 1. Results of the purchase price per user and the cumulative purchase price ratio (Facebook)

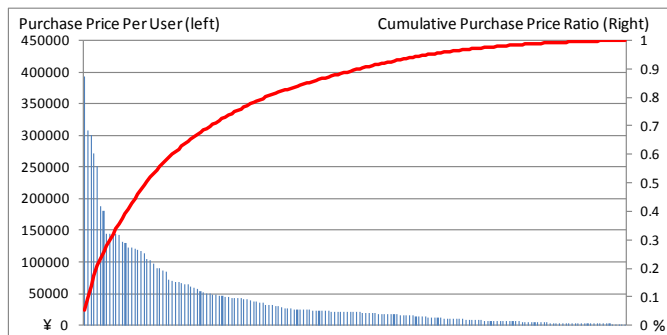


Fig. 2. Results of the purchase price per user and the cumulative purchase price ratio (Twitter)

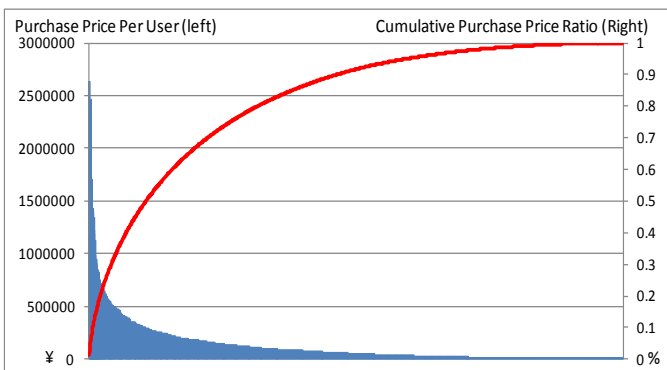


Fig. 3. Results of the purchase price per user and the cumulative purchase price ratio (General members)

TABLE II. RATIO AND THE NUMBER OF USERS OF EACH GROUP

	Data set 1 (Facebook)	Data set 2 (Twitter)	Data set 3 (General members)
High Group	355(22%)	41(24%)	366(22%)
Middle Group	646(39%)	70(40%)	616(37%)
Low Group	652(39%)	63(36%)	671(41%)
Total	1653(100%)	174(100%)	1653(100%)

The authors determined the number of purchases and purchase type per user during the relevant period for each data set. The authors derived the product types from the 49 separate classifications used by GDO; irons, outers (blouson, wind breaker, jacket), underwear, discount wear sets, wedges, calendars, caddy bags, socks, golf gear cases, golf gear sets, grips, gloves, competition gifts, sunglasses, shafts, shoes, drink cases, skirts, spikes, tees, drivers, travel covers, shorts, putters, videos/DVDs/tickets, fairway woods, vests, head covers, belts, balls, carry bags, utilities, small golf goods, repair goods, rainwear, long pants, one-piece dress, distance-measuring equipment, socks, health goods, umbrellas, books, mid-layer wear (sweater, trainer), long-sleeve shirts and polo shirts, electronics, short-sleeve shirts, polo shirts, hats, and practice goods.

Firstly, the difference between averages was tested for the number of purchases and purchase type in each set. The authors began by testing the Kolmogorov-Smirnov normality. The results showed that none of the data sets followed normal distribution for a significance probability of 0.01. Next, a Levene test was conducted to investigate the distribution ratio. The results showed that all of the data sets have unequal distribution for a significance probability of 0.01. Consequently, this study conducted the Kruskal-wallis test-one of the nonparametric testing methods-to test the distribution. Null hypothesis has the distribution of product purchases (purchase types) as equal for each data set. The results dismissed the null hypothesis and showed that the distribution is not equal to a significance probability of 0.01. Next, a multiple comparison using the Mann-Whitney U test was conducted to determine which data sets have a difference in distribution. In addition, Bonferroni correction was used for the multiple comparisons. The results showed that the distribution of product purchases and purchase types differed for all data set combinations for a significance probability of 0.01.

Secondly, linearization was conducted using the least-squares method to clarify the relationship between the number of purchases and purchase type for each data set. A fitted line was added to the scatter graph shown in Fig. 4, with the number of purchases as the x-axis and the product type as the y-axis. Users with the same number of purchases and purchase types are plotted on top of each other resulting in the high density seen in the bottom left of Fig. 4. The coefficient of determination was 0.7214 for general members, 0.7791 for Facebook, and 0.8693 for Twitter. Closer inspection of the fitted line clearly shows that its gradient is larger for both Twitter and Facebook than for the parent population of general members. This therefore clarified that users registered through social login have a strong tendency to increase the type of products purchased in proportion to the number of purchases in comparison to general members.

The following looks at the difference between each group using a classification based on the purchase price. Fig. 5, 6 and 7 are scattered graphs plotting users for each group with the addition of a fitted line and coefficient of determination where the number of purchases is the x-axis, and product type is the y-axis.

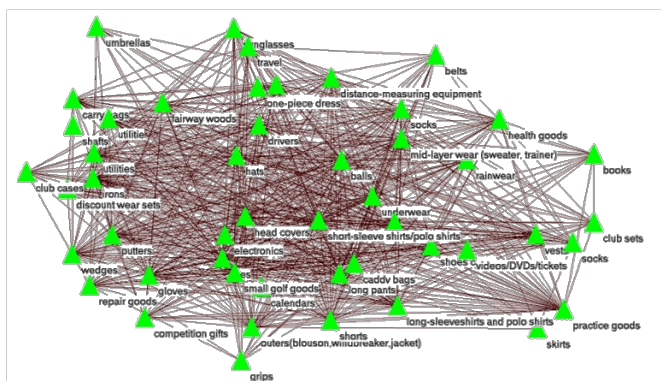


Fig. 9. A network graph created using the purchasing history (Twitter)

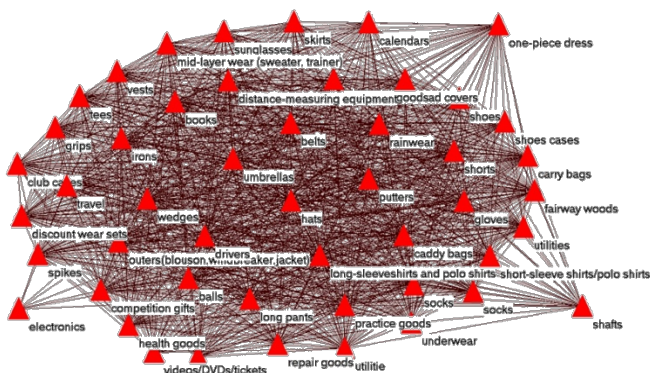


Fig. 10. A network graph created using the purchasing history (General members)

TABLE III. NETWORK INDEX OF EACH NETWORK GRAPHS

	Facebook	Twitter	General members
Users	1653	174	1653
Purchasing number of times	13999	997	28281
Edges	41247	2150	72475
Average Degree	33.74	1.79	60.37

Firstly, network graphs created using the purchasing history of every user in each data set are shown in Fig. 8, 9, and 10, and the network index of each network graph as TABLE III.

Closer inspection of each network graph reveals that nodes are equally spaced for general members in comparison to Facebook and Twitter users. This suggests that a lack of products that can be purchased simultaneously on the whole. On the other hand, when the authors look at Facebook and Twitter, nodes can be seen in proximity here and there. This suggests that these users have more products to purchase simultaneously, when compared with general members.

B. Analysis of Product Relationships with a Focus on the High Price Purchasing Group

Next, a more detailed analysis is conducted using the groups derived from cumulative purchase price ratio. The authors compared every possible combination of groups A, B, and C. The results showed a significant difference between the Facebook and general members of High Group (high price purchasing group). Analysis will herein focus on the Facebook and general members of High Group.

Firstly, the authors shown network graphs created using the purchasing history of users in both the Facebook and general member data sets in Fig. 11 and 12. Close inspection of Facebook’s High Group reveals a significantly greater deviation compared with the network graph (Fig. 8) created using all data. The authors thought that there is a strong relationship between specific products (more readily purchased together). Further examination of those combinations confirmed that winter apparel in the blue circle including sweaters, trainers, and long pants (center bottom right of Fig. 11), and summer apparel in the red circle including short (long)-sleeve shirts, polo shirts, hats, and shorts (center top of Fig. 11) tend to be more readily purchased simultaneously. Moreover, Golf Gears in the green circle including drivers, irons, wedges and grips (center bottom left of Fig. 11) tend to be more readily purchased simultaneously.

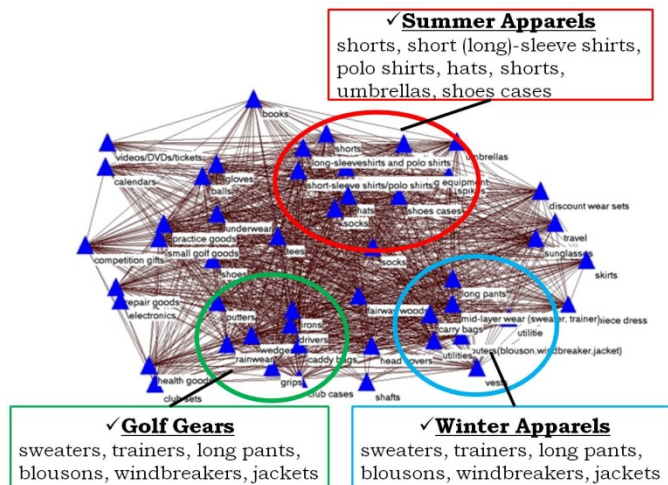


Fig. 11. A network graph created using the Facebook of High Group data

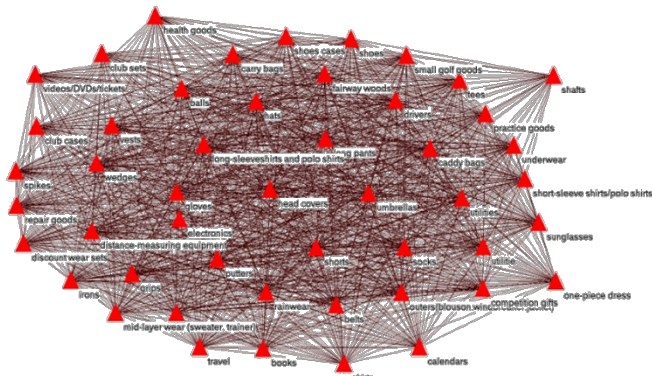


Fig. 12. A network graph created using the General members of High Group data

Next, the authors will focus on the general members of High Group. As with the network graph (Fig. 10) created using all data, a uniformly distributed graph was created with no deviation between products. From these results, it is thought that there are no specific products with a strong relationship (more readily purchased together) for general members.

Figure 13 and 14 are contour graphs of Fig. 11, and 12, respectively. These are created through 2-dimensional kernel

density estimating using coordinate data of nodes⁴. Comparing these graph, the authors can found that Fig. 13 which is the Facebook member have two peak regions. In these regions, some node concentrated. In this study, the authors set $\frac{1}{(80-1)^2} \times 1.2$ as a threshold for the peak. Then the authors found three peak regions in the graph. In Fig. 11, the authors depict three color circles (red, green and blue). The red region contains summer seasonal apparel and green are golf gears.

However, in Fig. 14, the authors cannot find an obvious peak. From this result, there are some different purchasing behavior between Facebook and general member. Moreover, Facebook members are easy to purchase simultaneously some specific categories.

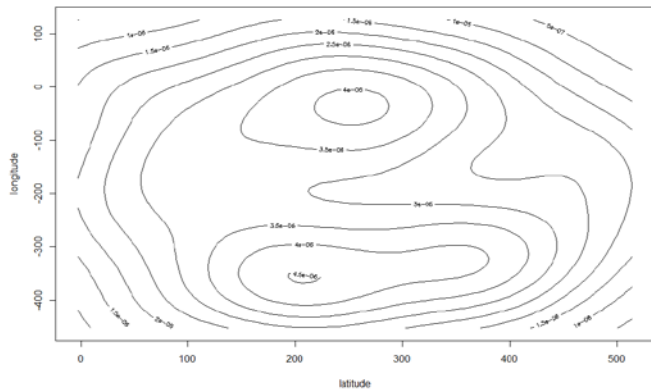


Fig. 13. Contour graph based on Fig. 11 (Facebook)

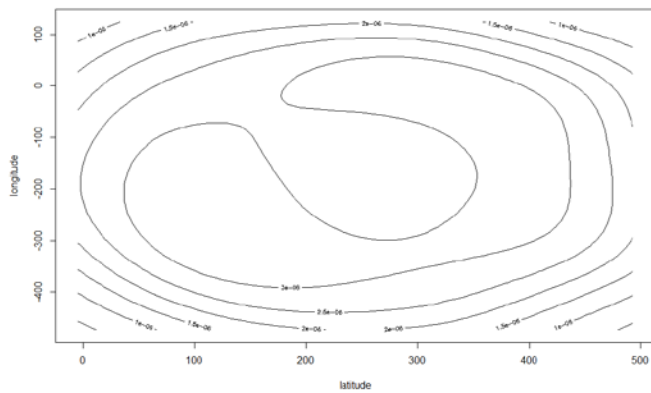


Fig. 14. Contour graph based on Fig. 12 (General Member)

Secondly, the authors compared each network tendency using a network index. This study compared network tendency using centrality. Centrality is an index used to identify the central node (therefore taking on a vital role) of a network. This study deals with the three separate index concepts of (1) degree centrality, (2) closeness centrality, (3) betweenness centrality. The three centralities were calculated for High Group of Facebook and general members and the five nodes with the highest centrality were extracted and compared.

The authors will first look at degree centrality. Degree centrality is the most basic method of calculating centrality in

which nodes where edges converge are deemed as having higher centrality. Therefore, the nodes degree is used as the centrality. The calculation results are shown in TABLE IV where the calculation equation is equation (1) when degree centrality is $dgc(v)$. The degree of node v is written as $deg(v) = |\Gamma(v)|$. Here, $\Gamma(v)$ is an adjacent node set of node v .

Closer inspection of products with a high degree of centrality shows that the same products appear in the same order up to the top 5 for both Facebook and general members. Therefore, it was clarified that the same products (quantitatively) were being purchased together with other products.

Next is a look at closeness centrality. Closeness centrality is a method of calculating centrality in which centrality is higher the closer the distance (therefore other nodes can be reached with a small step) is between the nodes. The calculation results are shown in TABLE V where the calculation equation is equation (2) when closeness centrality is $clc(v)$. Here, $d(v, u)$ is the step number between node v and node u .

Looking at products, it became clear that Facebook has a high centrality for golf gears. Therefore, it became clear that golf equipment was at the center of the network and strongly tended to resemble other products. On the other hand, it became clear that the distance between products such as clothes and bags tended to be small for general members.

Lastly is betweenness centrality. Betweenness centrality is a method of calculating centrality in which centrality is higher the more channels there are passing through a certain node. Therefore, this implies the importance of nodes acting as bridges between nodes in the network. The calculation results are shown in TABLE VI where the calculation equation is equation (3) when betweenness centrality is $bwc(v)$. Here, $\sigma_{s,t}$ is the minimal pass number between nodes s, t , and $\sigma_{s,t}(v)$ is the minimal pass number between nodes s, t passing through node v .

Looking at products, it became clear that there is a high centrality for consumable goods like balls, gloves, and small golf goods for general members, while Facebook showed a high centrality for golf gears like drivers, irons, and fairway woods.

$$dgc(v) = deg(v) \tag{1}$$

TABLE IV. CALCULATION RESULTS OF DEGREE CENTRALITY

Rank	Facebook High Group		General members High Group	
	Value	Id	Value	Id
1	2693	balls	4638	balls
2	2367	gloves	4268	gloves
3	2257	short-sleeve shirts / polo shirts	4236	short-sleeve shirts / polo shirts
4	2248	shoes	4080	shoes
5	2190	hats	4057	hats

$$clc(v) = \left(\sum_{u \in V, u \neq v} d(v, u) \right)^{-1} \tag{2}$$

⁴ The authors set 80 as band width

TABLE VI. CALCULATION RESULTS OF CLOSENESS CENTRALITY

Rank	Facebook_High Group		General mebmbers_High Group	
	Value	Id	Value	Id
1	0.9796	irons	0.9796	underwear
2	0.9796	fairway woods	0.9796	practice goods
3	0.9796	wedges	0.9796	hats
4	0.9796	drivers	0.9796	caddy bags
5	0.9592	caddy bags	0.9796	carry bags

$$bwc(v) = \sum_{s \in V} \sum_{t \in V} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}} \quad (3)$$

TABLE VII. CALCULATION RESULTS OF BETWEENNESS CENTRALITY

Rank	Facebook_High Group		General members_High Group	
	Value	Id	Value	Id
1	0.0074	drivers	0.0026	balls
2	0.0071	irons	0.0025	gloves
3	0.0070	fairway woods	0.0024	small golf goods
4	0.0063	balls	0.0023	short-sleeve shirts /polo shirts
5	0.0061	wedges	0.0021	outers(blouson, windbreaker,jacket)

VI. RESULTS REVIEW AND DISCUSSIONS

This chapter will summarize all aforementioned analyses, review results and discuss. Firstly, looking at the number of purchases (4-A, TABLE I) for data sets of the same period, social login registrants have a lower total number of purchases and a lower average number of purchases per user compared with general members. Moreover, examination of the average purchasing price shows general members at 125,601 (JPY), while Facebook is 65,430 (JPY). This therefore points to the fact that social login registrants are not always highly profitable consumers for companies.

On the other hand, detailed analysis (4-C) of the number of purchases and purchase type clarified that social login registrant's purchase a greater variety of products per purchase. It is speculated that this is due to users viewing pages across multiple product genres on GDO's EC site and purchase multiple products of their choice rather than purchasing several specific (predetermined) products. Moreover, social login registrants have better access to information regarding new products, sale items, and campaigns on GDO's Facebook page (in addition, the authors has confirmed that there were no social login specific sales or campaigns during the relevant period). It is speculated that while consumers have different values, in general they are likely to be enticed by sale items and bundle items with a lowered price, and that social login registrants are especially conscience of prices and price drops.

The interest social login registrants have in price drops can also be explained by network graph tendency. Network analysis (5-B) clarified the existence of specific product combinations (concentrated sets on the network) more readily purchased simultaneously by Facebook users than by general members.

It became clear that concentrated sets in the network are characterized by purchase behavior and a strong relationship between summer and winter apparel despite seasonal differences in sales numbers. Japan has 4 distinct seasons and

many EC sites hold sales in conjunction with seasonal changes. In particular, it is not uncommon for apparel to be heavily discounted to reduce the risk of dead stock due to constantly changing tendency. These results suggest that social login registrants are accustomed to shopping on EC sites. They are sensitive to GDO's sales such as seasonal price drops and campaigns and are shopping wisely.

On the other hand, network analysis clarified other strong relationship about the golf gears in sales numbers by Facebook users. Additionally, another characteristic was confirmed from the centrality derived through the network index analysis (5-B). Firstly, focusing on degree centrality produced exactly the same result for the most highly purchased products up to the top 5 for both Facebook and general members. On the other hand, golf gears came out on top for Facebook when focusing on closeness centrality. Furthermore, expensive golf gears came out on top even when focusing on betweenness centrality. Correspondingly, it became clear that the exact opposite is true for general members when focusing on betweenness centrality with inexpensive consumable goods like balls, gloves, and small golf goods coming out on top.

There are many companies in Japan selling golf products in retail stores (for example, Victoria Golf <http://www.victoria.co.jp/victoriagolf>, Niki Golf <http://www.nikigolf.jp/top/index.aspx>). Retail stores offer the chance to test swing golf gears, check form, and even consult real agents. In actual fact, many golfers test swing gears at retail stores and consult an agent when purchasing a golf gear. Golf gears are also some of the most expensive golfing items and many have reservations about purchasing them on an EC site. As such, companies operating EC sites like GDO are faced with the challenge of increasing golf gear sales on their sites.

The results of network analysis confirmed that social login registrants have less resistance to purchasing expensive products on an EC site compared with general members, that it is generally desirable to test products before purchasing (of course it is expected that some users test swing at retail outlets and then purchase on the net), and golf gears act as a bridge for purchasing.

The results of network analysis and analysis focusing on the number of purchases and purchase type showed that social login registrants and general members exhibit different purchasing tendency.

VII. CONCLUSION AND FUTURE WORKS

In this study the authors target social login registrants on the golf EC site and aims to clarify the difference between the purchasing tendency of social login registrants and general members by analyzing product purchasing history.

The authors analyzed the purchasing data comparing social login registrants with general members. It became clear that (1) the social login registrants and general members have different distribution regarding the number of purchases and purchase type, (2) the social login registrants have a larger range of purchase types per purchase and they are purchasing from a variety of genres. Based on the results of analysis of the purchasing data, the authors conducted network analysis focus on the relationship between products purchased. It became

clear that the existence of specific product combinations (concentrated sets on the network) more readily purchased simultaneously by Facebook users than by general members. Additionally, the authors compared each network tendency using a network index (degree, closeness and betweenness centrality). As the results, it became clear that social login registrants have less resistance to purchasing expensive products on an EC site compared with general members and golf gears act as a bridge for purchasing. From these results, the author considered that social login registrants and general members exhibit different purchasing tendency.

Future research will include a follow-up study of members analyzed in this study and a survey of changes in purchase behavior after registering as a social login user. Moreover, an analysis focusing on the attribute information of social login registrants will be conducted and their purchase behavior examined in comparison to general members with regard to social login limited sales.

ACKNOWLEDGEMENTS

The authors would also like to thank Golf Digest Online Inc for providing the data for this study.

REFERENCES

- [1] M. J. Shaw, C. Subramaniam, G. W. Tan and M. E. Welge, "Knowledge Management and Data Mining for Marketing," *Decision Support Systems*, Vol. 31, Issue. 1, pp. 127-137, 2001.
- [2] A. S.S.Reddy, P. Kasat and A. Jain, "Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining," *International Journal of Computer Application*, Vol. 56, No.1, pp.1-5, 2012.
- [3] M. Mizuno, Y. Takakai and N. Shinpo, "Communication with Consumers using Twitter Information: Conversation and Diffusion," *Journal of the Operations Research Society of Japan*, Vol. 58, No. 8, pp. 427-435, 2013.(In Japanese)
- [4] M. S. Yadav, K. Valck, T. H. Thureau, D. L. Hoffman and M. Spann, "Social Commerce: A Contingency Framework for Assessing Marketing Potential," *Journal of Interactive Marketing*, Vol. 27, Issue 4, pp. 311-323, 2013.
- [5] Y. Liu, "World of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue", *Journal of Marketing*, 70, pp.74-89, 2006.
- [6] G. Mishne and N. Glance, "Predicting Movie Sales from Blogger Sentiment", In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp.155-158, 2006.
- [7] N. Yoshida, A. Ishii and H. Arakaki, "Equation of the Smashing Success -Mathematize the Personal Influence Effect of Social Media-", *Discover 21 Inc*, 2007. (In Japanese)
- [8] H. Tsurumi, J. Masuda and A. Nakayama, "Relevance Analysis of the Communication on Twitter about Goods, and Sales Performance", *Journal of the Operations Research Society of Japan*, Vol. 58, No. 8, pp.436-441, 2013. (In Japanese)
- [9] H. Tsurumi, J. Masuda and A. Nakayama, "Possibilities and Limitations of Text Data Utilization on SNSs in Marketing Activities," *Japan Marketing Association, Marketing Journal*, Vol. 35, No. 2, pp. 38-54, 2015.(In Japanese)
- [10] M. Sato, E. Kato and Y. Matsuda, "Research of FSP Analysis in Food Supermarket," *UNISYS Technology Review*, Vol. 87, pp. 56-64, 2005.
- [11] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Information Processing Letters*, Vol.31, Issue 1, pp. 7-15, 1989.

Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus

Issa Atoum*

Faculty of Information Technology
The World Islamic Sciences & Education University
11947 Amman, Jordan

Ahmed Otoom

Independent Researcher,
Amman, Jordan

Abstract—Text similarity plays an important role in natural language processing tasks such as answering questions and summarizing text. At present, state-of-the-art text similarity algorithms rely on inefficient word pairings and/or knowledge derived from large corpora such as Wikipedia. This article evaluates previous word similarity measures on benchmark datasets and then uses a hybrid word similarity in a novel text similarity measure (TSM). The proposed TSM is based on information content and WordNet semantic relations. TSM includes exact word match, the length of both sentences in a pair, and the maximum similarity between one word and the compared text. Compared with other well-known measures, results of TSM are surpassing or comparable with the best algorithms in the literature.

Keywords—text similarity; distributional similarity; information content; knowledge-based similarity; corpus-based similarity; WordNet

I. INTRODUCTION

Text similarity is a field of research whereby two terms or expressions are assigned a score based on the likeness of their meaning. Short text similarity measures have an important role in many applications such as word sense disambiguation [1], synonymy detection [2], spell checking [3], thesauri generation [4], machine translation [5], information retrieval [6]–[8], and question answering [9].

There are three predominant approaches to compute text similarity. They can be categorized as corpus-based/distributional semantic models (DSMs), knowledge-based models, and hybrid methods. DSMs are based on the assumption that the meaning of a word can be inferred from its usage (i.e. its distribution in text). It is based on the following hypothesis: linguistic items with similar distributions have similar meanings [10]. Consequently, these models derive vector-based representations of the meaning of a word co-occurrence in a corpus. The vector-based representation is most often built from large text collections [5]. In this category, the latent Dirichlet allocation (LDA) assumes that each document is based on a mixture of topics, whereas a topic probabilistically generates various words [6], [11]–[13]. In the same category, the latent semantic analysis (LSA) is based on that the words that share similar meaning tend to occur in similar texts [6], [9], [14], [15].

TABLE I. TEXT SIMILARITY EXAMPLE

#	Sentence pairs	Human Score	LSA ^a	Li [16]	Mohler [17]
1	<i>The cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.</i>	0.01	0.19	0.33	0.45
59	<i>A cock is an adult male chicken. A rooster is an adult male chicken.</i>	0.86	1.00	0.83	1.00

^a Using TASA Space

The knowledge-based methods usually employ taxonomic information (e.g. WordNet) to estimate semantic similarity [18][19]. Sentence knowledge-based methods use semantic dictionary information such word relationships [19]–[21], information content [22], [23], parts of speech [18], [24], word senses [25], [26], and gloss definitions from a corpus [27], [28] to get the overall semantic score. These methods suffer from the limited number of general dictionary words, which are commonly used in general English literatures and may not suit specific domains.

Hybrid methods integrate various knowledge-based and/or corpus-based methods. They generally perform better [29]. In recent years, much of the work on lexical semantics has focused on distributional vector representation models [30], [31].

We have identified three cases where knowledge-based, corpus-based or *traditional* hybrid methods perform poorly. We illustrate these cases by examples. Table I shows two examples of two sentence-pairs taken from STS-65 benchmark dataset [16] that were compared using: LSA [15] (i.e. corpus-based method), [16] (i.e. knowledge-based method) and [17] (i.e. hybrid method).

The first case is as follows: methods that depend on a large corpus tend to overestimate relatively unrelated sentences or relatively related sentences (e.g., LSA). For the first sentence-pair, we obtained a similarity score of 0.19 (relatively high) for LSA measure, whereas the reported human similarity score

mean is 0.01. The LSA method depends on words' frequencies that tend to be relatively high in a large corpus (e.g., TASA). The second case is as follows: knowledge-based methods have the same drawback as the previously discussed method (LSA). The method of [16] depends on WordNet semantic relations (i.e. path and depth). This method can distinguish between general and specific concepts using WordNet but does not have information about words' distributions (or context). The third case is as follows: *traditional* hybrid methods that combine multiple measures over an average function generally perform poorly [17]. From [17], we determined that each sub-similarity method diverges in score compared to the overall similarity score. Each of the eight different measures has its strengths and weakness and thus will not get an acceptable semantic score in all cases. In many cases, one measure will have high similarity (e.g., >0.5 for LSA) and low similarity (e.g., <0.1 for path measure) over STS-65 dataset. In the second sentence-pair the same finding could be deduced. We deduced that the LSA and [17] measures overestimate the similarity score of the compared sentence-pair. Therefore, a similarity measure that use minimum data resources and get acceptable score is looked for.

Our work presents a hybrid-based text similarity measure that utilizes WordNet [32] information and a corpus[33]. The WordNet is a man-made ontology that shows promising results in the text similarity domain. The proposed method uses a small size word corpus, thereby eliminating the processing of large corpora. Using the weighted word similarity [34], a new text similarity measure is proposed. The proposed measure compares short text to long text and finds the maximum word similarity and the total exact matching words. The final similarity is calculated using the total similarity of the comparable words weighted by the text length in words.

First, the related works are summarized. Next, the proposed approach is presented and explained. Then, the proposed method is evaluated; finally, the article is concluded.

II. RELATED WORK

Sentence similarity methods (also called short text similarity) are used to measure word similarities in a sentence to reflect the overall semantic of the compared sentences. In general, sentence similarities can be categorized as corpus-based, knowledge-based, and hybrid methods.

A. Corpus-based Methods

Corpus models learn word co-occurrence from large corpora to predict the similarity of comparing text. Many models use information from internet sources such as: Wikipedia [35], Google Tri-grams [5], [36], and Search Engine documents [37]. These models can be categorized as DSMs and distributed vector representation models.

DSMs derive vector-based representations of the semantic meaning of patterns of word co-occurrence in corpora. In this category, LSA is based on that the frequency of words in certain contexts that could determine the semantic similarity of words to each other. That is, words that are similar tend to occur in similar texts [6], [9], [14], [15]. In latent Dirichlet

allocation (LDA) each document is based on a mixture of topics, whereas a topic probabilistically generates various words [6], [11]–[13]. The idea of the vector space model (VSM) [38] is to represent each document in a collection as a point in a space (a vector in a vector space). Points that are close together in the space are semantically similar, whereas points that are far apart are semantically different. The construction of a suitable VSM for a particular task is highly parameterized, and there appears to be little consensus over which parameter settings to use [39]. Moreover, many of these models are based on large corpora. The global vector model (GloVe) is an unsupervised learning model for word representation [40], which is trained on the non-zero elements in a global word–word co-occurrence matrix. The distributional model [41] combines visual features with textual ones, resulting in a performance increase. The explicit semantic analysis (ESA) represents the meaning of any text as a weighted vector of Wikipedia-based concepts [42]. Furthermore, the distributional method of LSA [43] is enhanced with WordNet semantic relations.

Distributed vector representation of words can capture syntactic and semantic regularities in language and help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words. The unified architecture of NLP [44] learns features relevant to the tasks at hand given very limited prior knowledge. This is achieved by training a deep neural network, building upon work by [30], [45]. Their models [44], [46] learn word representations in a binary classification task (related word to its context or not). They use the learned word representations to initialize the neural network models for other NLP tasks that also have word representation layers. One of the recent works on distributed representations is the work of [31] wherein they used probabilistic feed-forward neural network language model to estimate word representations in vector space. Align, disambiguate, and walk (ADW) model is a graph-based approach that has two steps; word transformation to the word senses (i.e. one of the meanings of a word) and disambiguation by taking context of compared words [47]. Based on WordNet, [48] exploit semantic representations of sentences using extracted features from a logic prover.

B. Knowledge-based Methods

Sentence knowledge-based methods use semantic dictionary information such word relationships [19]–[21], information content [22], [23], word senses [25], [26], and gloss definitions from a corpus [27], [28] to get word semantics. Based on human comprehension of sentence meaning, [49] proposed to measure the sentence similarity from three aspects that people identify in a sentence. People obtain information from a sentence on three aspects, or some of them: *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these objects. Consequently, they propose three similarities: objects-specified similarity, objects-property similarity, objects-behavior similarity, and overall similarity.

Some similarity models [19] measures the semantic relatedness between texts based on their implicit semantic links extracted from a thesaurus. Other models [25] measures sentence similarity based on word sense disambiguation and

WordNet synonym expansion. They build word sense disambiguation by using gloss interactions and expand it by synonyms. Then, the sentence is similarly calculated using cosine vectors. The reference [50] proposed a sentence similarity that used weighted word noun and verb vectors along with the order of words in a text.

In general, the knowledge-based approach is limited to the use of human-crafted dictionaries. Because of this, not all words are available in the dictionary and even though some word exists, they do not have full semantics.

C. Hybrid-based Methods

Hybrid-based methods are combinations of the previously mentioned methods. The reference [16] proposed a sentence similarity based on a non-linear function of WordNet path and depth, associated with information content from Brown Corpus, and sentence word orders. The reference [7] proposed a weighted similarity vector based on shortest path and term frequency to replace [16] semantic vector. They applied the similarity measure on photographic description data. The weighted textual matrix factorization (WTMF) model [11] is built on WordNet, Wiktionary, and Brown corpus. The reference [18] generated a semantic vector space using part of speech and WordNet. The reference [51] proposed a sentence similarity measure for paraphrase recognition and text entailment based on WordNet for existing words and an edit distance for proper nouns. The reference [24] proposed sentence similarity based on WordNet Information Content and part of speech tree kernels.

The reference [29] proposed a three-layer sentence measure: lexical layer, syntactic layer, and semantic layer. The overall sentence measure depends on the number of tokens, RDF triples that entail the semantic layer. In the same area, [52] combined the words meanings and phrase context in a sentence measure. The meaning words are implied by extracting words' lemma from a dictionary, whereas phrase context usage was extracted using a huge para-phrase alignment database [53].

Many hybrid methods are supervised models. They predict test sentence prevalence to training data. UNT model [54] uses regression machine learning based on hybrid text similarity methods of [17], [55], [56]. UKP system, which performed the best in the Semantic Textual Similarity (STS) task at SemEval-2012, uses the log-linear regression model to combine multiple text similarity measures of varying complexity. The reference [57] proposed the yiGou model. They used the support vector machine model with literal similarity, shallow syntactic similarity, WordNet-based similarity, and latent semantic similarity to predict the semantic similarity score of two short texts. The Takelab model [58] uses support vector regression model with multiple features measuring word-overlap similarity and syntax similarity to predict human sentence similarity. Each sentence is represented as a vector in the LSA model based on word vectors. Hybrid approaches show promising results on benchmark datasets.

III. PROPOSED METHOD

We highlighted the imperfections of word similarity

measures [34] that are either distance (knowledge)-based [16] or information content (IC)-based [22]. Distance-based methods suffer from the problem of having the same similarity value for words that share the same path or depth in a taxonomy such as WordNet. In contrast, the problem with IC measures is its limitation of available words in a corpus or getting the same similarity when the compared words has the same LCS ratio. We borrow the word similarity of [34] as shown in (1). Furthermore, we modified the word similarity factor of [34] as shown in (2).

$$Sim_{JDIC}(w_i, w_j) = \psi \cdot SimA \cdot SimB, \quad (1)$$

where $SimA = \log_2(Sim_{Li}(w_i, w_j) + 1)$, and

$$SimB = \log_2(Sim_{Lin}(w_i, w_j) + 1),$$

where w_i, w_j are compared words, $\psi \in [0,1]$ is a weighting factor that combines the IC of the pairs, and Sim_{Li} , Sim_{Lin} is the word similarity as in Li, Lin.

$$\psi = 1 - e^{-(\log_2(IC(w_i)+IC(w_j)+1))}, \quad (2)$$

where w_i, w_j are compared words, $\psi \in [0,1]$ is a weighting factor that combines the IC of the pairs, and Sim_{Li} and Sim_{Lin} is the word similarity as in Li [16] and Lin [22] respectively.

This article proposes a novel text similarity measure (TSM) that facilitates word similarity in (1). The TSM finds the maximum word similarity and the total exact matching words between compared sentences. Then, the total similarities of compared words are summed up and weighted by sentences' length and a logarithmic function.

The proposed maximum similarity of a word w and a text R is shown in (3).

$$Sim(w, R) = \arg \max_{1 \leq i \leq |R|} Sim_{JDIC}(w, R_i), \quad (3)$$

where R_i is the word i in text R and Sim_{JDIC} as defined in (1).

From [1], [33], [58], we inferred that compared text lengths and exact matches words have a direct effect on the final similarity score. The longer the compared text, the higher the chances of getting similar words.

The proposed TSM between two text fragments T, R is shown in (4).

$$\frac{\sum_{i=1}^{|T|} Sim(w_i, R) \cdot \text{Log}(2 \cdot \delta + 2 \cdot \text{Max}(|T|, |R|))}{|T| + |R|} \quad (4)$$

where, δ represents the exact word match between compared sentences. The Max function computes the maximum length between the compared sentences. The Sim function, as defined in (3), stands for the maximum similarity between a word and compared text fragment.

The application of (3) and (4) can be shown by the following sentence-pair taken from STS-65 dataset [16]:

S1: A boy is a child who will grow up to be a man.

S2: A rooster is an adult male chicken.

When we compare the two sentences using (3), the

maximum similar word-pairs from the sentence (S1) to the sentence (S2) are as follows: the word *boy* to the word *male* (0.282), the word *child* to the word *male* (0.153), and the word *man* to the word *adult* (0.786). The length of both sentences is 4 after stemming and removing stop words. Thus, applying (4) we got the similarity of 0.152. Compared to the reported human mean score (0.11), the proposed method got an acceptable similarity score.

IV. EVALUATION AND EXPERIMENTAL RESULTS

The evaluation of word and sentence measures are as follows.

A. Word Similarities

We evaluated the word similarity [34] on a relatively small benchmark datasets [60], [61]. Below, we extend the comparison to larger benchmark datasets: WordSim (WS)-353 [62], MEN dataset [63], and SimLex-999 [64]. The WordSimilarity-353 test collection contains two sets of English word pairs along with human-assigned similarity judgements. All the subjects in both experiments possessed near-native command of English. Their instructions were to estimate the relatedness of the words in pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words). The MEN test collection contains two sets of English word pairs (one for training and one for testing) together with human-assigned similarity judgments, obtained by crowdsourcing using Amazon Mechanical Turk via the CrowdFlower interface. The MEN data set consists of 3,000 word pairs, randomly selected on scales 1 (lowest) to 7 (highest) similarity. The SimLex-999 comprises 666 Noun-Noun pairs, 222 Verb-Verb pairs and 111 Adjective-Adjective pairs. SimLex-999 is challenging dataset for computational models to replicate. In order to perform well, they must learn to capture similarity independently of relatedness/association.

The Spearman correlation between different methods is shown Table II. The LSA, [65], [44], and VSM correlation were taken from [64]. We used Brown corpus and WordNet 3.0 for the JDIC measure, Li, and Lin measures. According to the results, both Li and Lin methods perform poorly which links to our initial hypothesis that a (corpus-based or knowledge-based) similarity method often does not perform well. In general, word similarity measures vary from one method to another depending on method features. Some methods use all word tags, while others use nouns only. Moreover, some methods support disambiguation or use additional domain information. The Spearman correlation of the JDIC method got the highest correlation for the SimLex-999 dataset. The JDIC approach looks for similar words and the SimLex-999 dataset is composed of similar words rather than related words. The WS-353 [62] list contains pairs that are associated but not similar in the semantic sense, for example: *liquid – water*. The list also contains many culturally biased pairs, for example: *Arafat – terror* [4]. Nevertheless, on average the borrowed method (JDIC) method achieved acceptable results compared with results of the state-of-the-art methods as shown in figure I. However, without a real system the comparison remains questionable.

We showed that the semantic similarity measures [66]

could play a major role in software quality detection. Therefore, we will confirm this finding in the next section by using JDIC in a new text similarity measure.

TABLE II. SPEARMAN CORRELATION OF WORD SIMILARITY MEASURES OVER DIFFERENT METHODS

Method/Dataset	MEN	SimLex-999	WS-353
Lin [22]	0.25	0.27	0.27
Li [16]	0.27	0.28	0.24
Huang [65]	0.30	0.10	0.62
VSM [39]	0.43	0.20	0.40
LSA[67]	0.48	0.23	0.40
Collobert [44]	0.57	0.27	0.49
Mikolov [68]	0.43	0.28	0.65
Islam [36]	0.72	0.33	0.62
JDIC [34]	0.56	0.53	0.61
Pennington [40]	0.66	0.40	0.67

B. Text Similarities

Table III shows the Pearson correlation of a list of text similarity methods over the benchmark dataset of Sem-Eval 2012 [69]. The dataset comprises pairs of sentences drawn from publicly available datasets that have been manually tagged with a number from 0 to 5:

- MSR-Paraphrase, Microsoft Research Paraphrase Corpus, 750 pairs of sentences.

<http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/>

- MSR-Video, Microsoft Research Video Description Corpus, 750 pairs of sentences.

<http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/>

- SMTeuroparl: WMT2008 development dataset (Europarl section), 734 pairs of sentences.

<http://www.statmt.org/wmt08/shared-evaluation-task.html>

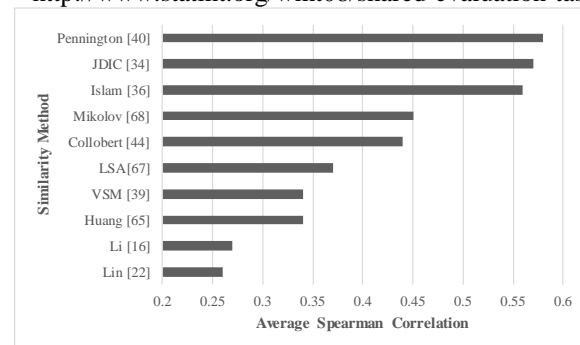


Fig. 1. Average spearman correlations over word similarity methods

Table III shows that the proposed method (TSM) was significant ($p < 0.01$) over all datasets except for a few. We implemented Li sentence measure while the Lin method was implemented based on the sentence measure proposed by [55]. The Google Tri-grams method [36] does not perform as well as for text similarity compared with its performance for word similarity measure (Table II). This finding shows that the word similarity on its own does not always lead to a good text similarity measure. It also supports our hypothesis that measures that use large collection of data could overestimate

unrelated sentences (shown in Table I). The low performance of Li measure was related to its inability to capture relatedness in compared sentences. Preliminary research showed that path and depth alone (Li measure) cannot give better semantic relatedness. Contrariwise, the Lin text similarity method

shows an average Pearson correlation of 0.51; thus, the information content gained better similarity scores. Further comparisons on the STS-65 dataset can be found at the work of [70].

TABLE III. PEARSON CORRELATIONS OF SEVERAL METHODS OVER THE SEM-EVAL 2012 DATA SETS

#	Method/Dataset	MSRvid	MSRpar	SMTeuoparl	Sur.OnWN	Sur.SMTnews
1	G. Tri-grams [36]	0.47	0.32	0.41	0.65	0.38
2	Li [16]	0.42	0.42	0.54	0.58	0.34
3	LDA	0.77	0.27	0.45	0.62	0.37
4	ESA [42]	0.75	0.43	0.38	0.62	0.33
5	Lin [22]	0.56	0.55	0.57	0.58	0.27
6	LSA [43]	0.66	0.36	0.57	0.66	0.39
7	ADW [47]	0.80	0.51	0.50	0.54	0.45
8	UNT [54]	0.88	0.54	0.42	0.67	0.40
9	WTMF [11]	0.84	0.41	0.51	0.73	0.44
10	TSM	0.83	0.58	0.45	0.66	0.42
11	yiGou [57]	0.84	0.51	0.48	0.67	0.48
12	TakeLab [58]	0.86	0.70	0.36	0.70	0.47
13	UKP [71]	0.87	0.68	0.53	0.66	0.49

We noted high performance of TSM (Pearson 0.66) on the dataset of OnWN because WordNet is one resource of TSM. The TSM performs better than methods (1–9) because each of them is considered to use one technique (knowledge-based or corpus-based) compared to TSM (hybrid). The application of TSM on the two sentence-pairs in Table I got the scores (0.002,0.80), thus our proposed TSM does not adhere to discussed drawbacks of knowledge-based and corpus-based measures. We found that the major performance of TSM was because of the proposed text similarity measure and the borrowed word similarity measure.

However, our method has some limitations. Compared with methods (11–13), it has lower performance. The main reason is that the top scoring methods tend to use most of the available resources and tools. For example, the yiGou 2015 adds the LSA features along with WordNet Similarity features. The TakeLab method uses multiple features that include syntax similarity which is not part of TSM. The UKP method uses a combination of approximately 20 features. These features include n-grams, ESA vector comparisons, and word similarity based on lexical-semantic resources. Furthermore, the TSM could not disambiguate words in different contexts. Therefore, we deduce that our method performance is accepted as it utilizes limited data resources. On average (figure II) our proposed TSM method got an acceptable Pearson correlation. The proposed method may be used in applications that do not require high accuracy such as in search engines or on systems that has low resources such as mobile applications.

V. CONCLUSION

This article presented a new text similarity measure based on previously proposed joint distance and information content word similarity measure, and the information content of compared words. The proposed text similarity is weighted based on comparable text length and the total exact word matches. The similarity measure outperforms much of the compared similarity measures and is significant at the 0.05 level. The reason behind the high achievement of our method

is due to the employment of additional information (corpus and information content) and the effectiveness of the borrowed word similarity measure. Although the proposed method has low performance compared to some compared models, it has less machinery and uses low information resources. In future, we plan to apply the proposed method on a real application of software quality.

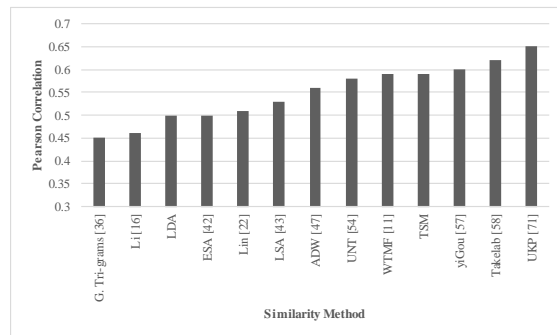


Fig. 2. Average Pearson correlations over text similarity methods

ACKNOWLEDGMENT

I would like to thank Prof. Dr. Narayanan Kulathuramaiyer for his valuable feedback and comments. I would like also to thank the anonymous reviewers for their comments.

REFERENCES

- [1] C. Ho, M. A. A. Murad, R. A. Kadir, and S. C. Doraisamy, "Word sense disambiguation-based sentence similarity," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, no. August, pp. 418–426.
- [2] P. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," in Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 491–502.
- [3] A. Islam and D. Inkpen, "Real-word Spelling Correction Using Google Web IT 3-grams," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3, 2009, pp. 1241–1249.
- [4] M. Jarmasz and S. Szpakowicz, "Roget's Thesaurus and Semantic Similarity," Recent Adv. Nat. Lang. Process. III Sel. Pap. from RANLP 2003, vol. 111, 2004.

- [5] A. Islam and D. Inkpen, "Unsupervised Near-Synonym Choice using the Google Web 1T," *ACM Trans. Knowl. Discov. Data*, vol. V, no. June, pp. 1–19, 2012.
- [6] B. Chen, "Latent topic modelling of word co-occurrence information for spoken document retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2009*, 2009, no. 2, pp. 3961–3964.
- [7] D. Croft, S. Coupland, J. Shell, and S. Brown, "A fast and efficient semantic short text similarity metric," in *Computational Intelligence (UKCI), 2013 13th UK Workshop on*, 2013, pp. 221–227.
- [8] S. Memar, L. S. Affendey, N. Mustapha, S. C. Doraisamy, and M. Ektefa, "An integrated semantic-based approach in concept based video retrieval," *Multimed. Tools Appl.*, vol. 64, no. 1, pp. 77–95, Aug. 2011.
- [9] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A Comparative Study of Two Short Text Semantic Similarity Measures," in *Agent and Multi-Agent Systems: Technologies and Applications*, vol. 4953, N. Nguyen, G. Jo, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2008, pp. 172–181.
- [10] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.
- [11] W. Guo and M. Diab, "A Simple Unsupervised Latent Semantics Based Approach for Sentence Similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012, pp. 586–590.
- [12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [13] J. Xu, P. Liu, G. Wu, Z. Sun, B. Xu, and H. Hao, "A Fast Matching Method Based on Semantic Similarity for Short Texts," in *Natural Language Processing and Chinese Computing*, Y. Zhou, Guodong and Li, Juanzi and Zhao, Dongyan and Feng, Ed. Chongqing, China: Springer Berlin Heidelberg, 2013, pp. 299–309.
- [14] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, no. 2–3, pp. 259–284, 1998.
- [15] S. Deerwester, S. S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, Sep. 1990.
- [16] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138–1150, Aug. 2006.
- [17] M. Mohler and R. Mihalcea, "Text-to-text Semantic Similarity for Automatic Short Answer Grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 567–575.
- [18] M. C. Lee, "A novel sentence similarity measure for semantic-based expert systems," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 6392–6399, 2011.
- [19] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, "Text relatedness based on a word thesaurus," *J. Artif. Intell. Res.*, vol. 37, pp. 1–38, 2010.
- [20] N. Seco, T. Veale, and J. Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet," in *Proceedings of the 16th European Conference on Artificial Intelligence*, 2004, no. Ic, pp. 1–5.
- [21] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," in *WordNet: An electronic lexical database*, vol. 305, C. Fellbaum, Ed. Cambridge, MA: The MIT Press, 1998, pp. 305–332.
- [22] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th international conference on Machine Learning*, 1998, vol. 1, pp. 296–304.
- [23] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95)*, 1995, vol. 1, pp. 448–453.
- [24] Y. Tian, H. Li, Q. Cai, and S. Zhao, "Measuring the similarity of short texts by word similarity and tree kernels," in *IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT)*, 2010, pp. 363–366.
- [25] K. Abdalgader and A. Skabar, "Short-text similarity measurement using word sense disambiguation and synonym expansion," in *AI 2010: Advances in Artificial Intelligence*, vol. 6464, J. Li, Ed. Adelaide, Australia: Springer Berlin Heidelberg, 2011, pp. 435–444.
- [26] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, 2009, pp. 190–199.
- [27] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in *Proceedings of the 5th annual international conference on Systems documentation*, 1986, pp. 24–26.
- [28] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation," in *Computational Linguistics and Intelligent Text Processing*, vol. 2588, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2003, pp. 241–257.
- [29] R. Ferreira, R. D. Lins, F. Freitas, S. J. Simske, and M. Riss, "A New Sentence Similarity Assessment Measure Based on a Three-layer Sentence Representation," in *Proceedings of the 2014 ACM Symposium on Document Engineering*, 2014, pp. 25–34.
- [30] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*, Springer, 2006, pp. 137–186.
- [31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.
- [32] C. Fellbaum, *WordNet: An electronic lexical database*. Dordrecht: Springer Netherlands, 1998.
- [33] W. N. Francis and H. Kucera, "Brown corpus manual," *Lett. to Ed.*, vol. 5, no. 2, p. 7, 1979.
- [34] I. Atoum and C. H. Bong, "Joint Distance and Information Content Word Similarity Measure," in *Soft Computing Applications and Intelligent Systems SE - 22*, vol. 378, S. Noah, A. Abdullah, H. Arshad, A. Abu Bakar, Z. Othman, S. Sahran, N. Omar, and Z. Othman, Eds. Kuala Lumpur: Springer Berlin Heidelberg, 2013, pp. 257–267.
- [35] L. C. Wee and S. Hassan, "Exploiting Wikipedia for Directional Inferential Text Similarity," in *Fifth International Conference on Information Technology: New Generations*, 2008, pp. 686–691.
- [36] A. Islam, E. Milios, and V. Kešelj, "Text similarity using google trigrams," in *Advances in Artificial Intelligence*, vol. 7310, L. Kosseim and D. Inkpen, Eds. Springer, 2012, pp. 312–317.
- [37] N. Malandrakis, E. Iosif, and A. Potamianos, "DeepPurple: Estimating Sentence Semantic Similarity Using N-gram Regression Models and Web Snippets," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012, pp. 565–570.
- [38] P. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, no. 1, pp. 141–188, 2010.
- [39] D. Kiela and S. Clark, "A Systematic Study of Semantic Vector Space Model Parameters," in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 2014, vol. 353, pp. 21–30.
- [40] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014, vol. 12, pp. 1532–1543.
- [41] A. Lopopolo and E. van Miltenburg, "Sound-based distributional models," in *Proceedings of the 11th International Conference on Computational Semantics*, 2015, pp. 70–75.
- [42] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis," in *International Conference on Artificial Intelligence*, 2007, vol. 7, pp. 1606–1611.
- [43] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese, "UMBC EBQUIITY-CORE: Semantic textual similarity systems," in

- Proceedings of the Second Joint Conference on Lexical and Computational Semantics, 2013, vol. 1, pp. 44–52.
- [44] R. Collobert and J. Weston, “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” in Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 160–167.
- [45] R. Collobert and J. Weston, “Fast semantic extraction using a novel neural network architecture,” in Annual meeting-association for computational linguistics, 2007, vol. 45, no. 1, p. 560.
- [46] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.
- [47] M. T. Pilehvar, D. Jurgen, and R. Navigli, “Align , Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity,” *Proc. 51st Annu. Meet. Assoc. Comput. Linguist.*, pp. 1341–1351, 2013.
- [48] E. Blanco and D. Moldovan, “A Semantic Logic-Based Approach to Determine Textual Similarity,” *Audio, Speech, Lang. Process. IEEE/ACM Trans.*, vol. 23, no. 4, pp. 683–693, Apr. 2015.
- [49] L. Li, X. Hu, B.-Y. Hu, J. Wang, and Y.-M. Zhou, “Measuring sentence similarity from different aspects,” in International Conference on Machine Learning and Cybernetics, 2009, 2009, vol. 4, pp. 2244–2249.
- [50] Y. Li, H. Li, Q. Cai, and D. Han, “A novel semantic similarity measure within sentences,” in Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, 2012, pp. 1176–1179.
- [51] G. Huang and J. Sheng, “Measuring Similarity between Sentence Fragments,” in 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2012, pp. 327–330.
- [52] M. A. Sultan, S. Bethard, and T. Sumner, “DLS@CU: Sentence Similarity from Word Alignment,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, vol. 2012, no. SemEval, pp. 241–246.
- [53] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, “PPDB: The Paraphrase Database.” in In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT ’13, 2013, pp. 758–764.
- [54] C. Banea, S. Hassan, M. Mohler, and R. Mihalcea, “UNT: A Supervised Synergistic Approach to Semantic Text Similarity,” *Proc. 6th Int. Work. Semant. Eval. conjunction with 1st Jt. Conf. Lex. Comput. Semant.*, pp. 635–642, 2012.
- [55] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” *Assoc. Adv. Artif. Intell.*, vol. 6, pp. 775–780, 2006.
- [56] S. Hassan and R. Mihalcea, “Semantic Relatedness Using Salient Semantic Analysis,” in Proceedings of the 25th AAAI Conference on Artificial Intelligence, (AAAI 2011), 2011, pp. 884–889.
- [57] Y. Liu, C. Sun, L. Lin, and X. Wang, “yiGou : A Semantic Text Similarity Computing System Based on SVM,” in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, no. SemEval, pp. 80–84.
- [58] F. Šarić, G. Glavaš, M. Karan, J. Šnajder, and B. D. Bašić, “Takeslab: Systems for Measuring Semantic Text Similarity,” *First Jt. Conf. Lex. Comput. Semant.*, pp. 441–448, 2012.
- [59] M. Lintean and V. Rus, “Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics,” in Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, 2012, pp. 244–249.
- [60] G. A. Miller and W. G. Charles, “Contextual correlates of semantic similarity,” *Lang. Cogn. Process.*, vol. 6, no. 1, pp. 1–28, 1991.
- [61] H. Rubenstein and J. B. Goodenough, “Contextual correlates of synonymy,” *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.
- [62] L. Finkelstein, E. Gabrilovich, and Y. Matias, “Placing search in context: the concept revisited,” *ACM Trans. Inf. Syst.*, vol. 20, no. 1, pp. 116–131, Jan. 2002.
- [63] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, “Distributional Semantics in Technicolor,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, 2012, pp. 136–145.
- [64] F. Hill, R. Reichart, and A. Korhonen, “SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation,” 2014.
- [65] E. Huang, R. Socher, C. Manning, and A. Ng, “Improving Word Representations via Global Context and Multiple Word Prototypes,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, 2012, pp. 873–882.
- [66] I. Atoum and A. Otoom, “Mining Software Quality from Software Reviews: Research Trends and Open Issues,” *Int. J. Comput. Trends Technol.*, vol. 31, no. 2, pp. 74–83, 2016.
- [67] T. K. Landauer and S. T. Dumais, “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge,” *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.
- [68] T. Mikolov, G. Corrado, K. Chen, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013, pp. 1–12.
- [69] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, D. Cer, and A. Gonzalez-Agirre, “Semeval-2012 task 6: A pilot on semantic textual similarity,” in Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, 2012, no. 3, pp. 385–393.
- [70] I. Atoum, A. Otoom, and N. Kulathuramaiyer, “A Comprehensive Comparative Study of Word and Sentence Similarity Measures,” *International Journal of Computer Applications*, vol. 135, no. 1. Foundation of Computer Science (FCS), NY, USA, pp. 10–17, 2016.

Trends of Recent Secure Communication System and its Effectiveness in Wireless Sensor Network

Manjunath B E
Research Scholar
Jain University
Bangalore, India

P.V. Rao
Prof & Head of R&D, Dept. of Electronics &
Communication Engg. RRCE
Bangalore, India

Abstract—Wireless sensor network has received increasing attention from the research community since last decade due to multiple problems associated with it. Out of many other significant problems e.g. routing, energy, load balancing, resource allocation, there is a lesser extent of effective security protocols towards solving security pitfalls in wireless sensor network. This paper studies the trend of research manuscript published in last six years about security problems to find that cryptographic techniques received more attention compared to non-cryptographic-based techniques. It also reviews the existing implementation towards addressing security problems and assesses its effectiveness by highlighting beneficial factor as well as limitations. Finally, we extract a research gap to identify the unexplored area of research, which is finalized to be implemented as a part of the future study to overcome the recent security issues.

Keywords—Wireless Sensor Network; Security; Cryptography; Encryption; Secured Routing

I. INTRODUCTION

A wireless sensor network consists of wireless sensor nodes which disperse evenly (in small scale deployment) or randomly (in large scale deployment) to capture the specific environmental information and forward it to the user using base station. This process is called as data aggregation in wireless sensor network [1] [2]. The complete success rate of data aggregation depends on how efficiently the routing among the nodes takes place in presence of uncertain traffic scenario. There are three types of routing protocols in wireless sensor network i.e. flat, hierarchical, and hybrid [3]. While performing communication, sensor nodes will require considering all forms of issues e.g. routing issues [4], load balancing issues [5], resource allocation issues [6], security issues [7], energy issues [8], etc. Majority of the sensor nodes works on the principle of 1st order radio-energy model which associate a direct relationship with radio (communication) and energy (i.e. battery). A sensor node is also known to possess very limited computational capability, restricted battery, and less memory. Due to this, it is quite a difficult task to run recursive algorithms in a sensor node, especially on those which require performing monitoring for longer duration of time without any human intervention. The majority of the standard routing protocols are meant for improving communication performance and not the security features. Although, there are multiple forms of secure routing protocols e.g. [9] [10], these routing protocols are not meant for protecting all dimensions of attacks e.g. Sybil attack, Denial-of-Service attack, Wormhole

attack, Sinkhole attack, node capture attack, statistical attack, replay attack, rushing attack, etc. There are various review papers discussing about existing security protocols in wireless sensor network [11]-[12], but the biggest challenges of those papers are i) they doesn't discuss the comparative analysis, which makes difficult to understand the best one, ii) majority part of the article has repetitive discussion of theory of sensor network and security which overshadows research contribution, and ii) they don't discuss explicitly the future work, which makes the reader vague to understand the contribution.

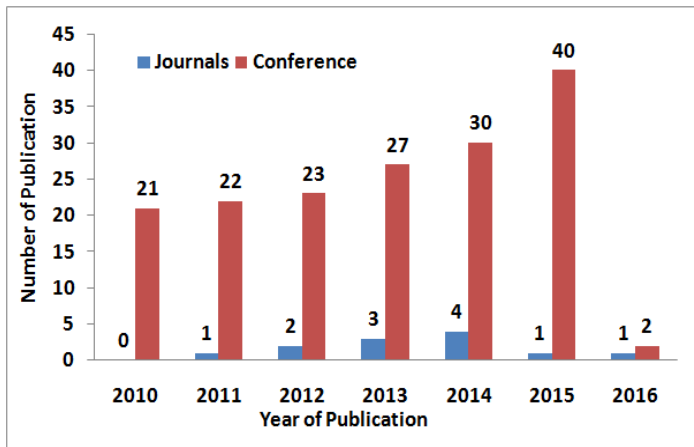
We actively comment that research on security problems in the sensor network is more than a decade old and still there are multiple underlying problems. There is a significant problem in existing techniques of Intrusion detection system [13] in WSN and identification of nodes are never studied in full fledged. There are multiple forms of possible nodes e.g. good nodes, attacker nodes, selfish nodes and partially defective nodes. It is said that differentiating all the above four categories of nodes is near to impossible for a given instant of time with heuristic routing data. It is because of the logic that when an attacker or malicious nodes intrudes a network by any means than it will never try to initiate an attack. Due to insufficient network and vital information from other nodes, the malicious nodes will not launch any form of attacks. In spite, it may start cooperating in data packet forwarding process just to accomplish more trust and reputation in the network. Because of this, differentiating a regular and malicious node is a challenging task. This is a very typical example to prove that existing techniques which are more inclined towards cryptographic usage, authentication mechanism, encryption, are not ultimately fruitful as in some point of time it could have missed identifying the intruders. Cryptographic techniques are quite expensive in implementation viewpoint, and existing methods don't bother about the practicality of the application of such technologies.

Therefore, the prime aim of this paper is to put forward the contribution of the research work being implemented during 2010-2016 towards thwarting the security issues in wireless sensor network. Section II discusses existing research trends on sensor network security followed by an explicit discussion of recent techniques in Section III. Research gap is explained in Section IV followed by the discussion of future work in Section V along with highlights of the tentative system architecture of it. The summary of the paper is discussed in Section VI.

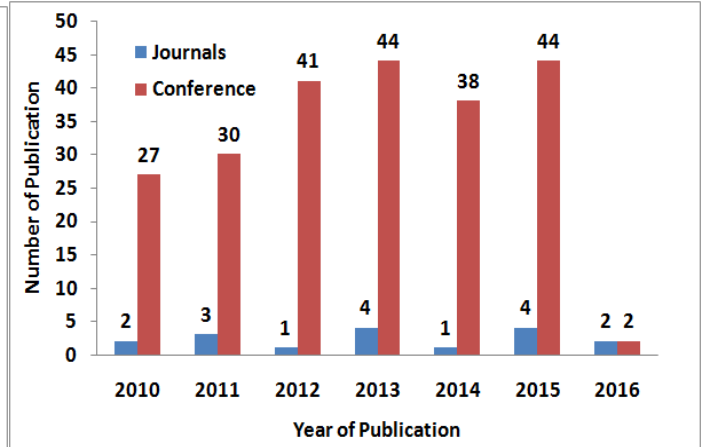
II. EXISTING RESEARCH TRENDS ON WSN SECURITY

At present, there are multiple review papers e.g. to prove that security problems in wireless sensor network have received a good number of attentions from the researchers in till date. Some of the important review papers e.g. [14]-[15] have descriptively discussed the contribution of the research field of security problems. However, as the security concerns are unsolved till date, we conclude that it was quite a difficult task to understand the effectiveness of the existing techniques. We strongly believe that security features can be incorporated in multiple ways and it is not necessary to include cryptographic techniques only. A closer look at the research trends shown in Fig.1 will highlights that journals (or Transaction papers) found in reputed site IEEE Xplore is extremely less. Fig.1.(a)-Fig.(c) are more or less related to cryptographic attention where we can see that good numbers of research papers certainly do exists. However, Fig.1.(d) is based on trust and reputation-based security techniques which have very less number of implementation studies to date. There are only 6 journals published in IEEE based on trust and reputation based approach to secure the communication system in wireless sensor network. A similar trend can also be seen in other research-based publishers e.g. ACM, Springer, Elsevier,

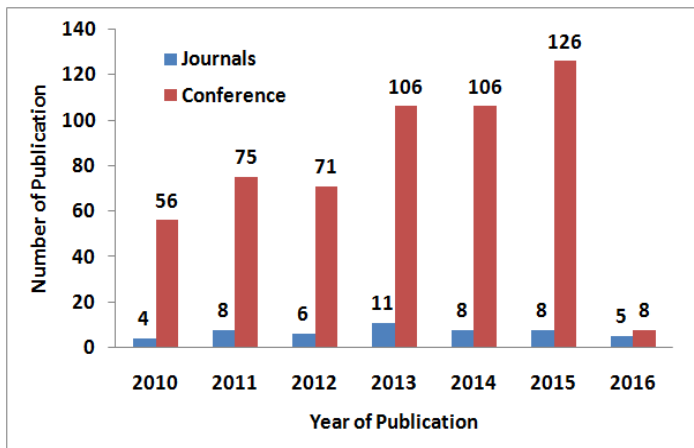
etc. Usually, the mechanisms that are designed based on Behavioral-based factors e.g. trust and reputations are much more light weight algorithms as compared to cryptographic techniques. However, it has received a poor attention. It should be known that wireless sensor node plays a chief role in advance technologies like Internet-of-Things (IoT) where existing security protocols are not that efficient. It will mean that IoT is a combination of cloud (Internet) and wireless sensor network which works on two different forms of security protocols. It will also mean that security protocols designs on the cloud are slightly incompatible to be executed over sensor nodes due to resource limitations. Hence, existing cryptographic protocols implemented on wireless sensor network will require an increasing attention in such cases. Implementing cryptographic applications will call for more resource consumption and larger management tools for associated security protocols which are expensive in nature. Hence, trust and reputation based techniques are the only solution of defense which doesn't require any additional resource or any dependency of complex key management or sophisticated encryption. Therefore, there is a need of investigating non-cryptographic algorithms for the optimal security keeping the future generation of sensor network usage.



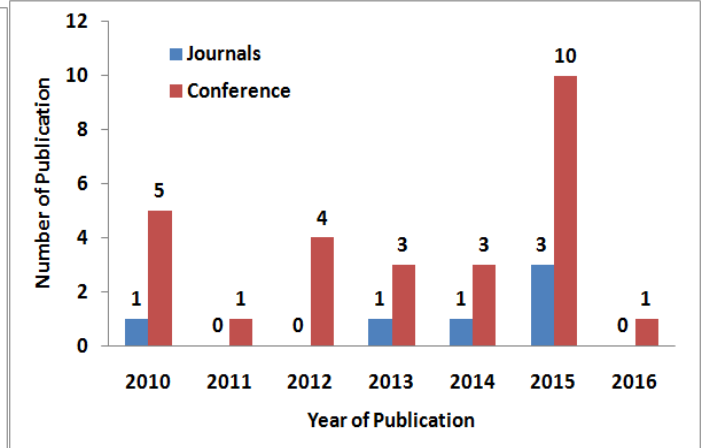
(a) Papers for keyword 'Secure routing in WSN'



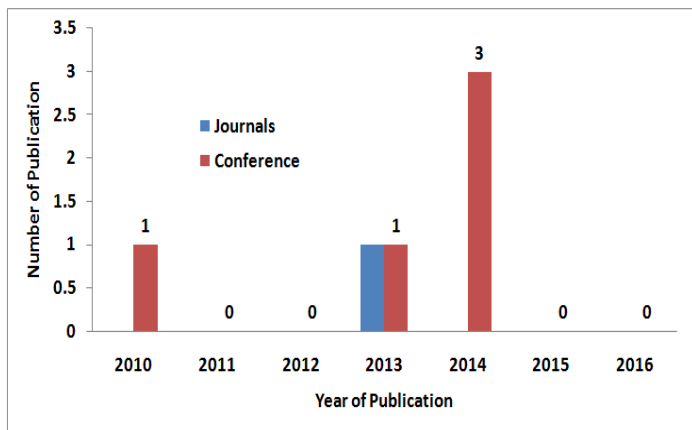
(b) Papers for keyword 'Encryption, WSN'



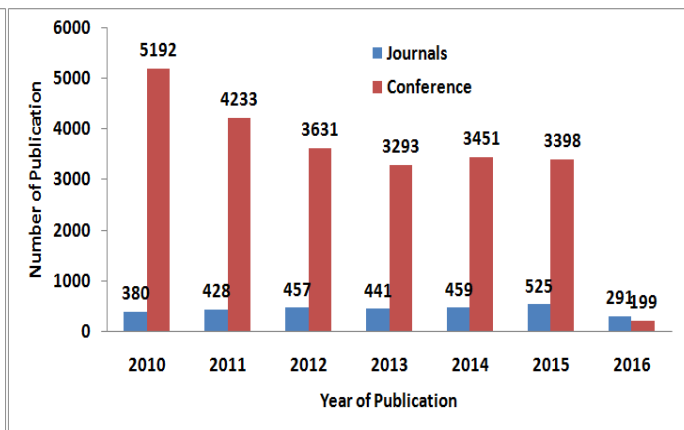
(c) Papers for keyword 'attacks, WSN'



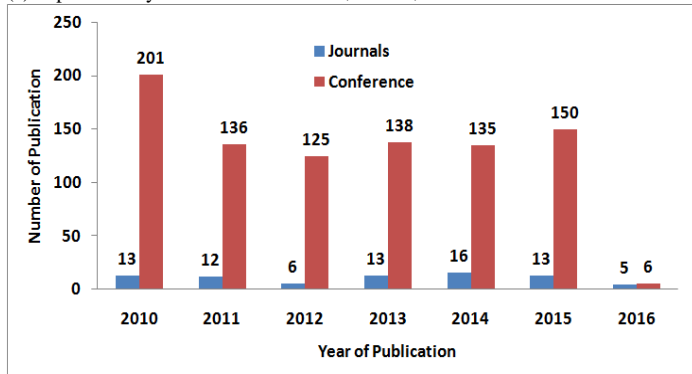
(d) Papers for keyword 'trust, reputation, WSN'



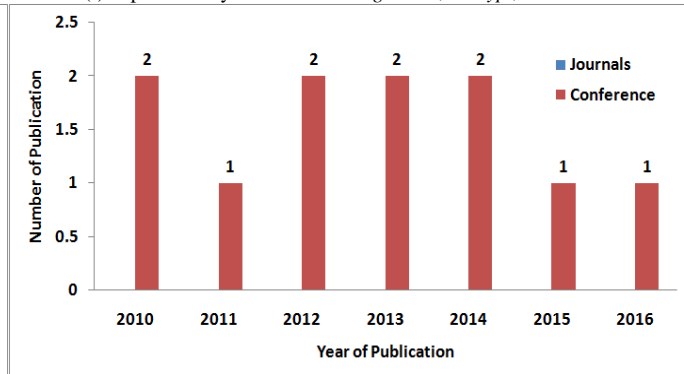
(e) Papers for keyword 'Neural Network, Secure, WSN'



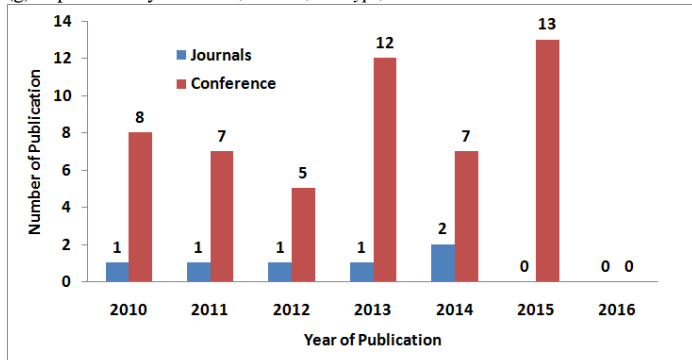
(f) Papers for keyword 'Genetic algorithm, encrypt, WSN'



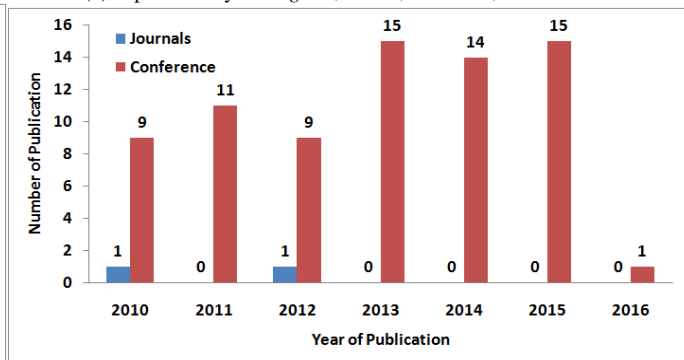
(g) Papers for keyword 'ant, swarm, encrypt, WSN'



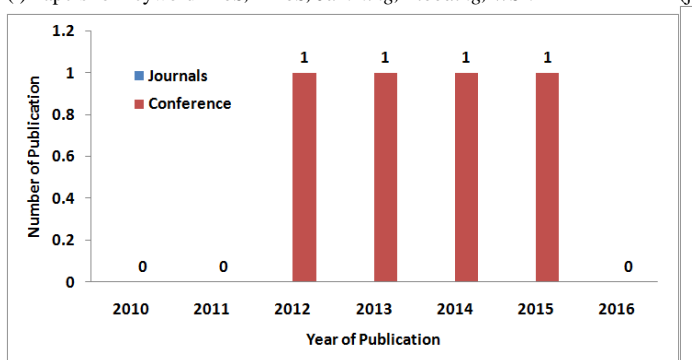
(h) Papers for keyword 'game, secure, malicious, WSN'



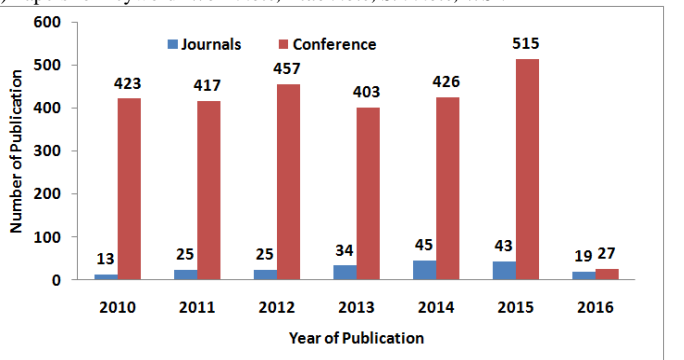
(i) Papers for keyword 'DoS, DDoS, Jamming, Flooding, WSN'



(j) Papers for keyword 'Wormhole, Blackhole, Sinkhole, WSN'



(k) Papers for keyword 'node capture attack, brute force, byzantine, WSN'



(l) Papers for keyword 'rushing, replay, routing, WSN'

Fig. 1. Trends of Research Publication for Security in Wireless Sensor Network between 2010-2016

III. RECENT TECHNIQUES

At present, there are dozens of research papers which has reviewed the existing security protocols in WSN. Hence, to avoid repetitive discussion, we discuss only the recent studies manuscript published during the year of 2010-2016. For productive discussion, we classify the techniques in three forms i.e. i) routing-based security methods, ii) cryptographic-based security techniques and iii) optimization-based security techniques. We only discuss the papers which are found to provide a solid base of security implementation in WSN most recently.

A. Routing-based Security Techniques

The routing based security techniques normally implement the Security features during performing routing operation. In such cases, the existing routing techniques are suitably modified to ensure security. Most recently, Das et al. [16] have presented a routing technique to resist wormhole attack as well as flooding attack in WSN. The method uses MAC scheme with the main management to ensure security. Nandu and Shekokar [17] have presented an authentication technique to resist DoS attack in WSN. The technique has used the concept of re-programming to do so. Re-programming is a mechanism of performing a selection of scope, decoding-encoding, versioning, etc. Henze et al. [18] have developed a secure access policy to protect sensitive sensor data over cloud environment. The study outcome shows minimization of key exchange time. The technique also uses AES encryption using 128 bit of key size as well as RSA with 2048 bit keys. Chen and Chen [19] have enhanced LEACH protocol suitable to incorporate security features. Considering the mobility of the nodes, the technique alters the clustering accordingly. The study outcome was compared with LEACH on energy, memory consumption, etc. Menaria et al. [20] have presented a unique algorithm for identifying the location of compromised node in WSN. The technique is used for isolating the compromised and misbehaved nodes from the network. The technique also uses Wiener index spanning tree to compute energy efficiency, and its outcome was evaluated on packet drop, energy, and throughput. Mengyao et al. [21] have developed a technique using ring-based grouping mechanism in WSN. The technique was meant for securing inter-cluster communication whose security depends on the trust factor. Tang et al. [22] have introduced a CASER (Cost-Aware Secure Routing) in WSN. The technique is developed based on energy stability factor and random walking model using probability theory. The study outcome showed better communication performance. Masdari et al. [23] have presented a technique to evaluate an effectiveness of Secured LEACH protocol in WSN. Ferng and Rachmarini [24] have introduced a method of the energy-efficient communication system over grid topology using simple key management techniques. Obaidat et al. [25] have presented a method for securing heterogeneous WSN. Using the concept of the trust factor, the presented technique evaluated link quality and node quality based on distance, energy, etc. Hence, it can be seen that there are various forms of routing technique to secure WSN communication system but the focus is more biased on energy efficiency.

B. Cryptographic-based Techniques

This is another frequently used technique to perform security. The cryptographic method usually performs encryption mechanism on the vulnerable links or susceptible sensors to secure communication. However, it should be known that majority of the cryptographic protocols are old and hence the existing researchers chooses either to enhance the older or to develop a new one. The most recent study presented by Shankar et al. [26] has used public-key cryptography to secure communication system in WSN taking the case study of the healthcare sector. The technique uses Elliptical Curve Cryptography to perform mutual authentication. Munivel and Ajit [27] have used public key infrastructure to perform encryption and key management. Soosahabi et al. [28] have introduced a probabilistic cryptographic approach to understanding the state of transmission (to be a harmless or harmful state). Al-Haija et al. [29] have presented a cryptographic technique that uses RSA for the smaller scale of the sensor network. The approach was also experimental-based, and the outcome was evaluated with respect to time. Kodali and Sarma [30] have used Elliptical Curve Cryptography along with Diffie-Hellman key exchange protocol in WSN. Xu and Dang [31] have presented a technique that is resistive against Denial-of-Service attacks in WSN. The method uses joint implementation of Elliptical Curve Cryptography with a digital signature. Yan and Shu [32] have used AES protocol to obtain energy efficient cryptographic operation in wireless body area network. The author has developed an analytical model, and its outcome was studied on energy only. Jeon et al. [33] have presented a scheme which uses free surrounding resources to incorporate encryption policy. The technique was claimed to offer lowered complexity and better modulation method. Huang et al. [34] have presented a simple encryption scheme to safeguarding data aggregation in a sensor network. Liu et al. [35] have illustrated a unique signature scheme based on the identity of nodes using experimental approach.

C. Optimization-based Techniques

Usage of optimization has seen increasing attention from the year 2010 onwards. Most recently, there are multiple techniques of optimization adopted by the researcher. Narad and Chavan [36] have used a neural network to formulate a new authentication mechanism in WSN. The author has used Shamir Secret Sharing to perform encryption. Karapistol and Economides [37] have adopted game theory to address jamming attack in WSN. The attack environment was modeling using Bayesian Stackelberg games where the outcome was studied on probability and utility factor. The same authors have presented a different work in the same year [38] for performing anomaly detection. Kumar et al. [39] have jointly used the neural network and game theory to formulate a novel defense mechanism in WSN. Alrajeh et al. [40] have presented a bio-inspired algorithm to maintain secured communication in WSN. The technique uses Ant colony optimization, and its outcome was testified using efficiency of data forwarding and packet loss. Branch et al. [41] have presented a technique of simple optimization of in-network to perform outlier detection. Another implementation of game theory was carried out by Ding et al. [42]. The method

establishes the relationship between the resource utilization and vulnerable situation. The technique is evaluated on probability of selfish node discovery. Ramesh et al. [43] have presented a study that uses the neural network to resist DoS attack in WSN. The study outcome was evaluated using computational energy required to perform ciphering, memory consumption and execution time. Marmol and Perez [44] have discussed a new bio-inspired algorithm. The technique uses both reputation and

trust factor using ant colony optimization. Estiri and Khademzadeh [45] have adopted game theory to perform intrusion detection. The method also assists in formulating defense strategies.

Hence, it can be seen that there are various techniques that call for inclusion of multiple methods for incorporating security features in WSN. The scale of the effectiveness of all the above-mentioned methods is tabulated in Table.1.

TABLE I. SUMMARIZATION OF RECENT TECHNIQUES OF SECURITY IN WSN

	Authors	Techniques	Advantage	Limitation
Routing-Based Security technique	Das et al. [16]	MAC-based authentication	-Resistive against Wormhole attack -energy efficient	-No discussion of computational complexity.
	Nandu [17]	Re-programming	-Faster Authentication	-No Benchmarking --No discussion of computational complexity.
	Henze et al. [18]	Re-programming, AES, RSA	Robust security over cloud	-Highly dependent on library -Storage / transmission overhead
	Chen [19]	Enhanced LEACH	-Energy Efficient	-Applicable to small networks -Not benchmarked with secure routing techniques.
	Menaria et al. [20]	Position Identification of compromised node, spanning tree	-Energy Efficient -Better communication	-No Benchmarking -No discussion of computational complexity.
	Mengyao et al. [21]	Ring-based clustering, trust, ant colony optimization	-Energy Efficient	-No Effective Benchmarking -No discussion of computational complexity.
	Tang et al. [22]	Cost-Aware Secure Routing	-Energy Efficient -Resistive against trace back attacks in routing	-No Effective Benchmarking -No discussion of computational complexity.
	Masdari et al. [23]	Evaluation of Secure-LEACH	-Supportability of extensive cryptographic mechanism -Supports broadcast authentication,	-less supportability of message freshness, and pairwise authentication except few of them
	Ferng and Rachmarini [24]	Grid-based, simple key management	-Energy Efficient	-Not applicable to dynamic networks -Not resilient against any major lethal threats in WSN
	Obaidat et al. [25]	Dynamic energy, heterogeneous WSN	-Energy Efficient	-Overhead discussion is not made
Cryptographic-Based Security technique	Shankar et al. [26]	Elliptical Curve Cryptography	-Smaller Key Size	-No Effective Benchmarking -No discussion of computational complexity.
	Munivel and Ajit [27]	Micro-Public Key Infrastructure	-Energy Efficient	-No Effective Benchmarking -No discussion of computational complexity.
	Soosahabi et al. [28]	Probability theory,	-Effective against statistical attacks	-No Effective Benchmarking -No discussion of computational complexity -Consumes Resources
	Al-Haija et al. [29]	RSA	Robust Encryption for smaller scale network.	-Not a lightweight encryption -Not applicable for large scale network
	Kodali and Sarma [30]	ECC, Diffie-Hellman	-Robust Encryption -Resilient against Brute-force attack	-Not Energy efficient
	Xu and Dang [31]	ECC, Digital Signature	-41% of energy minimization -resistive against DoS attack	-No Effective Benchmarking -No discussion of computational complexity
	Yan and Shu [32]	AES	-Energy Efficient	-No Effective Benchmarking -No discussion of computational complexity
	Jeon et al. [33]	Encrypted fusion rules	-Lowered error probability	-No Effective Benchmarking -No discussion of computational complexity
	Huang et al. [34]	Key-verification	-Simple authentication of key.	-Less Effective key management. -Lead to communication overhead -less Security strength
	Liu et al. [35]	Signature-based	-Supports both online and offline	-No Effective Benchmarking

			verification. -Suitable for large area.	-No discussion of computational complexity -Signature generation and validation doesn't conform to backward secrecy.
Optimization-Based Security technique	Narad and Chavan [36]	Neural network, Shamir Secret Sharing	-maintains message integrity	-No Effective Benchmarking -No discussion of computational complexity
	Karapistol and Economides [37]	Game theory	-resistive against jamming attack.	-No Effective Benchmarking -No discussion of computational complexity
	Karapistol and Economides [38]	Anomaly detection using ruleset	-Higher accuracy of attack detection	-Communication performance nor evaluated.
	Kumar et al. [39]	Game theory, neural Network	-60% accuracy in attack detection	-Theoretically sound but no evidence of practical implementation.
	Alrajeh et al. [40]	Ant colony optimization	-need less time to forward data -better communication performance	-No evidence of scalability -No Effective benchmarking
	Branch et al. [41]	Non-parametric optimization, outlier detection	-energy efficient	-Not secure against passive attacks
	Ding et al. [42]	Game theory	Detection of intrusion based on resource consumption	-The model doesn't have validation in uncertainty. -Low scope of utility function. -Less practical implementation
	Ramesh et al. [43]	Neural Network, symmetric key algorithm	-Resistive against DoS attack	-Leads to excessive iteration -Less effective classification of intrusion..
	Arnol and Perez [44]	Ant colony optimization, trust, reputation	-energy efficient	-Not compliant of space complexity
	Estiri and Khademzadeh [45]	Game theory	-better intrusion identification	-No evidence of scalability -No Effective benchmarking

IV. RESEARCH GAP

After reviewing the existing techniques and solutions offered by the researchers till date, it can be just inferred that existing techniques have both potentials and pitfalls. The potential beneficial factor of the existing techniques discussed in previous sections only shows they are more inclined to accomplish energy efficiency while implementing security protocols. However, sometimes the energy efficiency is obtained at the cost of overlooking security aspects in full dimensional of vulnerability. It was also observed that existing mechanism are too much symptomatic in nature on attacks. It will mean that solutions design for mitigating DoS attack is not even capable of identifying other forms of intrusion. It should be known that existing mechanism of routing and clustering in wireless sensor network involves mobility and dynamic topologies too which calls for multiple attacks with an uncertain or unpredictable pattern of intrusion. Hence, existing mechanism is not able to cater up to the first line of defense itself purely, which is just about identifying uncertain forms of attacks or hideous security breach. Moreover, existing solutions don't offer full fledge compliance towards integrity, privacy, anonymity, confidentiality, non-repudiation, availability, etc.

Moreover, we have seen that there are few benchmarked works published in 2010-2016 about security in wireless sensor network. We also find that studies are more on cryptographic usage ignoring the fact that a sensor can only process 48 kilobytes of physical memory. Existing optimization techniques are good attempt, but they are less practical in real-world notes. Optimization techniques based game theory sound good in theory, but its formulation cannot be judged to be potential until and unless there is any benchmarked work or designed using experimental study of large-scale sensor

deployment. The significant research gaps found after reviewing the existing system are briefed as follows:

- Ignorance to Explore Best Secure Route: The existing standard secure routing protocols (e.g. Sec LEACH, Sec Rout, and HEED), etc. are developed over a similar network, which has less supportability of the multipath routing technique. It was also explored that existing security technique also could not address the research gap among reliable routing with energy and quality of service. In this regards, multipath routing is the only way to ensure a better level of tolerance against any critical fault owing to any physical attacks in wireless sensor network. It is also known that usage of multipath routing supports heterogeneous networks as well as it also invites a massive problem owing to the presence of Duplicated Routing Data (DRD). An adversary can compromise such DRD using malicious eavesdropping and can easily insert its malicious code that is its replicated version. Such codes can easily go without any validation. The consequences of intrusions through multipath propagation in wireless sensor network will be quite collateral in nature. The existing techniques are not found to solve such issues most recently.
- Narrowed Scenario of implementation: Normally the wireless sensor network is always studied with respect to static nodes and mobility factor is concern only with the sink nodes. But there are many possibilities of upcoming applications where the sensors may go mobile. In such scenario, if a node in heterogeneous wireless sensor network go on mobile than there may be massive energy consumption owing to radio transmission. The scenario may turn more worst if there are selfish nodes as they will not be ready for spending

more energy in assisting other data packet forwarding. Such forms of scenario will easily welcome all sorts of routing attacks, replay attack, wormhole attack etc. Availability of an intermediate node is extremely important in heterogeneous network as compared to homogeneous networks. The allowance of such intrusion will further lead to a unstabilized network that may highly degrade the communication performance along with stealing of data. Hence, even after the knowledge of mobility factor, IoT, ubiquitous computing, the researchers have not considered broader spectrum of recent problems and its implementation requirements.

- Resiliency against Potential Threats: In the area of healthcare or nuclear plant monitoring system, if the transmission line is not secured, it may cost the life of somebody. There are various hazardous and adverse condition of network deployment that results in unauthorized access to private details of sensor nodes. Although, with an aid of the model description in Section 2.2, we have seen that usage of cryptography can render further security. But the fact still remains unsolved are that “*Is the system resistive enough against brute force attack?*” The question is quite difficult to crack as brute force attack may lead to decryption of ciphered information using malicious code with infinite computing resources, which is highly possible. We have already seen that existing system are more inclined to achieve energy efficiency and there is a bit of imbalance between energy efficiency and security potential.

V. FUTURE WORK

Based on the research gap discussion in previous section, our future direction of the work will be focused towards cost effective security protocols in wireless sensor network. Our future work will be towards designing a mathematical modelling for analyzing an availability of best secure route. The work to be carried out in this regards are as follow:

- Adversarial Model: The study considers essential physical attack (e.g. node capture attack) as an insider attack model. However, novelty is incorporated in this adversarial modelling, where we consider the facts of resource expenses too. Hence, adversarial node with less resource will perform different action compared to an adversarial node with high resource. We will also model for uncertain behaviour for the adversarial node.

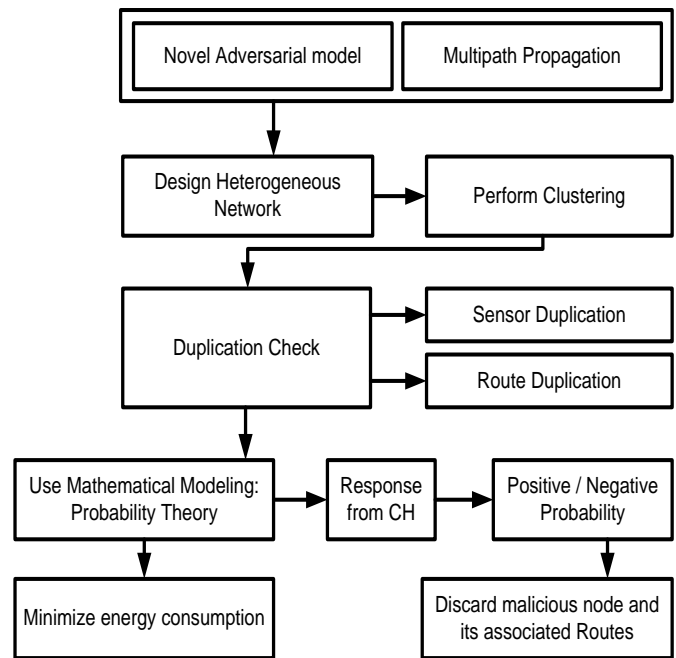


Fig. 2. Tentative System Architecture

- System Design: A simulation study of heterogeneous nodes with clustering process will be carried considering clusterhead and sensor node. The node distribution is carried out using random process. The novelty in the clustering process is that clustered area is not circular like in existing studies; they are more asymmetric in shape, which normally happens in real-time implementation. The system checks for duplication for both sensors and routes over multipath propagation. The design of the mathematical modelling will be entirely carried out by probability model. The system also checks for the positive / negative probability of a node / route being a malicious. Hence, a novel distributed election mechanism will be developed that can identify the malicious nodes. The mechanism will be initiated by random selection of neighbor nodes located near to monitored node. This process is carried out by all the clusterhead, which after computation shares its response using pairwise key. When maximum response is found to be negative for monitored node, it will be considered as malicious and hence will be discarded from the routing process. However, the process is quite bulky and may consume more energy, therefore, we will not be using conventional radio-

energy model and will chose to reformulate energy-based modelling where energy will be computed for performing security operations and not for normal routing and clustering operations. The tentative architecture is showcased in Fig.2

- Benefit: The benefits of the proposed system are as follows: i) the system doesn't use any forms of cryptography, which ensures that there is no complex computation process involved within a node. ii) The system uses probability theory for formulating a model in order to identify the suspicious sensors and routes along with less usage of energy of a node. iii) The identification of the response generated from CH is scaled positive / negative along with check for reliability that makes the detection and avoidance of the attackers highly precise.

VI. CONCLUSION

A better version of security protocols calls for inclusion of privacy, integrity, non-repudiation, anonymity, confidentiality factors. However, till date there is no single security algorithm in wireless sensor network which has maintained all of them in its security deployment among the nodes. One reason for it is basically the types of the nodes which are less capable of computational processing and other reason is the flawed design of security protocols. It was also seen that adoption of cryptographic technique is much high compared to non-cryptographic techniques. There is a benefit of using cryptographic technique which is robust authentication, but cryptographic implementation is generally recursive in nature and calls for multiple rounds of operation in order to perform encryption. Hence a good amount of memory and resource is required for this, which are the major pitfalls. Usage of non-cryptographic techniques is carried out mainly using probability theory and statistics. However, success rate of such usage depends upon the mathematical modelling. Last 6 years has found few implementations in the form of mathematical modelling. Therefore, our future work is dedicated to develop such a mechanism that will use a mathematical modelling and whose entire operation will permit identification of uncertain nodes and is capable to address the most difficult forms of adversaries in sensor network.

REFERENCES

- [1] S.Kaur, and M. Kaur, "Improvement In MAODV Protocol Using Location Based Routing Protocol", In MATEC Web of Conferences, vol. 57. EDP Sciences, 2016.
- [2] J. Ma, W.Lou, and X-Y. Li, "Contiguous link scheduling for data aggregation in wireless sensor networks", Parallel and Distributed Systems, IEEE Transactions, Vol.25, No. 7, pp.1691-1701, 2014.
- [3] A.G. Khan, A. Rahman, and N. Bisht, "Classification of hierarchical based routing protocols for wireless sensor networks", International journal of innovations in engineering and technology, pp.2319-1058, 2013
- [4] S. Anbumalar, S. Prabhadevi, "A Survey On Routing Issues And Routing Protocols In WirelessSensor Networks", International Journal Of Engineering And Computer Science, Vol. 4, Issue. 6, pp. 12927-12931, 2015
- [5] F. Bouabdallah, N. Bouabdallah, and R. Boutaba, "Load-balanced routing scheme for energy-efficient wireless sensor networks", In Global Telecommunications Conference, IEEE Globecom, pp. 1-6, 2008.
- [6] M. Chitnis, P. Pagano, G. Lipari, and Y. Liang, "A survey on Bandwidth resource Allocation and Scheduling in wireless sensor networks", In Network-Based Information Systems, 2009. NBIS'09. International Conference, pp. 121-128, 2009.
- [7] Q. Wang, and T. Zhang, "A survey on security in wireless sensor networks. Security in RFID and Sensor Networks",2009
- [8] F. Chen, F., L. Guo, and C. Chen, "A Survey on Energy Management in the Wireless Sensor networks", IERI Procedia, Vol. 3, pp.60-66,2012.
- [9] R. Selvam, and A. Senthilkumar, "Cryptography based secure multipath routing protocols in wireless sensor network: a survey", In Electronics and Communication Systems (ICECS), International Conference,pp. 1-5, 2014.
- [10] A.M. E-Semary, and M.M.A. Azim, "A two-tier energy-efficient secure routing protocol for Wireless Sensor Networks", In Information Assurance and Security (IAS), 2011 7th International Conference, pp. 331-337, 2011.
- [11] Y. Arfat, and R.A. Shaikh, "A Survey on Secure Routing Protocols in Wireless Sensor Networks",2016.
- [12] S. Renubala, and K. S. Dhanalakshmi, "Trust based secure routing protocol using fuzzy logic in wireless sensor networks", In Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, pp. 1-5, 2014.
- [13] S. Athmani, D.E. Boubiche, and A. Bilami, "Hierarchical energy efficient intrusion detection system for black hole attacks in WSNs", In Computer and Information Technology (WCCIT), 2013 WorldCongress, pp. 1-5, 2013.
- [14] V. Kumar, A. Jain, and P.N. Barwal, "Wireless sensor networks: security issues", challenges and solutions. International Journal of Information & Computation Technology, ISSN, pp.0974-2239,2014
- [15] Q. Yang, X. Zhu, H. Fu, and X. Che, "Survey of security technologies on wireless sensor networks", Journal of Sensors, 2015
- [16] A.K.Das, R.Chaki, and K.N.Dey, "Secure energy efficient routing protocol for wireless sensor network", Foundations of Computing and Decision Sciences, Vol. 41, No. 1, pp.3-27,2016.
- [17] P.Nandu, and N. Shekoker, "An Enhanced Authentication Mechanism to Secure Re-programming in WSN", Procedia Computer Science, Vol. 45, pp.397-406, 2015.
- [18] M. Henze, S. Bereda, R. Hummen, and K. Wehrle, "SCSlib: Transparently accessing protected sensor data in the cloud. Procedia Computer Science, Vol.37, pp.370-375, 2014.
- [19] L. Chen, "An Improved Secure Routing Protocol Based on Clustering for Wireless Sensor Networks", InMechatronics and Automatic Control Systems, pp. 995-1001, 2014 Publishing.
- [20] V.K. Menaria, D. Soni, A. Nagaraju,S.C.Jain, "Secure and energy efficient routing algorithm for wireless sensor networks", InContemporary Computing and Informatics (IC3I), International Conference, pp. 118-123, 2014
- [21] L. Mengyao, Y. Zhang, and X. Li, "Ring-based security energy-efficient routing protocol for WSN", In Control and Decision Conference, The 26th Chinese, pp. 1892-1897, 2014
- [22] D. Tang, T.Li, J. Ren, and J. Wu, "Cost-Aware Secure Routing (CASER) Protocol Design for Wireless Sensor Networks", Parallel and Distributed Systems, IEEE Transactions, Vol. 26(4), pp.960-973, 2015
- [23] M. Masdari, S.M. Bazarchi, and M. Bidaki, "Analysis of secure LEACH-based clustering protocols in wireless sensor networks", Journal of Network and Computer Applications, 36(4), pp.1243-1260, 2013.
- [24] H.W. Ferng and D. Rachmarini, "A secure routing protocol for wireless sensor networks with consideration of energy efficiency", In Network Operations and Management Symposium (NOMS), IEEE, pp. 105-112, 2012
- [25] M.S. Obaidat, S.K. Dhurandher, D. Gupta, N. Gupta, and A. Asthana, "DEESR: dynamic energy efficient and secure routing protocol for wireless sensor networks in urban environments",Journal of Information Processing Systems, Vol.6(3), pp.269-294, 2010
- [26] S.K. Shankar, A.S. Tomar, and G.K. Tak, "Secure Medical Data Transmission by Using ECC with Mutual Authentication in WSNs", Procedia Computer Science, Vol. 70, pp.455-461,2015

- [27] E. Munivel, and G. M. Ajit, "Efficient public key infrastructure implementation in wireless sensor networks", In *Wireless Communication and Sensor Computing, ICWCSC, International Conference*, pp. 1-6. IEEE, 2010.
- [28] R. Soosahabi, Naraghi-Pour, D. Perkins, and M.A. Bayoumi, "Optimal probabilistic encryption for secure detection in wireless sensor networks. *Information Forensics and Security*", IEEE Transactions on, 9(3), pp.375, 2014
- [29] Q.A. A-Hajja, A. Tarayrah, H. A-Qadeeb, and A. Al-Lwaimi, "A tiny RSA cryptosystem based on Arduino microcontroller useful for small scale networks. *Procedia Computer Science*, Vol.34,pp.639-646,2014
- [30] R.K. Kodali, and N.N. Sarma, "Energy efficient ECC encryption using ECDH", In *Emerging Research in Electronics, Computer Science and Technology Springer*, pp. 471-478, 2014
- [31] J. Xu, and L. Dang, "Multi-User Broadcast Authentication Protocol in Wireless Sensor Networks against DoS Attack", *Open Cybernetics & Systemics Journal*, Vol.8, pp.944-950, 2014
- [32] Y. Yan, T. Shu, "Energy-efficient In-network encryption/decryption for wireless body area sensor networks", In *Global Communications Conference (GLOBECOM)*, IEEE 2014, pp. 2442-2447, 2014
- [33] H. Jeon, J. Choi, S. W. McLaughlin, and J. Ha, "Channel aware encryption and decision fusion for wireless sensor networks", *Information Forensics and Security, IEEE Transactions*, Vol.8, No. 4, pp.619-625, 2013
- [34] S.I. Huang, S. Shieh, and J.D. Tygar, "Secure encrypted-data aggregation for wireless sensor networks. *Wireless Networks*, 16(4), pp.915-927, 2010
- [35] J.K. Liu, J. Baek, J. Zhou, Y. Yang, and J.W. Wong, "Efficient online/offline identity-based signature for wireless sensor network. *International Journal of Information Security*, vol.9(4), pp.287-296, 2010
- [36] S. Narad and P. Chavan, "Cascade Forward Back-propagation Neural Network Based Group Authentication Using (n, n) Secret Sharing Scheme", *Procedia Computer Science*, Vol. 78, pp.185-191,2016.
- [37] E. Karapistoli, and A.A. Economides, "Defending jamming attacks in wireless sensor networks using stackelberg monitoring strategies", In *Communications in China (ICCC)*, pp. 161-165, 2014
- [38] E. Karapistoli, and A.A. Economides, "ADLU: a novel anomaly detection and location-attribution algorithm for UWB wireless sensor networks", *EURASIP Journal on Information Security*, pp.1-12, 2014.
- [39] E.S. Kumar, S.M. Kusuma, and B.V. Kumar, "An intelligent defense mechanism for security in wireless sensor networks", In *Communications and Signal Processing (ICCSP)*, pp. 275-279, 2014.
- [40] N.A. Alrajeh, M.S. Alabed, and m.S. Elwahiby, "Secure ant-based routing protocol for wireless sensor network", *International Journal of Distributed Sensor Networks*, 2013.
- [41] J.W. Branch, C.Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks", *Knowledge and information systems*, Vol.34(1), pp.23-54, 2013
- [42] Y. Ding, X.W. Zhou, Z.M. Cheng, and F.H. Lin, "A security differential game model for sensor networks in context of the internet of things. *Wireless personal communications*, Vol.72(1), pp.375-388, 2013.
- [43] M.V. Ramesh, A.B. Raj, and T. Hemalatha, "Wireless Sensor Network Security:Real-Time Detection and Prevention of Attacks. In *Computational Intelligence and Communication Networks (CICN)*, 2012 Fourth International Conference, pp. 783-787, 2012
- [44] F.G. Mármol and G.M. Pérez, "Providing trust in wireless sensor networks using a bio-inspired technique. *Telecommunication systems*, Vol. 46(2), pp.163-180, 2011
- [45] M. Estiri, and A. Khademzadeh, "A game-theoretical model for intrusion detection in wireless sensor networks. In *Electrical and Computer Engineering (CCECE)*, 23rd Canadian Conference, pp. 1-5, 2010

Estimation Medicine for Diseases System to Support Medical Diagnosis by Expert System

Noor T. Mahmood¹

Computer Science Department
Al- Mustansiriyah University
Baghdad, Iraq

Abstract—Researches confirmed that 70 thousand cases of death, which happen yearly in the world, were because of the misprescribing of the drug itself or its dose (overdose or lower dose). Choosing the wrong alternative drug inspired the professionals in the healthcare field to the importance of assigning the best technologies to decrease the percentages of the therapeutic methods in giving the drug to prevent mistakes in prescribing the suitable drug. A system based on Rete Algorithm is proposed where the best-chosen medicine is offered through the suggested system. Selection of Estimation Medicine for Diseases (EMD) System is introduced where the diagnosis is made basically according to the symptoms and the medical history of the patient. This research aims to acquire a good model using this algorithm to obtain more accurate choices of medicine. The system (EMD) is tested by the doctors in Iraqi hospitals and it has been found that there is no other systems that can be compared to EMD system. The accuracy of estimating the appropriate medicine for heart diseases is approximately (87.26%).

Keywords—*Diagnosis; Disease; Medicine; Rete Algorithm; Expert System; Intelligent System*

I. INTRODUCTION

Many of the patients have some troubles when they visit the doctor for identifying the suitable treatment for them. After the diagnosis of the disease have been made by the doctor, another problem is appeared, the drug might be written in a nonspecific way. Since the choice of the drug depends on the disease history of the patient and whether he had any chronic diseases, consequently careful assessment should be done as the drug is given.

The Estimation Medicine for Diseases (EMD) is an intelligent system (Expert System) to identify the suitable treatment to the patient by knowing the disease or the case and the disease history of the patient and the symptoms that the patient is suffering. This process may prevent the mistakes that effects the wrong prescription of the drug which might not cause a significant health damage to many thousands of the patients only but also might lead to death to some of them.

What's done in the proposed approach is a process to enter the information into the system which identifies the correct diagnosis and obtains the accurate diagnosis of patient's diseases as quickly as possible with less cost. In our system selection, the best medicine is proposed for the patient and the doctor. It helps in determining medicine by disease or conditions(symptoms) according to some concepts, rules, and algorithms which are expressed in the next sections

As a comparative view between previous researches and that one; previous researches usually choose two or three side effects and then contrast them to reach drugs with less interference and are less risky for the patient. On the other hand, this research is tested by an expertise doctor in diagnosing the disease depending on the rules of choosing the right drug for the intended patient. Also to realize the side effects of each drug or medication in addition to the previous symptoms taken from the patient prohibits causing pharmacological interference. The system picks the selected medicine that takes less error ratio and higher priority of accuracy for the preferring drugs that have fewer side effects and less conflicts with other medicines.

Collecting information about both diseases and medicines along with the side effects of each drug is made by the assistance of expert doctors from Iraqi hospitals and from reliable international resources in [1][2][3][4].

II. LITERATURE SURVEY

The side effects of the prescribed medications are a common occurrence in the medical history of patients. Jenna, M. R., et. al [5] presented the opportunity to identify new side effects efficiently through electronic healthcare databases. A confirmation of concept method is suggested. It learns common associations and uses this knowledge to automatically refine exposure-outcome associations (i.e. side effect signals). A novel measure is determined. It is named the confounding-adjusted risk value, an accurate absolute risk value of a patient. They show that signals filtering might be possible at a patient level according to association rules acquired by taking into account patients' medical histories.

Chen, Jie, et. al [6] developed an association classification algorithm in which rules discovered can be used to alert medical practitioners when prescribing drugs. They propose two kinds of probability trees that can present clearly the risk of specific adverse drug reactions to prescribers.

Chen, Yu, et. al [7] used association rule mining approach to derive possible side effects due to exposure to multiple drugs at different durations of the pregnancy. They derived sequential temporal rules to discover new information that would not be detected by the traditional analysis method that is currently used by pharmacists.

The optimizing drug dose is a major factor in improving the sustainability of health care. Daughton, C. G., and Ruhoy, I. S. [8] presents the first critical examination of the multi-faceted role of drug dose in reducing the ambient levels of active

pharmaceutical ingredients (APIs) in the environment. Personalized adjustment of drug dose holds the potential for enhancing therapeutic outcomes while simultaneously lowering the incidence of adverse drug events and in lowering patient healthcare cost.

III. EXPERT SYSTEM

An expert system is an application that tries to react like a human expert on a special subject area. This system is used to advise non-experts in situations where a human expert is unavailable [9].

An expert system consists of three parts [9]: (see Figure 1)

A user interface - that accepts a non-expert user and it asks questions of the expert system, and is conferred with advice. The interface should be a simple and easy to use [9].

A knowledge base - It is a group of facts and rules. It is built with information that is supplied by human experts [9].

An inference engine - this acts as a search engine, inspecting the knowledge base for a particular arrangement that matches the user's query [9].

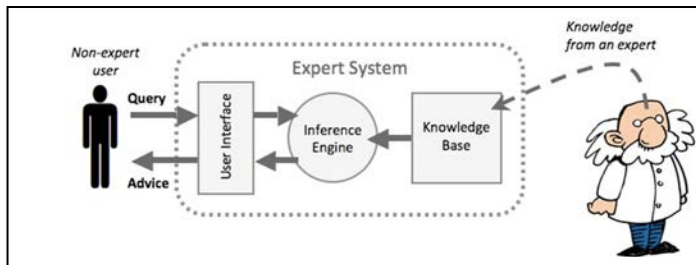


Fig. 1. An Expert system works in block diagram [9]

An example of expert systems is diseased diagnosis (the input would contain medical history information, the symptoms of the patient. Those might be the query, and the answer is a diagnose for the patient's mental or physical pain). However, Nowadays, "conventional" computer programs may achieve some of the typical expert system functionalities. When using, expert systems problems are solved using heuristics or approximate methods which, unlike traditional applications, are not guaranteed to present an optimal solution [9].

IV. GENERAL DESCRIPTION FOR ESTIMATING THE BEST MEDICINE SYSTEM

This section describes the proposed Estimation Medicine for Diseases (EMD) System framework, algorithms, and the applied rules. Figure (2) illustrates a block diagram for selection medicine for each patient's disease.

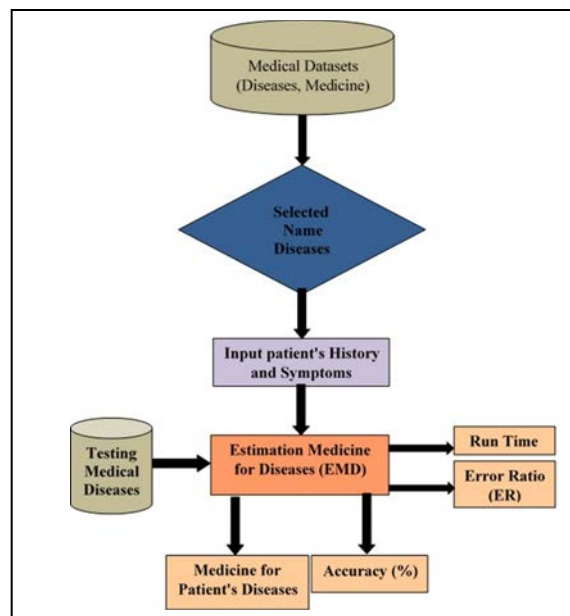


Fig. 2. Block Diagram Estimation Medicine for each Patient's Disease

The method used in the system makes use of dataset for heart diseases along with medicine dataset for medical data. The obtained medical data are utilized in the formation of the expert system. The system utilizes in a convenient way the concepts and algorithms introduced in the following subsections.

A. Reliable rules for safer drug use

This section explains how the doctor selects the best medicine for each patient's disease for conditions by some rules based on safer drug use; these rules are expressed in [10].

Those rules were compiled for more dependable drug use (or nonuse). They were listed particularly from the World Health Organization's General Prescribing Principles for the Elderly. Nevertheless, these rules might be applied to all ages. Doctors and patients involved in drug therapy should know them. The guiding principle is to use those rules as guidelines for patient treatment according to symptoms and patient's medical history. Among those rules are [10]:

- Additional drugs used may make the earlier drugs much more dangerous [10].
- The primary doctor ought to coordinate patient's care and drug use [10].
- The patient should make sure that drug therapy is needed [10].

- Important, believable, and adverse effects of the drug in addition to plausible drugs interactions should be foreseeable in advance [10].
- Any adverse drug reactions patient may have should be taken into consideration early [10].

B. Recognize-Act Cycle

Since the production system is a formulation, that the expert system’s inference engine can process efficiently, they can be used to generate new information. The basic recognize-act cycle consists of four steps [11]:

- 1) Recognize (get) matches.
- 2) Choose a match.
- 3) Act (perform the selected production).
- 4) Return to step (1).

It matches the premise patterns of the rules against elements in the working memory. If there is more than one rule that can be applied, use a conflict resolution strategy to choose one to apply. Early production systems spent over 90% of their time doing pattern matching, but there is now a solution to this efficiency problem which is the Rete algorithm [11].

C. Rete Algorithm

The traditional approach for recognize-act cycle matches all Elements (E) in working memory along with all premises (P) in all rules (R). It then checks E×P×R possibilities in each cycle. Nevertheless, the rules have structural similarity, since parts of their conditions are common. Where the application of any similar rule changes the working memory slightly (temporal redundancy) [11].

Rete Algorithm uses these facts to improve efficiency (‘rete’ is Latin for ‘net’). First, the initial conflict set is generated through the matching algorithm by connecting the network for all condition parts. After that, only the changed elements in working memory are fed into the network to distinguish the changes in the conflict set [11].

D. Estimation Medicine for Diseases (EMD) algorithm

The above concepts and foresight are consolidated into a single unified algorithm called Estimation Medicine for Diseases (EMD) which performs the main system functionality.

Algorithm: Estimation Medicine for Diseases (EMD)

Input: Diseases Table (D), Medicine Table (M), Disease Name for Doctor Diagnosis (DN).

Output: Medicines Name (MsN), Error Ratio(ER).

Begin:

- Input patient's History and Symptoms (PHS).
- For each (D)That has name (DN)
- Find Diseases Symptoms (DS)
- Find Similar Symptoms between(PHs) and (DS) and put it on new array (SS)
- (If SS is found in M Then MsN()→ Count +1)//Compare (SS) with (M) for (DN)

- Sort MsN()
- Select the minimum Error Ratio
- If the value of MsN is equal, then select the high priority from (M)

End

V. MEDICAL DATASET (DISEASES, MEDICINE)

The data was collected from an Iraqi hospital. Patients' files and information were extracted from the Statistics Division at Al-Kindi Teaching Hospital. The data for each disease consists of medical history, symptoms, laboratory analysis, etc. Those have been entered into the database. Each disease has ten medicines that were fed into the system. (See Table I). In this table, each column represents a single disease with its well-known ten medicines. The following are some of the heart diseases:

- High Blood Pressure (I10).
- Angina pectoris (I20).
- Myocardial Infarction (I21).
- Atrial Fibrillation (I48).
- Heart Failure (I50).

Note that I10, I20, I21, I48, and I50 represent the statistical identifiers of the diseases.

TABLE I. DISEASES AND MEDICINES

ID	I10(HighBlood Pressure)	I20(Angina pectoris)	I21(Myocardial Infarction)	I48(Atrial Fibrillation)	I50(Heart Failure)
1	Diazoxide	Glyceryl_trinitrate	Atenolol	Digoxin	Captopril
2	Hydralazine	Aspirin	Metoprolol	Dronedarone	Cilazapril
3	Sodium Nitroprusside	Clopidogrel	Acebutolol	Warfarin	Enalapril maleate
4	Methyldopa	Prasugrel	Propranolol Hydrochloride	Acenocoumarol	Lisinopril
5	Captopril	Heparin	Labetalol Hydrochloride	Phenindione	Valsartan
6	Lisinopril	Streptokinase	Pindolol	Diltiazem	Telmisartan
7	Moexipril Hydrochloride	Urokinase	Sotalol Hydrochloride	Verapamil	Olmesartan Medoxomil
8	Fosinopril Sodium	Tenecteplase	Timolol maleate	Propranolol Hydrochloride	Losartan potassium
9	Enalapril	Propranolol Hydrochloride	Carvedilol	Acebutolol	Irbesartan
10	Cilazapril	Atenolol	Bisoprolol	Atenolol	Eprosartan

A large amount of data is required to train the system since poor of data does not give high accuracy in diagnosis. Each represented disease has its specific table. Each table consists of some attributes. These attributes are different from one table to another. All data are taken from the patient's file, which was read by doctors, who helped me and taught me how to read and understand each patient's file.

Training datasets consist of two groups: the first is made of five tables where a table is specified for each medical case in

which the patient data are taken from the patient files. These are the symptoms and patient medical history. While the second category contains five tables, where a table is designated for each disease. This table has ten drugs commonly identified with that disease, yet each drug has a set of side effects. The name of the medicine is the scientific name as it is an international standard.

Most of the researches are different from this one because in these researches the acquisition of data is made from the internet and they are almost based on one disease. The results of these researches indicate that the disease may exist or not and they don't give the accurate medicine for the patient and as a result they cannot tell all the symptoms that the patient suffers from and hence they cannot prevent the misprescribing of drugs"

VI. TYPICAL SYSTEM OPERATIONS

One user interface of the system explained in this section is shown in Figure (3). This window is called as one of some sub-windows. Learning in the system depends on the data stored in the system (training dataset) and data tested by the system.

As an example, the system works to diagnose that the patient has heart diseases by using functional (EMS) strategies. The system deals with heart disease, which is diagnosed as the patient's disease. Doctor selects any heart disease, as shown in Figure (3) where the patient is suffering from one of them, but it is unknown for his doctor at this moment.

A sample medical history data is (patient's age: more than fifty years (age), Increase of blood cholesterol (cholesterol), Hypertension (HT), Respiration Rate (RespRate) and Blood Pulse High (BPH)). The symptoms are (Shortness of Breath, Dizziness, Weakness, Rapid heartbeats, Palpitation, Headache and General malaise).

The doctor in the hospital diagnosis that patient is suffering from High Blood Pressure and Myocardial Infarction. When his medical history and symptoms are entered into the system, the result would be selecting the best medicine for diagnosis to High Blood Pressure (I10) and Myocardial Infarction (I21). Hence, the doctor has selected medicine by the system (EMS) depending on all information and side effects of that medicine, symptoms, and history of this patient.



Fig. 3. A heart disease diagnosis and selected medicine by (EMD)

VII. EXPERIMENTAL RESULTS

This section illustrates the experimental results of the proposed approach described in the previous sections.

A. Settings

The experimental results also include Rete Algorithm for Medicine. The results are organized into a table for heart diseases. Brief descriptions and the diagnosis were given by the doctors and also by the proposed system—Estimation Medicine for Diseases (EMD) that offers selecting the best medicine for a given disease.

These approaches are running on a laptop (running Core i5 CPU@ 1.7MHz, RAM 4GHz on Windows 7 Home Basic), SQL Server 2012 for building Database and Visual Basic.NET 2012,2013 (VB.NET) environment as the programming language used in the implementation.

B. Results

The proposed system deals with many types of heart diseases. System experiments on the applied fed data. Each stored table is used to test different disease cases. The system was configured over five categories representing the selected heart diseases. Table I demonstrate this training dataset.

The system has been tested by entering symptoms and medical history of each patient. Data are entered into the built structure, then calculations of the probability of occurrences of the disease for each patient are made. This result is the likelihood of disease for each patient along with the selected medicine. The doctors can choose two of the top up possibilities of these medicines (i.e. the higher priority among them). The performance of the system is expressed in Table II. Where we can see twenty-two patients for heart diseases.

After that accuracy is calculated by (true (cases) / (true (cases) + false (cases))) to get the results for the system model of (EMD). The performance of the EMD can be seen for each disease. Hence, the accuracy rate of the given model of (EMD) is approximately (87.26%) in heart diseases. It can be deduced that the rate of accuracy of the system (EMD) is the best average results. These results are shown in table II while the running time is presented in Table III.

TABLE II. EMD PERFORMANCE FOR THE SELECTED MEDICINE

Disease code	Disease name	Number of patients used in testing	The selected medicine for disease	
			Accuracy (%)	Error Ratio (%)
I10	High Blood Pressure	22	81.8%	18.2%
I20	Angina pectoris	22	86.36%	13.64%
I21	Myocardial Infarction	22	90.9%	9.1%
I48	Atrial Fibrillation	22	90.9%	9.1%
I50	Heart Failure	22	86.36%	13.64%
Average			87.26%	12.74%

TABLE III. TRAINING DATASET RUNNING TIME FOR HEART DISEASES AND MEDICINE DATASET

Model for Heart Diseases	Time (Sec)
SBM	0.30

Training dataset running time for heart diseases and medicine dataset is (0.30) sec because it is only a time to choose the best medicine in the system yet it is not for diagnosing the disease. The system is an assistant to the doctor, and it is a complementary tool for the medical diagnosis.

Figure (4) demonstrates a chart indicating the Performance of (EMD) for Heart Diseases. Training datasets consist of two groups: the first is made of five tables where a table is specified for each medical case in which the patient data are taken from the patient files. These are the symptoms and patient medical history. While the second category contains five tables, where a table is designated for each disease. This table has ten drugs commonly identified with that disease, yet each drug has a set of side effects. The name of the medicine is the scientific name as it is an international standard.

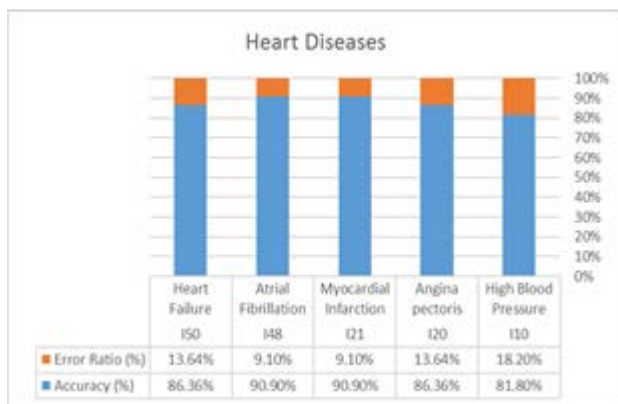


Fig. 4. The performance of (EMD) for Heart Diseases

VIII. CONCLUSIONS

In this paper, an approach based on Rete Algorithm is used where the best-chosen medicine through the system (EMD) is suggested. We want to reach into the best model using the algorithm to obtain a more accurate selection for an optimal medicine for diagnosis. The doctor can select medicine by the system (EMD) depending on all information and side effects of that medicine, symptoms and medical history for a patient. The system (EMD) is tested by the doctors in Iraqi hospitals. The accuracy of selecting the medicine to heart diseases is approximately (87.26%).

This accuracy is obtained because the principles of nominating drugs by a doctor for a given patient are based on two factors. The first factor is the symptoms experienced by the patient along with his medical history, while the second one is the chemical interactions that happen inside the patient's body. Our system depends on the first factor in which the selection of medicine is made to avoid the interfering with other medicaments for a certain disease. It doesn't concentrate on more of other suspicious diseases but simply treats the current symptoms that are experienced by the patient.

REFERENCES

- [1] U.S. Pharmacopeia National Formulary, "United States Pharmacopeia (U.S. Pharmacopeia National Formulary)," Us Pharmacopeia .Pck Slp edition, ISBN-13, January 1, 2015.
- [2] U K Stationery Office, "British Pharmacopoeia (BP) 2016," Stationery Office Books (TSO), ISBN-13, August 24, 2015.
- [3] Sid M. Wolfe, "Worst Pills, Best Pills: A Consumer's Guide to Avoiding Drug-Induced Death or Illness," Gallery Books; Revised ed. edition, ISBN-13, January 4, 2005.
- [4] BMJ Group , Pharmaceutical Press , "BNF 69: March 2015 - September 2015 (British National Formulary)," Pharmaceutical Pr, 1 edition, ISBN-13, March 31, 2015.
- [5] Jenna M. Repts, Uwe Aickelin, Jiangang Ma, Melbourne, Yanchun Zhang, "Refining Adverse Drug Reactions using Association Rule Mining for Electronic Healthcare Data," arXiv:1502.05943v1 [cs.DB] , 20 Feb 2015.
- [6] Jie Chen , Hongxing He ,JiuyongLi , Huidong Jin , Damien McAullay ,Graham Williams, Ross Sparks and Chris Kelman, "Representing Association Classification Rules Mined from Health Data," R. Khosla et al., Eds. Springer-Verlag Berlin Heidelberg, 2005, pp. 1225–1231.
- [7] Yu Chen, Lars Henning Pedersen, Wesley W. Chu, Jorn Olsen, "Drug Exposure Side Effects from Mining Pregnancy Data," ACM SIGKDD Explorations Newsletter - Special issue on data mining for health informatics, Vol. 9 Issue 1, June 2007 pp. 22 - 29 .
- [8] Christian G. Daughton , Ilene Sue Ruhoy, "Lower-dose prescribing: Minimizing "side effects" of pharmaceuticals on society and the environment," Elsevier B.V. , CC BY-NC-ND license,2012.
- [9] Nwigbo Stella N , Agbo Okechuku Chuks,"Expert System: A Catalyst in Educational Development in," Human Resource Management Academic Research Society, Proceedings of the 1st International Technology, Education and Environment Conference, 2011.
- [10] Brian R. Walker BSc MD FRCPE FRSE , Nicki R Colledge BSc (Hons) FRCPE , Stuart H. Ralston MD FRCP FMedSci FRSE , Ian Penman BSc MD FRCPE, "Davidson's Principles and Practice of Medicine: With STUDENT CONSULT Online Access, 22e (Principles & Practice of Medicine (Davidson's)) 22nd Edition," Churchill Livingstone, ISBN-13, 22 edition ,February 15, 2014.
- [11] Robert B. Doorenbos," Production Matching for Large Learning Systems," Robert B. Doorenbos, Thesis for the degree of Doctor of Philosophy, CMU-CS-95-113 ,January 31, 1995.

Context-Sensitive Opinion Mining using Polarity Patterns

Saeedeh Sadat Sadidpour

Department of Computer Science,
Faculty of ICT
Malek-Ashtar University of
Technology
Tehran, Iran

Hossein Shirazi

Department of Computer Science,
Faculty of ICT
Malek-Ashtar University of
Technology
Tehran, Iran

Nurfadhlina Mohd Sharef

Department of Computer Science,
Faculty of Computer Science and
Information Technology
Universiti Putra Malaysia
Malaysia

Behrouz Minaei-Bidgoli

Department of Computer Engineering
Iran University of Science and Technology
Tehran, Iran

Mohammad Ebrahim Sanjaghi

Department of Management and Soft Technology
Malek-Ashtar University of Technology
Tehran, Iran

Abstract—The growing of Web 2.0 has led to huge information is available. The analysis of this information can be very useful in various fields. In this regards, opinion mining and sentiment analysis are one of the most interesting task that many researchers have paid attention for two last decades. However, this task involves to some challenges that a very important challenge is the different polarity of words in various domain and context. Word polarity is an important feature in the determination of review polarity through sentiment analysis. Existing studies have proposed n-gram technique as a solution which allows the matching of the selected words to the lexicon. However, identification of word polarity using the standard n-gram method poses limitation as it ignores the word placement and its effect according to the contextual domain. Therefore, this study proposes a linguistic-based model to extract the word adjacency patterns to determine the review polarity. The results reflect the superiority of the proposed model compared to other benchmarking approaches.

Keywords—Opinion mining; Polarity patterns; Pattern matching; Context-sensitive; Politics domain

I. INTRODUCTION

In the past, people tried very to acquire data and knowledge. Whereas the appearance of web, and especially Web 2.0, brings on huge information is generated by users. Although, this is not desire because too much information leads to confused. Therefore, analysis, summarization and other related tasks are very useful and applicable results present for users and researchers. In this regard, opinion mining and sentiment analysis are one of the most important and helpful task that has been defined from two last decades.

Opinion mining is a field that its results can be used for different researches of various fields. In this regard, researchers have been attracted because of its application from the sociological and psychological analysis to the extraction of users' opinion in business field such as about products and services, or political discussions.

The first time, Wiebe was presented the most widely definition for subjectivity and opinion mining based on a linguist's idea in 1994. Also, she defined as the linguistically expression of opinions, sentiment, emotions, evaluations, beliefs, and speculations [1]. Hence, the goal of subjectivity analysis is determining the subjective or objective sentences [2] that Liu considered subjective and objective sentences against each other. In this respect, objective and subjective sentences define, respectively, as a fact information indicative about world and a person's emotions or beliefs indicative [3].

Next and in lower level, Zhang and Liu define opinion mining or sentiment analysis as "the computational study of opinions, appraisals, attitudes, and emotions about entities such as products, services, organizations, persons, events, and their different aspects" [4]. Also, they focus on recognizing the orientation and strength of polarity in different level using various methods and settings.

In this respect, various definitions in these years have caused to not be specific different between opinion mining and sentiment analysis and are used instead of each other. Of course, it is notable that some ones consider different levels for opinion and sentiment.

Regards, researchers investigated the various sides of opinion mining. For this purpose, some issues are defined such as how polarity is expressed, which level polarity is measured, or the polarity of words is fix or not.

Hence, primary researches on opinion mining determined the polarity as positive and negative [5]. However, Koppel and Schler proposed neural is also considered to be improved results [7 ,6]. In addition, opinion mining is defined as orientation (such as positive and negative) and strength (such as weak positive, middle positive, and strong positive) [8].

Meanwhile, researches have shown polarity is not fix and even changes in a domain [10 ,9]. Also, Ding expressed the polarity is fix only rather than a context [11].

scenario Give example to illustrate your idea.

What do you mean by 'replacement of words is not considered for its statistical calculation'? do you mean the vector space model/bag of words?

The change of polarity leads to context-sensitive opinion mining. For this purpose, some different solutions such as ontology are proposed for the context-sensitive problem. Although, the context-sensitive issue poses limitation to existing solution. For an example, the methods often consider the polarity of words as uni-gram. While words get together, they affect on the polarity of each other.

In contrast, n-gram is not a suitable method because the replacement of words is not considered for its statistical calculation; As a result, n-gram leads to the sparseness in repeating adjacent same words and lack of generalizability. For an instance, "supporting of terrorism" and "supporting of spies" are two different n-gram that their value of language model is very low. While the generality of them in pattern "supporting of (Neg. Exp.)" occurs more.

Thus, this paper proposes opinion mining is done using polarity pattern extraction based on language model helps to be possible words, which are synonym or have similar polarity, are replaced.

The polarity patterns categorize expressions which their polarity is same. For example, "supporting of terrorism" is negative; however, "supporting" is positive and "terrorism" is negative, and bag of word method often recognize the term as neutral. While the polarity pattern of "supporting of (Neg. Exp.)" leads to a negative result. Consequently, the results of context-sensitive opinion mining using polarity pattern matching express the significant improvement of accuracy rather than other methods.

In continue, section 2 presents a review on context-sensitive opinion mining. Then, context-sensitive opinion mining using polarity pattern matching and evaluation are discussed, respectively. Finally, conclusion is expressed.

II. A REVIEW ON CONTEXT-SENSITIVE OPINION MINING

Although many researchers have paid attention to opinion mining and sentiment analysis, the field poses several room of improvements to solve issues in data mining, web mining, and information retrieval. Therefore, different challenges are observed in the field such as how the polarity is calculated for word groups. So, the polarity is calculated using different methods [12] such as deep sentiment analysis [13], Markov logic [14], Fuzzy [16,15], LDA [17], and language model [18].

The primary opinion mining researches prepared opinion lexicons and calculated the polarity based on the lexicons. However, growing researches indicated that opinion mining faces to a big challenge. In this regard, the challenge is the polarity of words sometimes changes in different texts. Thus, some topics were considered such as domain-specific opinion mining [20,19,10,9], and context-sensitive opinion mining [21].

Contextual opinion mining expresses the polarity of words can be changed in a text, and words may have different polarities even in a domain [11]. In the meanwhile, co-

occurrence and conjunctions [22], top word collocations [23], Emoticons [24], opposite pairs [25], an approach based on holistic lexicon prepared using WordNet [11], and using features such as the characteristics of before and next words, modifiers, and dependency tree [26] are solutions for changing polarity in a text.

In addition, lexicon adapting is another strategy that finds context-sensitive polarity using contextual semantics and updates orientation and strength using a primary polarity lexicon and rule-base adapting [27].

However, the effectiveness of adjacent words is a significant criterion that has not been explored by any existing methods. Although language model can find adjacent words, it is typically rigid. Even co-occurrence and synonym words are not considered in statistical calculations. As a result, this paper proposes a method which is based on n-grams to categorize expressions that have similar polarity. Then, some polarity patterns are extracted that indicate for polarity language model. Finally, the polarity patterns are used to context-sensitive opinion mining and sentiment is calculated based on them.

III. CONTEXT-SENSITIVE OPINION MINING USING POLARITY PATTERN MATCHING

Different studies have used patterns for determining polarity. Pattern recognition is a helpful step for opinion mining tasks that detects text relations. In this regard, Riloff and Wiebe [28] extracted patterns to detect subjective and objective sentences. The patterns were extracted based on some templates and using corpus investigating.

Furthermore, the patterns for getting together words and language model are two another effective methods. Wiebe and others [2] reached to subjectivity language indicates adjacent words and language model accompanied with part of speech (PoS) patterns. Thereof, the aim of statistical language model is learning the probability function of word sequence. However, this problem needs many data that leads to sparseness. Therefore, words which operate same are considered as similar tokens and replaced to each other [30,29]. In the other hand, Tang and others [31] extracted possible bi-grams to calculate polarity. The determination of N in n-grams is another problem in opinion mining.

As a result, this paper proposes a method for opinion mining using polarity patterns which are the replacement of n-gram and solve its problems.

A. Polarity Pattern Extraction

Polarity pattern extraction proposed is based on n-gram method and creates a polarity language model for opinion mining. In according to the investigations of corpus, if X is the noun phrase of affect, three types of phrases, that can be shown as sub-sentences too, are effective on polarities:

- X + A: for example, "دشمنی تروریسم" means "hostility of terrorism" is negative when a negative affect is done by a negative effective.
- X + Conjunction + B: for example, "دشمنی با دین" means "hostility with religion" is negative when a negative affect is done rather than a positive affected.

- X + A + Conjunction + B: for example, “دشمنی تروریسم با دین” means “hostility of terrorism with religion” is negative when a negative affect is done by a negative effective rather than a positive affected.

X, A and B are sub-sections that can be positive or negative and led various phrases with different polarity. These three type of phrases are templates for finding polarity patterns. TABLE I is two examples of polarity of phrases.

TABLE I. THE EXAMPLES OF POLARITY PATTERNS

X		A	Conjunction	B	Polarity of phrase
Noun phrase of affect	translate				
دشمنی	Supporting		از of		
+		+		+	+
+		+		-	-
+		-		+	-
+		-		-	-
دشمنی	Hostility of		با with		
-		+		+	-
-		+		-	+
-		-		+	-
-		-		-	+

For this purpose, a rule-based method is used and polarity patterns are extracted based on pre-defined templates which were mentioned before. The steps of proposed method are shown in “Fig. 1”.

Firstly, natural language processing is done on corpus and tokens and their PoS are determined. Then, the corpus is parsed and its dependency trees are extracted. As follow, the dependency trees are adapted to templates; for each phrase group, if y_p and Y_q are the p th phrase group and q th desired phrase group, respectively, that will be obtained by (1).

$$Y_q = y_{p|DT_p=T_i} \quad (1)$$

that DT_p is the dependency tree of p th phrase group and T_i is different templates with $i = 1, 2, 3$. Thus, the desired phrase groups will be selected if the dependency tree of p th phrase group is matched with one of templates T_i . Also, the type of template, or i , which has been matched is determined.

In the next step, the polarity of sub-sections is determined. The sub-section X is a noun which is done an affect but A and B can be from a noun to a noun phrase which are an effective and an affected, respectively. For every Y_q , an adapted template AT_q is built based on matching among the desired phrase group, the polarity of sub-sections, and template T_{iq} in (2).

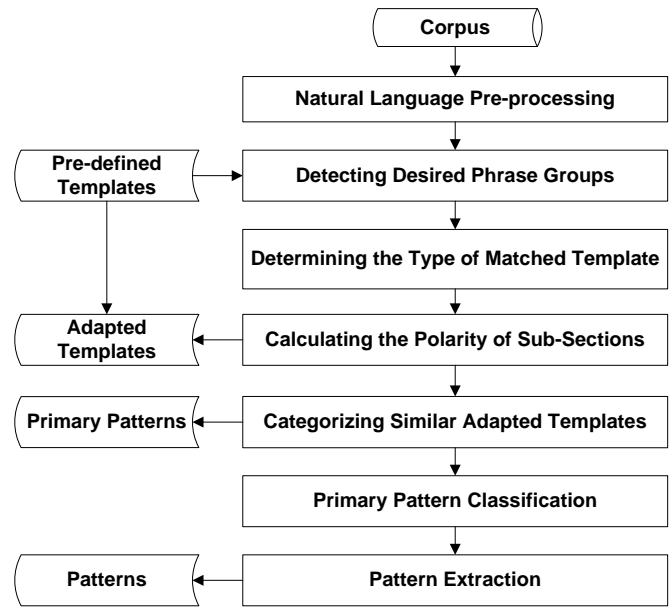


Fig. 1. The steps of proposed method for polarity pattern extraction

$$AT_q = Y_q \left\{ \begin{array}{l} AT_{aq} = \text{The polarity of sub-section } A_q \\ \text{if } A_q \in T_{iq} \\ AT_{bq} = \text{The polarity of sub-section } B_q \\ \text{if } B_q \in T_{iq} \end{array} \right. \quad (2)$$

that q index indicates q th phrase group. AT_{aq} , AT_{bq} and T_{iq} are the polarity of sub-sections A and B, and i th template, which correspond with Y_q .

After that, the similar adapted templates are categorized using the polarity of each existing A or B sub-sections to be constructed primary polarity patterns. In meanwhile, every primary polarity pattern is an index for the set of adapted templates that their “A” and “B” have been replaced by polarity and their X are a set includes the noun phrase of affect. According to (3),

$$PP_k = AT_h, \forall h, l \left\{ \begin{array}{l} AT_{ah} = AT_{al} \quad \text{if } A \in T_{iq} \\ AT_{bh} = AT_{bl} \quad \text{if } B \in T_{iq} \\ AT_{ah} = AT_{al}, AT_{bh} = AT_{bl} \quad \text{if } A, B \in T_{iq} \end{array} \right. \quad (3)$$

that PP_k is the k th primary polarity pattern. Also, $\forall h, l$ expresses only an adapted template is considered as PP for replacing all adapted templates which the polarity of their sub-sections except X is same.

In the next step, the set of X of every primary polarity pattern is classified to positive and negative classes. The classification is done rather than the polarity of “X”s. Then, the classified primary polarity patterns and the final polarity of phrases are investigated to be extracted polarity patterns. For this purpose, final polarity patterns are provided based on the different states of “A” and “B” in similar primary polarity patterns which their polarity of “X” is same. If P_m is the m th extracted polarity pattern, that will be presented based on (4)

$$P_m = \begin{cases} \{X^+\} + (PP_k - \{X\}) & \text{if } \{X\} \text{ is } + \\ \{X^-\} + (PP_k - \{X\}) & \text{if } \{X\} \text{ is } - \end{cases} \quad (4)$$

that $\{X\}$, $\{X^+\}$ and $\{X^-\}$ are the set of “X”s, and classified “X”s, respectively. According to the detected phrase groups in corpus, the corresponding polarity of every state is determined by experts, and as a result, the polarity patterns are prepared for opinion mining.

For generalization of polarity patterns, the sets of “X” are extended using semantic similarity strategies, because similar words can be replaced each other. For this purpose, the synonyms, hypernyms, hyponyms, co-occurrences, and co-locations of all sub-sets of “X”s are extracted. Then, every word would be added to its set of “X” if there was not. Finally, the extracted polarity patterns and the prepared sets of “X” are used for context-sensitive opinion mining.

B. Opinion Mining Using Polarity Pattern Matching

The goal of proposed method in this paper is to analyze the opinion of texts using polarity pattern matching. The main stages of context-sensitive opinion mining are indicated in “Fig. 2”.

The first stage is the pre-processing for context-sensitive opinion mining, which involves normalization, tokenizing, sentence detection, and Part-Of-Speech (POS) tagging. The dependency parser [32] is used for syntactic relations identification besides words and their POS. Following, the syntactic relations are used for polarity pattern matching.

In the next stage, which is main stage for the proposed method, the polarity identification of sections, which are matched with the polarity patterns, is begun. Firstly, the words of sentence are compared with the sets of “X”. If each word is matched, its related phrases as sub-sections are extracted based on its corresponded polarity pattern and the dependency tree. Then, the polarity of sub-sections is measured as direct or recursive using polarity patterns. In meanwhile, the direct polarity of words are calculated using SentiFarsNet. Thus, the polarity of phrase is considered as the result of polarity pattern.

Finally, opinion mining measures the polarity of every sentence. After that the polarity of all sections of sentence was calculated, the polarity of sentence would be determined dependent on the polarities of sections.

IV. EVALUATION

This paper implements the context-sensitive opinion mining using polarity pattern matching. For this purpose, the polarity patterns were extracted for politics domain. Also, the

SentiFarsNet that is the translation of SentiWordNet based on Farsnet is applied to find prior opinion.

A. Dataset

For this paper, we gathered data by a crawler from the political news. The news are collected from Mar. 2013 to Dec. 2015 on three news sites. Also, the crawled data contains 208,000 political news. After that, we selected our corpus from gathered data. The corpus includes about 14,000 news between 2KB and 10KB. Then, pre-processes were implemented on the corpus and dependency trees were obtained for opinion mining.

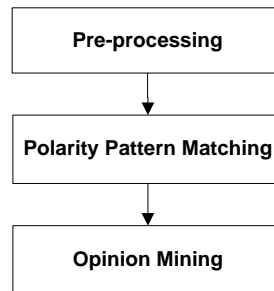


Fig. 2. The stages of context-sensitive opinion mining

B. Results

According to the proposed method of this paper, the polarity patterns are extracted that TABLE II is shown some samples. The investigation of corpus indicates 38% group nouns belong to the polarity patterns.

TABLE II. THE EXAMPLES OF POLARITY PATTERNS

X		A	Conjunction		B	Polarity of phrase
Noun phrase of affect	translate					
پشتیبانی	Supporting		از	of		
+		+			+	+
+		+			-	-
+		-			+	-
+		-			-	-
Other words of this polarity pattern						
اعتماد	Trust		به	of		
اطمینان	Confidence		به	of		
مقاومت	Resistance		در مقابل	against		

The first evaluation is calculating the exactness and completeness of polarity patterns. For this purpose, precision is defined as the fraction of obtained polarity patterns that are relevant to the content. In the other hand, recall is the fraction of the polarity patterns are relevant to the content that are successfully retrieved in documents. Finally, F-measure is an average of precision and recall that “Fig. 3” presents the quality of the extracted polarity patterns. Whereas the coverage of required polarity patterns for opinion mining is the most measurement, the value of precision, which is more than 90%, demonstrates the performance of the proposed method.

Of course, the precision of n-gram is more and better than the proposed method but in contrast, the recall of proposed method is better. As a result, the polarity pattern method needs less calculations and has more f-measure than n-gram.

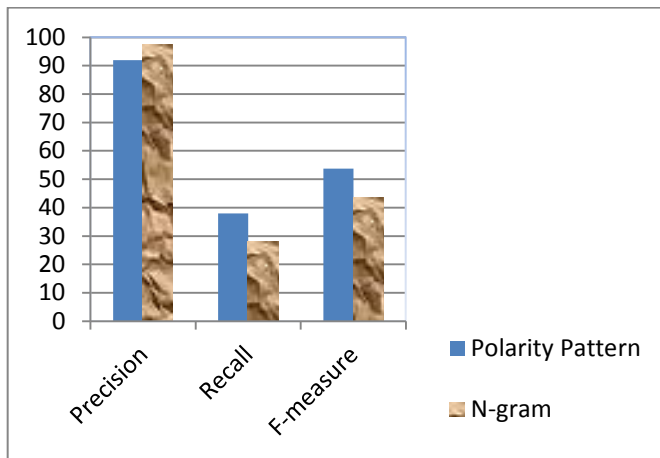


Fig. 3. The quality of the extracted polarity patterns and n-grams

Also, the accuracy of proposed method is compared with a base method that this paper uses bag of words and their polarity. The results are observed in “Fig. 4”. As the results show, the determination of polarity based on the polarity patterns leads to increase accuracy in opinion mining.

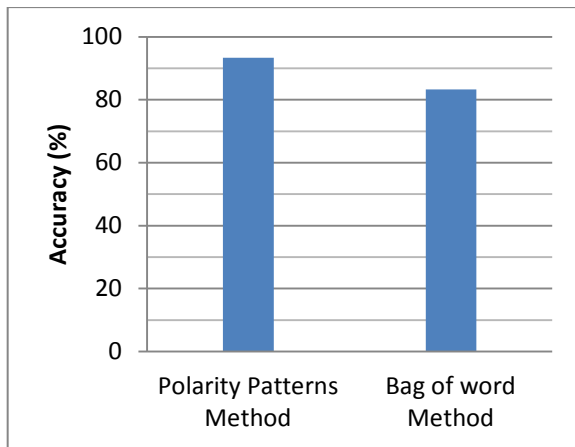


Fig. 4. The comparison of polarity patterns and bag of words usage

V. CONCLUSION

The growing of Web 2.0 has lead to extend interesting to opinion mining in research society. Whereas the aim of opinion mining is to recognize the positive or negative opinion of documents and opinionated sentences [33], the polarity of words and expressions is the most important issue. In the meanwhile, researches have shown polarity is not fix and even changes in a domain. Also, Ding expressed the polarity is fix only rather than a context [11]. Therefore, context-sensitive opinion mining is presented and indicates the primary polarity does not determine correct polarity.

Most methods often consider the polarity of words as uni-gram, but how words are placed adjacent to each other is very useful for opinion mining. Also, n-gram is used only as features for machine learning algorithms.

When words get together, they affect on the polarity of each other. In the other hand, n-gram is not a suitable method because the replacement of words is not considered for its

statistical calculation; As a result, n-gram leads to sparseness and lack of generalizability. Thus, this paper proposes opinion mining is done using polarity pattern extraction based on language model helps to be possible words, which are synonym or have similar polarity, are replaced.

The polarity patterns categorize expressions which their polarity is same. For example, “supporting of terrorism” is negative because “terrorism” is negative and the polarity pattern of “supporting of (Neg. Exp.)” leads to a negative result. Consequently, the results of context-sensitive opinion mining using polarity pattern matching show the significant improvement of accuracy the determination of polarity based on the polarity patterns rather than other methods.

REFERENCES

- [1] J. Wiebe, "Tracking point of view in narrative," *Computational Linguistics*, vol. 20, pp. 233-287, 1994.
- [2] J. Wiebe, et al., "Learning subjective language," *Computational linguistics*, vol. 30, pp. 277-308, 2004.
- [3] B. Liu, "Sentiment analysis and subjectivity," *Handbook of natural language processing*, vol. 2, pp. 627-666, 2010.
- [4] L. Zhang and B. Liu, "Aspect and entity extraction for opinion mining," in *Data mining and knowledge discovery for big data*, ed: Springer, 2014, pp. 1-40.
- [5] B. Pang, et al., "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79-86.
- [6] M. Koppel and J. Schler, "Using neutral examples for learning polarity," in *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 2005, p. 1616.
- [7] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," *Computational Intelligence*, vol. 22, pp. 100-109, 2006.
- [8] A. Esuli and F. Sebastiani, "Determining Term Subjectivity and Term Orientation for Opinion Mining," in *EACL*, 2006, p. 2006.
- [9] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," in *Proceedings of recent advances in natural language processing (RANLP)*, 2005, p. 2.1.
- [10] S. Mahalakshmi and E. Sivasankar, "Cross Domain Sentiment Analysis Using Different Machine Learning Techniques," in *Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO-2015)*, 2015, pp. 77-87.
- [11] X. Ding, et al., "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 231-240.
- [12] A. Ko, et al., "Ontology Supported Policy Modeling in Opinion Mining Process," in *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*, 2012, pp. 252-261.
- [13] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 355-363.
- [14] Z. Chen, et al., "Joint model for subsentence-level sentiment analysis with Markov logic," *Journal of the Association for Information Science and Technology*, 2015.
- [15] R. Y. Lau, et al., "Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis," *Decision Support Systems*, vol. 65, pp. 80-94, 2014.
- [16] L. Liu, et al., "Toward a fuzzy domain sentiment ontology tree for sentiment analysis," in *Image and Signal Processing (CISP)*, 2012 5th International Congress on, 2012, pp. 1620-1624.
- [17] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 815-824.

- [18] L. Garcia-Moya, et al., "Retrieving product features and opinions from customer reviews," *IEEE Intelligent Systems*, pp. 19-27, 2013.
- [19] A. Andreevskaia and S. Bergler, "When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging," in *ACL*, 2008, pp. 290-298.
- [20] P. Sanju and T. Mirnalinee, "Construction of Enhanced Sentiment Sensitive Thesaurus for Cross Domain Sentiment Classification Using Wiktionary," in *Proceedings of the Third International Conference on Soft Computing for Problem Solving*, 2014, pp. 195-206.
- [21] Y. Zhang, et al., "Joint naive bayes and lda for unsupervised sentiment analysis," in *Advances in Knowledge Discovery and Data Mining*, ed: Springer, 2013, pp. 402-413.
- [22] D. R. Rice and C. Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies," *Proceedings of NDATAD*, 2013.
- [23] S. A. Bahrainian, et al., "Fuzzy Subjective Sentiment Phrases: A Context Sensitive and Self-Maintaining Sentiment Lexicon," in *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, 2014, pp. 361-368.
- [24] N. Blenn, et al., "Context-sensitive sentiment classification of short colloquial text," in *NETWORKING 2012*, ed: Springer, 2012, pp. 97-108.
- [25] Z. Zhang, et al., "Context-Dependent Sentiment Classification Using Antonym Pairs and Double Expansion," in *Web-Age Information Management*, ed: Springer, 2014, pp. 711-722.
- [26] T. Wilson, et al., "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 2005, pp. 347-354.
- [27] H. Saif, et al., "Adapting Sentiment Lexicons Using Contextual Semantics for Sentiment Analysis of Twitter," in *The Semantic Web: ESWC 2014 Satellite Events*, ed: Springer, 2014, pp. 54-63.
- [28] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 105-112.
- [29] Y. Bengio, et al., "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137-1155, 2003.
- [30] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," *Decision Support Systems*, vol. 53, pp. 680-688, 2012.
- [31] D. Tang, et al., "A joint segmentation and classification framework for sentence level sentiment classification," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, pp. 1750-1761, 2015.
- [32] S. Huang, et al., "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowledge-Based Systems*, vol. 56, pp. 191-200, 2014.
- [33] B. Pang and L. Lee, "Opinion mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1-135, 2008.

Application of Intelligent Data Mining Approach in Securing the Cloud Computing

Hanna M. Said¹

¹Faculty of computer and information science,
Ain Shams University
Cairo, Egypt

Ibrahim El Emary²

²Information Science Department
King Abdulaziz University,
Jeddah, Saudi Arabia

Bader A. Alyoubi³

³Management Information Systems (MIS)
College of Business, University of
Jeddah, Jeddah, Saudi Arabia

Adel A. Alyoubi³

³Management Information Systems (MIS)
College of Business, University of
Jeddah, Jeddah, Saudi Arabia

Abstract—Cloud computing is a modern term refers to a model for emerging computing, where it is possible to use machines in large data centers for delivering services in a scalable manner, so corporations has become in need for large scale inexpensive computing. Recently, several governments have begun to utilize cloud computing architectures, applications and platforms for meeting the needs of their constituents and delivering services. Security occupies the first rank of obstacles that face cloud computing for governmental agencies and businesses. Cloud computing is surrounded by many risks that may have major effects on services and information supported via this technology. Also, Cloud Computing is one of the promising technology in which the scientific community has recently encountered. Cloud computing is related to other research areas such as distributed and grid computing, Service-Oriented Architecture, and virtualization, as cloud computing inherited their limitations and advancements. It is possible to exploit new opportunities for security. This paper aim is to discuss and analyze how achieve mitigation for cloud computing security risks as a basic step towards obtaining secure and safe environment for cloud computing. The results showed that, Using a simple decision tree model Chaid algorithm security rating for classifying approach is a robust technique that enables the decision-maker to measure the extent of cloud securing, and the provided services. It was proved throughout this paper that policies, standards, and controls are critical in management process to safeguard and protect the systems as well as data. The management process should analyze and understand cloud computing risks for protecting systems and data from security exploits

Keywords—Cloud computing; Cloud security issue; Data mining; Naive Bayes; multilayer perceptron; Support vector machine; decision tree (C4.5); and Partial Tree (PART)

I. INTRODUCTION

Security is a basic requirement for cloud computing [1]. This view is shared by many distinct groups, such as business decision makers [4], academia researchers [2, 3], and government organizations [5, 6]. The several similarities in these opinions illustrate a high concern on crucial legal and security obstacles for cloud computing including availability of service, confidentiality in data, provider lock-in and status

fate sharing [7]. The security is considered one of the major issues which reduces the growth of cloud computing and complications with data privacy and data protection continue to plague the market. The advent of an advanced model should not negotiate with the required functionalities and capabilities present in the current model. Cloud computing embraces cyber-infrastructure and builds on grid computing, utility computing, virtualization, distributed computing, networking, web and software services [11].

The illusion of infinite computing resources available on demand, the elimination of up-front commitments by cloud users, and the ability to pay for use of computing resources on a short-term basis are needed [3]. Cloud computing has been defined in many and diverse ways according to the point of view that deals with the cloud computing. One of these definition is adapted by [52] where cloud computing has been defined as a pool of highly scalable, abstracted, and managed infrastructure that is capable of hosting end-customer applications and billed by consumption. Another definition has been suggested by NIST [6], where cloud computing has been defined as a model for enabling on-demand network access to a shared pool of computing configurable resources in convenient way (e.g. servers networks, applications, storage, and services) that can be rapidly released and provisioned with minimal service provider interaction or management effort.

According to [42], a cloud is a pool of virtualized resources across the Internet that follows a pay per-use model and can be dynamically reconfigured to satisfy user requests via on-the-fly. Also, Cloud computing is known as a method for increasing the capabilities or adding capacity in dynamic manner without investing in new training new personnel, infrastructure or licensing new software, accordingly it works towards extending the existing capabilities of Information Technology (IT) [8]. Recently, cloud computing has been converted from being a concept of promising business to one of the fast growing segments of the IT industry. It is possible to say that the more the spread of cloud computing, the more the concern about the security and safety of the cloud computing environment. The security problem can be

considered as a barrier against the deployment of cloud computing in business environment [9].

In cloud computing, security is considered a critical and important aspect, where security in cloud computing has many problems and issues related to it. Both the cloud service consumer and cloud service provider should be sure that the cloud is sufficiently safe from external threats in order to make the customer avoid any problem such as data theft or loss of data [10]. The penetration through a malicious user into the cloud can be occurred through impersonating a legitimate user, thus infecting many customers.

The users of cloud service should understand the risks of data breaches in the new environment of cloud computing; as the architecture of cloud forms a threat to the existing technologies security when deploying such technologies in a cloud environment [11].

Data mining can be considered one of the most important for discovering the knowledge from large data. Different algorithms and techniques are available in data mining. For finding the mine rule, classification technique can be used in large data. In general decision tree technique can be utilized as efficient technique for classification; because it owns simple hierarchical structure for the decision making and user understanding. This paper evaluates and investigates the decision tree as data mining techniques and as an intrusion detection mechanism.

This paper focus on a survey about the risks and threats that faces cloud computing followed by deep analysis of cloud security major issues such as: trust, encryption, multi-tenancy and compliance and finally utilizing the intelligent data mining and attack classification methodology in analyzing the cloud security.

In the last few years, it was shown that Cloud Computing depends on gathering few new and many old concepts in several fields of research such as Service-Oriented Architectures (SOA), grid, and distributed computing as well as virtualization. This was a result of its high probability for substantiating advances in other technologies while presenting supreme advantage over the current under-utilized resources that are deployed at data centers [12].

It is possible to say that cloud computing can be understood as a new computing paradigm that give users the temporal ability to use computing infrastructure over the network, that is provided as a service by the cloud-provider at one or more than one of abstraction levels. Thus, several models of business are rapidly evolved to utilize this technology by introducing programming platforms, software applications, computing infrastructure, data-storage, and hardware as services. Their inter-relations have been ambiguous. The feasibility of enabling their inter-operability has been debatable while they refer to the core cloud computing services, taking into consideration that each cloud computing service has a unique interface and uses a different protocol of access [13-15].

This paper is organized as follows: Section 2 shows the Literature Survey. Section 3 describes the Cloud environment layers, Service Models and Deployment Models. In section 4, we cover the gaps and security issues in service models, Denial-of-service attack classification, and Decision tree: C4.5 as well as Performance evaluation. Finally, section 5 provides conclusions.

II. LITERATURE REVIEW

Nowadays, Small and Medium Business (SMB) organizations have been understood that simply by getting into the cloud computing, they can obtain fast access to best business applications or increasingly reinforce the resources of their infrastructure, all at negligible cost. Gartner [53] defined cloud computing as ‘‘a style of computing where massively scalable information technology- enabled capabilities are delivered ‘as a service’ to external clients using Internet technologies’’. [42] Mentioned that nowadays cloud providers obtain benefits from opportunity in the marketplace. The providers should ensure that they obtain the right security aspects; therefore they will bear the responsibility if things are wrong.

The cloud offers many benefits such as pay-for- use, fast deployment, scalability, lower costs, rapid elasticity, rapid provisioning, ubiquitous network access, greater resiliency, low-cost disaster recovery and data storage solutions, hypervisor protection against network attacks, on-demand security controls, real time detection of system tampering, and rapid re-constitution of services. The unique attribute of the cloud computing, poses several new challenges from security point of view [44]. The posed challenges include virtualization vulnerabilities, accessibility vulnerabilities, web application vulnerabilities such as SQL (Structured Query Language) injection, cross-site scripting, privacy and control issues arising from third parties having physical control of data, physical access issues, issues related to credential management, identity and issues related to data verification, tampering, integrity, data loss and theft, issues related to authentication of the respondent device or devices and IP spoofing.

III. CLOUD ENVIRONMENT LAYERS

Cloud computing attracts many managers and organizations. There are many similar terminologies that are usually utilized for describing cloud computing, these terms such as: distributed, grid, cluster, virtualization, on-demand, utility, and software-as-a-service. In other words, cloud computing refers to end-users connecting with applications running on sets of shared servers, often hosted and virtualized, instead of a traditional dedicated server.

For over thirty years client-server computing has provided applications that were assigned to specific hardware, often residing in on-premise data centers. On-demand cloud computing enables its end-users through allowing them using their selection of Internet-connected device, at any time [29].

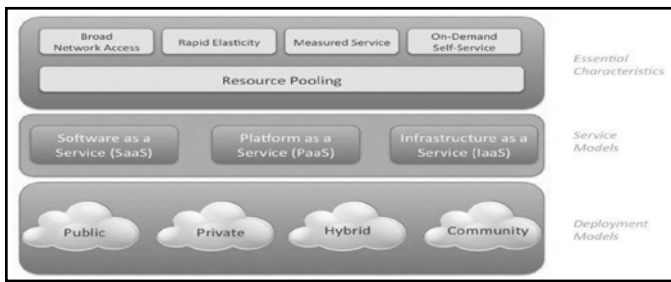


Fig. 1. Layers of cloud environment [19]

From Figure 1, it is shown that the lower layer of the cloud computing layers represents the different models of deployment for the cloud as follows: community, private, public and hybrid cloud models of deployment. The second layer above the deployment layer represents the different models of delivery that are used within a certain model of deployment. These delivery models are the IaaS (Infrastructure as a Service) delivery, PaaS (Platform as a Service) and SaaS (Software as a Service) models. These delivery models represent the core of the cloud and they show certain features like multi-tenancy, on-demand self-service, ubiquitous network, measured service and rapid elasticity that are illustrated in the upper layer. These basic elements of the cloud computing require security that depends and varies with respect to the used deployment model, the method of delivery and the character it shows. Some of the basic security challenges can be summarized in data transmission security, data storage security, security related to third-party resources and application security [19].

A. Service Models

It is possible to categorize cloud services into three categories namely: Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS).

- **Software-as-a-Service (SaaS)**, usually called on-demand software, is a software deployment and a model of subscription-pricing that gives an enterprise application as a managed service by a software vendor. The SaaS provider bears all responsibilities related to the implementation and maintenance of the system from the customer; this in turn, is useful when adding new hardware as it minimize the addition cost and complexity. One example of SaaS is the Salesforce.com CRM application. According to the Forrester study, "The State of Enterprise Software: 2009," security concerns are the most commonly cited reason why enterprises are not interested in SaaS. Consequently, addressing enterprise security concerns has emerged as the biggest challenge for the adoption of SaaS applications in the cloud [45].
- **Infrastructure-as-a-Service (IaaS)**, usually called on-demand infrastructure, it gives computing infrastructure as a utility service. IaaS users buy or rent software, servers, network equipment, data-center space etc. IaaS usually gives networking with immense possibility for extensibility, scale and raw storage. One example of (IaaS) is the Amazon web services [45].

- **Platform-as-a-Service (PaaS)** can be defined as a rich ecosystem that is used for database development along with other applications, programmer communities, and application development, as a solution stack or service. One of the major advantages of PaaS is that it empowers developers of an application to build their own applications on top of the platform. PaaS usually overcome the gaps in functional holes within a SaaS solution. An example of (PaaS) would be Google Apps (Cloud Security Alliance, 2011) [4, 13, and 14].

All models of cloud service (IaaS, SaaS and PaaS) should be strongly well-defined service level agreements (SLAs) that are able to protect the cloud user. According to the CSA, "security, governance, service levels, compliance, and liability expectations of the service and provider should be stipulated from contracting point of view, enforced and managed" (Cloud Security Alliance, 2011).

B. Deployment Models

The deployment models can be categorized into four categories namely Public cloud, Private cloud, Hybrid cloud and Community cloud [6]. These categories will be described in details as follows:-

- **Public Cloud**, this model is owned by an organization for selling the cloud services and the design of infrastructure is made in order to be available for industries, organizations and businesses.
- **Private Cloud**, this model is managed by the organization itself or by a third party. Private cloud may be either off or on premises. The major characteristic of this model is that the infrastructure of the cloud is private, in addition to its availability to a single organization.
- **Hybrid Cloud**, this model is similar to the private cloud as it is managed by third party or by organization itself and may exist off or on premises. But the cloud infrastructure may combine two or more clouds (public, private or community).
- **Community Cloud**, this model is similar to the previously mentioned private and hybrid cloud as the organization or third party are allowed to manage it and also exists off or on premises. But in community cloud, multiple organizations with common interests, requirements, or considerations share the infrastructure.

The security of the cloud needs testing, it is important for organizations that want to ensure the optimal product before distributing it. The results are used in finding out security weakness points and to patch them before the occurrence of penetration. However organizations' lack of time and resources, computer related crime is usually on the rise. Consequently penetration investigators (testers) have to reduce the amount of resources. This motivates testers to widely adopt automatic tools, as it is demonstrated by the continuous release of platforms finalized to automate this process, discovering gaps in compliance, verifying secure

configurations, finding holes now before somebody else does, Report problems to management and testing new technology.

IV. GAPS AND SECURITY ISSUES IN SERVICE MODELS

Although cloud computing has huge promising future but unfortunately it had not been adopted in enthusiasm and pace manner by the customers. This may refer to the reality of the existing gaps. The National Institute of Standards and Technology (NIST) [9] confirmed that security, portability and interoperability are the major barriers to wide adoption of cloud computing. Armbrust et al. [3] identified 10 major obstacles to cloud computing as follows: data lock-in, data confidentiality, availability of service, and audit ability, performance unpredictability, bugs in large distributed systems, data transfer bottlenecks, scalable storage, reputation fate sharing, scaling quickly, and software licensing. Ness [10] determined three major obstacles to cloud computing given by: first, cloud can break static networks; second, cloud is based on the new security approaches; and third, the criticalness of network automation.

Leavitt [11] mentioned six barriers as follows: latency and reliability; control; performance; vendor lock-in and standards; related bandwidth costs; security and privacy; and transparency. There may be many methods for defining gaps, and many parties are also embedded other than customers and cloud providers and. But, practically, what real situation at the end is that it refers to the customer whether he/she or his/her company is desire to join the cloud. The reputation of a company and the type of services expectations one is going to receive from a certain provider are the basic elements in selecting a cloud provider. According to [3, 7-11], it is possible to define cloud computing gaps as follows: The factors that slow down joining cloud computing from the current system are defined as gaps of cloud computing.

Fig 2 shows the gaps between expectations and perceived services by cloud customers' based on our understanding [2-5, 7-14]. Also a gap between customers' expectations and deliverable services has been witnessed. In our opinion, many of the potential clients are aware of this gap and consequently, they are waiting on the sidelines. Convincing these customers (clients) that the cloud will meet their expectation will encourage them to join the cloud computing. [2, 7, 8].



Fig. 2. Cloud Computing Risks [55]

According to the recent survey by Cloud Security Alliance (CSA) & IEEE, we can conclude that guaranteeing the security of corporate data in the "cloud" is difficult. There are different levels of security required by the different service models in cloud environment. **IaaS** represent the base of all

cloud service, upon which the **Paas** is built and thus **SaaS**, in turn, is built upon the **Paas**, this is shown in Figure 2. Tradeoffs should be taken into consideration for each model in terms of complexity and integrated features versus the security and extensibility. This means that the cloud service provider should take all aspects in account and should not concentrate only on security only at the lower part of the architecture of security as this may lead to make consumer more liable for managing and implementing the capabilities of securities [41, 42].

Each service has its own security issues [43]. The SaaS model provides the customers with important benefits, such as efficiency, reduced costs and improved operational. But according to the Forrester study, "The State of Enterprise Software: 2009," security concerns are the most commonly cited reason why enterprises are not interested in SaaS. Thus, enterprise security appears to the strongest challenge for adapting SaaS applications [45]. Regarding the security of IaaS, the only basic security measures introduced by IaaS are (perimeter firewall, load balancing, etc.) but these measures are not enough as applications, that move into the cloud, need higher levels of security that are supplied by the hosts [43-45]. Despite the various advantage of the PaaS layer, it has key disadvantage represented in that, these advantages itself can be used by the hackers to expose the PaaS cloud infrastructure to malware control, command and going beyond IaaS applications [46-48].



Fig. 3. Understanding cloud computing [8]

A. Denial-of-service attack classification

It is very important to understand and determine the most probable method used by attacker for attacking the application or a network [49, 50]. Determining the location of weakness points in network or application defenses is important as this will help in knowing how an attacker could use these weaknesses [38-40], this is shown in Figure 4.

Machine learning techniques can be used successfully for classification of any activity depending on predefined classes. Machine learning techniques are available from the computational intelligence community Figure 4. From the available list of algorithms in machine learning, we have selected Naive Bayes [32], multilayer perceptron [33], support vector machine [34], decision tree (C4.5) [35] and Partial Tree (PART) [36] for classifying our data. Naive Bayes is a probability-based technique, multilayer perceptron and support vector machine are function estimation based techniques, and decision tree and PART are rule-based machine learning techniques. All these techniques have been implemented in Weka [37], which is a Java-based popular

machine learning tool. Weka uses C4.5 [37] algorithm for decision tree implementation.

1) Tree Augmented Naive Bayes. This algorithm can be mainly used for the classification processes. It efficiently creates a simple Bayesian network model. The model is an improvement over the naïve Bayes model as it allows for each predictor to depend on another predictor in addition to the target variable. The main advantages of this algorithm are the accuracy of its classification and efficient performance compared with general Bayesian network models. The major disadvantage of this algorithm can be summarized in that although its simplicity; it generates more restrictions on the uncovered dependency structure among its nodes [32].

2) Multilayer perceptrons. The neurons are arranged in layers in such networks. Typically, one layer is assigned as input layer for the neurons, on the other hand, one or more layers are assigned for internal processing units that represent the hidden layers, and another one layer is assigned for neurons output that represents the output layer. There are interconnection between the different layers, e.g., in a network with an input layer, a single hidden layer, and an output layer, each neuron in the input layer should be connected to all neurons in the hidden layer, and each neuron in the hidden layer is connected to each one in the output layer. The strength of influence one neuron has on another can be determined through giving weights for the connections between neurons. The prediction generation can be obtained through information flows from the input layer through the processing layer(s) to the output layer. Adjusting the weights of connection during training leads to cope predictions to target values for specific records, the network “learns” to generate better and better predictions [33].

3) The Support Vector Machine (SVM) is a technique used for supervised learning that reproduces input-output functions for mapping from a group of training labeled data. The mapping function can be either a regression function or a classification function. For classification, nonlinear kernel functions are usually exploited to transform input data to a high-dimensional feature space in which the input data become more separable compared to the original input space [34].

4) Decision tree (C4.5); the voting for boosted C4.5 classifiers' algorithm is as follows, For each record, each composite classifier (decision tree or rule set) assigns a confidence and a prediction. The sum of confidence figures for each output value is computed, and the value with the greatest confidence sum is selected as the final prediction [35].

5) Partial Tree (PART): This algorithm provides only a partial specification of the data. A model of executable data should always contain the binary type for each field so, that the output and input data can be marshaled in correct manner. The sufficiently input model that is specified to allow a peer to compute plan of execution is called executable for that peer [36].

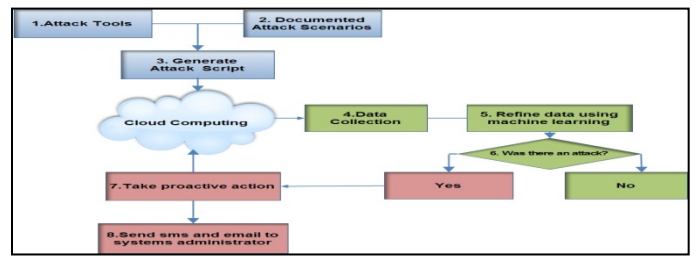


Fig. 4. Attack detection and proactive resolution in single cloud environment Using machine learning (modified by the authors) [8]

All the above algorithms have been implemented in Weka [37], which is a Java-based popular machine learning tool. C4.5 [37] algorithm has been used in Weka for implementation of decision tree [51]. At the beginning, some experimental tests have been carried out in order to determine the best-suited technique for classifying the attack. However, we suggest using C4.5 algorithm “this algorithm needs further explanation” in the cloud system, because it is a comparatively established algorithm and is computationally cheaper than PART. The next selection for our task “what do you mean by this” is multilayer perceptron.

TABLE I. CLASSIFICATION ACCURACY OF DENIAL-OF-SERVICE ATTACK

	Naïve Bayes	Multilayer perceptron	Support vector machine	part	Decision tree
Classification accuracy (%)	75	92.0	92.45	93	94
No. of unclassified instances	0	0	0	0	0
Model building time (s)	0.02	4.55	0.67	0.11	0.06
Model testing time (s)	0.01	0.01	0.01	0.01	0.00

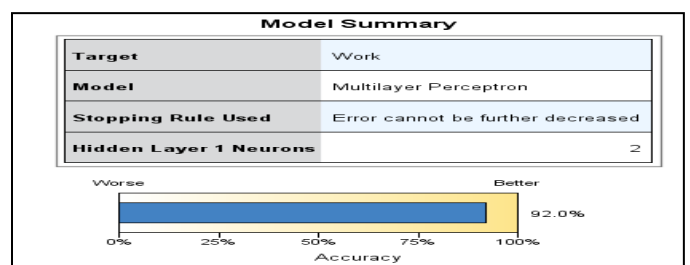


Fig. 5. Classification accuracy

At the beginning, we carried out some experimental tests to identify the best-suited technique for attack classification. The details of performances are available in Table 1. We primarily consider classification accuracy, number of unclassified instances and computational complexity. The classification accuracy calculated the percentage of activities that were classified correctly by the machine learning techniques. The number of unclassified instances basically

measured the technique's limitations, which means it failed to classify any attack as shown in Figure 5. We are also aware of the computational efficiency of the techniques and how well they learn because we are dealing with comparatively large data sets. Therefore, we observe the model building and testing time, which are listed in Table 2.

On the basis of the classification accuracy, number of unclassified instances and computational complexity, we found decision tree C4.5 could be a preferred choice for DoS attack classification in the cloud computing area. The classification accuracy and number of unclassified instances essentially summaries the average performances of the techniques for our attack classification task. So we tried to observe the details of performance about the attack classification scenario. As a result, we employed confusion matrix [38] analysis to see the details of the techniques' performance measures.

B. C4.5 Decision Tree Algorithm

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. [35, 40] has developed C4.5 algorithm. A large tree can be constructed by C4.5 taking into account all attribute values and finalizes the decision rule by pruning [38]. This algorithm uses a heuristic methodology for pruning, depending on the statistical significance of splits [39]. The process of tree construction essentially calculates information gain and the entropy to finalize the decision tree. Depending on this gain information, the C4.5 can determine the occurrence or non-occurrence of an attack. The expected information or entropy depends on the set partitioning process into subsets by the equation [1]:-

$$E(S) = - \sum_{j=1}^n f_s(j) \log_2 f_s(j) \quad (1)$$

Where:-

- E(S) is the subset information entropy (S);
- n is the number of different values of the attribute in S (entropy is computed for one selected attribute);
- $f_s(j)$ is the frequency (proportion) of the value j in the subset S; and
- \log_2 is the binary logarithm. Entropy of (0) defines a perfectly classified subset, whereas (1) indicates a completely random composition. Entropy is used for determining the next node to be split in the algorithm. This means that raising the entropy, leads to increase in the potential to improve the classification.

The encoding information that would be gained by branching on A is given by the following:-

- G(S, A) is the gain of the subset S after a split over the A attribute;

- E(S) is the information entropy of the subset S;
- M is the number of different values of the attribute A in S;
- $f_s(A_i)$ is the items frequency that possess A_i as a value for A in S;
- A_i is the i^{th} possible value of A; and
- S_{A_i} is a subset of S that contains all items, where the value of A is A_i .

Gain quantifies the entropy improvement through splitting over an attribute: higher is better. For to constructing the final decision tree, the algorithm computes the information gain of each attribute [40].

$$E = \frac{1}{N} \sum_{i=0}^N E_i \quad (2)$$

We build a model based on data mining for evaluating the security state of cloud computing through simulating an attack from a malicious source. This process involves identification and utilization of vulnerabilities in real world scenario which may occur in the cloud due to improper configuration, known or unknown weaknesses in software systems, or hardware, operational weaknesses or loopholes in deployed safeguards.

We will use how strategy of inferring and analyzing the data, searching for them in the cloud by one of the technology tools (data mining) this paper shows the vision of the insurance. and the general arrangement for extracting the required data, through the cloud, enabling fighting terrorism to limit the harms in advance by making the relief arrangements from the view of comprehensive security and through the analysis of the results for the data survey.

This process of assigning predictions to individual records is known as scoring. By scoring the same records used to estimate the model, we can evaluate how accurately it performs on the training data—the data for which we know the outcome. This example uses a decision tree model, which classifies records (and predicts a response) using a series of decision rules.

C. Testing and verification of security and Integrity using a simple decision tree model

Using a simple decision tree model Chaid algorithm security rating for classifying the data including the fields of entry or the variables, Decision tree is the structure of the tree on the shape of tree branches that represents sets of decisions. These decisions generate rules for classification of the set of the data. It includes limited forms for the branches of the branches, which includes the decision of classification, or decline, it includes the space of the automatic discovery of the mistakes.

TABLE II. FIELD NAME DESCRIPTION

Data security	N	Marginal Percentage
sec	Y	13
	No	12
variable	X1	6
	X2	11
	X3	8
Valid	25	100.0%
Missing	0	
total	25	
subpopulation	25(a)	
a. The dependent variable has only one value observed in 25(100.0%)		

Table 2. shows the Coding Input data (0, 1) and the independent variables [(x_1, x_2, x_3, x_4, x_5) & Y]

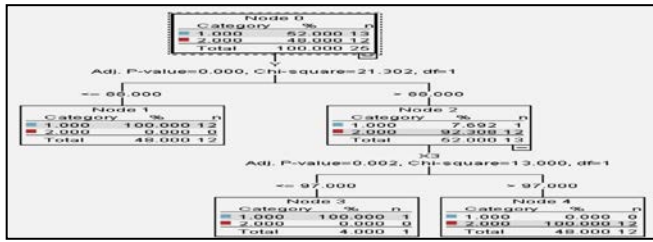


Fig. 6. Analyses of Decision Tree Model “by the authors”

Results and discussion: Figure 6. shows the upper part of the first node of the tree in C4.5, it summarizes all the records in the set of the data. We can find the rate of secured data is 48% that corresponds to 12 scores only in the cases of the set for the data sample representing the secured data (that has protection). While the rate of unsecured data is 52% that represents 13 scores for the data exposed to risk (unsecured), it needs to improve the performance for protection and security, it is exactly the first part of the analysis, so let us see if each tree can give us any evidence to what are the factors that may be responsible for the attack. Figure. 5 shows that the first division is according to the level of the input data. So, it will be possible to assign or to determine the scores on terms that the income level in allow class to (node 2) it is not surprising to see that this classification contains the highest rate of the unsecured data, it is a clear indicator for the data of this class, to contain high risks and needs a solution thus the rate of 52% for the data of this class represents a risk actually, if not supposedly, consequently, the prediction model practically cannot respond but that the model must be good and allow us to expect and to respond more likely for each score based on the available data by the same way . According to the analysis of node (2) shown in Figure 5, we can find that the vast majority (92.308%) appears unsecured that represents a risk and needs to set a new mechanism for security. So, the standards of security can be improved in this set of data to reduce the risk, accordingly we learned that each score is an indicator for this model. We will determine the points of weakness in the cloud through assigning particular node. The new predictions assignment (either good or bad), depending on the prevailing response for this node, this process is known for assigning the predictions of the individual scores as it is the aim, by recording the same scores that are used for assessing the model. According to the percentage, we can assess the extent of accuracy for the

training data. This model is used for the decisions tree that classifies the scores. It's expected the response by using a group of rules for taking the decision.

TABLE III. CASE PROCESSING SUMMARY

Field name	Description,
Input variable	(x_1, x_2, x_3, x_4, x_5) & Y
Security rating	Security rating : 0 = attack 1 = security
Data risk	Number of test range of security 1 = < 88.00 , 0 > 88.00

Table 3. Shows the cloud needs to be improved and to enhance its sufficiency and taking the necessary arrangements to raise the efficiency of the security. As the data is exposed for the occurrence of violations at the rate of 48.0% is no secure.

V. CONCLUSION AND FUTURE WORKS

Although cloud computing is a new emerging technology that introduces a number of benefits to the users, but unfortunately it faces lot of security challenges. In this paper data security challenges and solutions are provided for these challenges in order to overcome the risk included in cloud computing. In this paper, a review on cloud computing with the main focus on gaps that hinders cloud adoption has been undertaken, and at the same time, a review about threat remediation challenges has been mentioned.

The paper presented the performance of machine learning techniques used in attack identification in a cloud computing environment. A statistical ranking approach has been used for the final selection of a learning technique for the task. C4.5 technique's performance has been evaluated through different performance evaluation matrices that included the rigorous testing of 10-fold cross-validation, true positive rate, false positive rate, precision, recall, F-measure and the area of receiver operating characteristic. In another phase, we also counted computational complexity for our final selection.

Our experimental results showed that, using a simple decision tree model Chaid algorithm security rating for classifying approach is a robust technique that enables the decision-maker to measure the extent of cloud securing, show that the cloud needs to be improved and to enhance its sufficiency and taking the necessary arrangements to raise the efficiency of the security indicated the fact that C4.5 gives a better performance and the level of performance has acceptable standard. It is found that rule-depending technique (C4.5) is efficient technique for solve the problem of security. However, on the basis of the computational performance, we suggest C4.5 as the better technique for real-time attack protection in a cloud environment.

The paper presented some recommendation regarding the customers and vendors; however, to overcome the customer concerns about application and data security, vendors should deal with these issues head-on. In the future, concrete standards for cloud computing security should be improved. In the future, we will continue and follow up the study in this field through using search in data to be an active way in

decision making. It is expected that there will be several challenges related to operation and development of cloud computing system. The use of data mining techniques in cloud computing will be an effective tool that will help in securing the data.

REFERENCE

- [1] Schubert L, Jeffery K, et al. The future for cloud computing: opportunities for European cloud computing beyond 2010. Expert Group report, public version 2010; 1. <http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf>
- [2] Khorshed MT, et al. Trust issues that create threats for cyber attacks in cloud computing. In Proceedings of IEEE ICPADS, December 7–9, 2011, Tainan, Taiwan, 2011.
- [3] Armbrust M, et al. Above the clouds: a Berkeley view of cloud computing. EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009- 28, 2009.
- [4] Brunette G, Mogull R. Security Guidance for critical areas of focus in Cloud Computing V2. 1. CSA (CloudSecurity Alliance), USA. Disponible en: <https://cloudsecurityalliance.org/csaguide.pdf>, vol. 1, 2009.
- [5] Catteddu D, Hogben G. Benefits, risks and recommendations for information security. European Network and Information Security Agency (ENISA), 2009.
- [6] Mell P, Grance T. The NIST definition of cloud computing. National Institute of Standards and Technology 2009; 53(6): 50. <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>
- [7] Khorshed MT, et al. Monitoring insiders activities in cloud computing using rule based learning. In Proceedings of IEEE TrustCom-11, Nov. 16–18, Changsha, China, 2011.
- [8] Khorshed MT, et al. A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing. Future Generation Computer Systems 2012.
- [9] NIST. (2011, 21 May 2011). NIST Cloud Computing Program. Available: <http://www.nist.gov/itl/cloud/>
- [10] Ness G. (2009, 22 May 2011). 3 Major Barriers to Cloud Computing. Available: <http://www.infra20.com/post.cfm/3-major-barriers-to-cloud-computing>
- [11] Vouk, M.A., "Engineering of Telecommunications Software", High-Speed ... Journal of Computing and Information Technology, Vol 16 (4), 2008, pp 235-246.
- [12] Brodtkin J. Gartner: seven cloud-computing security risks, 2008. Available: <http://www.infoworld.com/d/security-central/gartner-seven-cloud-computing-security-risks-853> (Retrieved: 6th August 2012).
- [13] Archer J, Boehm A. Security guidance for critical areas of focus in cloud computing. Cloud Security Alliance 2009. Available: <https://cloudsecurityalliance.org/guidance/csaguide.v1.0.pdf>
- [14] Archer J, et al. (2010, 7 May 2011). Top Threats to Cloud Computing, Version 1.0. Available: <http://www.cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf>
- [15] Monfared AT. Monitoring intrusions and security breaches in highly distributed cloud environments. 2010.
- [16] Grosse E, et al. Cloud computing roundtable. Security & Privacy, IEEE 2010; 8: 17–23.
- [17] Wrenn G. (2010, 25 May 2011). Unisys Secure Cloud Addressing the Top Threats of Cloud Computing. Available: http://www.unisys.com/unisys/common/download.jsp?d_id=1120000970002010125&backurl=/unisys/ri/wp/detail.jsp&id=1120000970002010125
- [18] Grobauer B, et al. Understanding cloud-computing vulnerabilities. IEEE Security and Privacy 2010; 50–57. DOI: 10.1109/MSP.2010.115
- [19] Thomas Sommer, et al. The Conundrum of Security in Modern Cloud Computing. (2012) Communications of the IIMA: Vol. 12: Iss. 4, Article 2. Available at: <http://scholarworks.lib.csusb.edu/ciima/vol12/iss4/2>
- [20] Yildiz M, et al. A Layered Security Approach for Cloud Computing Infrastructure. Publisher IEEE, 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks, Kaohsiung, 2009; 763–767. DOI: 10.1109/I-SPAN.2009.157
- [21] Dahbur K, et al. A survey of risks, threats and vulnerabilities in cloud computing. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, ISWSA '11, ACM, New York, USA, 2011; 12.
- [22] Wang C, et al. Ensuring Data Storage Security in Cloud Computing. Publisher IEEE, 17th International Workshop on Quality of Service, 2009. IWQoS, Charleston, SC, 2009; 1–9.
- [23] Yan L, et al. Strengthen cloud computing security with federal identity management using hierarchical identity-based cryptography. Cloud Computing 2009; 5931: 167–177. DOI: 10.1007/978-3-642-10665-1_15
- [24] Ristenpart T, et al. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. In Proceedings of the 16th ACM conference on Computer and communications security Chicago, Illinois, USA, November 09–13, 2009; 199–212.
- [25] Chonka A, et al. Cloud security defence to protect cloud computing against HTTP-DoS and XML-DoS attacks. Journal of Network and Computer Applications 2010. DOI: 10.1016/j.jnca.2010.06.004
- [26] Danchev D. (2011, 31 May 2011). Dancho Danchev's Blog—Mind Streams of Information Security Knowledge. Available: <http://ddanchev.blogspot.com/>
- [27] Grossman J. (2011, 19 June 2011). Jeremiah Grossman. Available: <http://jeremiahgrossman.blogspot.com/>
- [28] Company H.-P. D. HP ProLiant DL380 G4 server - specifications, 2012. Available: <http://h18000.www1.hp.com/products/servers/proliantdl380/specifications-g4.html> (Retrieved: 6th August 2012).
- [29] Knorr, E., & Gruman, G. (2009). What cloud computing really means. Retrieved from <http://www.infoworld.com/d/cloud-computing/what-cloud-computing-really-means-031>
- [30] Corporation M. Windows 7, 2012. Available: <http://windows.microsoft.com/en-au/windows7/products/home> (Retrieved: 6th August 2012).
- [31] McDowell M. (2009, 21 June, 2011). Understanding Denial-of-Service Attacks. Available: <http://www.uscert.gov/cas/tips/ST04-015.html>
- [32] John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995; 338–345.
- [33] Lopez R, Onate E. A variational formulation for the multilayer perceptron. Artificial Neural Networks—ICANN 2006 2006; 4131: 159–168. DOI: 10.1007/11840817_17
- [34] Platt JC. Fast training of support vector machines using sequential minimal optimization. 1999; 185–208.
- [35] Quinlan JR. C4. 5: Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA, 1993.
- [36] Frank E, Witten IH. Generating accurate rule sets without global optimization. In Fifteenth International Conference on Machine Learning, 1998; 144–151.
- [37] Witten IH, et al. Data Mining: Practical Machine Learning Tools and Techniques (3rd edn). Morgan Kaufmann: San Francisco, 2011.
- [38] Kohavi R, Provost F. Glossary of terms. Machine Learning 1998; 30: 271–274.
- [39] Ali ABMS, Wasimi SA. Data Mining: Methods and Techniques. Thomson. 2007.
- [40] Quinlan JR. Induction of decision trees. Machine Learning 1986; 1: 81–106.
- [41] Ali ABMS, Smith KA. On learning algorithm selection for classification. Journal on Applied Soft Computing, Elsevier 2006; 6: 119–138.
- [42] Shafiullah G, et al. Prospects of renewable energy—a feasibility study in the Australian context. Renewable Energy, ELSEVIER 2012; 39(1): 183–197.
- [43] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner. A break in the clouds: towards a cloud definition, in: ACM SIGCOMM Computer Communication Review, 2008.p.50-55.

- [44] M.B. Mollah, K.R. Islam, and S.S. Islam. Next generation of computing through cloud computing technology, in: 2012 25th IEEE Canadian Conference on Electrical Computer Engineering (CCECE), May 2012.p.1-6.
- [45] Shucheng Yu, Cong Wang, Kui Ren, and Wenjing Lou. Achieving secure, scalable and fine-grained data access control in cloud computing, in: IN-FOCOM, 2010 Proceedings IEEE, 2010.p.1-9.
- [46] Kresimir Popovic and Zeljko Hocenski. Cloud computing security issues and challenges, in: MIPRO, 2010 Proceedings of the 33rd International Convention, 2010.p.344-349.
- [47] Akhil Bhel, Emerging Security Challenges in Cloud Computing. Information and Communication Technologies, in: 2011 World Congress on, Mumbai, 2011.p.217-222.
- [48] Farzad Sabahi. Cloud Computing Security Threats and Responses, in: IEEE 3rd International Conference on Communication software and Networks(ICCSN), May 2011.p.245-249.
- [49] Eman M.Mohamed, Hatem S Abdelkader, Sherif EI Etriby. Enhanced Data Security Model for Cloud Computing, in:8th International Conference on Informatics and Systems(INFOS), Cairo, May 2012.p.12-17.
- [50] Wentao Liu. Research on Cloud Computing Security Problem and Strategy, in: 2nd International Conference on Consumer Electronics. Communications and Networks (CECNet), April 2012.p.1216-1219.
- [51] Eystein Mathisen. Security Challenges and Solutions in Cloud Computing, in: International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011), 2011.p.208-212.
- [52] R. Velumadhava Raoa, K. Selvamanib , Data Security Challenges and Its Solutions in Cloud Computing , 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>). Conference Organized by Interscience Institute of Management and Technology, Bhubaneswar, Odisha, India ,Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)
- [53] J. Staten, "Is Cloud Computing Ready For The Enterprise," Forrester ... 17th International workshop on Quality of Service, pp.1-9, July 13-15, 2009.
- [54] Jay Heiser and Mark Nicolett, " An Engineering Process to Address Security Challenges in Cloud Computing" ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, May 27-31, 2014
- [55] Seccombe A, et al. Security guidance for critical areas of focus in cloud computing. Cloud Security Alliance, 2009.
- [56] Paula Kotzé , et al. Secure cloud computing: Benefits, risks and controls, ieeexplore.ieee.org/iel5/6017604/.../06027519.pdf

Identifying Green Services using GSLA Model for Achieving Sustainability in Industries

Iqbal Ahmed

Graduate School of Science and Engineering, Saga University, Japan

Hiroshi Okumura

Graduate School of Science and Engineering, Saga University, Japan

Kohei Arai

Graduate School of Science and Engineering, Saga University, Japan

Abstract—Green SLA (GSLA) is a formal agreement between service providers/vendors and users/customers incorporating all the traditional/basic commitments (Basic SLAs) as well as incorporating Ecological, Economical, and Ethical (3Es) aspects of sustainability. Recently, most of the IT (Information Technology) and ICT (Information and Communication Technology) industries are practicing sustainability under green computing domain through designing green services at their scope. However, most of these services only focused on power consumption, energy efficiency, and carbon emission. Moreover, the sustainability can not achieve without considering 3Es simultaneously. The recent development of sustainable GSLA are assisting to identify the missing green services under 3Es. This research attempts to design all missing green services for sustainability by using global informational model of Green SLA. All these newly identified green IT services could reside with other existing services in the industry. Additionally, the design and evaluation technique of these new green services could be used as a guideline for the ICT engineers and as well as other industries too. Moreover, the evaluation and monitoring of new green services are justified using general questionnaires design and analytical tools among the 20 startup ICT industries in Bangladesh and Japan. The proposed idea of designing new green services and their justification methods would be helpful for the ICT engineer to practice sustainability in their competitive businesses.

Keywords—GSLA; Green Services; GaaS; Sustainability; Informational model

I. INTRODUCTION

Currently, cloud and grid computing and many data centers act as most promising service providers. These computing and communication industry provides different services in compare to traditional computing with some scalability benefits. At the same time, cloud services are offered at various levels: Infrastructure, Platform and Software as a Service [1, 2]. At each level, they maintain a SLA and or GSLA with their parties [2]. Therefore, this shows the growth rate of GSLA in recent time as well as the need of introducing green services for sustainability [3]. Presently, the revolution of ICT and IT in average daily life has also resulted in the increase of Green House Gas (GHG), due to a continual increase in “carbon footprint” [4]. If ICT has a negative impact on the environment, it can use for greening the other human activities (logistic, city, industry, etc.) in the society [4, 5]. Indeed, the dimensions of Green Informatics contributions are: the reduction of energy consumption, the rise of environmental awareness, the effective communication of environmental issues and the environmental monitoring and

surveillance systems, as a means to protect and restore natural ecosystems potential [6]. At the same time, many ICT companies or service providers need to think about their business scope in the light of green and sustainable perspective [7, 8]. The GSLA research assists to understand the sustainability achievement from customers/users and service providers side for upcoming sustainable society [3, 9]. According to Gartner (2015), green IT services refer to the development of green IT to enable organizations in creation, management, and optimization of or access to information in the business process [10]. However, develop a standard green IT services under sustainability domain is still a challenging task for the ICT engineer. Now it is timely to conduct another layer of Green IT as a services (GaaS)with their existing service infrastructure considering 3Es of sustainability (Fig.1). According to figure 1, the interaction of 3Es of sustainability makes it difficult to design new green services on top of it. All these new green indicators of GSLA under ecology, economy and ethics are already identified by the previous study [3,9].

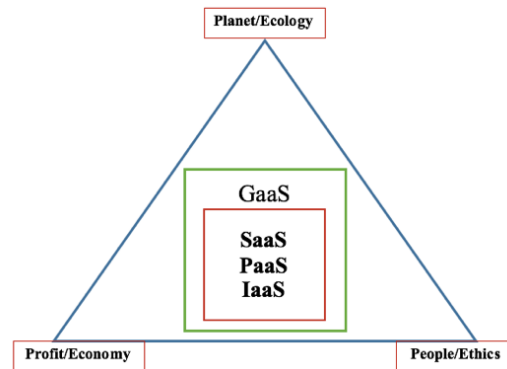


Fig. 1. Overview of Green IT as a Service (GaaS) under sustainability

This research identified new green services from a general model of sustainable GSLA [9] and GSLA is the only way to understand the importance of sustainable development in both sides (customers and providers). These identified seven new green services from the GSLA model [9] are shown in next Table 1. The table also depicts the cascading relationships (*direct, indirect important and indirect small effects*) between other parameters under 3Es of sustainable GSLA model [9]. All these newly identified green services are modeled using UML notation under GSLA hood, thereby interacting 3Es of sustainability. Moreover, this research attempts to justify “Total Recycling Services” using general questionnaires and their feedback analysis. The analytical results suggest that, new ICT industries are considering the green services for

achieving sustainability though there are still some interaction gaps with the proposed designed green services.

TABLE I. RELATIONSHIPS BETWEEN ALL SERVICES DEFINED FROM THE GSLA INFORMATIONAL MODEL[9]

Identified new green services	Relationships with other indicators under 3Es		
	Direct	Indirect Important Effects	Indirect Small Effects
Total Recycling	ICT Product Life Cycle; eWastage; Earth Pollution; Energy Consumption; GHG Emission; Energy Cost; Dismantling ICT Product.	ICT Radio Wave; ICT Toxic Material Usage; ICT Product Life Cost.	Comfort Pollution
Obsolescence Indication	ICT Product Life Cycle; ICT Performance; ICT Product Life Cost.	Pollution Level; Energy Consumption; GHG Emission	Ethics Pillar
GHG Emission	Total Recycling; Air Pollution; Non-renewable Energy; Carbon Taxation; Dismantling ICT Product; Energy Consumption.	Obsolescence Indication; ICT Toxic Material Usage	Comfort Pollution
Energy Consumption	Total Recycling; ICT Product Life Cycle; ICT Product Life Cost; Energy Cost; Carbon Taxation; GHG Emission; Renewable Energy; ICT Radio Wave.	Obsolescence Indication; Cooling Cost	Civil Engineering Cost
Pollution Level	ICT Product Life Cycle; Total Recycling; ICT Radio Wave; GHG Emission; Energy Type.	ICT Toxic Material; Obsolescence Indication; Energy Consumption.	Ethics Pillar
ICT Product Life Cycle	Energy Consumption; ICT Product Life Cost; Pollution Level; Obsolescence Indication; Total Recycling; Energy Cost.	GHG Emission	
Energy Cost	ICT Product Life; ICT Product Cost; Carbon Taxation; Energy Consumption;	Cooling Cost; Civil Engineering Cost;	Total Recycling;

The rest of the work is organized as follows- the next section elaborate the importance and designing of all newly identified green services under sustainability hoods. The evaluation and discussion section discovers the justification of proposed designed model of *Total Recycling Services* from various ICT startup industries from Bangladesh and Japan. In addition, this section also highlights the evaluation techniques of these non-technical green services. Finally, the conclusion gives a brief discussion about few challenges and plan of this green service designing and implementation in the industry.

II. DESIGNING GREEN SERVICES FROM GSLA INFORMATIONAL MODEL

The sustainable GSLA definition [3] and global informational model [9] helps to identify the complexity of managing all GSLA parameters by taking some of the important services from sustainability pillars and existing green computing practice. All these central entities have direct and indirect relationships for evaluating and assessing all existing performance parameters of the proposed global GSLA model [9]. Additionally, choosing central entities might

also help the ICT designer to view and design new services for the users. Moreover, these new services actually cover all the dependencies and respect all other existing and new indicators under three pillars of sustainability (Table I) and traditional green computing practice in IT industry. The rest of the work organizes all these services showing its direct relationships and indirect important and small effects with other entities using a UML notation. Therefore, this research identifies following central entities as new services in the future sustainable industry, - *Total Recycling, Obsolescence Indication, GHG Emission, Energy Consumption, Pollution level, ICT Product Life Cycle* and *Energy Cost*. All these new services need to consider for achieving sustainability as these are the missing services proved from previous studies [3, 9].

A. Total Recycling Services

Total Recycling has interrelationships with other existing and new indicators in the model (Fig.2). While recycling an item or ICT product, it could emit GHG directly into the atmosphere. Moreover, it has direct impact and relation with *Earth Pollution* entities (Air, water, and soil). For example, Cathode Ray Tube (CRT) used in computer monitors could emit lead, barium, and other heavy metals into the ground water and release toxic phosphor into the air [11]. Again, computer and networking wires could also recycle for extracting copper using open burning and stripping method, which creates hydrocarbon ashes released into the air, water and soil in the environment [11]. The *Air Pollution* entity is directly related with *GHG Emission* in proposed model. Recycling has a direct relationship with the *eWastage* entity. Recycling helps to the reduce global magnitude of e-waste as metals, plastics, glass and other materials could be recovered from ICT product through recycling procedure. *eWastage* entity has direct impact on *GHG Emission* and *Energy Consumption* entity in this model. Moreover, to recycle a product or equipment, it needs to consume energy or power and cost of energy need to consider. *Total Recycling* has direct relations to calculate existing *Energy Consumption, Energy Cost* indicators. *ICT Toxic Material Usage* makes recycling indicator more complex. Most of the toxic materials used in ICT industry have important indirect effects to the *Comfort Pollution* entity because the dumping or recycling procedure might irritate people's comfort through noise or visual pollution and also responsible for health hazards. For example, in fluorescent tubes, flat screen monitors, etc. mercury and its compound is used. This toxic material affects human health including sensory impairment, dermatitis, memory-loss, and muscle weakness [11]. Sulfur and lead are also commonly used in lead-acid batteries, might responsible for acute health problems such as liver, kidney, heart damage, behavioral disturbances, attention deficits and lower IQ [11]. Besides all these comfort level pollution, toxic material usages also have a direct relation with other earth pollution entity. Earth pollution of mercury and its substance affects plants, trees by reducing soil's fertility rate and thus slower their growth and development [11]. Also, when Sulfur released, it could create Sulphur dioxide, which is responsible for acid rain [10]. *ICT Radio Wave* could measure and monitor through standard value of EMF (Electromagnetic Frequency) or SAR (Specific Absorption Rate) value for some specific domain such as network or the internet etc. SAR is a measure of the rate at

which energy is absorbed by the human body when exposed to radio frequency; SAR also defines as power absorbed per

mass tissue and has units of Watts per kilogram (W/Kg) [12].

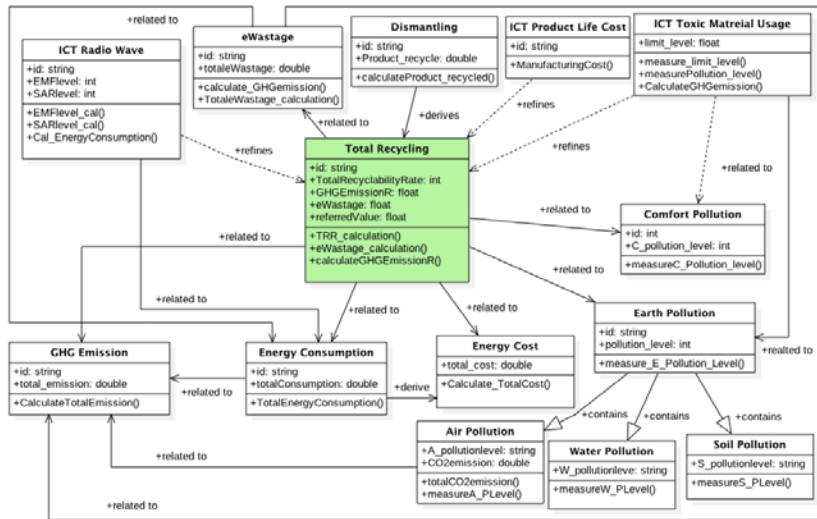


Fig. 2. Total Recycling Service

Also, *ICT Radio Wave* have direct relation with *Energy Consumption*, as to reduce the power of wave, it also requires installing more antenna; which consumes more energy and thus more money for that consumed energy and at the same time, more antenna increases the problem of equipment recycling. *Total Recycling* service has impact on *Economic Pillar* of sustainability as the more product would be recycled, the more money could gain. However, the cost of energy and other necessary costs to recycle ICT product also need to consider here. The *Dismantling* entity from *ICT Product Life Cycle* has direct relations with recycle, reuse or refurbish entity [3]. Additionally, *Manufacturing* entity of ICT product life refines total recycling entity as recycling helps to avoid extracting new earth resources as well as minimizes production cost to some extents. Thus, it has relation with economic indicator- *ICT Product Life Cost*. The main challenges to define this new *Total Recycling* services are to gather all necessary information and monitoring their effect. Most of non-technical parameters under sustainability pillar in this service need some laws and directives to derive exact information for the users. There are some standards available for recycling services in the USA (*Responsible Recycling (R2) Practices, e-Steward*) and also some directives such as *WEEE* or *D3E* from the European Union.

B. Obsolescence Indication Service

Obsolescence Indication could be another green service under the ecological pillar of sustainability [3, 9]. Minimum optimum obsolescence could be calculated using some mathematical model design for an ICT product and for the raw materials to produce that product [13,14,15,16]. Additionally, *Obsolescence Management* could also be used to find out the optimum indication for a product to be obsolete [15,16]. However, obsolescence is relative information estimated from other useful existing criteria. It could calculate from cost of energy, carbon/GHG emission, ICT product life cycle assessment, and or pollution level. There is an interesting relation between obsolescence and people. Therefore,

Obsolescence Indication entity has indirect relationship with *Ethics Pillar* entity of *GSLA* model [9]. There is an interesting relationship between existing *User Satisfaction* indicator with this entity. For example, people often change their mobile phone frequently because it might become old fashioned to use it. Moreover, to find out the optimum obsolescence of ICT equipment, the performance of that equipment should need to monitor and evaluate using classical/basic SLA parameters (availability, connectivity, bandwidth capacity, memory, uptime, and etc. for a switch). That's why, obsolescence indication entity need to incorporate the ICT equipment basic performance metric. The next Fig.3 shows the graphical notation of obsolescence indication service. There is still no available standard to define obsolescence indication. *Obsolescence management* of an ICT product could define according to design some regulatory lever, education/training for user behaviors and recycling practice in the society.

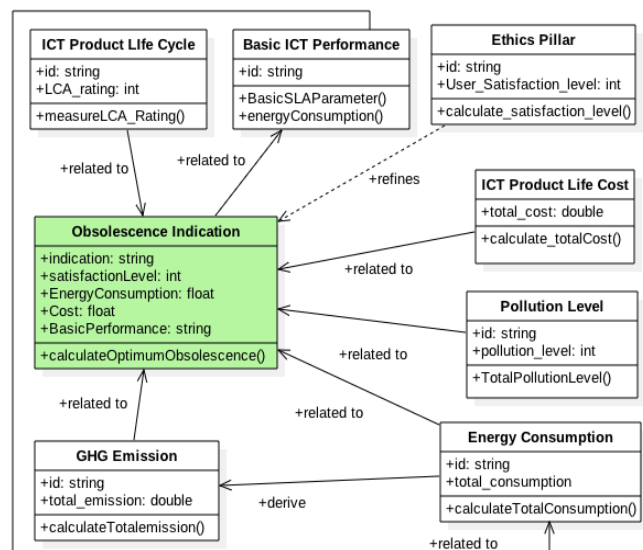


Fig. 3. Obsolescence Indication Service

C. GHG Emission Service:

Currently, Green House Gas (GHG) emission parameter exists in some previously defined GSLA for ICT industry[8]. This global service has a direct impact on the environment under sustainability lens. Fig.4 gives the idea of interrelationships and dependency of GHG Emission service. Air Pollution entity from ecological pillar has direct relationship with GHG Emission in the proposed model as the more air is polluted; the more carbon is emitted into the atmosphere. Additionally, the air is polluted because of carbon emission and this emission related with energy consumption issues.

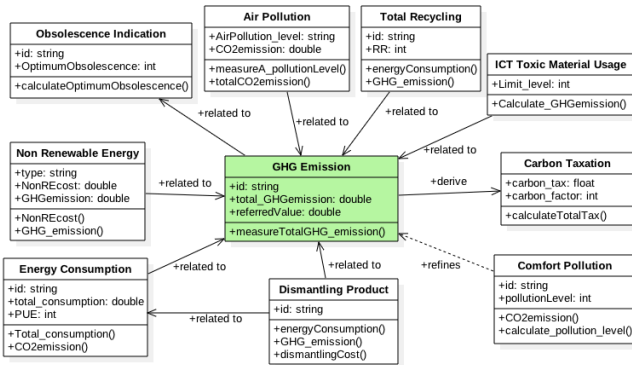


Fig. 4. GHG Emission Service

However, there are two types of energy currently used in most industries,- renewable energy and non-renewable energy. Non-renewable energy such as coal, oil, and natural gas is more responsible for producing GHG emission whereas renewable energy (Solar, Wind, Tidal, Nuclear, etc.) has negligible effects on GHG emission. Therefore, the GHG Emission entity has direct relationship with non-renewable energy type in the proposed model. The Comfort Pollution entity has an indirect small effect on GHG emission. The ICT product which is responsible for creating noise, light or visual pollution under comfort pollution level, might also emit carbon into the atmosphere. Moreover, GHG emission entity has direct relations with an economic entity in GSLA- Carbon taxation. Carbon Taxation usually derived after measuring total carbon emission and carbon factor in any facility/industry. Though carbon factor is varying according to different country’s government rules and regulation, it plays an important role to calculate total carbon tax.

D. Energy Consumption Service:

Fig.5 is demonstrating energy consumption service under green computing domain. This service has close relationships and dependencies to derive some existing indicators. For example, Power Usage Effectiveness (PUE) derived as the ratio between total energy consumption and IT energy consumption [17] and to completely find out total energy consumption; the Energy Consumption service is relating to all other entities in the model. Moreover, ITEE, ITEU [18] and Green Energy Coefficient (GEC) [19] indicators help to find Data center Performance Per Energy (DPPE) [17, 19] in existing green SLAs. GEC calculated from renewable energy source entity in the proposed model. Additionally, ICT Product Life is an important entity for GSLA on energy

consumption issues in the model. The manufacturing, transportation, usage and dismantling entity an ICT product requires energy in each stage. There are some EnergyWise standard ICT products from CISCO, which already used in many industries. The network engineer could easily monitor the real-time energy consumption of devices compatible with this EnergyWise [20] standard. Again, either for recycling, reusing or refurbishing procedure of an ICT product or equipment also needs energy. Therefore, ICT Product Life Cycle and Total Recycling services have direct and continuous dependencies for calculating energy consumption services. The more energy consumed; the more carbon emitted. The real-time energy consumption and carbon emission correlation observed during PERCCOM air quality project [20]. Thus, GHG Emission also has direct relationship with Energy Consumption entity. ICT Radio Wave has an important indirect effect because to reduce the power of radio, more antenna and other equipment are requires which also consume more energy.

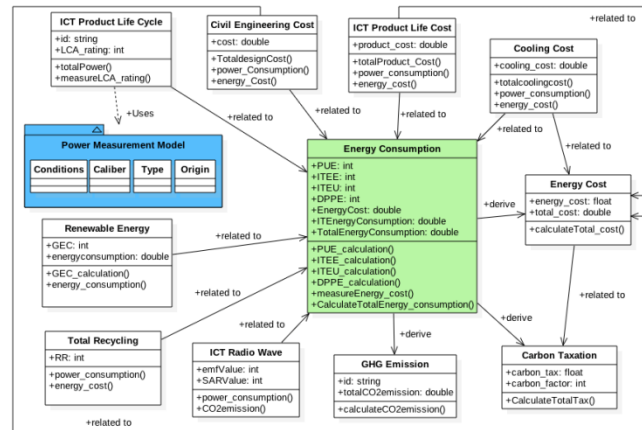


Fig. 5. Energy Consumption Service

Moreover, Energy Cost and Carbon Taxation, both economic entities have a direct impact on energy consumption in IT industry. Civil Engineering Cost, ICT Product Life Cost, and Cooling Cost have important indirect effects on this entity. In each stage of building or infrastructure design, it requires energy/power and also for installing ICT product in any facility needs energy too. In addition, the cooling techniques demand more energy than their ICT equipment in some ICT industry. All these economic entities depend on the calculation of energy consumption first and then their measurement according to different countries energy/power regulations. In addition, the power measurement model [21] could also be helpful to refine energy consumption service. The general relation between energy and power consumption of an ICT device could be derived as [21],

$$energyconsumption = \int_{time} powerconsumption$$

The main purpose of a power measurement model is to introduce a reference value for the power consumption of an ICT device during its life time. The model has relations with other entities, such as power measurement conditions, caliber, power/current type, and origin of power [21]. Here, in the

proposed energy consumption service, the integration of the power measurement model is not shown directly for simplicity. However, the model could be important for evaluating and assessing the total energy consumption of an ICT power enabled device.

E. Pollution Level Service:

Pollution level service is important from ecological aspects of sustainability. *ICT Product Life Cycle*, *ICT Toxic Material Usage*, *Total Recycling*, and *GHG Emission* have direct relationship with pollution level service whereas *ICT Radio Wave* and *Obsolescence Indication* entities have indirect important effects on both earth and comfort pollution. There is an interesting relationship between *Comfort Pollution* entities with ethics pillar in the GSLA model [9] as ethical pollution is mostly concerned with people’s comfort in their daily life. *Noise Pollution*, *Light Pollution* and *Visual Pollution* are the most three important comfort level pollution entities. *Noise Pollution* should need to calculate the standard level of noise in decibels (according to E-OSHA standard); *Light Pollution* might create Computer Vision Syndrome (CVS) [22, 23] on human health and this indicator should need some guideline and modeling to control CVS. *Visual Pollution* could monitor and control according to PAQ (Perception of Affective Quality) rating[24]. The earth pollution level entity is consists of three other entities, - Air, Water, and Soil Pollution and *Air Pollution* is directly responsible for *GHG Emission* in the atmosphere (Fig.2). Air, water, and soil pollution have direct relations with recycling, ICT toxic material usages and ICT radio wave entity. Pollution level central entity and its relationships are shown next Fig.6. Additionally, existing *Carbon Usage Effectiveness (CUE)* indicator is computed here as total carbon emission equivalent from the total energy consumption for any facility [19]. Therefore, this entity has relation to derive *CUE* using the formula, $CUE = CEF \times PUE$; whereas *CEF* is the carbon emission factor, which could vary according to different countries government rules and regulation [20, 25].

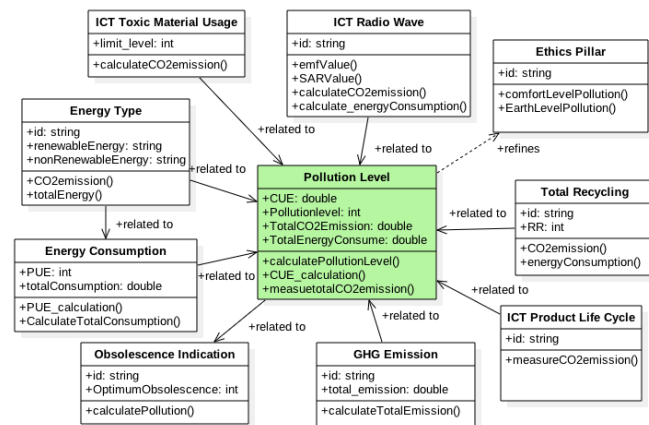


Fig. 6. Pollution Level Service

F. ICT Product Life Cycle Service:

The whole life cycle of an ICT product consists of following entities, - manufacturing, transportation, usage and dismantling entities [3]. All these entities should directly connect to *GSLA* entity to respect global analysis of proposed

model[3, 9]. The total GHG emission, total energy consumption and total costing of energy could not be estimated without considering all these product life cycles entities [26]. Therefore, *GHG Emission*, *Energy Consumption*, *Pollution Level*, *Total Recycling* entities of ecological pillar and *Energy Cost* of the economic pillar has direct relations with *ICT Product Life Cycle* service. Moreover, *ICT Product Life Cost*, which usually consider the production, usage level costs, and initial setup costing; have also an important indirect effect on life cycle’s entity. Fig.7 depicts this central service and shows its corresponding relationship with other entities.

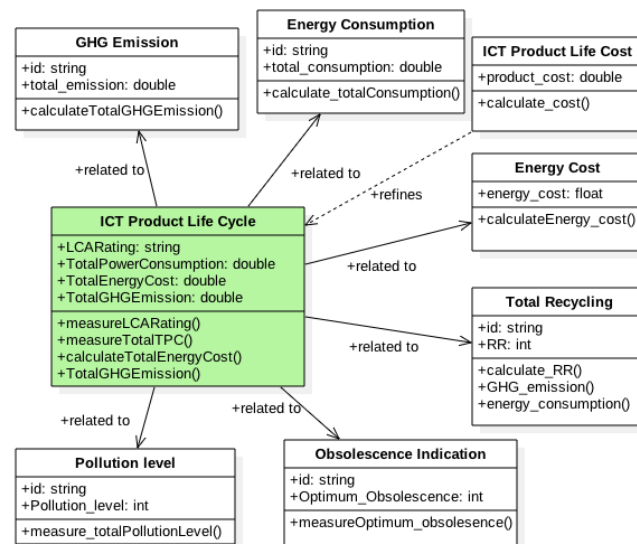


Fig. 7. ICT Product Life Cycle Service

Again, there is also an interesting relation between the dismantling entity of ICT product life cycle and *ICT Product Life Cost* as dismantling could be refined either with recycling, reuse or refurbish entity and the production cost will be reduced. For example, gold could retrieve after reusing old ICT equipment, which helps to reduce the production cost of new ICT product and also there might be no need to explore more earth resources. Therefore, *Dismantling* entity in ICT product life cycle has a direct relationship with *ICT Product Life Cost*. The power consumption model [21] could be used to refine *ICT Product Life Cycle* for evaluating total energy consumption in product’s whole life time, which already depicts in Fig.5.

G. Energy Cost Service:

Fig.8 demonstrates the analysis of Energy Cost service. This service has direct relations with *ICT Product Life Cycle*, *Energy Consumption*, *Carbon Taxation* and *Energy Type*. There are two types of energy is considered in the model, renewable, and non-renewable energy. The costing of energy depends on the types of energy sources used in the ICT facility. However, different types of energy cost generally depend on different countries government rules and regulations and their economic conditions. Again, the carbon tax calculated after retrieving energy cost according to government rules and regulation. At each stage of the life cycle for an ICT product, it demands energy and thus costing of these energy need to consider. Recycling procedure also

requires energy and money but also money could gain after reusing a recycled material for further use. Therefore, *Total Recycling* service also has a direct relation with energy cost entity in the model.

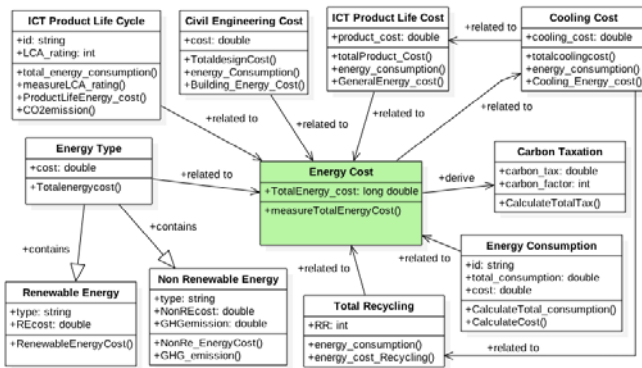


Fig. 8. Energy Cost Service

Moreover, *ICT Product Life Cost*, *Cooling Cost*, and *Civil Engineering Cost* has important indirect effects on energy costing issues. In future, the cost of depollution should need to consider here. *Cooling Cost* entity has important and interesting relations with recycling and ICT product cost issues. Currently, a huge amount of money spent on data center cooling facility, which actually motivates from the work on temperature management [27]. In some case, the idea to reduce the cooling cost is to increase the threshold of temperature acceptable for a server in the data center. However, temperature threshold has negative impacts on server reliability and performance [27]. The consequence of this scenario is that, it needs to change the server and other equipment prematurely in any data center facility. Therefore, the cooling cost entity might have negative effects on ICT product cost entity of economic pillar as it needs more money to buy and install a new server. At the same time, prematurely damaged servers and other equipment could be recycled for further use. Thus, cooling cost entity has direct relationships with the recycling service under the ecological pillar of sustainability [3, 9]. Additionally, the two types of cooling facility also need to consider for evaluating energy cost service in the model. For example, in the data center, natural cooling facility might be more cost-effective and environment-friendly than not a natural cooling facility.

The above figures (Fig.2 to Fig.8) depict the complexity of managing all the performance indicators to define new green services for achieving sustainability in the industries. All these designed services have different levels of relationships and interactions with other entities belonging to three sustainability pillars (Table 1). The next section discovers the real fact of evaluating new green services under sustainable development. This research aims to consider only one green service “Total Recycling” in this regard due to simplicity.

III. VALIDATION AND IMPLEMENTATION OF GREEN SERVICES

It is clearly evident, all newly identified services have a different level of interrelationships among them. It is important to mention that, all the relationships regarding

newly identified services (Table 1) are important to respect sustainable achievement.

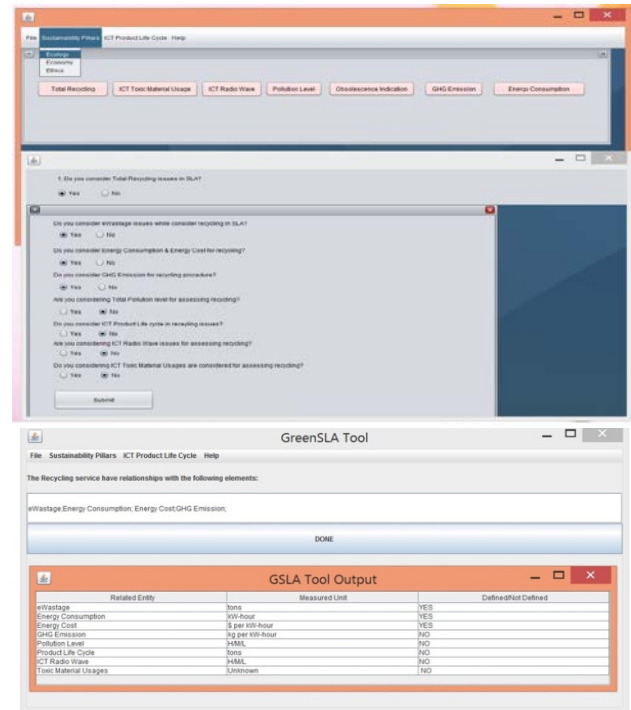


Fig. 9. Questionnaires generation under GSLA environment for recycling services

The ICT engineer should analyze their future green services by generating some questionnaires and then it's possible to evaluate their designed green services from the customer side. This is a small step to the way of justification of proposed designed services. The questionnaire is generated using Java (Eclipse Tool), following the UML model of *Total Recycling Service* (Fig.2) under sustainable GSLA environment. The generation of automated questionnaires represents in Fig.9. These questionnaires sent to 20 different ICT companies and as well as other companies in Bangladesh and Japan, who respect recyclability and sustainability in their business scope. Most of these industries are varied in sizes, - Large, Medium, Small; and their response to the questionnaires were completely unbiased. Among the 20 industries, 15 ICT based industries delivered their answers for further analysis. The feedback of these industries is then analyzed using SAS analytical tool (JMP 12.2.0) and represented in next fig. 10-11.

The analysis revealed, most of this ICT-based industries taking consideration of recycling services while designing, developing, implementing or providing services/product to their customers. Fig.10 shows the feedback analysis of 15 industries and among them 66% medium (red color) and large sized industries respect recycling services under sustainability. The small sized (green color) industries are usually using the slogan of sustainability but they are far behind of considering the proposed green services. In contrary, most of the large size industries (blue color) are practicing recyclability though they are not pretending to be a sustainable industry due to the lack of knowledge to design new green services in their scope.



Fig. 10. Analysis of recycling services in different sized industries

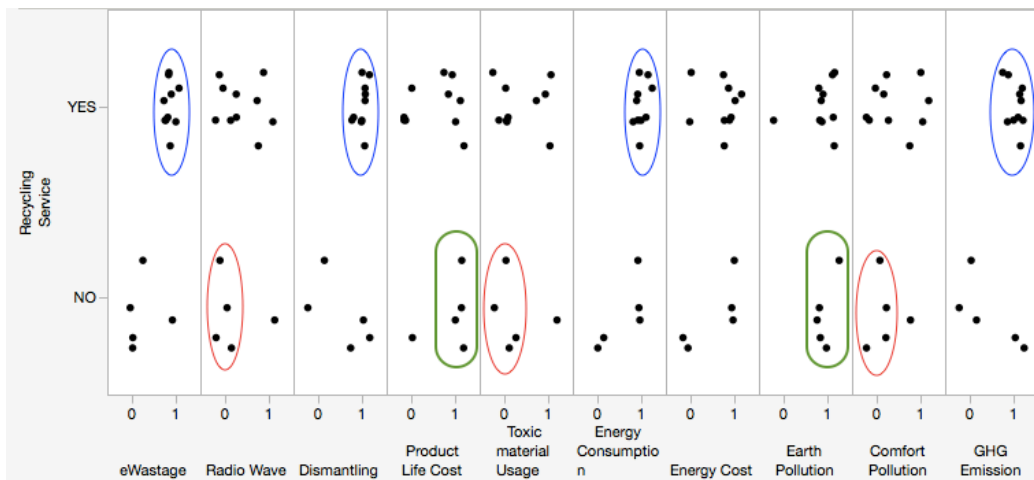


Fig. 11. Analysis of all entities of recycling services

However, these industries are unaware of designing their recycling services with the proposed designed model (Fig.2). It is evident from the fig. 11 that, these companies are just consider the interaction of *eWastage*, *energy consumption*, *energy cost*, *dismantling* and *GHG entities* in their scope while ignoring *Radio wave information*, *Toxic material*, *Comfort pollution* entities.

It is noticeable that, very few industries (only 01 company) are justifying all entities and their different levels of interactions (Table 1) in compare to the proposed designed *Total Recycling Services* (Fig.2). The interesting fact is that, most of these industries concentrate on *earth pollution and product life costing* while not considering recycling services in their business scope. It shows the importance and awareness of practicing sustainability in businesses to some extent.

IV. CONCLUSION

This research proposed new layer of green IT services on top of already existing services for the ICT industries to achieve sustainability. The identification and designation of all seven new green services (Fig.2 to Fig.8) derived from

sustainable GSLA model. Therefore, these seven green services could be satisfied from both parties (users/customers and providers) of the businesses. The justification of recycling services (Fig.9 to Fig.11) could be used as a guideline for the ICT engineer or other industries to develop future green services in their scope. However, this research still has some challenges to face as the informational model of sustainable GSLA; green service infrastructure are needed to be standardized by proper authority, rules or directives. The standardization of green indicators and green services is one of the main issues as mentioned by ITU-T report (2012). Also, further research is necessary on monitoring and evaluating green services for a viable Green IT framework design. Most of the interactions and different level relationships in the proposed designed services are non-technical parameters. Moreover, some entities have interaction with human behavior in their evaluation and validation. Thus, it might be the most challenging task of monitoring those newly designed services in future. The next steps of this research is to justify all green services with more feedbacks and develop a viable Green IT as a Service (GaaS) infrastructure under sustainability hoods for the industries.

ACKNOWLEDGMENT

The authors would like to show their gratitude and thanks to PERCCOM program (European Union) for giving the idea of GSLA research.

REFERENCES

- [1] SLA@SOI, Source: <http://sla-at-soi.eu/>, retrieved on April 2015.
- [2] R. Buyya, J. Broberg, and A. Goscinsk, "Cloud Computing: Principles and Paradigm," A John Wiley & Sons, Inc. Publication, ISBN: 978-0-470-88799-8, February 2011.
- [3] I. Ahmed, H. Okumura, and K. Arai, "Analysis on existing Basic SLAs and green SLAs to define new sustainable Green SLA", International Journal of Advanced Computer Science and Applications, Vol.6, No. 12, December 2015, pp. 100-108.
- [4] J. Mankoff, R. Kravets, and E. Blevis, "Some Computer Science Issues in Creating a Sustainable World," Computer, Vol. 41, No. 8, 2008.
- [5] SMART 2020 Report, "Enabling the low carbon economy in the information age," The Climate Group, GeSI, 2008.
- [6] Z. S. Andreopoulou, "Green Informatics: ICT for Green and Sustainability," Journal of Agriculture Informatics (EIFTA), Vol. 3, No. 2, 2012.
- [7] K. Van Wensen, W. Broer, J. Klein, and J. Knopf, "The State of Play in Sustainability reporting in the European Union", Executive Summary, The European Union's Programme for Employment & Social Solidarity, CREM/adelfphi, Amsterdam/Berlin, 2011.
- [8] J. Porritt, "Green IT: The Global Benchmark", A report on sustainable IT in the USA, UK, Australia and India, Fujitsu, 2012.
- [9] I. Ahmed, H. Okumura, and K. Arai, "An Informational Model as a Guideline to Design Sustainable Green SLA (GSLA)," International Journal of Advanced Computer Science and Applications, Vol.7, No.4, April 2016, pp. 302-310.
- [10] Gartner Report (2015), Source: <http://www.gartner.com/it-glossary/it-services>.
- [11] A. Chen, K. N. Dietrich, X. Huo, and S.-M. Ho, "Development Neurotoxicants in E-Waste: An Emerging Health Concern," Environmental Health Perspectives, Vol.119, No.4, April 2011.
- [12] J. Jin, "Electromagnetic Analysis and Design in Magnetic Resonance Imaging," CRC Press, ISBN 978-0-8493-9693-9, September 1998.
- [13] C. Tuppen, "Circularity and the ICT Sector," Advancing Sustainability LLP @ Ellen MacArthur Foundation, United Kingdom, September 2013.
- [14] P. Sandborn, "Software Obsolescence- Complicating the Part and Technology Obsolescence Management Problem," IEEE Transaction on Components and Packaging Technologies, Vol.30, No.4, December 2007, pp. 886-888.
- [15] D. A. Levinthal, and D. Purohit, "Durable Goods and Product Obsolescence," Marketing Science, Vol.8, No.1, printed in USA, Winter 1989.
- [16] P. Singh, and P. Sandborn, "Obsolescence driven design refresh planning for sustainment-dominated systems," The Engineering Economist, Vol.51, No.2, June 2006, pp. 115-139.
- [17] P. Mathew, S. Ganguly, S. Greenberg, and D. Sartor, "Self Benchmarking Guide for Data Centers: Metrics, Benchmarks, Actions," Report of New York State Energy Research & Development Authority (NYSERDA), July 2009.
- [18] T. Shiino, "Green IT by all Parties," PhD Presentation at Nomura Research Institute, Tokyo, Japan, March 2010.
- [19] A. Atrey, N. Jain, and Iyengar N. Ch. S. N, "A Study on Green Cloud Computing," International Journal of Grid and Distributed Computing, Vol.6, No.6, 2013, pp. 93-102.
- [20] E. Rondeau, F. Lepage, J. P. Georges, and G. Morel, "Measurements and Sustainability," Chapter 3, Green Information Technology, 1st Edition, A Sustainable Approach, Dastbaz & Pattinson & Akhgar, ISBN: 9780128013793, Elsevier Book, 304 pages, March 2015.
- [21] F. Beister, M. Draxler, J. Aelken, and H. Karl, "Power model design for ICT systems – A generic approach," Computer Communication Journal, Vol. 50 (Special Issue), September 2014, pp. 77-85.
- [22] T. R. Akinbinu, and Y. J. Mashalla, "Impact of Computer Technology on Health: Computer Vision Syndrome," Journal of Medical Practice and Review, Vol. 5 (3), November 2014, pp. 20-30.
- [23] Anonymous, "The Effects of Computer Use on Eye Health and Vision," White Paper, American Optometric Association, 2014.
- [24] P. Zhang, "Theorizing the Relationship between Affect and Aesthetics in ICT Design and Use Context," International Conference on Information Resource Management, Dubai, UAE, May 2009.
- [25] Anonymous, "Emission factors for Greenhouse Gas Inventories," EPA (2014) Inventory of U.S Greenhouse Gas Emissions and Sinks, White Paper, April 2014.
- [26] S. Naumann, M. Dick, E. Kern, and T. Johann, "The GREENSOFT Model: A reference model for green and sustainable software and its engineering," Sustainable Computing: Informatics and Systems Journal, Vol.1, No.4, June 2011, pp. 294-304.
- [27] N. El-Sayed, I. Stefanovici, G. Amvrosiadis, Andy A. Hwang, and B. Schroeder, "Temperature Management in Data Centers: Why Some(Might) Like It Hot," 2012 SIGMETRICS Conference, ACM Digital Library, London, UK, June 2012.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 33 books and published 510 journal papers. He is now Editor-in-Chief of IJACSA and IJISA.

Using a Cluster for Securing Embedded Systems

Mohamed Salim LMIMOUNI, Khalid BOUKHDIR, Hicham MEDROMI, Siham BENHADOU
Equipe Architectures des Systèmes (EAS), Laboratoire d'Informatique, Systèmes et Energies Renouvelables (LISER)
Hassan II University of Casablanca, Ecole Nationale Supérieure d'Electricité et de Mécanique (ENSEM)
Casablanca, Morocco

Abstract—In today's increasingly interconnected world, the deployment of an Intrusion Detection System (IDS) is becoming very important for securing embedded systems from viruses, worms, attacks, etc. But IDSs face many challenges like computational resources and ubiquitous threats. Many of these challenges can be resolved by running the IDS in a cluster to allow tasks to be parallelly executed. In this paper, we propose to secure embedded systems by using a cluster of embedded cards that can run multiple instances of an IDS in a parallel way. This proposition is now possible with the availability of new low-power single-board computers (Raspberry Pi, BeagleBoard, Cubieboard, Galileo, etc.). To test the feasibility of our proposed architecture, we run two instances of the Bro IDS on two Raspberry Pi. The results show that we can effectively run multiple instances of an IDS in a parallel way on a cluster of new low-power single-board computers to secure embedded systems.

Keywords—cluster; intrusion detection system; embedded system; security; parallel system

I. INTRODUCTION

Intrusion Detection Systems (IDS) were used for many years to protect networks and hosts. And since their design, they have not ceased to play a major role in the defense against intrusions and attacks. They allow analyzing and monitoring the activities on a network or a given machine to detect fraudulent use of resources, log, alert administrators and in some cases react and stop the threat to enforce the security policy.

Despite the progress made in intrusion detection, IDS remain limited in the protection of embedded systems against sophisticated attacks. This limit has prompted us to propose a new architecture to enable IDS to remain among the pillars of security solutions, especially in embedded systems security. We opted for the use of a cluster that offers high performance and better scalability.

A cluster is defined as a group of independent computers linked with a computational network and operating as a single computer [1]. In other words, a cluster is a collection of independent and cheap machines, used together as a supercomputer to provide a solution [1].

Using clusters has many advantages:

- The computers that form a cluster are cheap.
- You can add other nodes to the cluster as needed.
- On clusters, you can use open source software to reduce software costs.

- Clusters allow multiple computers to work together to solve several problems.

In this paper, a background is presented in the second section. The third section gives a short view on related works. The fourth section shows the proposed architecture. In the fifth section, we describe the implementation. The sixth section shows the obtained results.

II. BACKGROUND

A. Embedded systems

1) Definitions

An embedded system is a microprocessor-based system that is built to control a function or range of functions and is not designed to be programmed by the end user in the same way that a PC is [2].

It's also defined as a computing system which is designed for specific control functions and is embedded as part of the complete device which may include hardware and mechanical parts [3].

2) Reference model

The reference model of embedded systems as illustrated in Figure 1 [4] shows the main components of an embedded system. The Hardware Layer contains the physical components of the embedded system. The System Software Layer and the Application Software Layer contain the software being processed by the embedded system.

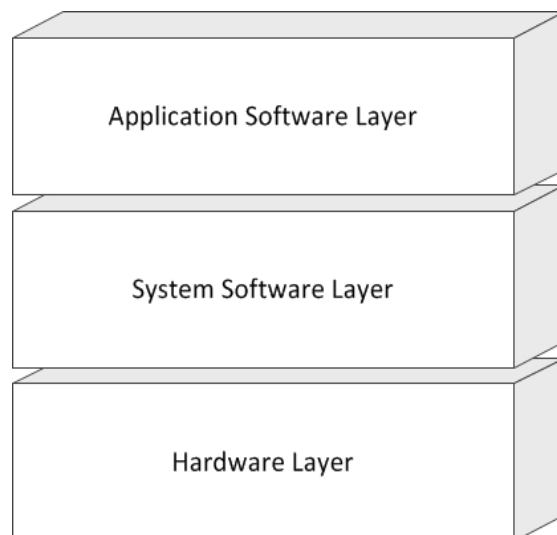


Fig. 1. Reference model of an embedded system

B. Intrusion Detection Systems

1) Definitions

An intrusion detection system is the intrusion alarm of the computer security field [5].

An intrusion detection system consists of an audit data collection agent that collects information about the observed system. This data is either stored or processed directly by the detector [5].

2) Classification

There are two main approaches used by IDS to detect intrusions:

- **Signature-based approach:** An IDS based on this approach monitors the packets in the network and compares them with a database of attributes or signatures of known vulnerabilities. This is similar to the way in which the antiviruses detect malwares. The problem with this approach is the time between the date of vulnerability discovery and the application of the signature associated with it. During this period, the IDS cannot discover attacks exploiting these vulnerabilities.
- **Anomaly-based approach:** An IDS based on the behavioral approach monitors traffic and compares it to an established standard (Baseline). This standard identifies what is "normal" - bandwidth use, used protocols, ports and machines - and alert the administrator or cancel the connection in the case of an anomaly or a different use.

C. Clusters

1) Definitions

A cluster is a single system comprised of interconnected computers that communicate with one another either via a message passing; or by direct, internode memory access using a single address space [6]. We can also define a cluster as a commonly found computing environment consisting of many PCs or workstations connected together by a local-area network [7].

2) Typical architecture

The typical architecture of a cluster is shown in Figure 2 [8]. A node of the cluster can be a single or multiprocessor system, such as a PC, workstation, or Symmetric MultiProcessor (SMP). The nodes must be connected via a Local Area Network (LAN) based on Ethernet, Myrinet or InfiniBand. The cluster middleware offers an illusion of a united system of the independent nodes. Parallel programming environments offer portable, efficient and easy-to-use tools for developing parallel applications.

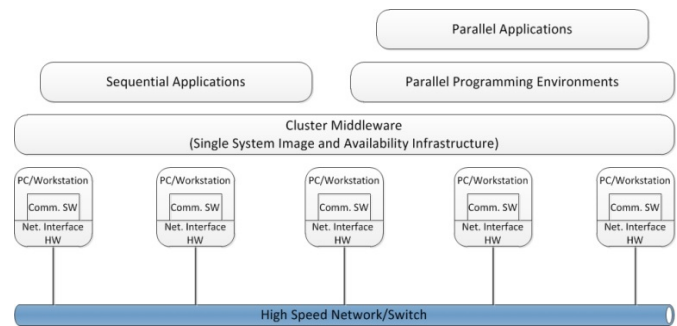


Fig. 2. Typical architecture of a cluster

3) Types

There are three varieties of clusters, each one offers different benefits for the user. These varieties are:

- **Load balancing clusters:** used to provide a single interface for a set of resources that can grow arbitrarily. We can imagine a web server that redirects client requests to another node when it has reached its limit of load. This is called "load balancing". Only the node that handles the distribution is visible from the outside.
- **High performance clusters:** they consist of a set of computers linked together to provide maximum power in solving a problem. The heart of these clusters is formed of compute nodes that will receive the code to execute. On smaller clusters we can count ten nodes, while the largest have more than 80 000. The network architecture used to communicate between nodes becomes very heavy, expensive and it limits its performance. You should know also that the ratio of the number of nodes and the performance of this type of clusters is not linear. It is necessary that the program executed is highly parallelizable and that it requires little communication between the computing units.
- **High availability clusters:** the High Availability clusters are built to provide a secure and fault tolerant environment. The redundancy is the most used method. It consists on multiplying the material that could be subject to failure. Server applications are installed the same way on the cluster nodes.

III. RELATED WORK

A. NIDS Cluster

"NIDS Cluster" [9] is a scalable solution based on a set of computers that analyze a traffic flow without sacrificing the detection accuracy. The nodes of this cluster run instances of Bro [10] [11] and share a low-level analysis state to compose a

comprehensive picture of the network activity. One of the main objectives of this solution is load balancing between the cluster nodes, one of the advantages of using a cluster of computers.

"NIDS Cluster" is a solution that takes advantage of the load balancing provided by clusters but it does not exploit all the performance offered by cluster nodes. It's also a solution for computer clusters and not for embedded systems.

B. Gnort

Graphics Processing Unit (GPU) have become very powerful and researches have begun to draw power from their ability to do intense calculations for highly parallel operations including intrusion detection [12], and cryptography [13]. Gnort [14] is a high performance intrusion detection system pulling power of graphic processors to accelerate search for patterns in network packets. This work uses the Single Instruction on Multiple Data (SIMD) instructions to turn the Aho-Corassik algorithm which allows the string search in a text, to achieve a maximum bandwidth of 2.3 Gbit/s. This IDS has been implemented on a NVIDIA GeForce 8 Series, which offers many advantages through Compute Unified Device Architecture (CUDA) which is currently the most widely used GPU programming toolbox. It includes a compiler for GPU cores development in an extended C language dialect [15].

Gnort is a solution that takes advantage of the use of GPU but cannot achieve the performance of the use of a HPC cluster as Gnort uses a single NVIDIA GeForce 8 Series. It's also a solution for GPU clusters and not for embedded systems.

C. Distributed platform for intrusion detection based on multi-agents system

As part of research conducted within the Equipe Architectures des Systèmes (EAS) team, a real-time distributed architecture for intrusion detection based on the multi-agent aspect was proposed in 2010 [16]. This architecture consists of two levels of analysis benefiting from agent's reactive and cognitive capabilities. Several agents are distributed at different network points with different roles to detect attacks and intrusions.

This solution takes advantage of using the multi-agent aspect capabilities, but it's a solution for computer networks and not for embedded systems.

IV. PROPOSED ARCHITECTURE

The proposed hardware architecture as illustrated in Figure 3 shows the different components of the cluster:

- **Device:** the device to protect from intrusions and attacks, such us drones, satellites, mobile robots, etc.
- **Manager Card:** the card which manages the other cards of the cluster using the network. The manager card can:
 - ✓ Start the other card's work,
 - ✓ Stop the other card's work,
 - ✓ Monitor the other cards,
 - ✓ Configure the other cards,

- ✓ Update the other cards,
 - ✓ Share information with the other cards,
 - ✓ Receive logs from the other cards,
 - ✓ Store logs,
 - ✓ Organize logs,
 - ✓ Generate statistics.
- **Worker Cards:** the cards running the IDS instances. The worker cards can:
 - ✓ Analyze the traffic on the secured interface (the interface being monitored),
 - ✓ Examining packets,
 - ✓ Send states to the manager card,
 - ✓ Send logs to the manager card.
 - **Network Interfaces:** the link between the cards and the network. They are the hardware that acts as a communication processor and which is responsible for transmitting and receiving packets of data between the cards of the cluster via a network switch.
 - **Network:** the Local Area Network (LAN) that connects the cluster nodes (cards).

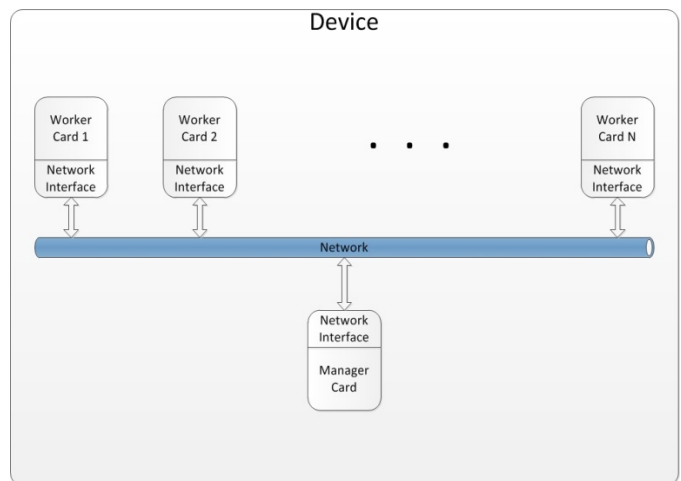
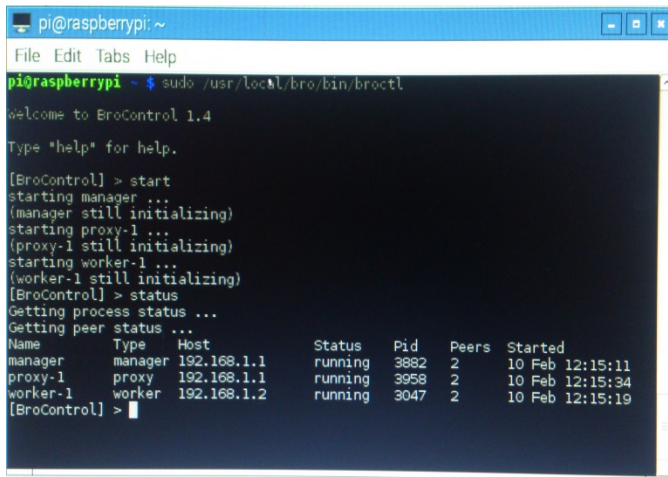


Fig. 3. Proposed hardware architecture

Figure 4 shows the layers of the proposed software architecture:

- **IDS:** the Intrusion Detection System parallely run on the cards of the cluster and acting as a singular and cohesive entity.
- **Compiler:** the necessaire compiler for running the code source of the chosen IDS.
- **OS:** the Operating System allowing us to exploit the cards.
- **Drivers:** the software allowing us to use the different components of the cards.



```
pi@raspberrypi: ~  
File Edit Tabs Help  
pi@raspberrypi ~$ sudo /usr/local/bro/bin/broctl  
welcome to BroControl 1.4  
Type "help" for help.  
[BroControl] > start  
starting manager ...  
(manager still initializing)  
starting proxy-1 ...  
(proxy-1 still initializing)  
starting worker-1 ...  
(worker-1 still initializing)  
[BroControl] > status  
Getting process status ...  
Getting peer status ...  
Name      Type      Host           Status  Pid   Peers  Started  
manager   manager  192.168.1.1    running 3882  2      10 Feb 12:15:11  
proxy-1   proxy    192.168.1.1    running 3958  2      10 Feb 12:15:34  
worker-1  worker   192.168.1.2    running 3047  2      10 Feb 12:15:19  
[BroControl] > |
```

Fig. 8. Results

VII. CONCLUSIONS AND FUTURE WORK

The growing need for powerful, faster and cheaper computers in the world increases the use of clusters. Today, clusters are used in different areas (commercial, scientific, etc.). The field of embedded systems security, an area in large changes, also needs to exploit the advantages of the use of clusters.

That is why we have proposed in this paper to use a cluster to secure and protect embedded systems from intrusions and attacks by running multiple instances of an Intrusion Detection System (IDS) on the different nodes of a cluster.

As future work, we would like to carry out detailed performance evaluations and test the realized implementation with real-life cases.

REFERENCES

[1] S. Aydin and O. F. Bay, "Building a high performance computing clusters to use in computing course applications", *Procedia - Social and Behavioral Sciences*, vol. 1, pp. 2396-2401, Feb. 2009.
[2] S. Heath, *Embedded Systems Design*, 2nd ed., Newnes, 2003.

[3] S. Mittal, "A Survey of Techniques For Improving Energy Efficiency in Embedded Computing Systems", *IJCAET*, vol. 6, no. 4, pp. 440-459, 2014.
[4] T. Noergaard, *Embedded Systems Architecture*, 2nd ed., Newnes, 2013.
[5] S. Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy", Chalmers University of Technology, Göteborg, Sweden, Technical Report 99-15, 2000.
[6] G. Bell and J. Gray, "What's Next in High-Performance Computing", *Communications of the ACM*, vol. 45, issue 2, pp. 91-95, Feb. 2002.
[7] J. Dongarra, "Trends in high performance computing: a historical overview and examination of future developments", *Circuits and Devices Magazine*, IEEE, vol. 22, issue 1, pp. 22-27, Feb. 2006.
[8] R. Buyya, "High Performance Cluster Computing: Programming and Applications", Prentice Hall PTR, NJ, USA, 1999.
[9] M. Vallentin, R. Sommer, J. Lee, C. Leres, V. Paxson and B. Tierney, "The NIDS Cluster: Scalable, Stateful Network Intrusion Detection on Commodity Hardware", in *Proc. RAID 2007*, 2007, p. 107-126.
[10] V. Paxson, "Bro: A System for Detecting Network Intruders in Real-Time", *Computer Networks*, Elsevier, vol. 31, issue 23-24, pp. 2435-2463, Dec. 1999.
[11] The Bro Network Security Monitor. [Online]. Available: <http://www.bro.org/>
[12] N. Jacob and C. Brodley, "Offloading IDS Computation to the GPU", in *Proc. ACSAC'06*, 2006, p. 371-380.
[13] D. L. Cook, J. Ioannidis, A. D. Keromytis and J. Luck, "CryptoGraphics: Secret Key Cryptography Using Graphics Cards", in *Proc. RSA Conference 2005*, 2005, p. 334-350.
[14] G. Vasiliadis, S. Antonatos, M. Polychronakis, E. P. Markatos and S. Ioannidis, "Gnort: High Performance Network Intrusion Detection Using Graphics Processors", in *Proc. RAID 2008*, 2008, p. 116-134.
[15] V.V. Kindratenko, J. J. Enos, G. Shi, M. T. Showerman, G. W. Arnold, J. E. Stone, J. C. Phillips and Wen-mei Hwu, "GPU clusters for high-performance computing", in *Proc. CLUSTER'09*, 2009, p. 1-8.
[16] D. Raoui, S. Benhadou and H. Medromi, "New distributed platform for intrusion detection based on multi-agents system", *Journal of Engineering and Technology Research*, vol. 2, issue 10, pp. 200-206, Oct. 2010.
[17] Raspberry Pi 1 Model B. [Online]. Available: <http://www.raspberrypi.org/products/model-b/>
[18] File:Drawing of Raspberry Pi model B rev2.svg. [Online]. Available: http://en.wikipedia.org/wiki/File:Drawing_of_Raspberry_Pi_model_B_rev2.svg
[19] FrontPage - Raspbian. [Online]. Available: <http://www.raspbian.org/>

Developing a Transition Parser for the Arabic Language

Aref abu Awad

Computer Information System, Zarqa University,
Zarqa, Jordan

Essam Hanandeh

Computer Information System, Zarqa University,
Zarqa, Jordan

Abstract—One of the most important Characteristics of the Arabic language is the exhaustive undertaking. Thus, analyzing Arabic sentences is difficult because of the length of sentences and the numerous structural complexities. This research aims at developing an Arabic parser and lexicon. A lexicon has been developed with the goal of analyzing and extracting the attributes of Arabic words. The parser was written by using a top-down algorithm parsing technique with recursive transition network. Then, the parser has been evaluated against real sentences and the outcomes were satisfactory.

Keywords—Natural language processing; Arabic parser; lexicon; Transition Network

I. INTRODUCTION

Natural language processing (NLP), which is considered a field of computer science, artificial intelligence, and computational linguistics, is dealing with the interactions between computers and natural languages. Accordingly, NLP is related to the area of human-computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input. Other challenges involve natural language generation. The history of NLP generally started in the 1950s, although studies can be traced from periods earlier than that a decade. In 1950, Alan Turing published an article entitled "Intelligence," which proposed what is now called the Turing test as a criterion of intelligence. Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. These algorithms are able to learn from data that have not been hand-annotated with the desired answers, or use a combination of annotated and non-annotated data. In general, this task is considerably more difficult than supervised learning and typically produces inaccurate results for a given amount of input data. However, an enormous amount of non-annotated data are available (including the entire World Wide Web content) often compensate the inferior results. Modern NLP algorithms are based on machine learning, particularly statistical machine learning. The machine learning paradigm is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls for using general learning algorithms, which are often grounded on statistical inference, to automatically learn such rules through the analysis of large corpora of typical real-world examples. A corpus (plural: corpora) is a set of documents (or individual sentences) that have been hand-annotated with the correct values to be

learned. The goal of the NLP group is to design and develop software that will analyze, understand, and generate languages that humans can use to address a computer and addressing another person [1]. Information retrieval is one of the natural language processing applications that appears in these definitions. Information retrieval is a field which deals with the structure, analysis, organization, storage, searching, and retrieval of information [2]. Moreover, information retrieval is a selective process by which the desired information is extracted from a store of information called a database [3].

II. RELATED STUDIES

Gilbert et al. [8] developed a bottom-up parsing strategy for summarizing an English text and integrated it with the Pruner and Redundancy Eliminator (PARE) system, replacing the old link grammar parser which was previously used. Constituency trees from our parser provide all parts-of-speech linkages as input to several other code modules in the PARE system. Our parser uses rules that are written in the Chomsky normal form, which is a specialization of a general context-free grammar. Updating the PARE system leads to an increase in the efficiency of the text summarization process [8].

Shalan et al. [10] developed an Arabic parser for modern scientific text. This parser is written in definite clause grammar and is targeted to be a component of a machine translation system. The development of the parser consisted of a two-step process. In the first step, we acquired the rules constituting the Arabic grammar that provided a precise account of what was considered a grammatical sentence. The grammar covered a text from the domain of the agricultural extension documents. The second step involved implementing the parser that assigns grammatical structure to the input sentence. An experiment on real extension document was performed, and the results observed were satisfactory.

Khufuet al. [11] recommended a method for Arabic parsing based on supervised machine learning. They used the support vector machines algorithm to select the syntactic labels of the sentence. Furthermore, we evaluated their parser following the cross validation method by using the Penn Arabic Treebank. The obtained results were substantially encouraging.

Al-Taani1 et al. [12] presented a top-down chart parser for parsing simple Arabic sentences, including nominal and verbal sentences within the specific Arabic grammar domain. We used context-free grammar (CFG) to represent the Arabic grammar. We first developed the Arabic grammar rules that

provided precise description of grammatical sentences. Thereafter, we implemented the parser that assigns grammatical structure to the input sentence. Experimental results showed the effectiveness of the proposed top-down chart parser for parsing modern standard Arabic sentences.

PARSIG METHOD

Parsing method involves revealing a structure in an input based on the external information about the elements of the input and their order. Generally speaking, external information comprises a lexicon, i.e., list of input words; and grammar to describe the structures that may be built from and implemented by the sequences of words [9]. Parsing has several definitions but most of them focus on the text structure. The common definitions of parsing are as follows. Parsing can be defined as the process of analyzing an input sequence in order to determine its grammatical structure regarding to a given formal grammar [5]. Parsing breaks a sentence down into its component parts of speech with an explanation of the form, function, and syntactical relationship of each part [6]. Parsing is also the process of converting text input into a data structure defining its syntactical structure and semantic meaning based upon a given formal grammar [8]. Parsing natural language is an attempt to discover a certain structure in a text (or textual representation) generated by a person [4]. A parser is a computational system that processes input sentences according to the productions of grammar, and builds one or more constituent structures that conformed grammatically. We consider grammar as a well-formed declarative specification, whereas a parser is a procedural interpretation of grammar.

III. LEXICON

Lexicography is the branch of applied linguistics concerned with the design and construction of lexica for practical use. Lexica can range from the paper lexica or encyclopedia designed for human use and shelf storage to the electronic lexica used in a variety of human language technology systems, such as word databases, word processors, and software for reading back (by speech synthesis in text-to-speech systems) and dictation (by automatic speech recognition systems). At a considerably generic level, a lexicon may be a generic lexicographic knowledge base from which these different types of lexica can be derived automatically [71]. Meanwhile, lexicology is the branch of descriptive linguistics concerned with linguistic theory and methodology for describing lexical information, and often focuses specifically on issues of meaning. Traditionally, lexicology has been mainly concerned with lexical collocations and idiom, lexical semantics, as well as the structure of words, meaning components and relationships between them.

IV. TRANSITION NETWORK GRAMMARS

Transition network grammar is considered as a formalism for representing grammars based on the concept of a transition network that comprises nodes and labeled arts. This formalism developed out from the transition network concept of a finite-state automaton. It is equivalent to push-down automata

because the arts, comprise the network of a transition network grammar and represent transcriptions of the rules of a context-free grammar [7]. Sentences generated by the grammar are accepted by a transition network grammar through the process of traversing the network comprising of these arcs.

Figure 1 shows the network called NP in which each art is labeled with a word category. Starting at a given node, one can traverse an art if the current word in the sentence is in the category on the art. If the art is followed, then the current word is updated to the next word. A phrase is a legal NP if a path from the node NP to a pop art accounts for every word in the phrase.

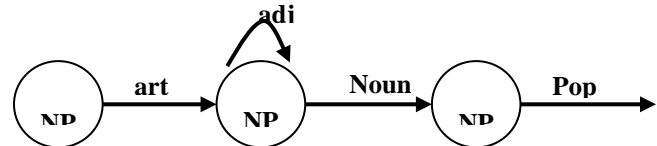


Fig. 1. Transition Network

V. SYSTEM EVALUATION

The objective of our experiment was to test whether the parser is sufficient for application to real Arabic sentences. We selected an unrestricted Arabic sentence, which is from the Arabic students' book.

VI. RESULTS

We discuss the experiment results whether the input sentence is parsable or not. Table (1) shows the results of the parser. These results are categorized into: parsable and unparsable sentences.

The parsable sentence is divided into two subcategories as follows.

1) Syntactically Correct: This subcategory led to a complete and successful parsing of the input sentence.

2) Syntactically Incorrect: This subcategory led to a complete parsing of the input sentence but the result, as can be seen, is a syntactically incorrect structure. The source of this error does not match in terms of attributes (e.g., gender, number) between words of sentence. For example, the input sentence

يذهب الطالبة إلى المدرسة

is not parsed by our parser. The subject (الطالبة) takes the female feature gender. However, the prefix (ي) of the verb (يذهب) of the sentence indicates that this feature value is for male. The syntactically correct sentence would be as follows:

تذهب الطالبة إلى المدرسة

The unparsable sentence can be divided into three subcategories:

1) Lexical Problem: The parser does not find out the word in the lexicon.

2) Incorrect Sentence: This subcategory has failed to parse because the input sentence is incorrect:

يلعب يدرس الطالب الشيط .

3) Failure: The sentence is not identified by linguists according to Arabic grammar rules. An example is the following input sentence:

الطالب النشيط يدرس.

TABLE I. RESULTS OF THE PARSER

		Number of Sentences	Percentage
Parsable Sentence	Syntactically Correct	77	87.1 %
	Syntactically Incorrect	2	2.6 %
Unparsable Sentence	Lexical Problem	4	4.8 %
	Incorrect Sentence	2	2.4 %
	Failure	5	5.8 %
Total		93	100 %

The number of sentences used in the test was 93 and the length of each sentence was 6 words. The result shows that the number of successfully parsed sentences were 77 (87.1%) and 2 sentences were syntactically incorrect (2.6%). The number of sentences that were not parsed (i.e., has lexical problem) were 4 (4.8%). The number of sentences that were not parsed (incorrect sentence) were 2 (approximately 2.4%). The number of sentences that were not parsed (i.e., not recognized by linguists according to Arabic grammar rules) were 5 (approximately 5.8%).

VII. ANALYSIS OF RESULTS

1) Analysis of the Syntactically Incorrect Sentences

Recall that the number of syntactically incorrect sentences were 2 sentences. The parser assigned the incorrect result to the input sentence. Hence, the parser completed the sentence parsing, but the result is incorrect. This result was due to an incomplete agreement between word attributes (e.g., gender, number).

2) Analysis of the Unparsable Sentences

Recalling that the number of unparsable sentences were 11; the parser failed to identify any rule to the input sentence. These are classified into three categories as follows.

a) *Lexical Problem*: The parser fails to recognize any rule to the input sentence and this is because certain parts of the sentences are unavailable in the lexicon. Thus, the parser does not obtain the attributes of these parts.

b) *Incorrect Sentence*: The parser fails to produce a rule for the input sentence because of the incorrect syntactic form of the sentence. Hence, determining an equivalent role in the sentential form in the parser is impossible.

c) *Failure*: The parser fails to produce a rule for the

input sentence because the syntactic form of the sentence is excluded in the grammar. Thus, failure may result when the sentence structure is correct.

VIII. CONCLUSION

Our contribution in this paper is to design, build and Evaluate system for parsing Arabic sentences and Determine if these sentences syntactically correct or not. In addition, the proposed system builds a lexicon for Arabic sentences.

The Arabic language lacks parsing systems for analyzing Arabic sentences. Parsing systems are crucial in natural language processing because they are used as a first step in most natural language processing applications. Moreover, this system can be extensively used for educational purposes.

In the natural Arabic language processing, predefined forms, exist for analyzing sentences, make parsing problematic. The Arabic sentence is complex and syntactically ambiguous because of the frequent usage of grammatical relationships, conjunctions, and other constructs.

The methodology we adopted in this study based on analyzing the Arabic language grammar conforming to gender and number, formalization of rules using CFG, representation of the rules using transition networks, constructing a lexicon of words that will be in the sentences structure, implementing the recursive transition network parser, and evaluating the system using real Arabic sentences. Finally, the current analysis was effective and provided good results

REFERENCES

- [1] Preeti1, and B. Sidhu, 2013. NATURAL LANGUAGE PROCESSING. Int.J.Computer Technology & Applications, Vol 4 (5),751-758.)
- [2] T. Strzalkowski, F. Lin, J. Wang, J. Perez-Carballo, 1999. Evaluating Natural Language Processing Techniques in Information Retrieval. TREC, Volume 7, pp 113-145.
- [3] J. allan, J.Aslam, N. Belkin, 2003. Challenges in Information Retrieval and Language Modeling. ACM SIGIR Forum, 37(1):31-47.
- [4] Taboada, Maite, and William C. Mann. "Applications of rhetorical structure theory." *Discourse studies* 8.4 (2006): 567-588.
- [5] Kübler, Sandra, Ryan McDonald, and Joakim Nivre. 2009 Dependency parsing. *Synthesis Lectures on Human Language Technologies* 1.1 pp. 1-127..
- [6] Weise, D. Neal. 2007. Method and apparatus for improved grammar checking using a stochastic parser. U.S. Patent No. 7,184,950. 27
- [7] Budanitsky, Alexander, and G. Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness." *Computational Linguistics* vol.32.pp 13-47.
- [8] Gilbert, Nathan, E. Welborn, and S. Thede. 2005 PARSING ENGLISH TEXTS IN PARE.
- [9] Bird, Steven, and M. Liberman, 2001.A formal framework for linguistic annotation. *Speech communication*, pp. 23-60.
- [10] Shaalan, Khaled, A. Farouk, and A. Rafea,1999.Towards an Arabic parser for modern scientific text. *Proceeding of the 2nd Conference on Language Engineering*.
- [11] Elarnaoty, Mohamed, S. AbdelRahman, and A. Fahmy, 2012. A machine learning approach for opinion holder extraction in Arabic language.*arXiv preprint arXiv:1206.1011..*
- [12] T. Ahmad, M. Mohammed, and A. Sana, 2012."A top-down chart parser for analyzing arabic sentences." *Int. Arab J. Inf. Technol.* 9.2,pp. 109-116.

Multi- Spectrum Bands Allocation for Time-Varying Traffic in the Flexible Optical Network

KAMAGATE Beman Hamidja

Laboratoire de Recherche en Informatique et
Télécommunication (LARIT)
Ecole Doctorale Polytechnique de l'INP-HB
Yamoussoukro, Côte d'Ivoire

Michel BABRI

Laboratoire de Recherche en Informatique et
Télécommunication (LARIT)
INPH-HB
Yamoussoukro, Côte d'Ivoire

GOORE Bi Tra

Laboratoire de Recherche en Informatique et
Télécommunication (LARIT)
INP-HB
Yamoussoukro, Côte d'Ivoire

Souleymane OUMTANAGA

Laboratoire de Recherche en Informatique et
Télécommunication (LARIT)
INP-HB
Yamoussoukro, Côte d'Ivoire

Abstract—The flexible optical networks are the promising solution to the exponential increase of traffic generated by telecommunications networks. They combine flexibility with the finest granularity of optical resources. Therefore, the flexible optical networks position themselves as a better solution than conventional WDM network. In the operational phase, traffic of connections fluctuates. In fact, the user's need is not the same during day periods. Such traffic may experiment evidence of rising working hours, end of months or years and decreases during the night or on holidays. This variation requires the expansion or contraction of the number of frequency slots allocated to a connection to match the exact needs of the moment. The expansion of the traffic around the reference frequency of connection may lead to blockage because it must share frequency slots with neighboring connections in compliance with the constraints of continuity, contiguity, and non-overlapping. In this study, we offer a technique for allocating frequency slots for time-varying traffic connections. We share out the additional traffic load on different spectrum paths by respecting the constraint of time synchronization related to the differential delay to reduce the blocking rate due to traffic fluctuation.

Keywords—Spectrum band; Multi-spectrum bands; time-varying traffic; elastic optical network

I. INTRODUCTION

The flexible optical network constitutes an efficient alternative facing the exponential rise of traffic. It compensates for the shortcomings of conventional WDM network which are the rigidity of the frequency spectrum and resources waste. Based on the Optical-Orthogonal Frequency Division Multiplexing (O-OFDM) technology, it provides a finer granularity and allows the use of several parallel subcarriers or frequency slots for transmitting the traffic of a connection from a source node to a destination node on a definite physical path. The set defined by subcarriers and the physical path is the optical channel or spectrum path. The bandwidth of the optical channel is the sum of bandwidth of adjacent frequency slots that form it. Each frequency slot with

a bandwidth typically set to 12.5GHz. The number of frequency slots allocated to a connection depends on the flow of its traffic and transmission distance [1]. The allocation of frequency slots is subject to constraints. The first is the contiguity constraint, according to which frequency slots allocated to the same connection must be adjacent. The second constraint is the continuity of the frequency slots allocated to each link of the optical path, i.e., the utilization of the same frequency slots on all links which compose optical path, based on the assumption that there is no conversion. The last constraint is the constraint of non-overlapping, i.e. two different connections cannot simultaneously use the same frequency slots.

The management of these constraints and the multiplicity of frequency slots is a challenge in the determination of optimal spectrum path for a connection. That problem of optimal spectrum allocation is an Integer Linear Program (ILP) problem optimization. It includes two aspects that are the determination of physical path from the source to the destination of the connection and the determination of the frequency slots on that physical path which respects the constraints of flexible optical networks. Besides, this problem known as Routing and Spectrum Allocation (RSA) is reckoned as np-hard [2]. It is an extension of the well-known RWA problem in the conventional WDM networks [3]. The resolution of this problem is even more complex in the case of dynamic traffic variables in time. In fact, the fluctuating traffic of connection leads to resource sharing problems with neighboring connections. One has to ensure that different connections do not simultaneously seek the same frequency slots because it can cause the rejection of the request. This paper aims to contribute at solving that problem by developing a spectrum allocation mechanism which takes into account the possibility of sharing out traffic over different optical paths to reduce the blocking rate when the traffic of a connection undergoes a fluctuation. This mechanism is the multi-spectrum bands allocation; the frequency slots allocated to the connection are on different spectrum bands. A spectrum band

is a portion of the spectrum consisted of several contiguous frequency slots. So here we exploit the possibility of distribution traffic among several spectrum bands of the different path by respecting the differential delay constraint due to Group Velocity Delay (GVD), propagation delay and latency in the nodes.

This study is structured as follows. Firstly Section II, we shall present the characteristics of flexible optical networks as well as solving approaches for time-varying traffic of slots at to varying dynamic traffic time base connection. Secondly, Section III is devoted to our contribution which is an approach based on multi-spectrum bands allocation. Thirdly in Section IV, we evaluate the performance of our proposal by comparing it to existent works. Then, we shall finish with a conclusion and provide perspectives for future works.

II. SPECTRUM ALLOCATION FOR TIME-VARYING TRAFFIC IN FLEXIBLE OPTICAL NETWORKS

A. Feature of flexible optical networks

According to the recommendation G.694.1 of ITU-T, in the flexible optical networks, optical spectrum is divided into subcarriers called also frequency slot. Each frequency slot has a central frequency F_c (in THz) defined by $F_c = 191.3 \times n + 0.125$; n is relative numbers. Furthermore, the spectrum band assigned to a connection consists of several subcarriers which width is $S_B = 12.5 \times m$ [4] with m the number of frequency slots. One of the advantages of flexible optical networks is its capability to adapt the bandwidth defined by the number of frequency slot to the flow of traffic and modulation format according to the transmission distance. The formula in (1) indicates the number of frequency slots required by a connection of capacity C in Gb/s.

$$m = \left\lceil \frac{C}{M \times F_{slot}} \right\rceil \quad (1)$$

$M \times F_{slot}$ represents the capacity of a frequency slot in Gb/s. M ($b/s/Hz$) is the modulation level in bits per second and represents the efficiency of the selected modulation format. F_{slot} is the bandwidth of a frequency slot in GHz. M can take the values 1, 2, 3 or 4 depends on the modulation format is BPSK, QPSK, 8-QAM or 16-QAM [5]. The use of OFDM technology allows simultaneous transmission on parallel subcarriers. However, differential delays that have two sources in the flexible optical network must be taken into account[6]. The first source is the differential delay induced by Group Velocity Delay (GVD) which is related to the fact that the subcarriers of different frequencies have different transmission speeds, even when they are adjacent. That is to say; they form a monolithic block. Thus maximum differential delay caused by GVD related to B spectrum bands on a path length L is given by the formula in the equation (2.1).

$$\Delta d_{max} \approx D(f_c) \times \sum_{1=1}^B (f_{max} - f_{min}) \times L \quad (2.1)$$

Where $D(f_c)$ is the fiber dispersion at the level of central frequency, and f_{max} , f_{min} represent respectively the maximum and minimum frequency of each spectrum band. Let's also note that the fiber dispersion of an SMF (Single Mode Fiber) is $17ps/nm/km$. The second source is the differential delay due to propagation delay and latency in the nodes. The propagation delay of the signal per kilometer is $r_1 = 5\mu s/km$ and maximum latency in a node $r_2 = 25\mu s/km$. Thus the delay of transmission of traffic on a path P of distance L and having N links by adding the GVD is given by (2.2).

$$D_p = L \times r_1 + (N + 1) \times r_2 + \Delta d_{max} \quad (2.2)$$

And the formula (2.3) gives the transmission differential delay (DD) between two physical paths P_i and P_j .

$$DD = \left| D_{P_i} - D_{P_j} \right| \quad (2.3)$$

According to ITU-G 709, the tolerable threshold of DD is 250 μs . But there are commercial products that achieve a tolerable DD up to 128 ms by adding the SDRAM memory Off-Chip [8, 9].

B. Spectrum allocation for time-varying traffic

In the operational phase of the network, traffic is dynamic and fluctuates. There are essentially two types of dynamic traffic. There is in one hand, traffic of connections that have some finite amount of time, and their spectrum paths are released after this short time. The frequency slots belong to these spectrum paths become available for another connection. In the other hand, there is traffic of permanent connections whose traffic fluctuate. That kind of connections that corresponds more to reality, and we call them time-varying traffic connections. The traffic fluctuation of these type of connections translates to an additional demand for frequency slots in the case of traffic expansion or release of some frequency slots in traffic contraction. The policy of dynamically adapt the required frequency slots of a connection to traffic fluctuation is called Spectrum Expansion/Contraction (SEC) policy [10].

In this study, we assume that the request of frequency slots for each connection occurs at a rate λ according to a Poisson process. Moreover, the duration follows an exponential distribution parameter μ equal to 1. Therefore, the load of each connection under this fluctuation is λ/μ . Based on the calculation of λ in [11] with a fixed modulation format, the traffic load ρ is obtained through the following: $\rho = \lambda = M \times m$, where m is the number of frequency slots and M the modulation efficiency of each modulation format, according to principle of "half law" [5]. The values respectively taken by M can be 1, 2, 3 or 4 belong to BPSK, QPSK, 8-QAM or 16-QAM modulations. A time-varying traffic connection has a reference frequency F_R and a number m_{t_i} of frequency slots allocated at time slot t_i . This number of frequency slots allocated to connection R at the time slot t_i is composed by $m_{t_i}^H$ and $m_{t_i}^L$ which are respectively a number of frequency slots used on the lower side and the upper side of the connection's reference frequency. That number $m_{t_i} = m_{t_i}^H + m_{t_i}^L$ is assumed to be constant during each time slot t_i . Figure

1 shows spectrum occupancy by connection R and its neighboring connections.

The spectrum allocation to time-varying traffic connections as indicated by several studies [10, 12] is an ILP optimization problem. It is also known to be np-hard when the network size increases. Heuristics are most common in resolving this problem. These heuristics method of frequency slots allocation to time-varying traffic connections can be classified into two categories as follows: the expansion/contraction frequency slots within the constraints defined in [10] and the combining of expansion/contraction with defragmentation or reconfiguration policies. The first studies of spectrum resources allocation to time-varying traffic connections deals with the methods of expansion/contraction of frequency slots, based on the constraint that two adjacent connections must not use the same frequency slots simultaneously in the same time slot. In this way, the paper [12] provides three policies of spectrum expansion/contraction depending on the fluctuation in traffic. These policies differ from one to other by the limitations suffered by the reference frequency and the allocated spectrum band width. The first of these is the fixed allocation in which the reference frequency and allocated spectrum band width remain fixed and reserved. This spectrum band is exclusively allocated to a particular connection and cannot be used by another connection. Even if, the full width of the spectrum band has is unused. That leads to a waste in case of contraction and request blocking when the request of the connection resulting from the expansion goes beyond the width of the allocated spectrum band. The second policy of spectrum expansion/contraction taking into account is called semi-elastic allocation, in this case, the reference frequency is fixed and can not be moved from its initial position. But, the spectrum bandwidth is allowed to extend or contract according to the traffic fluctuation and the limits of frequency slots occupied by the adjacent connections for each time slot. Frequency slots can be shared out between adjacent connections in different time slots, based on changes in traffic of connections. The third one is the elastic allocation that gives the way to move the reference frequency and extend or restrict the frequency slot block according to traffic variations of the connection. The elastic allocation allows more flexibility in the spectrum managing than the others mentioning above. Three other methods of SEC policies are developed in [10]. There are Constant Spectrum Allocation (CSA), Dynamic High-Low Expansion Contraction (DHL) and Dynamic Alternate Direction (DAD). In the CSA, physical path and a reference frequency are allocated to every connection. The connection has exclusive use of spectrum band located between the reference frequency and the reference frequency of the adjacent connection located above. This method does not allow the sharing out of frequency slots with adjacent connections, even if there are available. As for DHL, it enables the sharing out of frequency slots between adjacent connections. With DHL, a connection that needs to expand its flow first scans the frequency slots which are located on the higher side as its reference frequency to achieve frequency slots already used by the adjacent connection higher reference frequency. Then, if it needs additional frequency slots, it explores the frequency slots on

the side below the reference frequency. Concerning to DAD, it also allows the sharing of frequency slots between connections. However, for the DAD, SEC takes place alternatively in both directions. These methods on-cited define the policy of expansion or contraction of the connections. According to fluctuations in traffic, the increase is limited by the constraint of non-overlapping. Formula (3) defines this constraint of non-overlapping that induces two inequalities.

$$0 \leq m_{t_i}^H \leq \min_e (F_{U_e} - m_{U_e}^L) - F_R - G, \text{ and} \quad (3)$$

$$0 \leq m_{t_i}^L \leq F_R - \max_e (F_{B_e} + m_{B_e}^H) - G$$

U_e and B_e represent the highest and, the lowest adjacent connections of connection R on link e that include in path p of the network represented by a graph (V, E) . V is the set of nodes which can be bandwidth variable transponders (BVTs) or Optical Cross-connect (OXC)s and E is the set of fiber links. F_{B_e} and $m_{B_e}^H$ are respectively the reference frequency and the higher frequency slots used by connection B_e on link e . F_{U_e} and $m_{U_e}^L$ are also respectively the reference frequency and the lower frequency slots occupied by connection U_e on link e . Figure 1 inspired by, thus, provides in [10] shows spectrum occupancy of this adjacent connection on links e and e' .

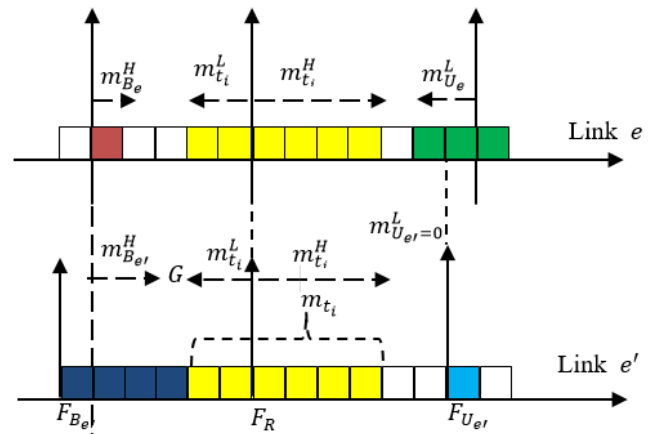


Fig. 1. Spectrum occupancy of connection R and its neighbor connections U_i and B_l on two links e, e' at time t_i

The rise in transmission rate leads to the block of requests when the above SEC policies have been implemented. Even, if the sharing out of frequency slots among adjacent connections are allowed in the SEC policy, blocking will occur certainly when the connection, due to his traffic expansion has already used the available frequency slots in the both side of his reference frequency. In other words, when the equations (4.1) and (4.2) occur and the connection still needs additional frequency slots to satisfy traffic increase.

$$m_{t_i}^H = \min_e (F_{U_e} - m_{U_e}^L) - F_R - G \quad (4.1)$$

$$m_{t_i}^L = F_R - \max_e (F_{B_e} + m_{B_e}^H) - G \quad (4.2)$$

To cope with this blocking of connections, defragmentations policies are implemented. Defragmentation consist in reallocating or rerouting currunt running connections to make enough spectrum paths for a new one or

future arrival connections. These defragmentations policies explore other areas of the spectrum to release necessary contiguous and adjacent frequency slots required by traffic expansion [11]. The first one is the Shift Blocking Neighbors (SBN) which tries to reallocate the adjacent connections without touching the reference frequency. The second method of defragmentation is the Float Blocking neighbors (FBN), unlike the SBN provides the ability to move the reference frequency and connections, therefore, have the opportunity to float in the spectrum. When a new request arrives for a connection, in the absence of frequency slots in the expansion authorized area, the FBN tries to move the connections to have a maximum expansion zone. This tests initially to release the spectrum occupied by its neighbor connection in the direction that maximizes the minimum number of frequency slots available among all its neighbors. If this is not enough, so it tries to move the connection that is in the other direction. However, the approach based on defragmentation, even if it brings a significant reduction in the blocking rate is expensive for the operator because it can cause some interruption of existing connections or require the use of additional transponders. So, to avoid the reverse effects of defragmentation, we propose an allocation strategy based on the multi-spectrum bands.

III. MULTI-SPECTRUM BANDS ALLOCATION IN FLEXIBLE OPTICAL NETWORKS

In this section, we propose an optimization model with the objective of minimization blocking rate. We introduce in this model the concept of multi-spectrum bands and constraint of differential delay, due to: GVD, propagation delay and node latency. We also take into account usually constraints in flexible optical network; those are contiguity, continuity and non-overlapping spectrum. As the model is np-hard; a heuristic method is proposed for solving it.

A. Optimization model

We address the problem of spectrum allocation to time-varying traffic connections with the distribution of traffic load during a specific time slot on multiple paths. This allocation have a priority objective for reducing the blocking rate. The blocking rate is calculated by dividing the number of rejected requests by the total number of requests. It is true that traffic suffers fluctuations during a certain period composed of time slots. We assume that each time slot t demand is constant. Here, is the mixed integer linear programming model. The model relies on the concept of spectrum path which is set of a specific physical path composed by fiber links and spectrum band. Spectrum band is some amount of adjacent frequency slots. Before the optimization model, we give the notations and variables (Table 1).

TABLE I. NOTATIONS AND VARIABLES

Symbol	Description
$G(V,E)$: A graph represents a flexible optical network with nodes in set V and edges in set E .
e	: $e \in E$: link of fiber
s_i	: A frequency slot with index i
S	: Set of frequency slots
t_i	: Time slot of period T
$P_b(T)$: Blocking rate during period T
$N_b(t_i)$: The number of requests blocked during time slot t_i
$N_{tot}(t_i)$: Total number of requests during time slot t_i .
Δ	: Set of all permanent connections in traffic fluctuates
\mathcal{R}	: A particular connection of Δ
D	: Set of all requests resulting from changes in connections that belong to Δ at time slot t_i .
d	: A particular demand in D
$\mathcal{L}(d)$: Set of allowable spectrum path of demand d
$\mathcal{L}(D)$: Set of all allowable spectrum path $\mathcal{L}(D) = \cup_{d \in D} \mathcal{L}(d)$
ζ	: Number of allowable spectrum path in $\mathcal{L}(d)$
ℓ	: A particular spectrum path in $\mathcal{L}(d)$
d_i	: A split of demand d on particular allowable spectrum path ℓ
m_i	: Number of frequency slots allocated to spectrum path ℓ
p_i	: Physical path of spectrum path ℓ
$ p_i $: Number of link e in physical path p_i
$\eta_{\ell\ell'}$: Differential delay between spectrum path ℓ and ℓ'
$x_{d_i\ell}$: Boolean, equal to 1 if d_i use spectrum path ℓ , else it equals to 0
$x_{i,l,l'}$: Boolean variable equals to 1 if physical paths p_i and p_i' share at least one edge, else it is 0
$x_{i,l,l',e}$: Boolean variable equals to 1 if physical paths p_i and p_i' share the same edge e , else it is 0
$x_{i,l,e}$: Boolean variable equals to 1 if physical path p_i used edge e , else it is 0
$x_{i,l}$: Boolean variable equals to 1 if spectrum path l use frequency slot s_i share at least one edge, else it is 0

Objective function

Minimize

$$P_b(T) = \sum_{t_i \in T} \frac{N_b(t_i)}{N_{tot}(t_i)} \quad (5.1)$$

Equation (5.1) is the objective function of the overall block rate during the period T .

the following constraints must be met at every slot time.

Constraints

$$d = \sum_i d_i \quad (5.2)$$

Equation (5.2) is a constraint on the distribution of the flow of traffic during the time slot. Indeed, traffic distribution on the different spectrum paths should equal the overall flow traffic requested by a connection. A traffic demand of particular existent connection is split in several d_i flows.

$$\sum_{d_i} \sum_{l \in \mathcal{L}(d)} d_i l \leq \zeta \quad (5.3)$$

The constraint (5.3) indicates that the number of spectrum paths l in which request d is splitting must be lower or equal the spectrum path that respect the following differential delay constraint (5.4).

$$\forall \ell, \ell' \in \mathcal{L}(d) : \eta_{\ell\ell'} \leq \tau \quad (5.4)$$

$$\sum_{e \in p_l} \sum_i x_{p_l, e, i} = |p_l| \times m_l \quad (5.5)$$

Equation (5.5) expresses the constraint of contiguity or adjacency on each link of spectrum path. In each link, m_l frequency slots forming the spectrum band must be contiguous. To ensure the continuity, the contiguity constraint of the same m_l frequency slots must be met in all the $|p_l|$ links of the path p_l .

$$\forall \ell, \ell' \in \mathcal{L}(D), l \neq l', e \in E : x_{l, \ell'} \leq \sum_e x_{l, \ell', e} \quad (5.6)$$

$$\forall \ell, \ell' \in \mathcal{L}(D), l \neq l', e \in E : x_{l, \ell', e} \leq x_{l, e} \quad (5.7)$$

$$\forall \ell, \ell' \in \mathcal{L}(D), l \neq l', e \in E : x_{l, \ell', e} \leq x_{l', e} \quad (5.8)$$

$$\forall \ell, \ell' \in \mathcal{L}(D), l \neq l', s_i \in S : x_{l, i} + x_{l', i} + x_{l, l'} \leq 2 \quad (5.9)$$

Constraints (5.6), (5.7), (5.8) and (5.9) ensure the non-overlapping constraints. One frequency slot can not be used simultaneously by two different spectrum paths that share at least one link. The variable $x_{l, l'}$ is a boolean that takes value 1 when spectrum paths l and l' share at least one link, otherwise it takes value 0. His value is determined by constraints (5.6), (5.7) and (5.8). As to the constraint (5.9), it ensures frequency slot s_i can not be allocated to spectrum paths l and l' , if they have a link sharing. If frequency slot s_i is used simultaneously by spectrum paths l and l' then $x_{l, i} + x_{l', i} + x_{l, l'} = 3$ or that is impossible because whatever values, this constraint (5.9) must be less than or equal to 2.

This problem output is the set of spectrum paths allocated to a set of demands when permanent connections traffics fluctuate. As in each time slot, the novel request of the spectrum is constant. We can say that this problem can be seen as offline routing and spectrum assignment in each time slot [13]. So, this problem is np-hard. When the network size grows the time to find an optimal solution also grows exponentially as several studies have already shown it [2, 11, 12]. In that condition, the suitable heuristic can be used to obtain near optimal solution within a reasonable time. With this end, in the following section, we shall propose a heuristic based on multi-spectrum bands concept in which it contributes to avoiding the worst effect when defragmentation or reconfiguration policies are used when blocking occur.

B. Heuristic multi-spectrum bands allocation

Our heuristic relies on the principle of combining the implementation of the DAD with the multi-spectrum band allocation. This concept of Multi-Spectrum Band Allocation (MSBA) means that traffic is progressively distributed over different spectrum paths. Each of these spectrum paths is formed by a physical path and a spectrum band so that the sum of all spectrum bands can support all traffic. After running the DAD, in the case of dissatisfaction of the entire request, another optical path is used to bear the remaining traffic and so on until there is no traffic. As regards the distribution of traffic on the spectrum bands, each spectrum band used to carry some traffic must have a minimum width. This precaution must be taken to avoid a proliferation of spectrum bands taken into account the transmission of traffic. Indeed, the proliferation of spectrum bands causes waste of frequency slots. It is necessary to associate a guard band to reduce interference between different adjacent spectrum paths. For the allocation spectrum band to traffic supplement after the DAD, we choose a spectrum path which allows allocating the entire traffic supplement. In the absence of such a path, we divide the traffic supplement on the k-shortest paths of the connection whose traffic fluctuates. We assumed that the k-shortest paths of each connection are determined by Yen's algorithm [14] and an initial path with a reference frequency is associated with each connection. The modulation format used is chosen according to the transmission distance and follow the "half-distance low"[5]. The heuristic can be reflected by the following steps:

Step 1: Run the Dynamic Alternate Direction (DAD) as a policy of expansion/ contraction of the spectrum when the traffic of connection fluctuates.

Step 2: In case of dissatisfaction,

Select all the paths respecting differential delay constraints. If there is one which supports all the traffic, make the allocation on this path.

Otherwise, go to step 3

Step 3: Select the first path that offers the largest spectrum band

Step 4: Make the allocation of traffic supplement, if not the entire satisfaction. Return to Step 3 to find another path to allocate the remaining traffic. Repeat this until there is no more traffic or no more path.

The algorithm below presents the pseudo-code that reflects the MSBA heuristic. We assume that at the end of time slot t_i , traffic request for time slot t_{i+1} is known. Here are the meaning of the variables of the pseudo-code:

- $\mathcal{C}(t_{i+1}, s, d)$: the flow of traffic required at time t_{i+1} by the connection from the source s to destination d .
- $SA(t_{i+1}, s, d)$: the allocation of frequency slots to a connection at time t_{i+1} from source s to destination d .
- SD: number of frequency slots corresponding to traffic $\mathcal{C}(t_{i+1}, s, d)$ after choosing an appropriate modulation format.

- SDAD: Number of frequency slots available around the reference frequency.
- $C_r(t_{i+1}, s, d)$: Traffic flow remaining after execution of the Dynamic Alternative Direction.
- SSUP: additional frequency slots corresponding to $C_r(t_{i+1}, s, d)$
- SDISP: frequency slots available on a particular path P_{diff}
- DAD: Dynamic Alternative Direction
- P_{diff} : The subset of k-shortest path connecting that respect the differential delay.

Algorithm of MSBA heuristic

Input: Network $G(V,E)$, Set Δ of connections
 $SA(t_i, s, d)$
 $C(t_{i+1}, s, d)$

Output: $SA(t_{i+1}, s, d)$

BEGIN

```

1:  $SD \leftarrow \frac{C(t_{i+1}, s, d)}{M \times F_{slot}}$ 
2: Run dynamic alternative direction SEC policy
3:  $C_r(t_{i+1}, s, d) \leftarrow (SD - SDAD) \times M \times F_{slot}$ 
4: While ( $C(t_{i+1}, s, d) \neq 0$  and  $P_{diff} \neq \emptyset$ ) do
5:   For all path in  $P_{diff}$  do
6:      $SSUP \leftarrow \frac{C_r(t_{i+1}, s, d)}{M \times F_{slot}}$ 
7:     If SSUP is available in current  $P_{diff}$  path then
8:       Assign spectrum on this path to the
9:       connection
9:       Break for
10:    Else
11:      Select among the path in  $P_{diff}$  the one
12:      which have the highest SDISP
13:      Assign spectrum with size SDISP on this
14:      path
14:      Extract this path from  $P_{diff}$  set
14:       $C_r(t_{i+1}, s, d) \leftarrow (SSUP - SDIP) \times M \times F_{slot}$ 
15:    End for
16:  End while
17:  If  $C_r(t_{i+1}, s, d) = 0$  then spectrum assignment is
18:  accomplished
18:  Else Spectrum assignment is blocked
END

```

IV. SIMULATION AND ANALYSIS OF RESULTS

The simulations to evaluate the performance of our proposal were carried out on two network topologies. The NSFNET network (14 knots, 22 links) well-known network is using to simulate most of the work in this area. The second network is the national Backbone of Ivory Cost from the will of the State of Ivory Cost to mesh all Ivorian territory with a fiber optical transport (28 nodes and 46 links) as shown in Figure 2.

A guard band is $G = 1$ slot, and as regards the choice of modulation format, each path has a format according to the principle of "half distance law" [5]. Furthermore, we assume that each connection is associated with a reference frequency and 4-shortest paths determined by the Yen's algorithm [14]. It takes into account the initial path in which the connection is running before the fluctuation in traffic occurs. The differential delay is set at $\tau = 250\mu s$ as recommended by the ITUT.709. As for the distribution of traffic load, the minimum number of frequency slots corresponding to an eligible distribution is three frequency slots. It means the minimum width of spectrum band is three frequency slots. As for the dynamic traffic, each connection frequency slots request follows the Poisson process with a transmission flow λ and duration which has an exponential distribution parameter $\mu = 1$ as described in (Section II.B).

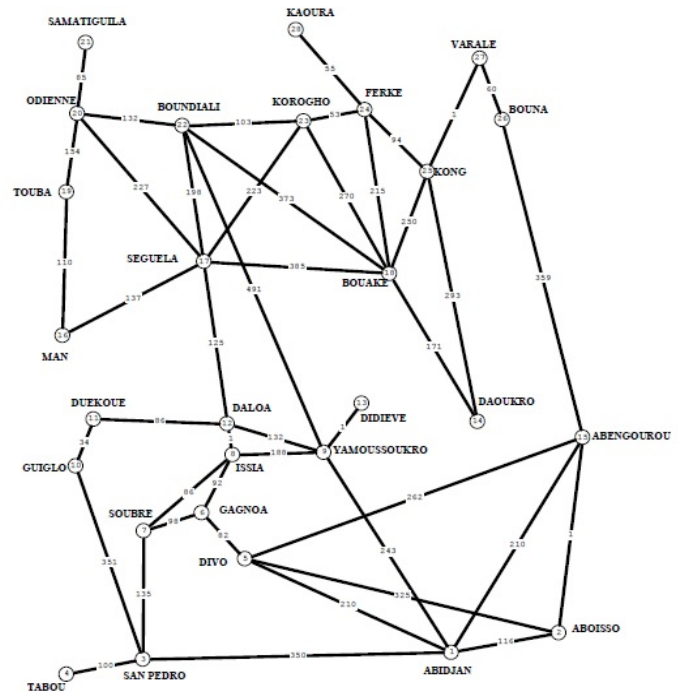


Fig. 2. Ivory Coast National Backbone Network

The simulations carried out in Java on a PC with a Dual Core processor (2.16 GHz) and memory 4GB. The aim was to observe the impact of traffic load and spectrum available on the blocking rate. Dealing with the impact of the traffic load, we vary the load from 200 erlangs (E) to 1600 E by keeping the number of frequency slots per link at 100 frequency slots. Then, we evaluate its effect on the blocking rate. As for the impact of the number of frequency slots available by link, we vary the number of frequency slots from 100 to 240 frequency slots maintaining the traffic load at 1000 E.

The observed results are in the graphs of Figures 3 and 4. Figure 3(a) shows the performance of the Dynamic Alternate Direction policy (DAD), Float Blocking neighbors (FBN), and our Multi-Spectrum Band Allocation policy (MSBA) on the blocking rate depending on the network load for the topology NSFNET.

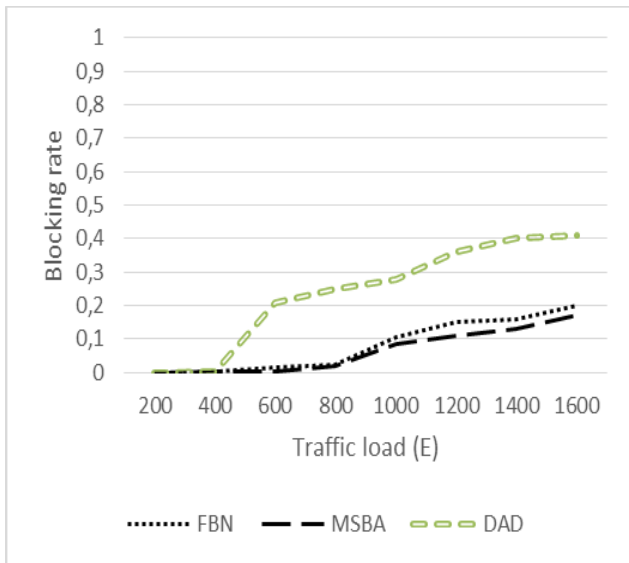


Figure 3(a)

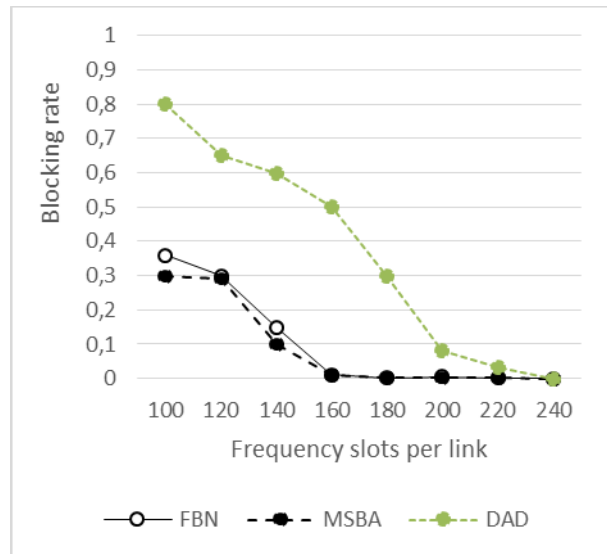


Figure 4(a)

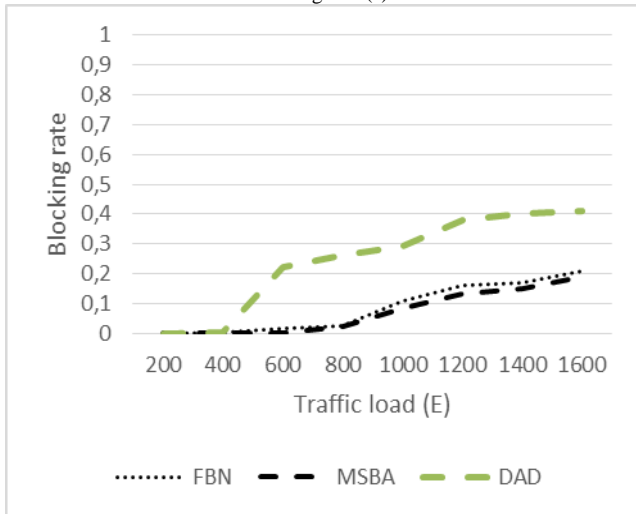


Figure 3(b)

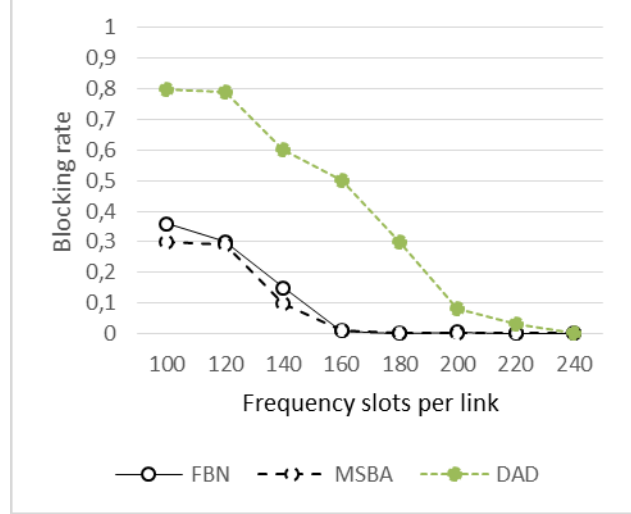


Figure 4(b)

Fig. 3. (a) Blocking rate as a function of Traffic load in the NSFNET. (b) Blocking rate as function of Traffic load in the Ivorian National Backbone

Fig. 4. (a) Blocking rate as a function of available slots in the NSFNET. (b) Blocking rate as a function of available Fiber link of the Ivorian National Backbone

The number of frequency slots per fiber equal to $\delta = 100$ when the network load varies from 200 to 1600 E. The MSBA has better performance than the DAD from 400E. But, regarding the FBN, method based on defragmentation, gives the same rates of blockages like MSBA under 1000E. After reaching 1000E, slight improvement of the blocking rate is noted.

Figure 3 (b) reproduces the same trends observed with the NSFNET topology on another national backbone topology Ivory Coast. In these simulations, it appears that MSBA reduces the blocking rate of about 16% by varying the traffic load than DAD method. Regarding the performance of MSBA compared to FBN, we remark a reduction in the blocking rate of about 3%.

Figure 4 (a) shows the performance of the Dynamic Alternate Direction (DAD), Float Blocking neighbors (FBN), and Multi-Spectrum Bands Allocation method (MSBA) on the block rate, based on the number of frequency slots with NSFNET topology. It sets the total traffic load to 1000 Erlangs. The number of link frequency slots varies from 100 to 240 frequency slots. As expected, the blocking rate decreases when the number of frequency slots increases. Also, compared to other methods MSBA provides a blocking rate significantly lower than the DAD. Nevertheless, beyond 160 frequency slots. The MSBA and FBN have almost the same performance. We observe the same trends with the national Ivory Coast backbone topology.

V. CONCLUSION

In this work, we have dealt with the problem of dynamic allocation of frequency slots in the flexible optical networks. The situation tackles the connections that have undergone fluctuation. We have offered a heuristic based on the distribution of the additional traffic load on strips of different spectrum paths. This method has simulated on the NSFNET network and Ivorian National Backbone topologies, allows a far interesting reductions on blocking rate when the fluctuation of traffic demand leads to an additional frequency slots. MSBA heuristic exploits the advantages of flexibility of modulation offered by flexible optical networks. Moreover, it appears as an interesting policy of blocking reduction for time-varying traffic connections. By avoiding harmful effects related to defragmentation method that often bring about the interruption of existing connections.

Our future reflections will focus on the recurring character of the traffic fluctuations, that has been observed in certain types of traffic, for instance Internet traffic; one can introduce predictions methods in spectrum allocation knowing the recurrence of traffic.

REFERENCES

- [1] JINNO, Masahiko, TAKARA, Hidehiko, KOZICKI, Bartlomiej, et al. Spectrum-efficient and scalable elastic optical path network: architecture, benefits, and enabling technologies. *Communications Magazine, IEEE*, 2009, vol. 47, no 11, p. 66-73.
- [2] LINKOWSKI, Miroslaw et WALKOWIAK, Krzysztof. Routing and spectrum assignment in spectrum sliced elastic optical path network. *IEEE Communications Letters*, 2011, vol. 15, no 8, p. 884-886.
- [3] ZANG, Hui, JUE, Jason P., MUKHERJEE, Biswanath, et al. A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks. *Optical Networks Magazine*, 2000, vol. 1, no 1, p. 47-60.
- [4] ITU-T, Spectral grids for wdm applications:Dwdm frequency grid, Recommendation G.694.1, February 2012
- [5] A. Bocoi, M. Schuster, F. Rambach, M. Kiese, C.-A Bunge, and B. Spinnler, "Reach-dependent capacity in optical networks enabled by OFDM," OFC/NFOEC 2009, Paper OMQ4, March 2009.
- [6] CHEN, Xiaomin, JUKAN, Admela, et GUMASTE, Ashwin. Optimized parallel transmission in elastic optical networks to support high-speed Ethernet. *Journal of Lightwave Technology*, 2014, vol. 32, no 2, p. 228-238.
- [7] Interfaces for the optical transport network (OTN),"ITU-T Recommendations.[Online]. Available: <http://www.itu.int/rec/T-REC-G.709/e>
- [8] Intelixf19301datasheet."[Online]. Available: <http://www.intel.com/design/network/products/optical/tsp/IXF19301.htm>
- [9] Cisco xsl-series.[Online]. Available: <http://www.cisco.com/en/US/products/hw/modules/ps2710/ps5479/index.html>.
- [10] CHRISTODOULOPOULOS, Konstantinos, TOMKOS, Ioannis et VARVARIGOS, Emmanouel. Time-varying spectrum allocation policies and blocking analysis in flexible optical networks. *Selected Areas in Communications, IEEE Journal on*, 2013, vol. 31, no 1, p. 13-25
- [11] STIAKOGIANNAKIS, Ioannis, PALKOPOULOU, Eleni, KLONIDIS, Dimitrios, et al. Dynamic cooperative spectrum sharing and defragmentation for elastic optical networks. *Journal of Optical Communications and Networking*, 2014, vol. 6, no 3, p. 259-269
- [12] KLINKOWSKI, Miroslaw, RUIZ, Marc, VELASCO, Luis, et al. Elastic spectrum allocation for time-varying traffic in flexgrid optical networks. *IEEE journal on selected areas in communications*, 2013, vol. 31, no 1, p. 26-38.
- [13] CHATTERJEE, Bijoy Chand, SARMA, Nityananda, et OKI, Eiji. Routing and spectrum allocation in elastic optical networks: a tutorial. *IEEE Communications Surveys & Tutorials*, 2015, vol. 17, no 3, p. 1776-1800.
- [14] YEN, Jin Y. Finding the k shortest loopless paths in a network. *management Science*, 1971, vol. 17, no 11, p. 712-716.

Robust Image Watermarking using Fractional Sinc Transformation

Almas Abbasi

Department of Computer Science and software engineering
International Islamic University
44000 Islamabad, Pakistan

Chaw Seng Woo

Faculty of Computer Science and Information Technology
Department of Artificial Intelligence
University of Malaya
50603 Kuala Lumpur, Malaysia

Abstract—The increased utilization of internet in sharing and dissemination of digital data makes it is very difficult to maintain copyright and ownership of data. Digital watermarking offers a method for authentication and copyright protection. Digital image watermarking is an important technique for the multimedia content authentication and copyright protection. This paper present a watermarking algorithm making a balance between imperceptibility and robustness based on fractional calculus and also a domain has constructed using fractional Sinc function (FSc). The FSc model the signal as polynomial for watermark embedding. Watermark is embedded in all the coefficients of the image. Cross correlation method based on Neyman-Pearson is used for watermark detection. Moreover fraction rotation expression has constructed to achieve rotation. Experimental results confirmed the proposed technique has good robustness and outperformed another technique in imperceptibility. Furthermore the proposed method enables blind watermark detection where the original image is not required during the watermark detection and thus making it more practical than non-blind watermarking techniques.

Keywords—*Fractional Calculus; fractional Sinc; image Watermarking; robust*

I. INTRODUCTION

Receiving and transmitting digital data has instigated its wide appearance and storage. The use of digital data on a broader scale has brought a lot of ease in different aspects, but it is not without side effects such as tempering and copyright protection issues. With digital data so widely used, watermarking are mostly used to address these issues. Watermarking is a method to embed a message while stenography is the art of hidden communication. The purpose of watermarking is to keep the message secret whereas data hiding is general term and covers a vast range of problems related to making information confidential.

In general watermarking techniques necessitates certain properties. These properties are required for all kinds of multimedia data such as audio, video and images. However, the significance of these properties varies with the purpose and application of watermarking. In case of copyright protection application the watermark should be robust enough to resist any attempt for its removal. While for the authentication applications, robustness is not required. The Fundamental watermarking system properties are: Imperceptibility and robustness.

Watermarking schemes are usually based on special domain methods [1, 2, 3] as well as transformed domain techniques such as DCT, DFT, and wavelet, etc. Watermarking in spatial domain is straight forward and easy to implement as compared to watermarking in transformed [4, 5, 6] domain. Historically spatial domain watermarking was the first watermarking scheme the researcher had investigated upon. However, it has low robustness compared to transformation domain as the watermark easy be obliterated by lossy image compression techniques. Most common example of spatial domain method is Least significant bit (LSB). The watermark embedded in the least bit is least perceive by human eye. However the spatial domain methods are not robust i-e. They cannot survive various kinds of attacks as the least bit are usually removed or destroyed in these attacks. Transform domain methods embed the watermark into the transformed coefficients of the original image. Transform based methods are most popular as they are generally more robust to malicious attacks. In transform domain method the transformed coefficient of the original image is altered to embed watermark. Researchers have considered many image transformations. Most popular transform domains are discrete cosine transformation (DCT), discrete wavelet transforms (DWT), and discrete Fourier transforms (DFT). Moreover, polynomial transformation based watermarking techniques are proposed by researchers to obtain improved imperceptibility and robustness. Researchers have also used polynomial based transformation for watermarking [7-9].

Most of the above mentioned techniques embed watermark in selected areas to get robust watermarking. In the proposed technique watermark is embedded in the whole image while in order to achieve balance between imperceptibility and robustness new model based on Fractional order of Sinc is used.

II. FRACTIONAL CALCULUS

Fractional calculus is a mathematical discipline which deals with derivatives and integrals of arbitrary real or complex orders. The history of fractional calculus is 300 years old and genuine expansion and progression has seen in 19th century. During the last decade fractional calculus has been applied to different fields most noticeable among them are physics, biology, science, engineering, image processing and other fields. Fractional Calculus generalized the ideas of integer

order differentiation and n-fold integration. Fractional order differentiators and integrators are example of fractional order systems. Fractional order systems are defined by fractional order differential equations. Fractional order differentiators and integrators are used to compute the fractional order time derivative and integral of the given signal [10-15]. Fractional derivatives introduce an excellent instrument for the description of general properties of various materials and processes such as signal processing and image processing [16]. One of the advantages of fractional calculus is that it can be well thought-out as a super set of integer order calculus. Thus fractional calculus has the potential to realize what the integer order calculus cannot. One of these famous operators is the Riemann-Liouville operator (differential and integral), which defined in definition 2.1

A. The Sinc function and its properties

1. The limit of sinc:

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1.$$

i-e lim of sinc(x) approaches to Heaviside function.

2. Expression for universal constant π :

$$\sin c(x) = \left(1 - \frac{x^2}{1^2 \pi^2}\right) \left(1 - \frac{x^2}{2^2 \pi^2}\right) \left(1 - \frac{x^2}{3^2 \pi^2}\right) \dots \mathcal{D}^\alpha t^\mu = \frac{\Gamma(\mu+1)}{\Gamma(\mu-\alpha+1)} t^{\mu-\alpha}, \mu > -1; 0 < \alpha < 1$$

3. **Fourier analysis:** The sinc function appears frequently in Fourier analysis. This appearance of sinc explains its occurrence in areas of engineering that rely on the analysis of signals such as communication theory. Fourier analysis allow us to expand the function in a trigonometric series of the form:

$$f(x) = a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \dots$$

Where a_i and the b_i are constants.

4. **Shannon entropy:** for a function f , one can interpolate it as follows:

$$C(x) = \sum_{n=-\infty}^{\infty} f(nh) \sin c \pi \left(\frac{x}{h} - n \right).$$

Where $n = 0, \pm 1, \pm 2, \pm 3, \dots = 0$, and $h > 0$. [17]

5. **Jordan's inequality:**

$$|\sin c(x)| \leq 1.$$

This lead to all the derivations of sinc(x) are also less than 1.

6. sinc is an even function:

$$\sin c(-x) = \sin c(x).$$

7. **Basic function:**

We call a function $\phi_n(t)$ basic functions of signal expansion:

$$\delta(t) = \sum_n b_n \phi_n(t).$$

Definition 2.1. The fractional (arbitrary) order integral of the function of order $\alpha > 0$ is defined by

$$I_a^\alpha f(t) = \int_a^t \frac{(t-\tau)^{\alpha-1}}{\Gamma(\alpha)} f(\tau) d\tau.$$

Definition 2.2. The fractional (arbitrary) order derivative of the function of order $0 < \alpha < 1$ is defined by

$$D_a^\alpha f(t) = \frac{d}{dt} \int_a^t \frac{(t-\tau)^{-\alpha}}{\Gamma(1-\alpha)} f(\tau) d\tau = \frac{d}{dt} I_a^{1-\alpha} f(t).$$

Remark 2.1. From Definition 2.1 and Definition 2.2, we have

$$I_a^\alpha t^\mu = \frac{\Gamma(\mu+1)}{\Gamma(\mu+\alpha+1)} t^{\mu+\alpha}, \mu > -1; \alpha > 0.$$

This paper investigate the utilization of the Sinc basis functions

Definition 2.3. The function Sinc(t) defined by

$$\text{Sinc}(t) = \frac{\sin(t)}{t}, \quad t \neq 0 \tag{1}$$

For fractional Sinc(t) of order α , (SC_α), defined by [10].

$$SC_\alpha(t) = \frac{\sin_\alpha(t)}{t} = \sum_{n=0}^{\infty} \frac{(-1)^n t^{(2-\alpha)n+1}}{\Gamma((2-\alpha)n+2)}, \tag{2}$$

$$S = s + (w_b \times \kappa) \tag{3}$$

Where

$$\sin_{\alpha}(t) = \sum_{n=0}^{\infty} \frac{t^{n-\alpha}}{\Gamma(n-\alpha+1)} \sin\left((n-\alpha)\frac{\pi}{2}\right), \quad \Gamma$$

is the gamma function, t is the variable and $\alpha \in (0,1)$ is a constant. From Definition 2.3, the following coefficients are obtained

$$\phi_0 = \frac{1}{\Gamma(2)}$$

$$\phi_1 = \frac{(-1)}{\Gamma(4-\alpha)}$$

$$\phi_2 = \frac{1}{\Gamma(6-2\alpha)}$$

⋮

$$\phi_n = \frac{(-1)^n}{\Gamma((2-\alpha)n+2)}$$

III. PROPOSED TECHNIQUE

The watermarking domain used in this technique is transform domain based on FSc. The watermark embedding and extraction is performed in FSc domain. Further the watermarking technique used in this study is blind watermarking technique. In blind watermarking technique original image is not required for the watermark detection and is more practical than non-blind watermarking techniques.

A. Balanced watermarking

The basic requirement of a Watermarking technique is to achieve high imperceptibility and robustness. These two properties are inverse to each other. When the imperceptibility increase, robustness decrease and vice versa. A balance between imperceptibility and robustness is required for a watermarking technique. In this technique Fractional calculus transformation has been used to achieve the balance.

B. Watermark embedding

Let $I(i, j)$ represent the original gray scale image of size $M \times N$ pixels and $w(i, j)$ is the watermark pattern to be embedded using additive embedding technique. Generally, additive embedding is implemented by using $I' = I + k w$, where I' is the watermarked image, and k is the embedding strength. The size of the watermark w is equal to the size of the image selected for watermark embedding. The strength of watermark k is kept as a constant value.

For the current implementation, the watermark signal consists of $\{+1, -1\}$ bits. The watermark embedding process can be represented by the following equation:

where S represents the watermark signal, s is the original signal (i.e. coefficients in our case), w_b is the watermark bit. The parameter κ is a constant value.

The proposed watermarking embedding technique includes the following steps:

- Read the grayscale images of size 512×512 .
- Apply fractional Sca transformation:
 - a. Set the parameter $\alpha > 0$.
 - b. Fix the value of the variable t .
 - c. Calculate the fractional order of Sca polynomial using Definition 2.3
- Embed the watermark using the Eq. (3).
- Calculate inverse fractional Sca by taking transpose of the resultant image.
- Perform above steps for each image.

C. Watermark detection

Watermark detection is in fact the reverse of the embedding processes. Watermarked image transformed using FSc. The correlation between the watermarked coefficient and the watermark to be tested for the existence is computed using the following expression:

$$\rho = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} I'_0(i, j) w(i, j) \quad (4)$$

where $I'_0(i, j)$ is the FSc coefficient and $w(i, j)$ represents the watermark. MN is the size of the image. The computed value of ρ is then compared to the threshold value T_ρ calculated as follows:

$$T_\rho = 3.97 \sqrt{2\sigma_{\rho\beta}^2} \quad (5)$$

where $\sigma_{\rho\beta}^2$ represents the variance and

$$\sigma_{\rho\beta}^2 \approx \frac{1}{(MN)^2} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (I'_0(i, j))^2 \quad (6)$$

For the detailed derivation of these equations, please refer the interested readers to [6]. The existence of the watermark will be confirmed when $\rho > T_\rho$.

The proposed technique watermark extraction steps are summarize in the following pseudo code:

- Read the watermark/attacked image.
- Apply FSc transformation on the image.
- Compute ρ and T_ρ using equation 4 and 5 respectively.
- Compare ρ and T_ρ , If $\rho > T_\rho$ then watermark is detected otherwise not.
- Repeat above steps for all images.

IV. EXPERIMENTS ANALYSIS

In this section, the performance of the proposed watermarking technique has evaluated by considering robustness and imperceptibility. Five standard test images from the USC-SIPI dataset [18], namely, Baboon, Cameraman, Lena, Peppers, and Sailboat, are considered for evaluation purposes. These images are each of dimensions 512×512. As a proof of concept, the algorithm is coded by using Matlab and checkmark software is deployed for testing the robustness against different set of attacks. The watermark signal is presented by the sequence of +1 or -1.

TABLE I. PSNR, MSE AND SSIM VALUE OF SAMPLE TEST IMAGES IN THE PROPOSED FSC DOMAIN

	Lena	Baboon	Cameraman	Sailboat Boat	Peppers
PSNR	42.02	41.80	42.52	41.54	41.37
MSE	4.09	4.30	3.64	4.56	4.74
SSIM	0.93	0.98	0.92	0.95	0.94

A. Imperceptibility

The watermarked images obtained by using the proposed technique where total 262144 bits of watermark is embedded. By visual inspection, the watermarked images appear perceptually similar to their original counterparts. To quantify the transparency of the embedded watermark, the peak signal to noise ratio (PSNR) and structure similarity index measure (SSIM) are considered, which are commonly used by the watermarked community. The results are recorded in Table 1. It is observed that the PSNR and SSIM values range from 41.37 to 42.52dB and 0.92 to 0.98, respectively. These readings suggest that the watermark image generated by the proposed method is of high perceptual quality.

B. Robustness

Fig 1(a-e) summaries the different set of attacks performed for the evaluation of our proposed watermarking technique. Cross correlation method based on Neyman-Pearson is deployed to detect the embedded watermark and its formula is detailed in [6].

In our experiment, the watermarked images have gone various types of attack to investigate the robustness of the proposed technique. In particular, each watermarked image is distorted using different geometric and image processing attacks, namely: (1) rescaling attack with scaling factor ranging from 0.5 to 2; (2) JPEG compression with quality factor ranging from 50% to 90% with increment of 10; (3) row and column removal attack with the number of rows and column removal varying from 1 to 17; (4) change in aspect ratio in the x and y directions; (5) Gaussian filtering with kernel size of 3×3 and 4×4 pixels; (6) sharpening attack. (7) Cropping attack with 10, 20 and 50% cropped relative to the size of image. Fig 3.(a- t) shows the result of applying these attacks.

The cross correlation computation based on Neyman-Pearson criterion and the threshold value are considered to test the presence of the embedded watermark.



Fig. 1. Robustness of Proposed Technique against different types of attacks. (a) Comparison of average correlation values of five standard images after JPEG compression attack with quality factors changes from 50 to 90, against dynamic threshold. (b) Comparison of average correlation values of five standard images after scaling attack of different scaling factors i-e from 0.5 to 2, against dynamic threshold. (c) Aspect ratio attack of 2:7 in relation to x and y axis. (d) Row Column removal attack (17, 5). (e) Sharpening attack



Fig. 2. (a-c) Original test images, (d-f) Attacked watermark images for different cropped size such as 10%, 20%, and 50% of the image size respectively

The scaling attack using various factors ranging from 0.5 to 2.0, remains well above the (i.e., threshold) considered as suggested by Fig. 1(b). Robustness against JPEG compression with quality factor ranging from 50% to 90% are shown in Fig 1(a). The presence of watermark is detected in all these cases, as the Correlation values lies quite above the threshold value. Moreover, common image processing operations such as, aspect ratio change Fig 1(c), row and column removal attack Fig 1(d), sharpening and Gaussian filtering Fig 1 (e) are also applied on the watermarked image. For all cases, the dynamically computed value always stays above the threshold , i.e., 100% successful detection. Therefore, the results suggest that the proposed method is robust against the commonly considered watermarking distortion attacks. Fig 3 shows the images after the above mentioned different set of attacks.

Fig 2. Represent cropping attack performed on the watermark image to test the robustness against different size of cropped attempt. Fig 2(a-c) represent original test images while Fig 2(d-f) represent the cropping using 10%, 20%, and 50% of image size. This shows that the proposed watermarking scheme is robust against the cropping attack.

V. FRACTIONAL MATRIX EXPRESSION FOR ROTATION

A new rotation expression has derived using fractional trigonometric functions. Fractional rotation is a special case of fractional sinc. Now let us examine rotation of images by utilizing method based on the fractional derivative D^α of $\sin(t)$ and $\cos(t)$. The following result are obtained.

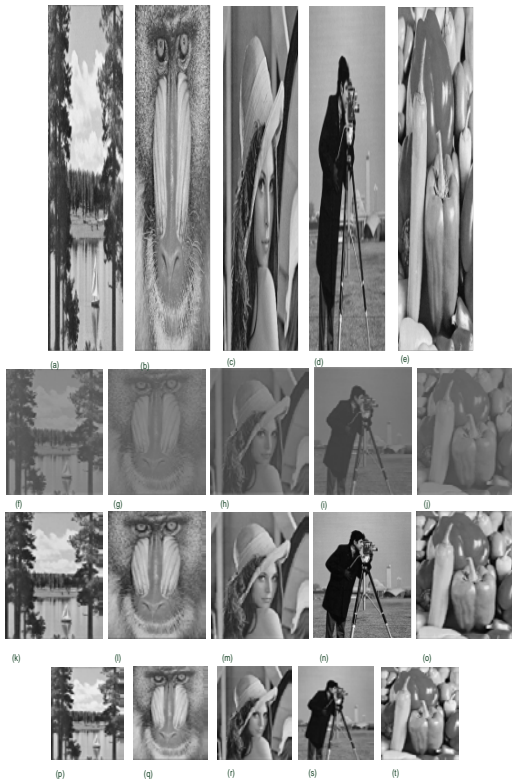


Fig. 3. Experimental results:(a) to (e) show the watermarked test images manipulated by using aspect ratio of 2:7 in relation to x and y axis; (f) to (j) represent watermarked images corrupted by using sharpening attack;(k) to (o) illustrate the watermarked images are compressed with JPEG compression

attack with quality factor 50;(p) to (t) display the scaled down to 50% of the image size

Lemma 5.1 Let $\cos_\alpha(t)$ and $\sin_\alpha(t)$ defined respectively by

$$\cos_\alpha(t) = \sum_{n=0}^{\infty} \frac{t^{n-\alpha}}{\Gamma(n-\alpha+1)} \cos((n-\alpha)\frac{\pi}{2}) \quad (7)$$

And

$$\sin_\alpha(t) = \sum_{n=0}^{\infty} \frac{t^{n-\alpha}}{\Gamma(n-\alpha+1)} \sin((n-\alpha)\frac{\pi}{2}) \quad (8)$$

The fractional rotation obtained are:

$$R_\alpha = \begin{pmatrix} \cos(\alpha\pi/2) & \sin(\alpha\pi/2) \\ -\sin(\alpha\pi/2) & \cos(\alpha\pi/2) \end{pmatrix}.$$

Proof. By using Definition 2.2 together with Remark 2.1, we obtain that

$$D^\alpha \sin(t) = \sin(\alpha\pi/2) \cos_\alpha(t) + \cos(\alpha\pi/2) \sin_\alpha(t) \quad (9)$$

and

$$D^\alpha \cos(t) = \cos(\alpha\pi/2) \cos_\alpha(t) - \sin(\alpha\pi/2) \sin_\alpha(t) \quad (10)$$

A geometrical interpretation of the derivative relations in Eq 9 and Eq 10 can be found by imposing the matrix form

$$R_{D^\alpha} = \begin{pmatrix} D^\alpha \cos(t) & D^\alpha \sin(t) \end{pmatrix} \quad (11)$$

Collecting the coefficients of R_{D^α} , the following fractional rotation matrix obtained:

$$R_\alpha = \begin{pmatrix} \cos(\alpha\pi/2) & \sin(\alpha\pi/2) \\ -\sin(\alpha\pi/2) & \cos(\alpha\pi/2) \end{pmatrix} \quad (12)$$

By applying rotation (8) of different degrees such as 5, 10, 15, 20, 25, 30, 35, 40 to the proposed scheme achieve fractional rotation. Result are shown in Fig 4.

Remark 5.2 One can use the transpose of R_{D^α} to get good result as well.

$$R_\alpha^T = \begin{pmatrix} \cos(\alpha\pi/2) & -\sin(\alpha\pi/2) \\ \sin(\alpha\pi/2) & \cos(\alpha\pi/2) \end{pmatrix}$$

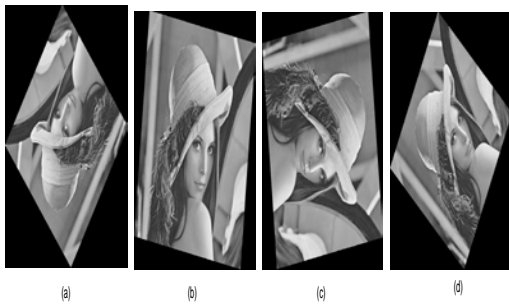


Fig. 4. (a-d) Rotation achieved using fractional rotation expression having angle 15, 25, 30 and 45 respectively

VI. CONCLUSION

Digital image watermarking is an important technique for the multimedia content authentication and copyright protection. Some watermarking techniques are extremely robust but they suffer poor imperceptibility. This paper introduces a watermarking algorithm balanced between imperceptibility and robustness based on fractional calculus. A new domain has been constructed using fractional order of Sinc (FSc). Moreover, fractional rotation expression has been designed to achieve rotation. The FSC model the signal as a fractional polynomial for watermark embedding. Watermark is embedded in all the coefficients of the image. Cross correlation method based on Neyman-Pearson is used for watermark detection. Experimental results confirmed the proposed technique is robust and imperceptible.

ACKNOWLEDGMENT

Authors would like to thank all those who provide technical help and correction in writing this article.

REFERENCES

[1] Nikolaidis, N. and Pitas, I., 'Robust image watermarking in the spatial domain', *Signal processing*, 66, pp.385-403,1998.
[2] Darmstaedter, V., Delaigle, J.-F., Nicholson, D. and Macq, B., 'A block based watermarking technique for MPEG2 signals: Optimization and validation on real digital TV distribution links', in *Multimedia Applications, Services and Techniques—ECMAST'98*: Springer, pp. 190-206,1998.

[3] Wolfgang, R.B. and Delp, E.J., 'A watermark for digital images', in *Image Processing, 1996. Proceedings. International Conference on*: IEEE, pp. 219-222, 1996
[4] Hernandez, J.R., Amado, M. and Perez-Gonzalez, F., 'DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure', *Image Processing, IEEE Transactions on*, 9, pp.55-68,2000.
[5] Kundur, D. and Hatzinakos, D., 'Digital watermarking using multiresolution wavelet decomposition', in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*: IEEE, pp. 2969-2972,1998.
[6] Barni, M., Bartolini, F. and Piva, A., 'Improved wavelet-based watermarking through pixel-wise masking', *Image Processing, IEEE Transactions on*, 10, pp.783-791,2001.
[7] Le, L., Krishnan, S. and Ghoraani, B., 'Discrete Polynomial Transform for Digital Imagewatermarking Application', in *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 1569-1572,2006.
[8] Krishnamoorthi, R.M., P. D. Sheba Kezia., 'Image Adaptive Watermarking with Visual Model in Orthogonal Polynomials based Transformation Domain', *International Journal of Signal Processing*;2009, Vol. 5 Issue 2, pp.146, April,2009.
[9] Al-Qaheri, H., Mustafi, A. and Banerjee, S., 'Digital watermarking using ant colony optimization in fractional Fourier domain', *Journal of Information Hiding and Multimedia Signal Processing*, 1, pp.179-189,2010.
[10] Podlubny, I., *Fractional Differential Equations*: Academic Press, San Diego - New York - London,1999.
[11] K.B. Oldham, J. Spanier, *The Fractional Calculus*, Academic Press, New York, 1974.
[12] K.S. Miller, B. Ross, *An Introduction to the Fractional Calculus and Fractional Differential Equations*, John Wiley Sons, 1993.
[13] B.J. west, M. Bologna, P. Grigolini, *Physics of Fractal Operators*, Springer Verlag, 2003.
[14] Tofighi, A. and Pour, H.N., 'e-expansion and the fractional oscillator', *Physica A: Statistical Mechanics and its Applications*, 374, pp.41-45,2007.
[15] Ortigueira, M.D., 'An introduction to the fractional continuous-time linear systems',2008.
[16] Abbasi A, Woo CS, Ibrahim RW, Islam S Invariant Domain Watermarking Using Heaviside Function of Order Alpha and Fractional Gaussian Field. *PLoS ONE* 10(4): e0123427. doi:10.1371/journal.pone.0123427,2015.
[17] Gearhart, W. B., & Schultz, H. S., 'The function sin(x)x', *The College Mathematics Journal*, 2,pp. 90-99,1990.
[18] USC-SIPI. (2014). USC-SIPI image database. On-line: <http://sipi.usc.edu/database/>.

A Semantic Approach for Mathematical Expression Retrieval

Zahra Asebriy

Dept of Applied Mathematics and Computer Science
Laboratory (LAMAI), Cadi Ayyad University, Marrakesh,
Morocco

Said Raghay

Dept of Applied Mathematics and Computer Science
Laboratory (LAMAI), Cadi Ayyad University, Marrakesh,
Morocco

Soulaimane Kaloun

SAEED, Higher School of Technology, Cadi Ayyad
University, Essaouira, Morocco

Omar Bencharef

SAEED, Higher School of Technology, Cadi Ayyad
University, Essaouira, Morocco

Abstract—Math search or mathematical expression retrieval has become a challenging task. Mathematical expressions are very complex, they are highly symbolic, and they have a semantic meaning that we should respect. In this paper, we propose a similarity search method for mathematical expression based on a multilevel representation of expressions and a multilevel search. We used the K-Nearest Neighbors with three types of distances to evaluate relevance between expressions. In the experimental level, the proposed system significantly outperforms statistical algorithms.

Keywords—Mathematical expression; Retrieval information; MathML; Semantic similarity

I. INTRODUCTION

The fast expansion of information technology and the spread of digital libraries in different domains make search engines necessary to help users to share and to retrieve any information from the web or from numerical libraries.

Now days, search engines can search all kinds of documents including text, image, audio and video. Therefore, documents containing special data such as mathematical expressions, tables, diagrams and drawings cannot be retrieved by classical search engines.

As distinct from text retrieval, retrieving math expressions have been researched for several years. Now it's still in the research stage. There are a few researches in the field of Mathematical Expressions Retrieval (MER) and a few number search engines dedicated to this subject, like MathFind [1], Active Math [2] Wolfram search, Wikipedia search formulas and MathWeb search. Most of these number search engines are based on text retrieval techniques.

Mathematical expressions are highly symbolic and they have their own structures. For example [3]:

- The order of elements in the mathematical expressions has semantic meaning, for example, $\sum \sin(\exp(x))$ and $\sum \exp(\sin(x))$ are two completely different expressions with the same elements but do not have the same orders. So it is important to respect the order of the

elements in the mathematical search to retrieve the right expression.

- if there are two math expressions $(a + b)$ and $\sqrt{(a + b)}$, the role of sub-expression $(a + b)$ in each expression is different, and if the query is to find the square root of $(a + b)$, the system must consider this particularity and all relevant expressions $\sqrt{(a + b)}$ should be strongly ranked.
- Mathematical equations can be written with different notation but they can have the same semantic meaning for example $(a + b)$ and $(x + y)$ are the same expressions.

Retrieving mathematical formulas with all these constraints requires a system based on semantic representations of the query math expressions. As examples of these representations, There are several common Mathematical Markup Languages: Latex [4] OpenMath [5], ASCII [6] and MathML[7].

MathML is an application of XML (Extensible Markup Language) for encoding notational and semantic structure of mathematical expressions. Actually, it is used by many systems for retrieving math expressions on the web [8, 9, 3].

In this paper, we propose an algorithm to extract features vectors of mathematical expressions represented on MathML and a multilevel search algorithm based on K-Nearest Neighbor's. We are going to use a variety of distances measure, first to evaluate the efficiency of our system and to find the best one for our system.

II. RELATED WORK

Retrieving mathematical equations has attracted much attention from researchers in the past decade, and several related systems or methods on this task have been reported. Currently, several researches have been realized to develop and improve retrieving mathematical equations from the web or in digital library.

The system proposed by Yokoi et al. [10] was a new similarity search scheme for mathematical expressions. They

started by introducing a similarity measure based on Subpath Set and proposed a MathML conversion that is apt for it. The aim of this method used is to return similar equations by measuring the similarity using tree matching techniques and by reforming the structure of content based MathML. Based on their First experiences, they believe that their proposed system has the potential to provide a flexible interface for searching mathematical expressions on the web.

Tam T. Nguyen et al. [3] presented a lattice-based approach for mathematical search using Formal Concept Analysis (FCB) which is a powerful data analysis used for information retrieval [11]. This approach involves several phases. In the first time, they extract features from code MathML representation. These features are used to construct a mathematical lattice construction. At the query retrieval phase, the query expression is processed and inserted into math concept lattice, which matches with math expressions concept in the concept lattice to rank the relevant math expressions. The results have shown that the proposed approach has performed better than the conventional best match retrieval technique. Another important advantage of the proposed lattice-based approach lies in its support for the visualization and navigation of search results via a dynamic graph.

In their work [12], S-Q Yang and X-D Tian tried to research and develop special retrieval method. They proposed a maintenance algorithm of mathematical expression index based on Formula Description Structure (FDS), which includes the index item searching, inserting and deleting operations. Moreover, S.Q Yang et al. designed a matching model of mathematical expressions based on Formula Description Structure (FDS) index [13]. For realizing exact matching, the math retrieval attributes were embedded an index in three query modes called global query mode, local query mode and operational query mode.

L. Gao et al. [14] proposed a semantic enrichment technique to retrieve mathematical formulae from web pages and PDF documents with a novel query input interface, which allows users to copy formula queries directly from PDF documents without using formulas with Markup languages. They used a novel indexing and matching to search similar mathematical expressions based on both textual and spatial similarities. The proposed system achieves better performance compared with two representative mathematics retrieval systems.

MathSearch [15] is a formula-based search engine for mathematical information on the internet. In this system, Mathematical formula Query Language (MQL) [16] was designed for expressing and processing query. MQL contains two forms: a character string form (MQLS) and XML form (MQLX). By MQLX and MQLS, semantics query wildcard and combination query can be accomplished in MathSearch.

WikiMirs [17] is a tool to facilitate mathematical formula retrieval in Wikipedia. This system involves several phases. In the first phase, this system normalized Latex formulas of Wikipedia into a unified mode. Then terms were extracted from the normalized presentation tree. These terms reflect the

features of the expression through series of processors such as presentation tree parser normalize and term extractor. These extracted terms used to establish an inverted index. In the last step, users query math expressions in latex form were processed with the above steps and retrieved.

III. FEATURE EXTRACTION

A. Processing of Mathematical expressions

The processing of mathematical expressions as uniform mathematical representation plays an important role in the area of math search systems and digital libraries.

In this research, we use MathML for encoding notational and semantic structure of mathematical expressions (show examples 1&2). Currently, it is used by many systems for retrieving math expressions on the web [8, 9, 3], and by other applications like code MathML translator to Nemeth Braille code [18].

Example 1: Code MathML of math expression $x + y = 2$

```
<math>
<mrow>
<mi>x</mi><mo>+</mo><mi>y</mi><mo>=</mo><mn>2<
/mn>
</mrow>
</math>
```

Example 2: Code MathML of math expression $\frac{2x^2 + \sqrt{1+x}}{x}$

```
Tapez une équation ici.
<math>
<mrow>
<mfrac>
<mrow>
<mn>2</mn><msup>
<mi>x</mi>
<mn>2</mn>
</msup>
<mo>+</mo><msqrt>
<mrow>
<mn>1</mn><mo>+</mo><mi>x</mi>
</mrow>
</msqrt>
</mrow>
<mi>x</mi>
</mfrac>
</mrow>
</math>
```

B. Extraction process

To find a structural and semantic similarity between mathematical expressions in a big data base or on the web scale level we need a reduced and efficient representation that respects the structural and semantic specification of mathematical formula. First we choose to act with a simple algorithm that counts the number of occurrence of each operator and Math function. In the second algorithm we propose to use a multilevel representation of expressions.

1) Statistical algorithm

It extracts through the MathML code the number of each operator (+, -, *, /), variables, constants, and functions (log, sin, cos, exp...) and stores them in a vector (Fig. 1).

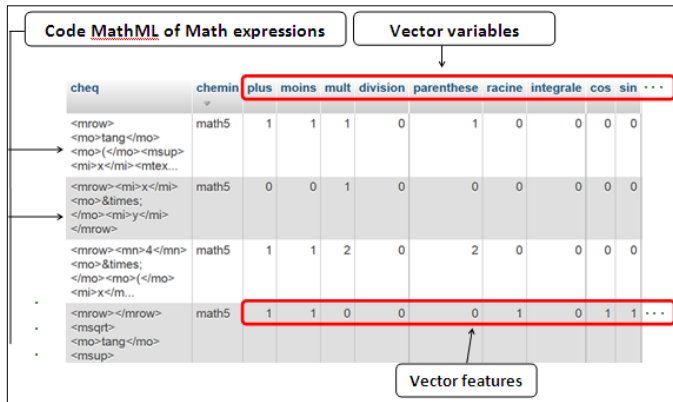


Fig. 1. Example of extracted vectors applied to dataset examples, using algorithm 1

The statistical approach is fast and reduced and in many cases it can detect real similarity between mathematical expressions. On the other hand using just the number of occurrence of each expression can produce false similarity detection like the case in examples 1 and 2 in Fig. 2.

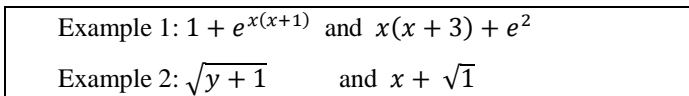


Fig. 2. Example of false similarity obtained using statistical algorithm

It's clear that there is no semantic similarity in the two examples of Fig. 2. As a result we need an algorithm also fast and more efficient regarding semantic similarity.

2) Proposed method

In order to define different levels for each math formula, we need to convert all mathematical equations into MathML code.

The first level was established by searching all main operators (+, -, *, /) linking all brackets and functions (trigonometric, logarithmic and algebraic) which all expressions into brackets and arguments of functions were defined and replaced by the term "exp". For example:

$$\sqrt{x+1} + x^2 \ln(x^2+1) \text{ Become } \sqrt{\text{exp}} + \text{exp} \ln(\text{exp})$$

The values of each "exp" are stored to be used in the second level.

The vector of level 1 is the outcome of the statistical algorithm applied to the reduced expression. What gives: one $\sqrt{\quad}$, one \ln , and 3 exp. So the representative vector of level one becomes:

$$(3,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0)$$

These features were extracted and stored in the vectors V_{l1} . In the 2nd level, we treated each "exp" by repeating the same procedure used in the first level and we stored the extracted features in the vectors V_{l2} . We continued applying this method

until obtained all levels and all vectors $V_{li}(V_{l1}, V_{l2}, \dots, V_{ln})$ for each equation. In practice we use only 3 levels.

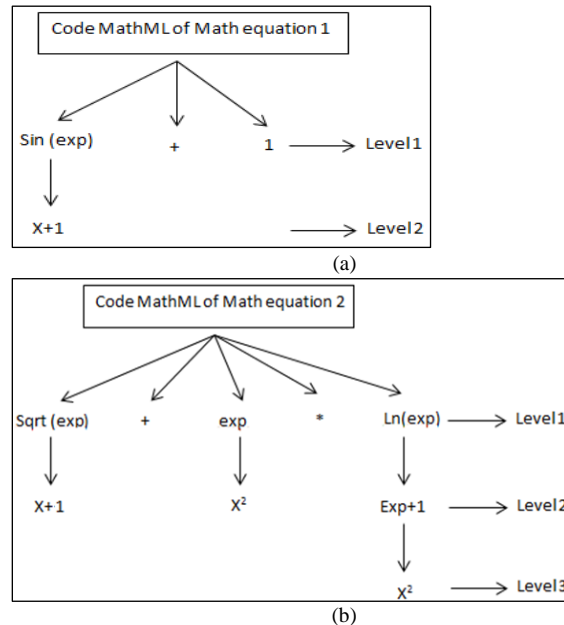


Fig. 3. (a) Code MathML of Math equation $\sin(x+1) + 1$ by levels, (b) Code MathML of Math expression $\sqrt{x+1} + x^2 \ln(x^2+1)$ by Levels

IV. RETRIEVING RESULTS

The proposed method was based on the structural and semantic multilevel similarity between mathematical equations. The similarity degree was obtained based on all representation levels (Fig 3a, 3b).

After defining the expression levels and vectors V_{li} , in the first time we retrieve all expressions that have similar vectors V_{l1} to our math query. Then we move to the second level for only these equations already retrieved in the first level. The same procedure, used in the first level, was repeated to recover expressions that have similar vectors V_{l2} . We continue with the same procedure in level 3.

The K-Nearest-Neighbor (KNN) algorithm was used in this phase to retrieve math equation. It is a non-parametric lazy learning algorithm [19] and is an instance-based learning algorithm that uses a distance function of pairs of observation. KNN was based on the measurement of the distance to search a similarity between the query math equation and those of database. This distance is calculated using one of the following measures:

- Euclidean distance:

$$d(V_{lx}, V_{ly}) = \sqrt{\sum_{i=1}^n (V_{lx_i} - V_{ly_i})^2}$$

Where: V_{lx} and V_{ly} are two features vectors of two mathematical equations;

n is the number of attributes (vector size)

- Minkowski distance:

$$d(Vlx, Vly) = \left(\sum_{i=1}^n |Vlx_i - Vly_i|^p \right)^{\frac{1}{p}}$$

Euclidean distance is a special case for p=2 of Minkowsky distance

- P=1 we obtain the Manhattan distance:

$$d(Vlx, Vly) = \sum_{i=1}^n |Vlx_i - Vly_i|$$

- P=∞ we obtain the Chebychev distance:

$$d(Vlx, Vly) = \max_i |Vlx_i - Vly_i|$$

- Correlation distance:

$$d(Vlx, Vly) = 1 - \frac{\sum_{i=1}^n (Vlx_i - \overline{Vlx})(Vly_i - \overline{Vly})}{\sqrt{\sum_{i=1}^n (Vlx_i - \overline{Vlx})^2 (Vly_i - \overline{Vly})^2}}$$

V. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed system we create a dataset of mathematical expression. The dataset is constructed using MathType. MathType is an interactive tool for authoring mathematical material, In the Microsoft Word or Power Point. There is MathType Ribbon Tab to facilitate editing, inserting and math equations creation. The dataset elements can be easily converted to MathML or Latex. In this set we have created 6925 mathematical expressions using symbols from five languages Latin, Arabic, Tifinagh [20,21], Hebrew and Japanese(Fig. 4 a, b, c, d, e). For each language, we have written 1385 different types of math expressions such as polynomial, algebraic, statistic, trigonometric and logarithmic.

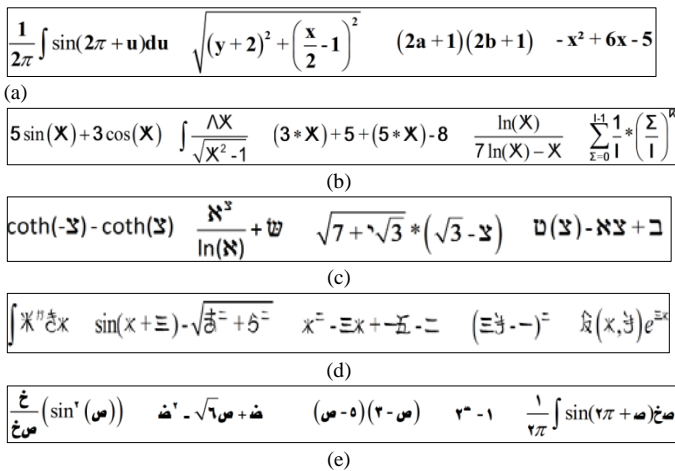


Fig. 4. (a) Math expressions in Latin; (b) Math expressions in Tifinagh; (c) Math expressions in Hebrew; (d) Math expressions in Japanese; (e) Math expression in Arabic

In this subsection, we present the results of the proposed system using Euclidean distance, Minkowski distance and Correlation distance. We compared our results to the statistical

approach.

We found difficult to evaluate results using recall and precision evaluation using a similarity measure as proposed by T.T Nguyang et al. [3], in their paper the similarity measure between two expressions E1 and E2 represented by their attribute sets M(E1) and M(E2) is :

$$sim(E1, E2) = \frac{|M(E1) \cap M(E2)|}{|M(E1) \cup M(E2)|}$$

When our goal is to evaluate the semantic similarity, not only the number of similar components, we choose to evaluate manually different type of test queries by different level of similarity:

- Identical similarity, like: $\sqrt{x+y}$ and $\sqrt{a-b}$
- Sub expression similarity: $\frac{2}{\sqrt{x+y}}$ and $\sqrt{x+y}$
- Categorical similarity: $\int x^2 dx$ and $\int (x + x^3)^2 dx$ are both Integrals with polynomial functions.

We decide to give a score of tree points for identical similarity, two points for sub expression similarity and categorical similarity and zero for non-relevant expressions.

Table I, shows the performance results of the proposed system using Euclidean distance, Minkowski distance, and Correlation distance compared to the statistical approach. The score is based on the top 10 results of 10 test queries. A perfect score is 120 points.

TABLE I. PERFORMANCE RESULTS

Approach	PS (Euclidian)	PS (Minkowski)	PS (Correlation)	Statistical
Score/120	100	113	102	81
Score (%)	0.83	0.94	0.85	0.68

The experiments show that results obtained using the proposed system outperforms the statistical approach. Using Minkowski distance our system become more efficient (a score of 94%) Table II and III show test queries with their relevant expressions in the dataset using simultaneously the proposed system with Minkowski distance and the statistical approach.

TABLE II. RELEVANT EXPRESSIONS OF EACH ONE OF THE TEST QUERIES USING THE PROPOSED SYSTEM WITH MINKOWSKI DISTANCE

Query	Relevant expressions
$\int x^2 + 1 dx$	$\int x^2 + 1 dx$, $\int \rho^2 + 2 d\rho$, $\int (x+3)^2 dx$, $\int x^3 + 2x^2 + 2 dx$, $\int x^3 + 2x^2 + x + 1 dx$
$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$	$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$, $\frac{-b - \sqrt{b^2 - 4ac}}{2a}$, $\frac{x + \sqrt{x^2 - 2}}{x - 2}$, $\frac{\sqrt{a^2 - b^2}}{a}$, $\frac{1 - x^2}{\sqrt{x^2 - 2x}}$
$\sqrt{x+1}$	$\sqrt{x+1}$, $\sqrt{a+2}$, $2\sqrt{a+b}$, $\sqrt{x+x^2+1}$, $\sqrt{x+1} + 1$
$\sin x \cos^2 x$	$\sin x \cos^2 x$, $\frac{\sin x}{\cos^2 x} \sin x + \cos x$, $\int \sin x \cos^2 x dx$, $\cos^2 x + 1$,
$x^3 + 2x^2 + 3$	$x^3 + 2x^2 + 3$, $2x^3 + x^2 + 1$, $x^3 + x^2 + x + 1$, $(x+1)^3 + x$, $y^4 + y^3 - y^2 + 1$

TABLE III. RELEVANT EXPRESSIONS OF EACH ONE OF THE TEST QUERIES USING STATISTICAL APPROACH

Query	Relevant expressions
$\int x^2 + 1 dx$	$\int x^2 + 1 dx \int \rho^2 + 2 d\rho, \int \frac{1}{y^2} dy, \int (x + 3)^2 dx,$ $\int \cos(x^2 + 1),$
$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$	$\frac{-b + \sqrt{b^2 - 4ac}}{2a}, \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \frac{x + \sqrt{x^2 - 2}}{x - 2}, \frac{\sqrt{a^2 - a}}{a}, x^2 - 2xy +$ $\sqrt{\frac{x}{2y}}$
$\sqrt{x + 1}$	$\sqrt{x + 1}, \sqrt{a + 2}, x + 1 + \sqrt{2}, \sqrt{x + 2}, \sqrt{x + 1} + 1$
$\sin x \cos^2 x$	$\sin x \cos^2 x, \frac{\sin x}{\cos^2 x},$ $\sin^2 x + \cos x, \int \sin x \cos^2 x dx, x^2 \sin x$
$x^3 + 2x^2 + 3$	$x^3 + 2x^2 + 3, 2a^3 + a^2 + 1, x^3 + x^2 + x + 1,$ $\frac{x^3}{x^2 + 1}, y^4 + y^3 - y^2 + 1$

Our system with Minkowski distance allows a better detection of categorical similarity. We notice that the most of relevant expressions returned by the proposed system exactly match the query. For the statistical approach the absence of a semantic input can generate wrong outputs like relevance between $\sin x \cos^2 x$ with $\frac{\sin x}{\cos^2 x}$ and $\sqrt{x + 1}$ with $x + 1 + \sqrt{2}$.

VI. CONCLUSION

In this paper, we proposed a semantic approach to retrieve mathematical expressions. Based on MathML, we extract a multilevel representation for each mathematical expression. We used KNN with different types of distances to measure similarity between each representation level. In the light of our experiments we can conclude that the results are encouraging. Our system outperforms significantly the statistical approach and the implementation of Minkowski distance allows a better detection of categorical similarity. In our future work we have two goals, first to take into consideration more types of math expressions and second to evaluate our system on the web scale level.

REFERENCES

[1] R. Munavalli, R. Miner. "Mathfind: A math-aware search engine". 29th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA ACM. p735, (2006).

[2] P. Libbrecht, E. Melis. "Methods to access and retrieve mathematical content and activemath" The 2nd international congress on mathematical software. ICMS'06 Castro Urdiales, SPAIN, 1-3 September (2006).

[3] T. T. Nguyen, S. H. Cheung, K. Chang. "A lattice-based approach for mathematical search using Formal Concept Analysis". In Expert Systems with Applications 39. 5820-5828. (2012)

[4] D. Waud. "What is next?" Available at: <http://www.tex.uk/cgi-bin/texfaq2html>. (2003)

[5] S. Buswell, O. Caprotti, D. P. Carlisle, M. C. Dewar, M. Gaetano, and M. Kohlhase, "The open math standard" version 2.0.(2004)

[6] P. Jipsen, "Translating ASCII math notation to mathml and graphics". Available at: <http://www.chapman.edu/jipsen/mathml/asciimath.html>. (2007)

[7] R. Ausbrooks, S. Buswell, S. Dalmas, S. Devitt, A. Diaz, R. Hunter, B. Smith, N. Soiffer, R. Sutor, S. Watt. "Mathematical markup language" (mathml)version 2.0. (2000).

[8] M. Nghiem, G. Yoko, Y. Matsubayashi, A. Aizawa. "Automatic Approach to understanding mathematical expressions using Mathml Parallel Markup corpora". The 26th Annual Conference of the Japanese Society for Artificial Intelligence, (JSAI2012) Yamaguchi, June (2012)

[9] B. R. Miller, A. Youssef. "Augmenting presentation mathml for search". The 7th international conference on mathematical knowledge management MKM '08, pp. 536-542, (2008).

[10] K. Yokoi, A Aizawa. "Towards Digital Mathematics Library" pp. 27-35, (2009).

[11] K. S.Cheung, D. Vogel,, "Complexity reduction in lattice-based information retrieval". Information retrieval, 8, pp 285-299, (2005).

[12] S.Q. Yang, X. D. Tian: "A maintenance algorithm of FDS based mathematical expression index". International Conference on machine learning and Cybernetics, Lanzhou, 13-16 July, (2014).

[13] S. Q. Yang, X. D. Tian, B. T. Yu, F. Yang. "A matching model of mathematical expressions with FDS based index". International Journal on machine learning and Cybernetics. 6, pp 993-1004, (2015).

[14] X. lin, L. Gao, X. Hu, Z. Tang, Y. Xiao, X. Liu. "A mathematical retrieval system for formulae in layout presentations". The 37th international ACM SIGIR conference on research & development in information retrieval (SIGIR 14) ACM, pp 697-706, (2014).

[15] MathSearch, <http://wme.lzu.edu.cn/mathsearch/index.html>

[16] W. Guo, W. Su, L. Li, N. L. Cui. "MQL: a Mathematical Formula Query Language for Mathematics Search". The 4th IEEE International conference on Computational Science and Engineering (CSE2011). Dalian, pp 245-250, (2011).

[17] X. Hu, L. Gao, X. Lin, J. B. Baker. "WikiMirs: a mathematical information retrieval system for Wikipedia". The 13th ACM/IEEE-Cs joint conference on digital libraries ACM, pp 11-20, Indianapolis, IN, USA, 22 - 26 July, (2013).

[18] P. B. Stanley, A. I Karshmer. " Translating MathML into Nemeth Braille Code". The 10th International Conference, ICCP 2006, LNCS 4061, LNCS 4061, pp. 1175-1182, Linz, Austria, (2006)

[19] B. V. Dasarathy, "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques", Mc Graw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, pp. 217-224, (1991).

[20] M. Oujaoura, R. El Ayachi, B. Minaoui, M. Fakir, B. Boukhalene, O. Bencharef. "Invariant descriptors and classifiers combination for recognition of isolated printed Tifinagh characters." International Journal of Advanced Computer Science and Applications 3.2: pp22-28, (2013).

[21] O. Bencharef, Y. Chihab, N. Moussaid, M. Oujaoura. "Data set for Tifinagh handwriting character recognition." Data in brief 4: pp11-13, (2015).

Variability of Acoustic Features of Hypernasality and its Assessment

Shahina Haque*

Department of CSE
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh
Permanent address:

Department of ETE, Daffodil International University, Dhaka-1207, Bangladesh

Md. Hanif Ali

Department of CSE
Jahangirnagar University
Savar, Dhaka-1342
Bangladesh

A.K.M. Fazlul Haque

Department of ETE
Daffodil International University
Dhaka-1207
Bangladesh

Abstract—Hypernasality (HP) is observed across voiced phonemes uttered by Cleft-Palate (CP) speakers with defective velopharyngeal (VP) opening. HP assessment using signal processing technique is challenging due to the variability of acoustic features across various conditions such as speakers, speaking style, speaking rate, severity of HP etc. Most of the study for hypernasality (HP) assessment is based on isolated sustained vowels under laboratory conditions. We measure the variability of acoustic features and detect HP using vowel /i/, /a/ and /u/ in continuous read speech with gradually increasing severity of HP of CP speakers. Linear predictive coding (LPC) method is used for acoustic feature extraction. In first part of our study, we observe the variation in acoustic parameters within and across vowel category with gradually increasing HP. We observe that inter-speaker variability in spectral features among CP subjects for vowel /i/ is 0.96, /a/ has 1.13 and vowel /u/ has 2.05. The inter-speaker variability measurement suggests that high back vowel /u/ is mostly affected and has the highest variability. High front vowel /i/ is least affected and has the lowest variability with HP. In the second part, ratio of vowel space area (VSA) of hypernasal and normal speech is calculated and used as a measure for HP detection. We observe that VSA spanned by CP subjects is 0.65 times less than isolated uttered Bangla nasal VSA and 0.43 times less than read speech uttered English oral VSA.

Keywords—Speech analysis; Acoustic feature; Hypernasality; Cleft palate; Velopharyngeal opening; Vowel space area; Read speech

I. INTRODUCTION

Speech is the acoustic end product of the thoughts which is originated in the brain. Disordered speech in which speech quality is reduced may hamper normal communication. Due to physical or neurological impairment the speech quality may be reduced which is a challenge in professional and social

activities [1]. A specific example of a vocal tract dysfunction that reduces the speech quality is defective VP mechanism [2] which can be caused by physical defects (CP) [3]. CP is an incomplete soft or hard palate formation that separates nasal and oral cavities, generating speech disorder such as HP and is the second most frequent congenital malformation worldwide [4]. HP is the most common pathology suffered by patients with CP. The research community is becoming more and more interested in the development of techniques for its detection and evaluation [5-8].

Voice pathologies are usually diagnosed by invasive techniques using different instruments which may bring discomfort to the patients. This is also not recommended by health physicians as they can produce psychological stress in patients. One of the non-invasive techniques for voice pathology is based on acoustic analysis of voice. Acoustic features contain information regarding voice source and vocal tract behavior. Any abnormality in speech arising from physical defect may be assessed appropriately using speech features of vocal tract or voice source. Fig.1 shows how the vocal tract transfer function varies for vowel /i-/ and /i/ (normal oral, various degree of HP). Consequently, it is convenient to apply signal processing techniques in determining the effectiveness of speech features. The identification of severity of HP using speech processing techniques can be a useful contribution to the diagnosis of CP speakers which aids to decide the severity of HP and what support (surgery or speech therapy) is to be provided to the CP speakers by the physicians. Most of the study on HP assessment using acoustic features concentrates on sustained isolated vowel. This study explores the variation in acoustic parameters and HP assessment in read speech with gradually increasing VP opening of CP speakers.

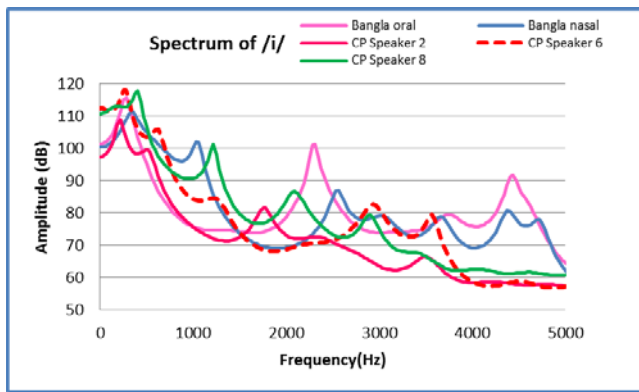


Fig. 1. Vocal tract transfer function for /i-/ and /i/for normal and CP speakers

The rest of the paper is organized as follows:

Section II presents a description of previous work. Section III describes about speech materials, Section IV discusses about used method and results obtained from experiments and the observations. Section V discusses about the analysis of the result. Section VI concludes the paper.

II. ACOUSTIC ANALYSIS OF SPEECH

Since 1970, researchers have studied abnormal changes in the acoustic features of voices. In 1971, Fujimura, *et al.* [9] made a detailed analysis of the variation in the voice tone. Fant [10] found that for a nasalized vowel, oral-nasal coupling introduces an additional pole-zero pair into the oral vowel. HP detection is performed by many researchers by analyzing the disordered speech, synthesized hypernasal speech, and nasalized vowels of normal speech. Signal processing based techniques for the assessment of HP is carried out by finding the deviation of the spectrum of hypernasal vowels from the non-nasalized vowels.

Main features of nasalization are changes in the low-frequency regions of the speech spectrum, where there is a very low-frequency peak with wide bandwidth along with the presence of a pole-zero pair due to the acoustic coupling was shown by Hawkins and Stevens [5] and Glass and Zue [7]. Chen [6] and Hawkins [5] showed that nasalization gives rise to changes in the spectrum in the high frequency region in addition to introduction of a new pole-zero pair in the first formant region. However, these changes are not as consistent across speakers and vowels as those in the low frequency region. Sensitivity of the Teager energy operator for multicomponent signals, to detect HP was used by Cairns *et al.* [11]. Presence of zeros in spectrum is used as a cue for the detection of HP by Rah *et al.* [12]. From the literature, it is observed that the acoustic cues of hypernasal speech are additional formants, antiformants, formant bandwidth broadening. Study on perceptual analysis was carried out on the vowel sounds by adding nasal formants in spectrum [13-14], showed that formant at 250 Hz plays important role in the nasalization of vowels. Group delay function was used successfully for hypernasality detection using acoustic features of speech in CP speakers [14-15].

VSA refers to the two-dimensional area bounded by lines connecting first and second formant frequency coordinates ($F2/F1$) of vowels [16]. VSA has been used for various purpose such as studying vowel identity, speaker characteristics, speech development, speech disorder, vowel distinctiveness and assess intelligibility that influences vowel production [17-19]. VSA computation is done by measuring $F1/F2$ values for several utterances for each of the three point vowels, /a, i, u/ for plotting vowel triangle. The mean $F1/F2$ value for each of the corner vowels is then used to compute the area of the triangle formed by the corner vowels. As frequencies of the first and second formants is related to the size and shape of the cavities created by mouth opening ($F1$) and tongue position ($F2$), the VSA reflects the dynamics of the articulators. In general, studies have shown that VSA is larger in speech that is clearer and more intelligible than speech associated with smaller VSAs. This is because if the articulatory excursions are greater it results in more distinct acoustic vowel targets. Thus, the VSA related to vowel distinctiveness have been quite successful in the study of speaking style and languages. As abnormal vowel formant change (centralization) is a common feature of speech production deficiency, VSA estimations. is used for characterizing speech motor control, including speech development, speech disorders. In a study with large database, automatic assessment of VSA was done and is reported to have good result than the traditional method of VSA measurement [20]. In another study it was found that psychological distress and depression reduces the VSA [21]. Hypernasal vowel speech near a plosive of CP children were analyzed and proposed an objective measure. It was found that mean falling and rising slopes of the amplitude in the nasalized vowel are smaller than those of the oral vowel [22].

Most of the study on acoustic analysis of HP is based on isolated sustained vowel for HP. There has not been much contribution on variation in speech parameters with defective VP opening causing hypernasal speech in continuous read speech sentences which exhibit greater complexities with respect to speech intelligibility, which formed the motivation for this study. This study aims to investigate how useful the extracted speech parameter information from read speech rather than isolated sustained vowel to reflect the impact of the underlying movement disorder in terms of VSA for the assessment of HP. This report presents the study on the VP opening variability on speech features and the relationship between HP assessments using VSA in read speech.

Nasality has similarity with hypernasality in production which is reflected in acoustic features. Most of the study is concerned with nasalization of vowels near a nasal consonant to make a comparison with HP. In this study Bangla nasal vowels are used to make a comparison of HP with nasality. In Bangla, all the seven vowels have their nasal counterpart. Bangla is a language in which nasality is phonemic. Thus to make a comparison between nasality and HP, VSA of Bangla oral-nasal vowels are taken into account. Previous work was carried out to explore vowel space of Bangla oral-nasal vowel pairs [23]. Acoustic categorization of Bangla oral-nasal vowel pair was done and was shown that VSA of nasal vowel

shrunk within the oral VSA.

III. SPEECH MATERIALS

HP is usually observed in vowels and voiced oral consonants. As vowels can be sustained for a relatively longer duration as opposed to consonants, only vowels are considered for the current study. Among all the vowels only the vowels /i/, /a/ and /u/, are considered to represent the three categories of vowels namely, front, mid, and back. The aim of this section is to describe how the speech samples are acquired.

For the purpose of this study, a database is recorded by three male speakers, aged around 25-27 and native non-CP Bangladeshi Bangla speaking. For the acoustic analysis and detection of HP speech data are collected from 7 male speakers with CP and 4 normal non-CP. Three types of data used are:

1) English vowels (/a/, /i/ and /u/) obtained from read speech of eight speakers from normal non-CP (EO) to gradually increasing severity of HP CP speakers

2) Isolated Bangla oral (BO) (/i/, /a/, and /u/) and Bangla nasal (BN) vowels (/i-/, /a-/, /u-/) obtained from three non-CP speakers. The best one is selected for the work.

The experimental part consists of recording each of the isolated vowels at a normal speaking rate three times in a quiet room in a DAT tape at a sampling rate of 48 kHz and 16 bit value. The best one of these three speakers sample data is used for the study. Speech data for three English vowels /i/, /a/ and /u/ of normal and CP speakers with gradually increasing severity of HP are obtained from read speech data of American Cleft Palate Craniofacial Association. A stable portion is cut from each of the selected vowel for the purpose of our work. These digitized speech sound are then downsampled to 22050Hz and normalized for the purpose of analysis. Vowels uttered by non-CP speakers are used as reference.

IV. ACOUSTIC ANALYSIS OF NORMAL AND HYPERNASAL SPEECH AND RESULTS

In this section, preprocessing and the method which is chosen to extract the acoustic features from speech signal is discussed.

A. Preprocessing of the Speech Signal

Speech signal is non-stationary in nature, but it can be assumed to be stationary over short duration called frames by windowing for the purpose of analysis. Speech signal is analyzed frame-wise, with a frame-rate of 50-100 frames/sec, and for each frame the duration of speech segment is taken to be 20-30 msec. A new frame is obtained by shifting the Hamming windowing function by 10msec to a subsequent time. After normalization and windowing, the speech samples are ready to be used for analysis.

B. LPC Analysis Technique

LPC analysis decomposes digitized speech signal into its fundamental frequency (F0 and its amplitude i.e. loudness of the source) and the vocal tract is represented by all pole filters, which can be modeled by a number of coefficients known as LPC order. The vocal tract system is excited by an impulse

train for voiced speech or a random noise sequence for unvoiced speech. Thus, the parameters of this model are: voiced/unvoiced classification, pitch period for voiced speech, gain parameter G , and the coefficients $\{a_k\}$ of the digital filter. Eq. 1 expresses the transfer function of the filter model in z-domain, where $V(z)$ is the vocal tract transfer function. G is the gain of the filter and $\{a_k\}$ is a set of autoregression coefficients called Linear Prediction Coefficients. The upper limit of summation, p , is the order of the all-pole filter.

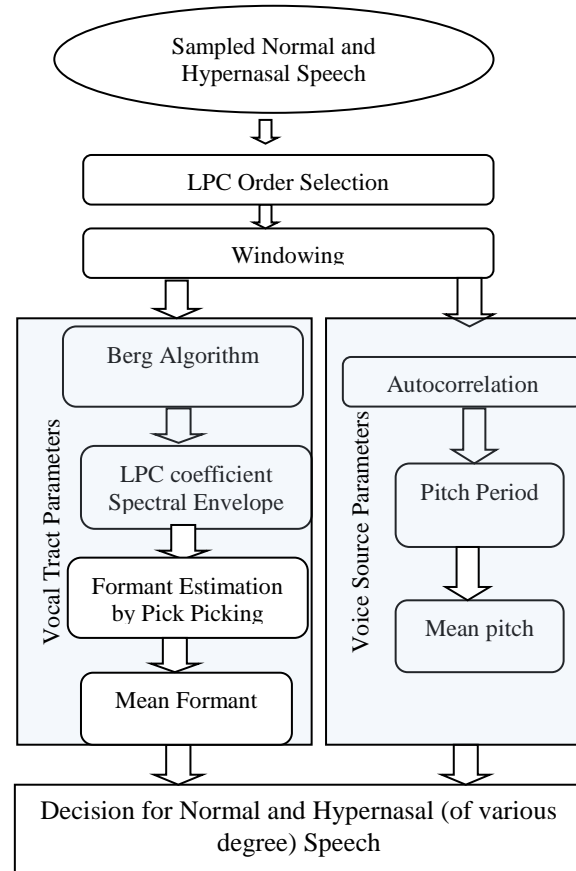


Fig. 2. Process of LPC analysis for speech parameter extraction

$$V(z) = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (1)$$

C. Acoustic Analysis

This paper examines the variability of CP speakers characteristics using LPC based acoustic features for HP assessment measured by VSA in read speech. The formant analysis is carried out for particular selected speech data. The utterances made by 2 normal subjects as explained in section III are analyzed and reference level is considered for each selected vowel phoneme. Fig. 3 shows the GUI used for acoustic feature extraction. The LPC spectrum is studied for formant analysis. The speech data is analyzed to check the disorders due to various VP opening. The acoustic analysis

(e.g., vocal tract parameters, voice source parameters, vowel triangle area) is conducted on speech data.

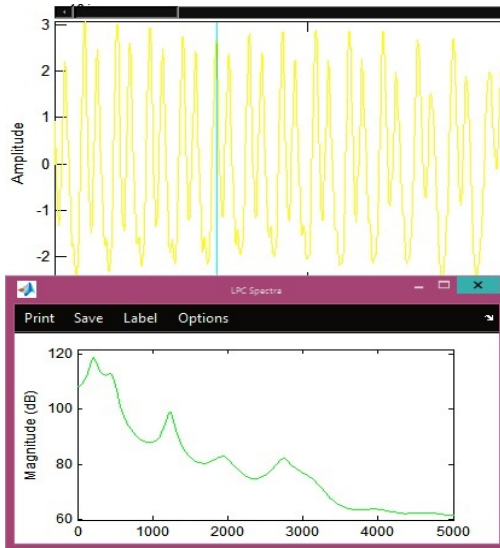


Fig. 3. GUI of obtaining LPC spectrum from speech waveform for /i/ of CP Speaker 8

If the prediction order is not chosen properly LP-based formant extraction technique may produce ambiguous result for the detection of HP. If the order of analysis of LP spectrum is too low then it fails to resolve two closely spaced formants and if the higher order LP analysis is chosen, it may introduce many spurious peaks in the resultant spectrum. Speech samples are analyzed by LPC method using LPC order 28. Fig. 2 shows the block diagram of procedure of LPC analysis for procuring speech parameters. The selected speech samples are windowed using hamming window of 20ms at 10ms interval. Acoustic parameters (vocal tract parameters, voice source parameters) for the three types of selected speech data (non-CP Bangla oral-nasal vowels, non-CP English oral vowels, CP English hypernasal vowels) are calculated. Acoustic features are extracted from a stable portion of segmented vowels of read speech and a part of data is tabulated in Table 1.

V. VARIABILITY OF ACOUSTIC FEATURES WITH INCREASING VP OPENING AND ASSESSMENT OF HP

In order to study the variation of acoustic features with HP and assess HP within CP speakers, scatter plots and VSA (Isolated Bangla oral, read English, isolated Bangla nasal, read CP English) are plotted among the three types of speakers and language. Fig. 4 shows the block diagram of the working procedure. As discussed in section II, VSA is the acoustic space area which contains information regarding diagnosis and treatment of defects of speech organs and its function. Simplest relationships between vocal tract configurations and formant frequencies take place for vowels which is utilized for this study.

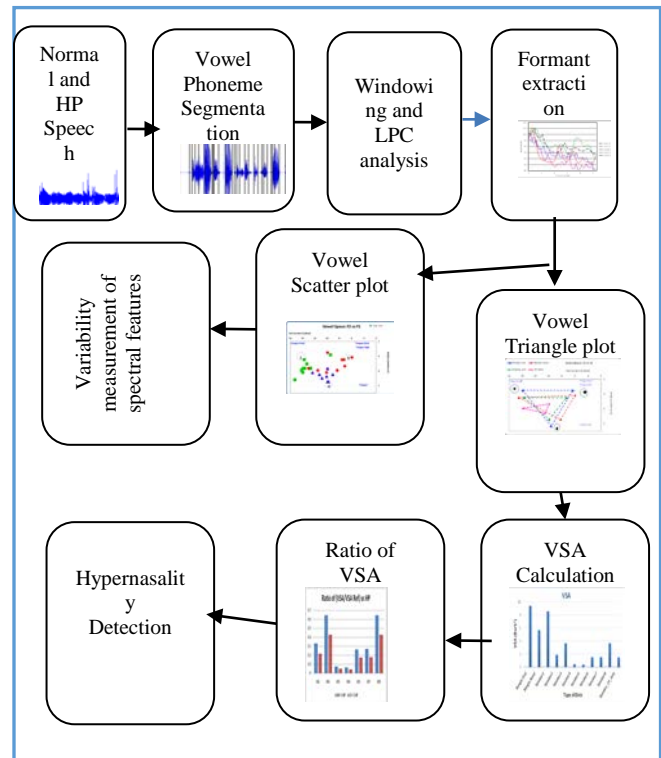
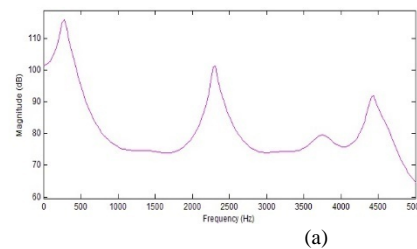


Fig. 4. Block diagram of working procedure for spectral variability and ratio of VSA calculation

A. Variability of Acoustic Features

Vocal tract transfer function for various degree of HP is plotted in Fig. 5. As the severity of hypernasality increases, it is reflected in the spectrum and changes are visible as compared to the oral vowel and nasal vowel spectrum. New spectral peaks are visible at near about F1 and F2 noticeably around 200Hz, 500Hz, 1kHz, 1.5kHz depending on the vowel.

Fig. 6 plots the variation of F1, FN1 (first nasal formant in hypernasal /i/) and FN2 (second nasal formant in hypernasal /i/) against HP. It is observed that F1 has increasing tendency in vowel /i/. This indicates highness and frontness of this vowel reduces, as VP opening increases and more air flows into the nasal tract resulting in reduction. $FN1 < FN$ (Nasal formant of BN vowel /i/) is located around 700 Hz) and FN1 has a decreasing tendency as HP increases. As $FN1 < FN$, FN1 should increase with increasing HP. $FN2 > 1100\text{Hz} > FN$. As $FN2 > FN$, FN2 should increase with increasing HP. FN lies between FN1 and FN2.



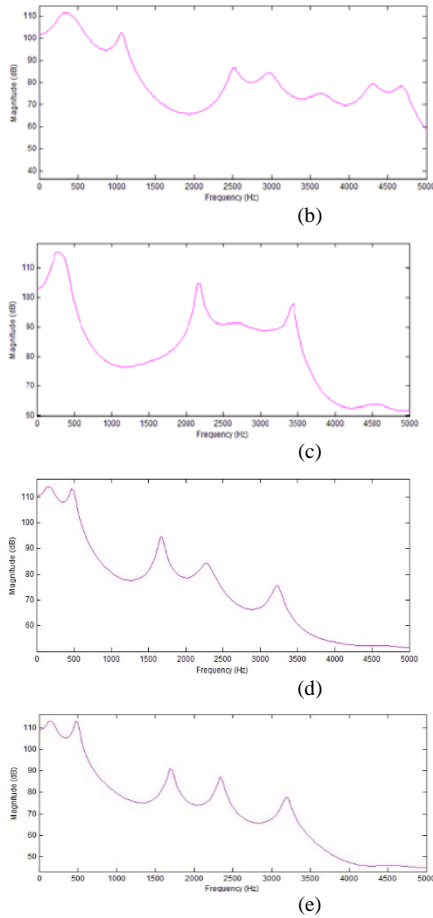


Fig. 5. LPC spectrum for /i/ (a) Isolated Bangla oral (b) Isolated Bangla nasal (c) Isolated English oral (d) CP Speaker 5 (e) CP Speaker 8

F1 and F2 values in Hz for each vowel for all speakers are converted to Bark scale which provides more appropriate frame of reference. Each vowel is represented by their F1 and F2 values displayed on scatter plots as shown in Fig. 7 for hypernasal speech. The scatter plot of F1 and F2 for vowels reflects the inter-speaker variation within vowel and across vowels making them useful for differentiation and identification of vowels. Fig. 7 plots the individual and mean formant values (F1 and F2, in Bark) in vowel space for the vowels /i a u/ measured for all speakers in the selected speech data described in section II. Each point represents the mean of three formant measurements per speaker.

TABLE I. VOCAL TRACT PARAMETERS (MEAN VALUES OF THREE OBSERVATION) FOR VOWEL /i/ FOR NON-CP SPEAKERS

	Pitch (F0) (Hz)	Spectral Peaks /i/ (Hz)	Spectral Peaks /a/ (Hz)	Spectral Peaks /u/ (Hz)
Bangla Oral (BO)	120	250	188	250
		2281	813	688
		3063	1188	2813
		3723	3000	
		4449	4073	
English Oral (EO)	121	290	625	375
		2219	1219	844
		3412	1906	1531
			3212	3091
Bangla Nasal (BN)	142	281	250	313
		688	688	625
		2621	938	2250
		3051	1331	2903
		3723	1922	4073
		4435	3000	4731
		4745	4086	

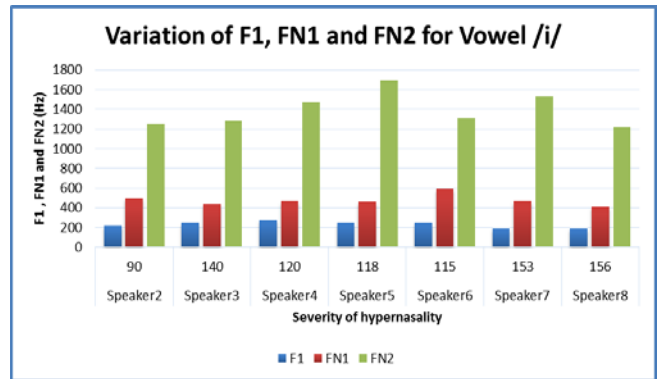


Fig. 6. Variation of F1, FN1 and FN2 with VP opening

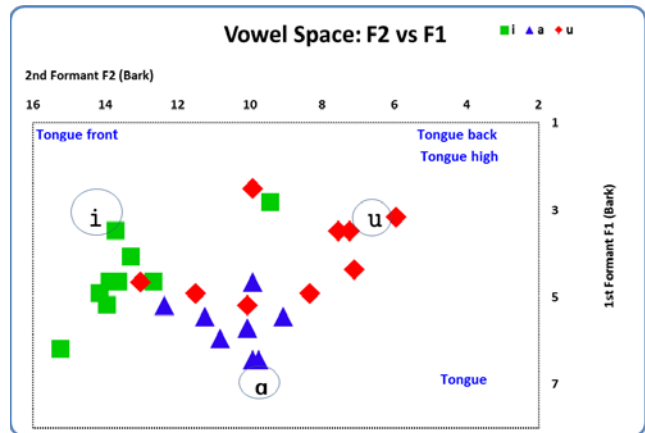


Fig. 7. Scatter plot of F1 and F2 of /i/, /a/ and /u/

In the considered utterance context, various degrees of inter-speaker variability is measured in terms of standard deviation about the mean. Variability of acoustic features among CP speakers are calculated to be different depending on the vowel. The inter-speaker variability among CP subjects for /i/ is 0.96 with mean (4.9,13.82). For pronouncing normal /i/ articulator is characterized by semi-openness, and has the highest front position among the vowels. /i/ and /u/ has the lowest F1 among the vowels as observed from Fig. 6. Among vowels, /u/ has the highest back position. During the production of /u/, articulators are characterized by lip-rounding, closeness, backness. Vowel /u/ shows the highest variability among speakers in the concerned speech data reflecting to differences in articulatory openness for some speakers. The standard deviation of /u/ is given by 2.05 with mean (4.58,9.64). The amount of inter-speaker variability of CP speakers in the high front vowels /i/ is less than open vowel /a/ is 1.13 with mean (5.65,10.46) and high back vowel /u/ which is 2.05.

B. Variability of VSA and HP Assessment

Fig. 8 (a) shows average VSA of all speakers spanned by mean values for the three repetitions for each of the three vowels. The differences between isolated oral VSA, isolated nasal VSA, read speech average VSA of CP speakers is investigated. Four types of VSA are obtained. VSA of isolated oral vowel marked by blue triangle has the highest area. Isolated nasal vowel VSA marked by red triangle has the second highest area. Read speech vowel of non-CP speakers marked by green triangle has the third highest area. The lowest VSA marked by magenta triangle is obtained for average of CP speakers read speech. The results show that

$$VSA_{\text{isolated oral}} > VSA_{\text{isolated nasal}} > VSA_{\text{read oral}} > VSA_{\text{HP}}$$

Fig. 8(b) plots VSA against degree of HP graph for all speakers as well as average VSA of CP speakers concerned in this study. This graph shows that as the degree of HP increases VSA changes. Isolated uttered oral vowel has the highest VSA, read speech has the second highest, isolated uttered nasal VS is fronted, shrinks and has VSA smaller than isolated oral vowel which is according to the previous study [20]. As the VP opening of CP speakers increases gradually VSA changes, but no gradual change is observed. This may be partially due to particular vocal tract characteristics of individual. Ratio of vowel space ($VSA_{\text{ind}}/VSA_{\text{ref}}$) of individual's VSA (VSA_{ind}) and the reference VSA (VSA_{ref}) is calculated to characterize how large the individual's vowel space of CP speakers is to the reference (BN and EO) VSA. Fig. 9 shows the vowel space ratio obtained across various VP opening conditions by taking BN and EO as the reference. Highest vowel space ratio came out to be 0.65 and 0.43 while taking BN vowel and EO vowel as reference respectively. These measures may be used as threshold for determining HP. Therefore, VSA of CP subjects is at least 0.65 times less than isolated uttered BN VSA and 0.43 times less than read speech uttered EO VSA. The significant reduction in VSA appropriately reflects the effect of HP of CP speakers across various conditions of severity showing the centralization tendency of articulators while pronouncing the vowels leading to reduction of speech clarity. Therefore as observed from this study, VSA of connected read speech of CP speakers is

suitable for detecting HP of CP speakers across various conditions of severity.

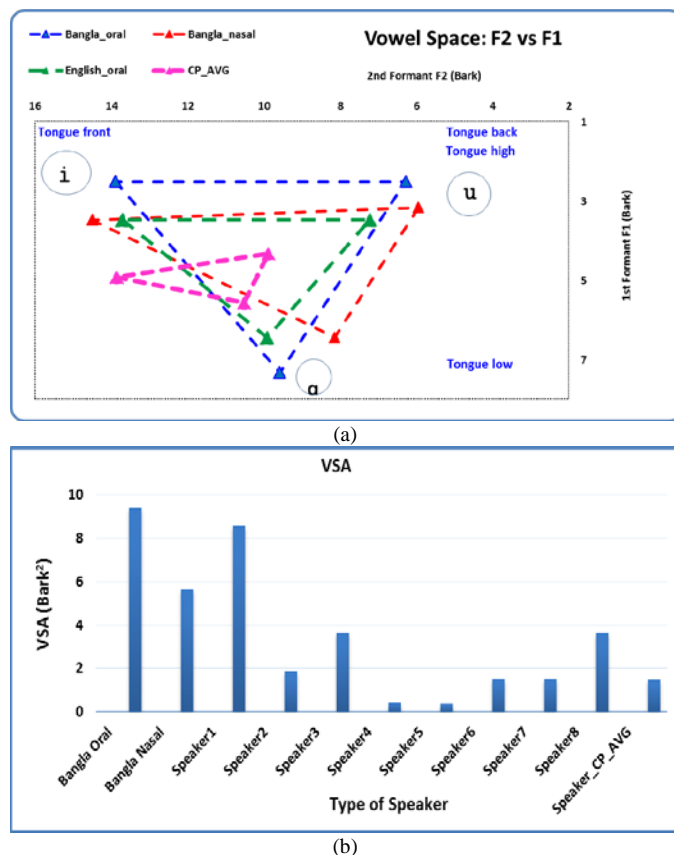


Fig. 8. (a) F1 vs F2 plots for VSA measurement for various speakers (b) VSA in Bark² versus type of speaker

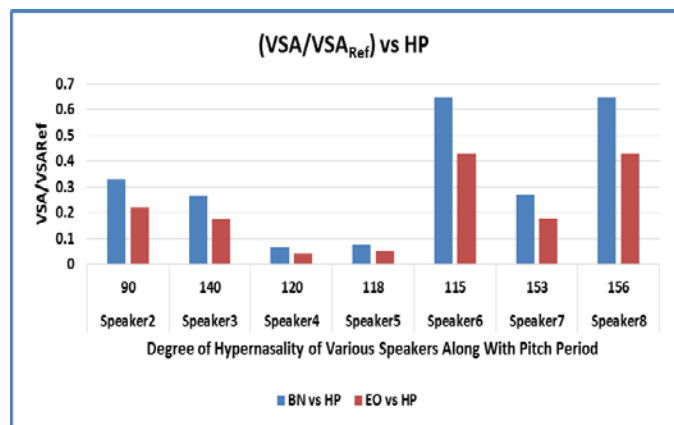


Fig. 9. Vowel space ratio across various VP opening conditions

VI. CONCLUSION

The study brings together the isolated Bangla oral-nasal vowels and read English vowel for normal and CP speakers. This leads to study the variations of extracted vocal tract features and their comparison in terms of VSA for determining the status of speech disorder of CP patients. The main objective of this work is to study the variation of acoustic features for CP speakers due to various VP opening and assessment of HP. For this purpose, an estimate of the

VSA for each of the 9 speakers utilizing the speech data per speaker are computed. The evolution of VSA with the 7 degrees of hypernasal articulation is analyzed. This triangle consists of the three vowels /a/, /i/ and /u/ represented in the space of the two first formant frequencies F1 and F2. The first main conclusion is that interspeaker variability of HP among CP speakers is measured by calculating mean and standard deviation of the selected vowels and /u/ shows the most variability among speakers. Second main conclusion is the significant reduction of the vocalic space as speech becomes less articulated. As the articulatory boundaries are less marked, the resulting acoustic targets are less separated in the VSA.

This study can be further be extended to make a comparison with sustained isolated hypernasal vowel data of CP speakers. Reduced intelligibility in hypernasal speech may partially be explained by this. Read speech of CP speakers is articulated differently from isolated vowel speech and reading proficiency might be a factor for which further investigations are required.

REFERENCES

- [1] Ruben, "Redefining the Survival of the Fittest: Communication Disorders in the 21st Century", *Laryngoscope*, Vol. **110**, No. 2, pp. 241-245, 2000.
- [2] D. Cairns, J. Hansen, and J. Kaiser, "Recent Advance in Hypernasal Speech Detection Using the Nonlinear Teager Energy Operator," in *ICSLP-96: Inter. Conf. on Spoken Language Processing*, vol. 2, 1996, pp. 780-783.
- [3] Rullo, R., Di Maggio, D., Festa, V.M. and Mazzarella, N., "Speech assessment in cleft palate patients: a descriptive study", *Int. J. of Pediatric otorhinolaryngology*, 73(5):641-644, 2009.
- [4] "Congenital malformations worldwide". International Clearing house for Birth Defects Monitoring Systems, Amsterdam, Holland. Tech. Rep., 1991.
- [5] Hawkins, S., and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels," *J. Acoust. Soc. Amer.*, vol. 77, no. 4, pp. 1560-1574, Apr. 1985.
- [6] Chen, M. Y., "Acoustic parameters of nasalized vowels in hearing impaired and normal hearing speakers," *J. Acoust. Soc. Amer.*, vol. 98, no. 5, pp. 2443-2453, Nov. 1995.
- [7] J. R. Glass and V. W. Zue, "Detection of nasalized vowels in American English," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1985, pp. 1569-1572.
- [8] Vijayalakshmi, P., Nagarajan, T. and Jayanthan Ra, V., "Selective pole modification-based technique for the analysis and detection of hypernasality", *Proc. of IEEE TENCON*, pp. 1-5, 2009.
- [9] O. Fujimura and J. Lindqvist, "Sweep-tone measurements of the vocal tract characteristics", *J Acoustical Soc. Am.*, vol. 49(2), pp. 541-58, 1971.
- [10] G. Fant, *Speech Sounds and Features* (MIT Press, Cambridge, 1973).
- [11] D.A. Cairns, J.H.L. Hansen and J.E. Riski. Detection of hypernasal speech using a nonlinear operator. *Proc. of IEEE Conf. on Engineering in Medicine and Biology Society*, pp. 253-4, 1994.
- [12] D. K. Rah, Y. I. ko, C. Lee, and D. W. Kim, "A noninvasive estimation of hypernasality using a linear predictive model," *Ann. Biomed. Eng.*, vol. 29, pp. 587-594, 2001.
- [13] P. Vijayalakshmi and M. R. Reddy, "Analysis of hypernasality by synthesis," in *Proceedings of Int. Conf. Spoken Language Processing*, Jeju island, South Korea, Oct. 2004, pp. 525-528.
- [14] P. Vijayalakshmi, M. R. Reddy and Douglas O'Shaughnessy, "Acoustic analysis and detection of hypernasality using group delay function", *IEEE Trans. Biomedical Engineering*, vol. 54, no. 4, pp. 621 - 629, Apr. 2007.
- [15] Vijayalakshmi, P., and M. R. Reddy, "The analysis of band-limited hypernasal speech using group delay based formant extraction technique" in *Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 665-668.
- [16] A. Bladon, "Two-formant models of vowel perception: Shortcomings and enhancement," *Speech Commun.* 2(4), 305-313 (1983).
- [17] A. T. Neel, "Vowel space characteristics and vowel identification accuracy," *J. Speech Lang. Hear. Res.* 51(3), 574-585 (2008).
- [18] S. Skodda, W. Grönheit, and U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease," *PloS ONE* 7(2), e32132 (2012).
- [19] L. B. Leonard, S. E. Weismer, C. A. Miller, D. J. Francis, J. B. Tomblin, and R. V. Kail, "Speed of processing, working memory, and language impairment in children," *J. Speech Lang. Hear. Res.* 50(2), 408 (2007).
- [20] S. Sandoval, V. Berisha, R. L. Utianski, and J. M. Liss, "Automatic assessment of vowel space area", *J. Acoust. Soc. Am.* **134**, EL477 (2013).
- [21] S. Scherer, L. Morency, J. Gratch, and J.P. Pestician, "Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4789-4793, 2015.
- [22] Marziye Eshghi, Mohammad Mehdi Alemi and Mohammad Eshghi, "Vowel nasalization might affect the envelop of the vowel signal by reducing the magnitude of the rising and falling slope amplitude", *J. Acoust. Soc. Am.* **137**, 2304 (2015).
- [23] S. Haque, T. Takara, "Bangla Oral-nasal Vowel Pairs: Acoustic Categorization and Comparative Study of Feature Extraction Methods", Volume 3, Issue 7, *Journal of Computing*, July 2011, (ISSN 2151-9617), NY, USA.

Optimization of Dynamic Virtual Machine Consolidation in Cloud Computing Data Centers

Alireza Najari^{1,*}

Department of Computer Engineering
Ahvaz Branch, Islamic Azad University
Ahvaz, Iran

Department of Computer Engineering
Khuzestan Science and Research Branch Islamic Azad University
Ahvaz, Iran

Seyed EnayatOllah Alavi²

Department of Computer Engineering
Shahid Chamran University of Ahvaz
Ahvaz, Iran

Mohammad Reza Noorimehr³

Department of Computer Engineering
Ahvaz Branch, Islamic Azad University
Ahvaz, Iran

Abstract—The present study aims at recognizing the problem of dynamic virtual machine (VM) Consolidation using virtualization, live migration of VMs from underloaded and overloaded hosts and switching idle nodes to the sleep mode as a very effective approach for utilizing resources and accessing energy efficient cloud computing data centres. The challenge in the present study is to reduce energy consumption thus guarantee Service Level Agreement (SLA) at its highest level. The proposed algorithm predicts CPU utilization in near future using Time-Series method as well as Simple Exponential Smoothing (SES) technique, and takes appropriate action based on the current and predicted CPU utilization and comparison of their values with the dynamic upper and lower thresholds. The four phases in this algorithm include identification of overloaded hosts, identification of underloaded hosts, selection of VMs for migration and identification of appropriate hosts as the migration destination. The study proposes solutions along with dynamic upper and lower thresholds in regard with the first two phases. By comparing current and predicted CPU utilizations with these thresholds, overloaded and underloaded hosts are accurately identified to let migration happen only from the hosts which are currently as well as in near future overloaded and underloaded. The authors have used Maximum Correlation (MC) VM selection policy in the third phase, and attempted in phase four such that hosts with moderate loads, i.e. not overloaded hosts, liable to overloading and underloaded, are selected as the migration destination. The simulation results from the Clouds framework demonstrate an average reduction of 83.25, 25.23 percent and 61.1 in the number of VM migrations, energy consumption and SLA violations (SLAV), respectively.

Keywords—Cloud Computing; Dynamic Consolidation; Energy Consumption; Virtualization; Service Level Agreement

I. INTRODUCTION

According to the definition provided by NIST [1] "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage,

applications and services). It can be rapidly provisioned and released with minimal management effort or service provider interaction". Infrastructure as a Service (IaaS) is among the services provided by cloud computing that offers processing resources to users as services. The clients rent the equipment from infrastructure providers as a service and only pay for the amount of service they really consume [2, 3].

The ever-increasing growth and wide applications of cloud computing, as well as the extensive usage of cloud services in all scopes, have caused a growing trend in energy consumption of cloud computing data centres. Therefore, the operational costs in these centres are intensively increasing due to the electric energy used [4].

According to the reports published by Microsoft [5], the consumed energy used by physical resources can account for 45 percent of the operational costs in a data centre. This amount has multiplied in the last five years [6]. Therefore, to maintain their business and to remain in the market competition, service providers need to minimize energy consumption to cut the excessive operational costs in a way that the integrity and quality of service remain intact [7]. Hence, two challenging tasks in IaaS are management and optimized allocation of resources, to the extent that the success of cloud services heavily relies on this issue.

The present study recognizes the problem of dynamic virtual machine Consolidation using virtualization, live migration of VMs from underloaded and overloaded hosts and switching idle nodes to the sleep mode, as a very effective approach for utilizing resources and accessing energy efficient cloud computing data centres [8-11].

The challenges faced are the consolidation of VMs and their allocation and placement on physical service providers in a way that they minimized energy consumption in the entire data centre and the number of active hosts as well as SLA violation, which is a contract between the clients and the providers. Many studies [11-13] have reported that fully idle

hosts consume as much as 70 percent of the energy used in maximum utilization mode. Therefore, approaches should be taken to minimize the number of underloaded hosts; transfer hosted VMs to other hosts and switch idle hosts to the sleep mode to further decrease energy consumption.

Live VM migration technique transfers a VM from one host to another without any interruptions with the minimum downtime [6, 14-16]. As authors have pointed out in [11, 15], every VM migration process can cause performance degradation, which can roughly be considered 10 percent of CPU utilization. This finding indicates that each VM migration can lead to SLA violation and unnecessary VM migrations may impose extra management costs and consequently extra energy consumption. Therefore, it is necessary to minimize the number of VM migrations and therefore minimize SLA violation and energy consumption as well.

In most research studies conducted in this area [6, 8, 17-21], the necessary decisions are made based on current utilization of hosts and VMs are immediately migrated from hosts which researchers currently identify them as overloaded [8, 15]. The proposed algorithm in the present work attempts to predict CPU utilization using Time-Series Method and SES technique and identifies a VM as overloaded or underloaded based on current and predicted utilization values and their comparison with dynamic upper and lower thresholds. The rest of the paper is structured as follows: it presents Related Work in Section 2 and explains the proposed algorithm in Section 3. Then it provides the results from simulating the proposed algorithm in Section 4, along with their analysis and evaluation. Ultimately, in Section 5, the paper discusses conclusions and suggestions for future works.

II. RELATED WORK

Beloglazov and Buyya [8] proposed Mean Absolute Deviation (MAD) and Interquartile Range (IQR) methods to determine the dynamic upper threshold. In their study, they considered a host as overloaded if the current utilization was greater than the upper threshold. They suggested LR and LRR methods to forecast future loads on the host. This approach recognizes a host as overloaded if its predicted utilization is 100 percent or higher.

To solve the consolidation problem, in addition to VMs energy consumption, authors in [21] also investigated energy consumption in intercommunication networks at data centres. The generated solutions using the genetic algorithm (GA) were significantly better than those of the first-fit decreasing algorithm. However, the computation time in GA was linearly proportional to the number of VMs and hosts.

Gao et al. [18] used Multi-Objective Ant Colony Optimization (MOCO) algorithm for resource allocation with energy efficiency and resource wastage as the two objectives. They used a modification of ant colony algorithm (ACO) in which pheromone updates, definition and accumulation were modified to suit multi-objective problems better. Ultimately, the ACO-based method outperformed GA algorithm.

By continuing the work in [21], authors presented a Hybrid GA (HGA) in [21] for solving the consolidation problem. They

used an infeasible solution repairing procedure, in which by gradual resolving of constraint violations it converts an infeasible solution to a feasible one, along with a local optimization procedure which quickly improved the solutions. As compared with GA, HGA yielded more promising results and was able to find local optimums more efficiently in a new search space. However, the workload in HGA increased after implementing the two procedures.

Singh and Shaw [15] employed a load forecast model to determine the necessity of migration and identify appropriate destination hosts. They utilized a dynamic upper threshold and incorporated Time-Series prediction method and Dynamic Exponential Smoothing (DES) and SES techniques. According to their algorithm, a host is considered overloaded if the values of current and predicted CPU utilizations exceed the upper threshold.

In [6], authors presented a novel selection policy called MP, in which they used the dynamic upper and lower thresholds as well as a variable to determine the degree of resource satisfaction. They suggested a new placement policy called MCC, which relocates a migratable VM to a host with minimum correlation to the VM.

Arianyan et al. [17] proposed a holistic resource management procedure and a heuristic intelligent technology method based on multi-criteria decision-making method to determine underloaded hosts for placement of migratable VMs. They presented a multi-criteria method known as Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) by focusing their work on methods for determining the time to consider a host as underloaded and by finding a new location for placement of the VMs selected among underloaded and overloaded hosts.

Joseph et al. [19] introduced a Parallel GA model known as Family GA (FGA) with the aim to generate an optimized mapping between the set of hosts and VMs. This model divides the entire population into a number of families on which it performs genetic operations to overcome the GA limitations. They used a self-adjusting mutation operator to prevent premature convergence of the individuals in the population, which makes the probability of mutation dynamic.

In [22], the authors attempted to solve the consolidation problem by presenting a type of self-adjusting mutation operator as well as considering current and future resource demands and based on the k-Nearest Neighbor (K-NN) regression/prediction model. They proposed the K-NN model in their previous study [23].

III. THE PROPOSED ALGORITHM

Since the problem of dynamic consolidation of VMs in cloud computing data centres is wide extent, it is broken down into the four following phases [4]:

- Phase 1: Identification of overloaded hosts.
- Phase 2: Identification of underloaded hosts.
- Phase 3: Selection of VMs to migrate from overloaded hosts.

- Phase 4: Determining appropriate destinations for migration.

This study presents algorithms and approaches for each of the phases:

A. Phase 1: Identification of Overloaded Hosts

In this phase, similar to [15], the researchers identified a host as overloaded if it is currently and in near future overloaded; however, unlike [15], they used the Time-Series Prediction method as well as SES techniques to predict CPU utilization in near future. The reader may refer to [24-26] for further information on this topic. Moreover, the study proposed a more optimized, efficient equation for determining dynamic upper threshold. It is noteworthy that the study used DES method for obtaining the best results in [15], therefore, it compared our proposed algorithm with these best results from [15].

Algorithm of Phase 1: Identification of Overloaded Hosts

```
1.  Input: host      Output: migration decision (true/false)
2.  flagP = flagF = false
3.  Find current Utilization of host h
    Utilization = total requested MIPS/h.getTotalMips ()
4.  data [] = h.getUtilizationHistory ()
5.  Calculate MAD and find UpperThreshold using MAD
    UpperThreshold = 1- MAD
    If (Utilization > UpperThreshold) then
        |  flapP = true
7.  If(data.length < 12 and flagP == true) then
        |  OverUtilizedHosts.add (h)
        |  Return True
8.  Find future Utilization using SES Technique
    futureUtilization = getHostFutureLoad (data)
9.  If (futureUtilization > UpperThreshold) then
        |  flagF = true
10. If (flagF == false and flagP == true) then
        |  currentOverUtilizedHosts.add (h)
11. If (flagF == true and flagP == false) then
        |  predictedOverUtilizedHosts.add (h)
12. If (flagF == true and flagP == true) then
        |  overUtilizedHosts.add (h)
        |  Return true
    Else
        |  Return false
```

In the first phase of the proposed algorithm, a host is received as input, and depending on the status of the current and predicted CPU utilizations, the researchers categorized it in one of the three lists. Similar to [15], they used the two flags flagF and flagP. True values of flagF and FlagP for a host

indicate an overload in near future and at the present, respectively.

The basis of the decision-making in this phase is on the upper threshold. this paper proposed Equation (1) for Calculation of dynamic upper threshold. This Equation is inspired by the method presented in [8] and by employing Median Absolute Deviation (MAD) method [8].

$$\text{UpperThreshold} = 1 - \text{MAD} \quad (1)$$

If the current utilization is greater than the upper threshold, the researchers consider the host as overloaded and they set flagP to True (Step 6, Phase 1). [15] and [8] used 10 and 12 data values from the CPU utilization history, respectively, to predict a host overload. Our study considered 12 history values according to our investigations. If this value is lower than 12 and the value of flagP is True, we place the host in the OverUtilizedHosts list, and the algorithm terminates; otherwise, it continues operation (Step 7, Phase 1).

As it was noted earlier, CPU utilization in near future is predicted and calculated using SES method (Step 8, Phase 1). Predicted values higher than the upper threshold mean a host overload in near future will occur; therefore, the researchers set flagF to True (Step 9, Phase 1).

Similar to [15], the study considered three different categories for overloaded hosts. The first category includes currently overloaded hosts (True values for flagP), but are predicted not to remain overloaded in near future (False values for flagF). The researchers added such hosts to the current Over Utilized Hosts list. Since these hosts will not be overload in future and to decrease unnecessary migrations, VMs will not migrate from this category of hosts (Step 10, Phase 1).

The second category includes hosts which are not currently overloaded (False values for flagP), but the study predicts that they will overload in near future (True values for flagF). The researchers will add such hosts to the predicted Over Utilized Hosts list. Since these hosts are not currently overloaded, VMs will not migrate from them (Step 11, Phase 1). Third category includes hosts which are currently and in near future overloaded (True values for both flagP and flagF). The researchers added such hosts to the over Utilized Hosts list, and some of their hosted VMs are selected and migrated to decrease their load (Step 12, Phase 1).

Among the categories mentioned above, the VMs hosted on the third category are certainly considered overloaded, and some of them will migrate from the host to normalize its load.

B. Phase 2: Identification of Underloaded Hosts

In this phase, the following algorithm is presented to identify the underloaded hosts. In the second phase of the proposed algorithm, the researchers received a list of hosts as the input and a list of underloaded hosts is returned. Decision making in this phase is performed based on the lower threshold. If current CPU utilization of the host is below the lower threshold, the host is considered currently underloaded, and similarly, if the predicted CPU utilization of the host in near future is below the lower threshold, the host is known as

underloaded in near future. Inspired by the dynamic upper

Algorithm of Phase 2: Identification of Underloaded Hosts

```
1.  Input: hostList   Output: List of underUtilizedHosts
2.  For each host h in HostList
3.      LowerThreshold = 1
4.      MAD = 0
5.      data[] = h.getUtilizationHistory ()
6.      If (data.length >= 10)
           Calculate MAD and find LowerThreshold using MAD
           LowerThreshold = 0.25 + MAD
       Else
           LowerThreshold = 0.25
7.      Find current Utilization of host h
           Utilization = total requested MIPS/h.getTotalMips ()
8.      Find future Utilization using SES Technique
           futureUtilization = getHostFutureLoad (data)
9.      If (Utilization > 0 and Not (all VMs are migrating
           from host or any VM migrating to host))
10.         If (data.length < 10 and Utilization < LowerThreshold)
                UnderUtilizedHosts.add (h)
                continue
11.         If (futureUtilization < LowerThreshold and Utilization
                < LowerThreshold)
                UnderUtilizedHosts.add (h)
12.      End for
```

threshold method in [8] and using the MAD method, the study proposed an optimized equation for determining the dynamic lower threshold. The study requires at least 10 data values from CPU utilization history for predicting an underloaded host.

In this Paper, With 10 or more data values, the dynamic lower threshold is calculated according to (2); otherwise, the lower threshold assumes a constant value of 0.25 (Step 6, Phase 2).

$$\text{LowerThreshold} = 0.25 + \text{MAD} \quad (2)$$

CPU utilization in near future is predicted and calculated using SES method (Step 8, Phase 2). Before identification of an underloaded host, the researchers investigated two conditions. First, CPU utilization of the host should be larger than zero, and second, no VMs should be in the process of migrating from and to the host (Step 9, Phase 2). If they met conditions, then they will do the investigation to find that if the data length of the host utilization history is lower than 10 and if the current CPU utilization of the host is below the lower threshold. If the conditions hold true, given that the sufficient data for prediction of CPU utilization are not available, and the host is currently underloaded, it is added to the list of underloaded hosts and program control flow makes a jump to Step 2 of Phase 2 (Step 10, Phase 2).

If the condition in Step 10 is not met, Step 11 will evaluate current and predicted CPU utilizations of the respective host. If both values are below the lower threshold, the host is considered currently and in near future as underloaded and hence should be added the list of underloaded hosts. The control program flow then jumps to the beginning of the loop to check the conditions for the next host.

C. Phase 3: Selection of VMs to migrate from overloaded hosts

In this phase, unlike [15], in which the proposed minimum utilization (MU) policy of [8] was used, our proposed algorithm employs maximum correlation (MC) policy introduced in [8] due to its superior performance. The main idea behind MC policy was presented by [27]. The basis of this fact is that the more the correlation between the resource consumptions by the running applications on the host, the higher the possibility of overloading. According to this theory, the researchers will select the VMs on a host which have the maximum correlation with other VMs in consumption of processing resources for migration [8].

D. Phase 4: Identification of appropriate destinations for migration

In this phase, to identify the appropriate destinations of migration, the work in [15] is optimized by excluding underloaded hosts from the list of migration destinations. In [15], the researchers exclude only the three categories mentioned above including overloaded and prone to overload hosts from the list of appropriate hosts as destinations of migration, and efforts were made to select underloaded hosts and hosts with moderate loads as the destination of migration. In our study, in addition to excluding the overloaded and/or prone to overload hosts, underloaded hosts were also excluded from the list of appropriate destinations for migration. According to the made decisions, the effort was to select the VMs among those with moderate loads. This way, selection of destination hosts was optimized, the number VM migrations dropped significantly and they prevented from the underloaded machines to remain switched on, which could be turned off to significantly decrease energy consumption in the data centre.

IV. SIMULATION RESULTS AND ASSESSMENT OF THE PROPOSED ALGORITHM

This section provides a simulation of the algorithm and its assessment. Then it compares proposed algorithm with MAD-MU algorithm in [8] the proposed algorithm in [15] and the results were analyzed and examined. Clouds framework [28] was used to simulate the proposed algorithm.

A. Experiment Settings

In the present study, a data centre with 800 heterogeneous physical hosts was simulated using Clouds framework. Half of the hosts are HP ProLiant ML110 G4 (Intel Xeon 3040, two cores \times 1860 MHz, 4 GB) and the other half are HP ProLiant ML110 G5 (Intel Xeon 3075, two cores \times 2660 MHz, 4 GB). The data centre includes 4 types of single-core virtual machines: High-CPU Medium Instance: 2500 MIPS, 0.85 GB; Extra Large Instance: 2000 MIPS, 3.75 GB; Small Instance:

1000 MIPS, 1.7 GB and Micro Instance: 500 MIPS, 0.633 GB. The proposed algorithm aims to improve the proposed algorithms in [8, 15]. Therefore, to be able to carry out a performance comparison, settings similar to those of [8] and [15] were applied to our proposed algorithm.

B. Workload Data

Since most of the reviewed studies used the workload data from the CoMon project, which is a monitoring infrastructure associated with PlanetLab, we also used the same data in our study for assessing the proposed algorithm and for its comparison with its counterpart algorithms. For more realistic results, a CPU utilization dataset was used, the data of which were measured in 5-minute time intervals and were collected from more than thousands of operational VMs in over 500 locations around the world. To carry out a reasonable and appropriate comparison, the researchers used a workload data collected from 10 days in March and April 2011. These data are available in the Clouds framework at the moment.

C. Performance Metrics

Six parameters were used to assess and compare the proposed algorithm with those of other studies. These metrics included the number of VM migrations from overloaded and underloaded hosts, the total energy consumption of physical resources, performance degradation due to VM Migration (PDM) [8], SLA Violation Time per Active Host (SLATAH) [8], which can be defined as the percentage of the period when the host experiences a CPU utilization of 100%, the researchers calculated the combined metric SLAV by multiplying PDM with SLATAH [8] and indicated the duration in which the allocated resources to the host is lower than the required amount, ESV combined metric which is calculated by multiplication of total energy consumption with SLA violation [8] and is used to measure the simultaneous improvements in both metrics and indicates the trade-off between them.

D. Simulation Results

Figs. 1 to 6 demonstrate simulation results of the compared algorithms in for different metrics, and a detailed discussion is presented for each metric as follows. From now on, the study will refer to the proposed MAD-MU algorithm in [8] and the proposed Shaw and Singh Algorithm in [15] as MM and SSA for brevity respectively. Since our proposed algorithm is a modification to optimize SSA, we will call it Optimized SSA, and refer to it as OSSA for brevity.

Authors of [8] implemented MM which is currently in Clouds framework and it employs MAD technique to determine the upper threshold, and MU method to select the VMs for migration. SSA, which depends on MM to select VMs for migration as well as determining the upper threshold, used DES for its best CPU utilization prediction in future. Our proposed algorithm attempts to optimize SSA using the presented methods in Section III.

A comparison between MM, SSA and OSSA is demonstrated in Fig. 1 regarding the number of VM migrations. OSSA achieved 86.83, and 79.65 percent decreases as compared with MM and SSA, respectively. Increased accuracy in calculation of the upper threshold and consequently increased accuracy in identification of

overloaded hosts is among the reasons for the significant reduction in the number of migrations in OSSA algorithm as compared with the other two. Therefore, migrations only take place on the VMs which are more accurately identified as overloaded. Another reason for the reductions are the presentation of a new algorithm for identification of underloaded hosts. Through this, underloaded hosts are more accurately identified and the entire hosted VMs are more accurately migrated. As the third reason, by optimizing the procedure of finding appropriate destinations of migration, unnecessary VM migrations to inappropriate hosts are eliminated to a great extent.

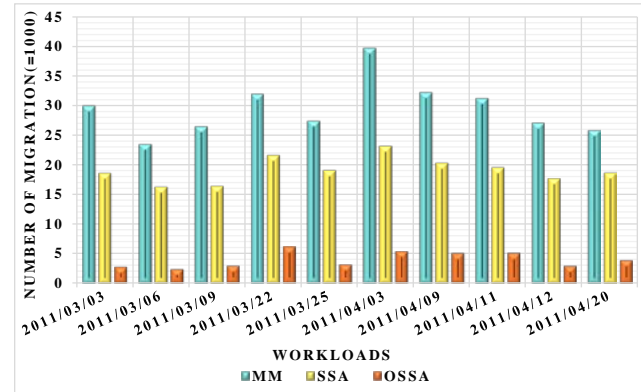


Fig. 1. Comparison of Number of VM Migrations metric against workload

Fig. 2 demonstrates a comparison between MM, SSA and OSSA from the energy consumption perspective. OSSA achieved 32.25 and, 18.2 percent decreases as compared with MM and SSA, respectively.

The main reason for these significant reductions is that OSSA uses a lower threshold for optimized selection of hosts with low utilization levels to prevent energy dissipation by switching them off. Another reason for this improvement is the use of an optimized upper threshold by OSSA which leads to more efficient and effective utilization of processing resources on the hosts by VMs. This improvement gives the opportunity to switch more hosts off to further decrease energy consumption.

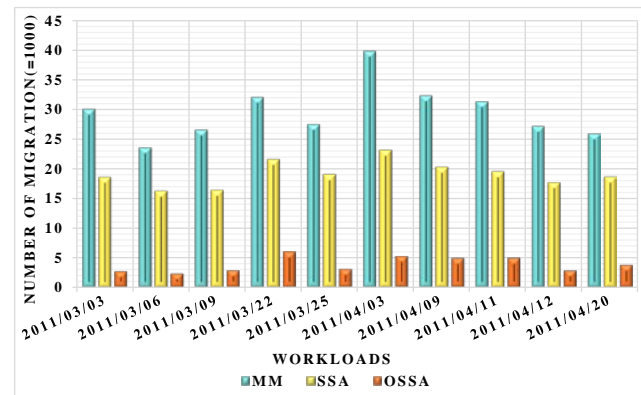


Fig. 2. Comparison of Energy Consumption metric against workload

Fig. 3. Demonstrates a comparison between MM, SSA, and OSSA on PDM metric. OSSA achieved 71.37 and 61.83 percent decreases as compared with MM and SSA,

respectively. The main reason for these significant reductions is the significant decrease in the number of migration in OSSA as compared with the other two algorithms.

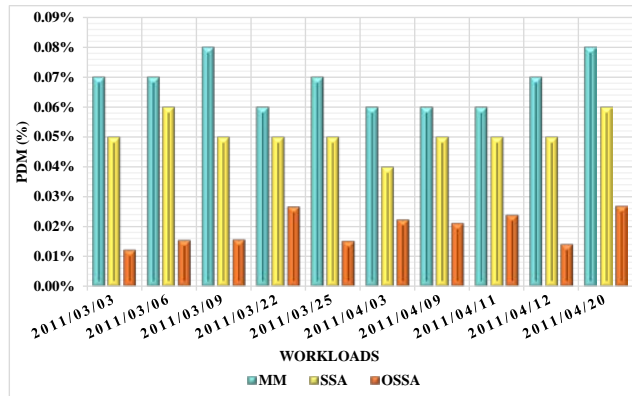


Fig. 3. Comparison of PDM metric against workload

Fig. 4 shows a comparison between MM, SSA and OSSA from the SLATAH metric point of view. As can be seen from this figure, OSSA demonstrated a poor performance in most cases as compared with the other two algorithms. The main reason for the poor performance can be associated with the attempts made by OSSA to achieve maximum host utilizations. However, since SLAV metric is calculated by a multiplication of PDM with SLATAH metrics, poor SLATAH performances may be neglected against the good PDM performances.

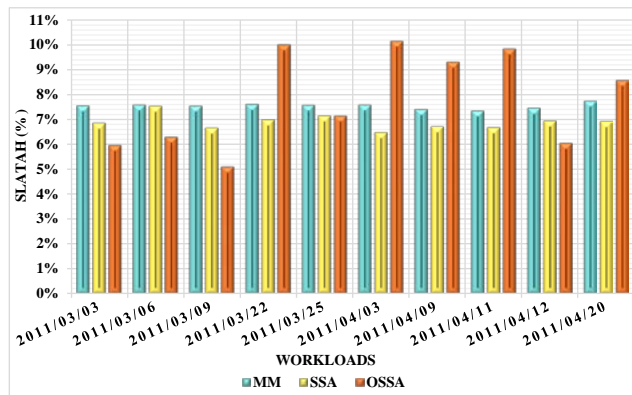


Fig. 4. Comparison of SLATAH metric against workload

A comparison between MM, SSA and OSSA with respect to SLAV metric is given in Fig. 5. OSSA achieved 68.06 and 54.13 percent decreases as compared with MM and SSA. This significant improvement, which is a result of the significant decrease in PDM metric, confirms that the poor performance of SLATAH metric in OSSA could in effect be neglected.

Fig. 6 demonstrates a comparison between MM, SSA and OSSA with respect to the combined ESV metric. OSSA achieved 77.49 and 60.47 percent decreases as compared with MM and SSA. Considering that ESV is calculated from multiplication of the two metrics of energy consumption and SLAV, therefore, the reason for this significant decrease is improvements in both mentioned metrics. There is a good trade-off between the two metrics in OSSA.

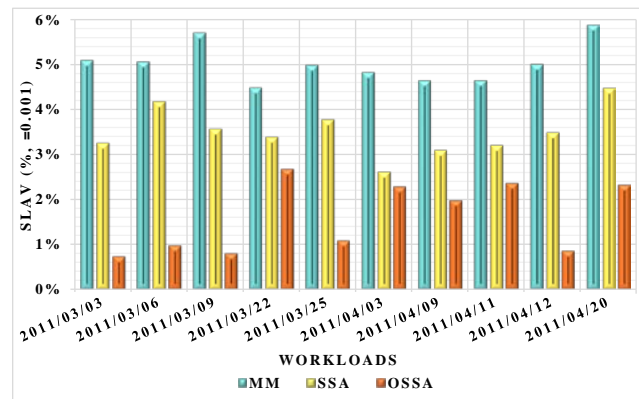


Fig. 5. Comparison of SLAV metric against workload

V. CONCLUSION AND FUTURE WORKS

This study investigated 4 phases of dynamic virtual machine consolidation problem, and for each, presented proper solutions. Also proposed an optimized equation for calculating the dynamic upper threshold and utilized maximum CPU capacity. SLA violation was decreased by eliminating unnecessary migrations, since migrations only took place on actually overloaded hosts. Use of maximum host processing power while maintaining SLA violation in an acceptable level led to increased number of VMs on the hosts, which consequently resulted in better conditions for switching off idle hosts and for decreasing energy consumption.

The study presented an optimized algorithm for identification of underloaded hosts and proposed an equation for calculation of the dynamic lower threshold. Using this threshold, VMs were migrated from underloaded hosts more accurately, allowing them to be switched off. This way, the researchers eliminated unnecessary migrations and decreased SLA violation, and on the other hand, optimized switch offs resulted in decreased energy consumption in the entire data centre.

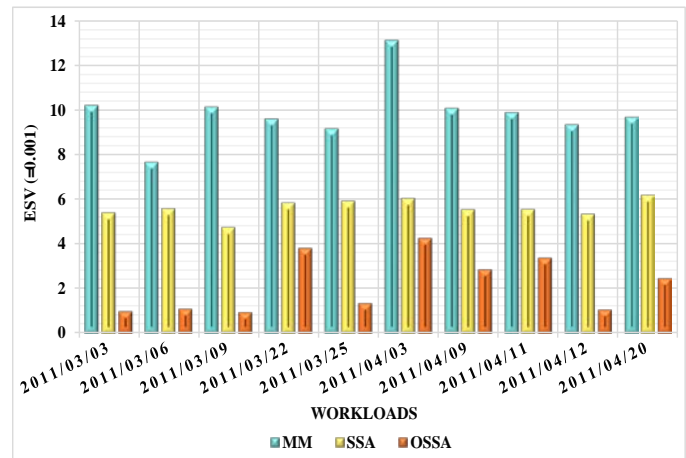


Fig. 6. Comparison of ESV metric against workload

To determine appropriate hosts as the migration destination, all hosts who were considered currently and in

near future overloaded, as well as underloaded hosts, were excluded from the list of migration destinations. The list helped to migrate VMs to destinations of higher quality, and by prevention of unnecessary migrations, SLA violation was decreased. By employing this policy, underloaded hosts were excluded from the list of appropriate destinations of migration, hence preventing VM migrations to this category of hosts. Therefore, opportunities to switch off hosts were protected, leading to further decreases in energy consumption.

OSSA, as compared with MM and SSA, were able to respectively achieve 86.83 and 79.65 percent decreases in the metric of number of migrations. It also can achieve 32.25 and 18.25 percent decreases in energy consumption metric, 71.37 and 61.83 percent decreases in PDM metric, 68.06 and 54.13 percent decreases in SLA violation metric, and 77.49 and 60.47 percent decreases in ESV metric. It also achieved a good trade-off between energy consumption and SLA violation.

It is suggested for future works to further investigate the poor performance of the proposed algorithm in SLATAH metric, since achieving improved SLA metrics leads to increases in the quality of the proposed algorithm. The performance of the proposed algorithm in real infrastructures are yet to be known. Therefore, for a real world performance evaluation, use of software packages such as OpenStack are suggested.

REFERENCES

- [1] P. Mell, and T. Grance, "The NIST definition of cloud computing," *Communications of the ACM*, vol. 53, no. 6, pp. 50, 2010.
- [2] R. Jeyarani, N. Nagaveni, and R. V. Ram, "Design and implementation of adaptive power-aware virtual machine provisioner (APA-VMP) using swarm intelligence," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 811-821, 2012.
- [3] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of internet services and applications*, vol. 1, no. 1, pp. 7-18, 2010.
- [4] X. Zhu, D. Young, B. J. Watson, Z. Wang, J. Rolia, S. Singhal, B. Mckee, C. Hyser, D. Gmach, and R. Gardner, "1000 islands: an integrated approach to resource management for virtualized data centres," *Cluster Computing*, vol. 12, no. 1, pp. 45-57, 2009.
- [5] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data centre networks," *ACM SIGCOMM computer communication review*, vol. 39, no. 1, pp. 68-73, 2008.
- [6] X. Fu, and C. Zhou, "Virtual machine selection and placement for dynamic consolidation in Cloud computing environment," *Frontiers of Computer Science*, vol. 9, no. 2, pp. 322-330, 2015.
- [7] J. Dong, X. Jin, H. Wang, Y. Li, P. Zhang, and S. Cheng, "Energy-saving virtual machine placement in cloud data centres," *Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference*, pp. 618-624, 2013.
- [8] A. Beloglazov, and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centres," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397-1420, 2012.
- [9] A. Beloglazov, and R. Buyya, "Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centres under quality of service constraints," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 7, pp. 1366-1379, 2013.
- [10] S. Esfandiarpour, A. Pahlavan, and M. Goudarzi, "Structure-aware online virtual machine consolidation for datacentre energy improvement in cloud computing," *Computers & Electrical Engineering*, vol. 42, pp. 74-89, 2015.
- [11] A. Beloglazov, and R. Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centres," *International Workshop on Middleware for Grids, Clouds and e-Science*, 2010
- [12] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer." pp. 13-23, 2007.
- [13] D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing environments via lookahead control," *Cluster computing*, vol. 12, no. 1, pp. 1-15, 2009.
- [14] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines." pp. 273-286, 2005.
- [15] S. B. Shaw, and A. K. Singh, "Use of proactive and reactive hotspot detection technique to reduce the number of virtual machine migration and energy consumption in cloud data centre," *Computers & Electrical Engineering*, vol. 47, pp. 241-254, 2015.
- [16] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centres for cloud computing," *Future generation computer systems*, vol. 28, no. 5, pp. 755-768, 2012.
- [17] E. Ariyanan, H. Taheri, and S. Sharifian, "Novel energy and SLA efficient resource management heuristics for consolidation of virtual machines in cloud data centres," *Computers & Electrical Engineering*, vol. 47, pp. 222-240, 2015.
- [18] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230-1242, 2013.
- [19] C. T. Joseph, K. Chandrasekaran, and R. Cyriac, "A Novel Family Genetic Approach for Virtual Machine Allocation," *Procedia Computer Science*, vol. 46, pp. 558-565, 2015.
- [20] M. Tang, and S. Pan, "A Hybrid Genetic Algorithm for the Energy-Efficient Virtual Machine Placement Problem in Data Centres," *Neural Processing Letters*, vol. 41, no. 2, pp. 211-221, 2014.
- [21] G. Wu, M. Tang, Y.-C. Tian, and W. Li, "Energy-Efficient Virtual Machine Placement in Data Centres by Genetic Algorithm," vol. 7665, pp. 315-323, 2012.
- [22] F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, and H. Tenhunen, "Utilization Prediction Aware VM Consolidation Approach for Green Cloud Computing." pp. 381-388, 2015.
- [23] F. Farahnakian, P. Liljeberg, and J. Plosila, "LiRCUP: Linear Regression Based CPU Usage Prediction Algorithm for Live Migration of Virtual Machines in Data Centres." pp. 357-364, 2013.
- [24] C. Chatfield, *Time-series forecasting*: CRC Press, 2000.
- [25] C. Chatfield, *The analysis of time series: an introduction*: CRC press, 2016.
- [26] M. Natrella, "NIST/SEMATECH e-handbook of statistical methods," 2010.
- [27] A. Verma, P. Ahuja, and A. Neogi, "pMapper: Power and Migration Cost Aware Application Placement in Virtualized Systems." pp. 243-264, 2008.
- [28] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, vol. 41, pp. 23-50, 2011.

A Light Weight Service Oriented Architecture for the Internet of Things

Omar Aldabbas

Faculty of Engineering
AL Balqa' Applied University
Al-Salt, Jordan

Abstract—Internet of Things (IoT) is a ubiquitous embedded ecosystem known for its capability to perform common application functions through coordinating resources distributed on-object or on-network domains. As new applications evolve, the challenge is in the analysis and implementation of multimodal data streamed by diverse kinds of sensors. This paper presents a new service-centric approach for data collection and retrieval, considering objects as highly decentralized, composite and cost-sufficient services. Such services are constructed from objects located within close geographical proximity to retrieve spatiotemporal events from the gathered sensor data. To achieve this, we advocate coordination languages and models to fuse multimodal, heterogeneous services through interfacing with every service to accomplish the network objective according to the data they gather and analyze. In this paper we give an application scenario that illustrates the implementation of the coordination models to provision successful collaboration among IoT objects to retrieve information. The proposed solution reduced the communication delay before service composition by up to 43% and improved the target detection accuracy by up to 70% while maintaining energy consumption 20% lower than its best rivals in the literature.

Keywords—Internet of Things; wireless sensor networks; sensing services; information extraction; data mining

I. INTRODUCTION

The key function of the Internet of Things (IoT) is to offer users access to information of interest from the big data gathered by smart devices connected over the Internet. Mining IoT data is a multi-phase procedure during which end users apply a structured approach to discover and retrieve the information encapsulated in data [1]. The authors of [2] reviewed IoT data mining approaches and categorized them into three classes based on the elements that motivate information requests; these are event-driven, periodic, and request-response interaction. Based on extensive literature review and investigation, it is clear that implementing one of these above approaches as a tool for mining IoT data is impractical due to its limited ability to extract a full picture of the current state of the ecosystem [2]. Based on this limitation, there is a requirement for a hybrid data mining approach that integrates two or more of the above mentioned data mining classes.

The heterogeneity of operating systems, hardware platforms, sensors and programming models in IoT devices makes data mining a challenging mission. The data streaming nature of some devices further raises the difficulty of

implementing an integrated data mining approach. Methods of IoT data mining are also categorized by [3] into three classes. Resource-centric methods are constraint-based aiming at optimizing the usage of the restricted resources at single IoT objects. Nevertheless, the decentralized kind of IoT data makes it hard to extract network-level information by resource centric methods. The second data mining scheme is the network-centric approach. Such methods, unlike the one proposed in [4] [27], are characterized by constrained knowledge of the semantics of the network topology and the application. The third approach is the data-centric that relies on data identifiers and pre-specified object locations. The cost of obtaining data locations, particularly in large-scale applications, restricts the suitability of such approaches. Accordingly, there is a requirement for autonomous hybridization and optimization of data mining techniques for IoT applications. Particularly, there is a need for hybrid autonomous adaptation approach to provision inter-object coordination and communication among different data mining approaches.

The amalgamation of methods from different data mining classes offers a comprehensive solution to optimize information retrieval to current user's needs and adaptations initiating from the aforementioned discussed concerns. This aim is not merely a network self-configuration challenge; the core problem here is the heterogeneity of the data mining approaches and the data extraction and analysis methods. From the perspective of an integrated system behavior, theoretically rudimentary tasks, e.g., data extraction, need multifaceted programming skills, particularly when working with constraint object and network resources. For example, the data streamed by objects cannot be correctly analyzed, and the object cannot communicate with neighbors when the data mining model on each object is different.

Additionally, the majority of the current hybrid data mining methods have intrinsic limitations that constrain their deployment including tailoring for particular applications [5]; have weak spatiotemporal data correlation abilities [5, 6]; utilize high bandwidth and drain power resources [7]; several approaches compromise the quality and amount of extracted information for energy consumption [8]; some approaches do not provide high-level interfaces to the user to configure thresholds and generate queries [5]; and the high dependency among particular data mining approaches -i.e., static composition - applications, protocol stack and hardware platforms preventing making code reuse [5, 6]. The

nonexistence of IoT data mining development platforms imposes handling information retrieval from the ground up for each new application. Such problems restrict the applicability of the implemented IoT hybrid data mining approaches, causing them to be complicated to use on anything other than the application developed for.

There is an increasing interest to modify the way for using the limited resources and capabilities in IoT ecosystems, to abstract and simplify them, converting them into communicating services of the network rather than capabilities of individual objects. Commonly, many objects interconnect to respond to a user request, linking their sensing, memory, and processing together. A decentralized query processing approach combines these resources and allocates jobs and data to participating object, so that the effectiveness of the query handling and the quality of the extracted data to be maximized. Consequently, different objects abilities together constitute the capability of the ecosystem as a whole. For instance, when an object equipped with sensors, memory, and processing resources is introduced to the ecosystem, it expands the data resources, on-board processing and memory capabilities to the ecosystem. This object is likely to improve the overall accuracy of the extracted data and reduce information retrieval. Alternatively, if an object leaves the ecosystem, it strips the ecosystem of its data, processing and memory resources.

This research addresses the development of a scalable, adaptive, and energy efficient hybrid IoT data mining approach, which integrates the strengths of three different data mining classes to increase the quality and amount of the extracted information, while reducing resource consumption. This approach embraces a service oriented view for the implementation of modular and adaptive applications. We introduce a service-oriented and semantics layer to the current network stack. The IoT is constructed at various layers of abstraction. At every layer, a group of services is specified. This service-oriented approach implements heterogeneous data mining approaches as lightweight services. To enable uniform associations between different services, located on the same or different objects, the desired technique utilizes "coordination" models [9]. Coordination models offer methods to combine heterogeneous services by interfacing with each service to achieve the application goals using the data they gather and analyze. This integration platform provides the opportunity for quick and accurate information extraction as it allows parallel and asynchronous data processing.

The rest of the paper is organized as follows: Section 2 presents the related works focusing on their limitations. Section 3 gives the architectural components of the proposed service-oriented data mining approach. Section 4 shows the service abstractions on individual objects. Section 5 presents a case study to show how to use coordination rules to compose a hybrid data mining service. Section 6 concludes the paper.

II. RELATED WORK

There has been continuous research into Service-Oriented Architecture (SOA) for highly decentralized systems, such as Internet of Things (IoT) over the last decade [10]. The authors

of [12] proposed an SOA-based application development model, a standardized interface to retrieve network data, and a group of configurable service components to provision the implementation of applications and to manage the network behavior at runtime. Another SOA platform, called Atlas, with middle-ware designed around the theory of self-integrative and programmable ubiquitous space [11]. Atlas's service utilities are run on a centralized control server, and power consumption is not a high priority. Blumenthal and Timmermann introduced the Resource Aware Service Architecture (RASA) [13], which structures software modifications by injecting services at runtime. The regular transmission of code results in high communication overhead.

Recently, in [14], a vision of a future IoT system architecture, which is based on service discovery across each layer of IoT is presented. This architecture provides mechanisms for on-demand discovery and integration of devices, cloud storage and computing resources, as well as application integration services, which can be dynamically chosen and orchestrated to create IoT applications. More recently, the authors of [15] an SOA to address the scalability issues leveraging the Path Computation Element (PCE) model. PCE proved to be an efficient technology to separate the control tasks from the sending objects, which has a great impact on scalability growth. For a broad and recent state-of-the-art, we refer the interested to the reader to [16] and the references therein.

The reviewed SOA approaches concentrated on developing architectures for the inter-object collaboration and communication, particularly, remote service access and service orchestration. Nevertheless, there has been little focus on object local service composition.

III. THE SERVICE-CENTRIC FRAMEWORK DETAILS

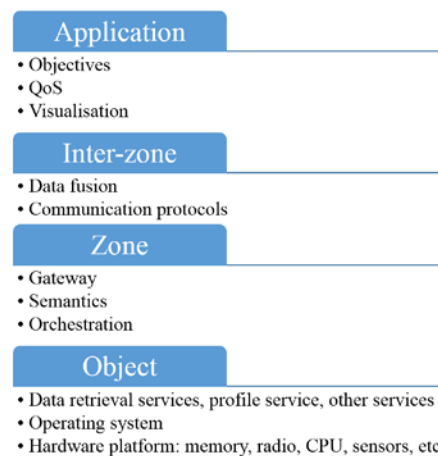


Fig. 1. Service stack of the proposed SOA collaborative platform

In the design of our IoT service-oriented application platform, we build on notions from numerous SOA systems [3] in the domain of Wireless Sensor Networks (WSN) and current approaches for coordinating the functions of large numbers of parallel active objects [17]. In Figure 1, we give a component diagram that shows the service stack of the suggested collaborative platform.

The top layer, Application Layer, designates the application aims or the high-level business processes of the IoT ecosystem. The goal of an IoT ecosystem specifies the information of interest from the user perspective and the kind of services required to retrieve it. The application aims are what defines the nature of data to be extracted, its frequency, scope, QoS, amongst other service or aim specific metrics. Even though the desired platform attempts to minimize the cost of retrieving data, some applications require specific end-to-end quality-of-service guarantees. For instance, for objects that are part of and self-driving car, it is critical that they respond in real-time to ensure the safety of passengers and other road users. Consequently, the desired platform must provide mechanisms to satisfy various QoS needs to ensure suitable utilization of object resources to satisfy the application objectives. The link between application aims and numerous layers and network components can be utilized to accomplish the required QoS. Eventually, data visualization is an essential component to deliver information in a way that end users can easily use and understand. It is also essential for applications to be able to show extracted information in various formats, e.g., maps or graphs, such that the format of information visualization strengthens its structure and content. Querying, executing, and responding to service requests between objects, potentially inter-zone communication. Finally, system managers utilize the profile-building interface as a tool to interact with the Profile services, i.e., to enter data mapping requests, configure coordination rules, and set other application configuration.

The Inter-zone Layer hosts the semantics of macro programs to deliver decentralized data flow. It fundamentally offers multi-hop communication with optional QoS feedback and control. Services at this layer utilize underlying protocols, which describe how services process and parse packets through description metadata instead of coding calls to each other in the source code. Communication protocols improve the interoperability among various services created by numerous providers by messages over specified message channels. This helps in reducing the complexity of the end application, hence, supporting application developers to concentrate on the application features. Most IoT systems utilize programming models that are application dependent, in-network abstractions, which are utilized in data processing. The systems published in [18] and [19] are instances of neighborhood-based abstractions, which deal with many objects cooperatively and a group of operations on it to allow the programmer to retrieve information on the state of the zone. Domain specific requirements, e.g., specific data semantic provision and data fusion, are held by the application layer to support the composition of application-specific information from raw data processing. Data fusion is a technique to enhance bandwidth utilization, power consumption, and information accuracy [20][21]. It integrates and merges inter-zone and multi-sourced data to produce comprehensive and higher accuracy information responses. To generate comprehensive information responses, the data fusion technique fuses data from different zones that are related to the same demand for information.

The Zone Layer is an abstraction layer to permit programmers to handle sensor objects as a spatially distributed entity rather than single isolated objects. This ultimately aims at reducing the amount of required processing and communication operations to retrieve the desired information. The cost of gathering raw data from the network is minimized by choosing a subgroup of sensor objects holding data related to the required information from a bigger group. Discounting extraneous data from an information request handling is also expected to increase the precision of the retrieved information. The Watershed procedure autonomously clusters sensor objects into uniform network clusters based on their topological associations and their soft-state, i.e., sensing-values. Clusters are then utilized as programming abstractions over which various information requests are processed. An information query could end up in creating a group of information retrieval services; everyone involve all active sensor objects in its zone. Such services are formed of a group of services, which fit in multiple zones. Service associated activities are considered as autonomic. At this layer, all zone semantic abstractions are given. Such abstractions are identified as contract expressions. The advantages of design-by-contract are voiced equally in software engineering [23] and programming languages [22] literature. Contracts are transparent to the object's operating system and are used on dataflow paths in a service composition. Composing a complex information retrieval service is regarded as the amalgamation of two distinct operations, the real processing operations with several operations that are part of extracting data and the coordination operations for the collaboration and communication between operations. Consequently, coordination is utilized to distinguish between the processing tasks of asynchronous and decentralized information retrieval from the communication ones, permitting the incorporation of these two key operations. The data extraction element in the Zone Level offers semantics of the sensor data, whereas the data computation elements offer service oriented access to data retrieval services.

A parametric orchestration, which incorporates a set of distinct services, is introduced to allow the system to adapt to various information requests. Every service is autonomously activated by allocating parameters values to meet the required information. For the desired system to function more effectively, a lower layer of access to physical devices is introduced. The lower level purpose is to stop the dependency of information extraction operations on the device operation and communication protocols. Classical layered protocol stack method limits the network software modularity. The majority of the current approaches comprise just the platform and application levels, leading to bounding the software to the attributes of the deployed platform. Therefore, sensor objects with different hardware/software platforms well as diverse methods for data sensing, storage, representations, filtering, processing, and routing cannot cooperate effortlessly. Eliminating the protocol layer helps to reduce the excessive overhead using its composite levels of abstraction and offers a simple method to alter and implement different software elements.

The Object Layer is concerned with the physical characteristics of sensor objects such as their OS, CPU power, memory size, power sources and wireless transceivers. It also includes the actual layout of sensors (physical topology) and network protocols. This layer is accountable for the interfacing with the operating systems and for the administration of resources of sensor objects. Moreover, this layer enables the coordination of physical resource allocation based on the application requirements as defined in the upper layers. Furthermore, services offered by the upper layers will potentially need some resource allocation provision. This level contains many operations related to information retrieval services. It specifies of three fundamental operations, i.e. trigger-based, periodic, and request-response service. Nevertheless, the overall performance, zone or inter-zone, which defines the organization of a hybrid information extraction services is platform-independent and is situated at higher layers of the defined stack. This level provision probing various sensor objects based on the dynamic topological, physical location and logical relationships between the services engaged in the application. The Profile service primary purpose is to log users of an information retrieval service. Generally, the Profile service stores the configuration regarding the utilization of data in a cooperative sensing manner. This configuration is essentially a set of coordination rules, which define the application business logic. Additionally, the Profile service stores other parameters, which are configured manually by system administrators to specify crucial and inherently static settings to help in the composition a complex-hybrid data retrieval service, which is suitable for a specific objective. Such settings cover service clients details (e.g., ID, type), data retrieval configuration (e.g., types of sensed data, parameters of gathered data, time stamp), environmental events, thresholds, etc. In addition to sensing the environment, IoT commonly has task-driven necessities related to hybrid services integration. Services integration necessitates information on the network topology, the state of objects, their position, and density, amongst others. All sensor objects host services to deliver such information.

IV. TRANSFORMING THE SENSOR OBJECT TO A SERVICE

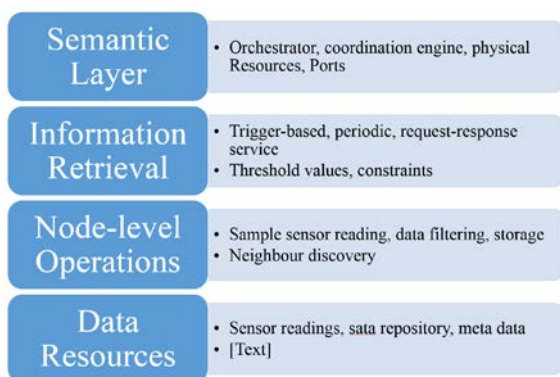


Fig. 2. An abstract representation of a sensor object as service

Semantic Layer: This is the core layer to transform the sensor object into a service that abstracts its capabilities, resources, hardware and software platform details. At the

Semantic Layer, the object retrieves the configured coordination rules, interprets, analyze and apply the coordination constraints to build a complex-hybrid information retrieval services at the individual object scale. At the Zone level, this layer also governs the provisioning of inter-object service composition, i.e., services located on different objects. Any object can advertise its services into one or more logical groups. In practice, service composition is implemented via ports. A port specifies events and commands offered by an object. This layer has parameters, which can be modified to change the behavior of the object. Furthermore, it applies non-blocking concurrency approaches that are investigated in [24]. To avoid any potential issues related to power consumption, clock distribution, and dense object distribution; the author implements asynchronous and asynchronous ports the local and inter-object exchange of information. The asynchronous ports permit remote running of commands on networked objects without disturbing the execution of local object processes, which initiated the remote call. The synchronous ports communications interrupt the execution of local process and pause the process termination on the remote host. The Orchestrator component provides a lightweight orchestration mechanism to provision the composition of services according to the defined coordination rules. Coordination rules define the resources to be assigned at the logical and physical levels and how to orchestrate them to ensure the desired service composition is achieved. The orchestration mechanism is designed to handle static-service-composition because the system administrator manually enters the coordination rules. The Orchestrator utilizes the object-defined services to discover the accessible resources at the local and inter-zone levels.

Information Retrieval: A service encapsulates some function, such as producing data, validating a transaction, or providing simple analytical services. The usability of any service is based on the functions offered through its interfaces, e.g. data extraction, logging, filtering, exchange, etc. Service interfaces are the primary tool that enables services to interact with each other. An information retrieval service wraps a nested functionality, which utilizes an ordered approach to find unstructured data and identify particular patterns encapsulated in the gathered data. In the presented solution, three types of information retrieval services, these are (1) Trigger-based: An object produces a notification message to report events, such as rising temperature or a moving object. At the basic level, a simple event is when a predefined threshold is exceeded. (2) Periodic: An object reports about the monitored environment on regular intervals. The system designer can set the reporting frequency, or it can adapt to application conditions. (3) Request-response: System users issue requests for information using a suitable query language and application interfaces. Queries can be distributed to network zones that carry data that is relevant to the user query. The query can be issued at regular intervals to gather data in reactive mode about the monitored environment. Queries can provide an effective solution to retrieve data from particular sections of the network.

Object-level operations: In the model advocated in this paper, an object is made up of a set of services, which offer

various functions including, an interaction service deals with the exchange of data messages, a fault-tolerance service attempts to heal software/hardware faults, a sensor calibration service to calibrate sensors, and a neighborhood service to discover neighbors and setup zones. The majority of object-based services offer applications and information retrieval services the capability to interact with the object hardware, namely sensors, actuators, battery level or the radio unit, by calling OS primitives.

Data Resources: The object Data Resources layer is a crucial layer of the object as a service stack. At this layer, the utilization of shared data resources is managed and tracked, a log and a register for available resources are upheld, and buffer pools are grouped. Logs and reports are produced at this layer and offered to the whole system. The logs contain events read/write, in sequential order and specify the services which are detecting the event, and the notification messages presents detected events in a concise method to the whole system. IoT ecosystems are characterized by composite data structures because of the multimodal sensing and decentralized applications implemented on heterogeneous objects. To construct an effective service, this layer offers a data abstraction service, which interprets the imperfect data setting. Information abstraction advances the capability to influence data analysis, irrespective of its structure, because of logical and new schema's, which is present in middleware. For instance, the field characteristics of sensing data generated by an object at this layer are ID, Object, Sensor Samples, Time Stamp and Status.

V. APPLICATION SCENARIO OF COORDINATED INFORMATION RETRIEVAL

This section of the paper presents how the primitives of a rule-based grammar can be applied to provide effective coordination between IoT objects to retrieve information. Such primitives are the primary phase on the way to the incorporation of rule-based grammar in IoT objects information retrieval tasks. Moreover, it shows an application scenario, which demonstrates how various information retrieval services can be coordinated. Coordination languages offer means of integrating two or more information retrieval services through interfacing with every service to create a unified service, which can run on a highly decentralized large-scale IoT ecosystem. The author addresses events, which are triggered by several units targets, e.g. a group of soldiers in a battlefield surveillance application. In this IoT ecosystem, an event-based information retrieval service is utilized to notify the periodic-based information retrieval service-driven service to enhance the gathered data accuracy. Objects send their sensor samples periodically to the end user to specify the position of the soldiers. In addition to periodic transmission of data, some objects run a trigger-based service. Once an object senses an event in the form of movement in the monitored environment, it adjusts its periodic information extraction service to escalate the information frequency. Then, the object informs its neighbors in the same logical zone about the event in their region to adapt their periodic data collection services.

In this scenario, the Coordination Language Facility (CLF) [25] is chosen as the coordination layer on top of the IoT

ecosystem infrastructure. In CLF, the utilization of rules to manage the operation of different or single vendor services is established on a proactive approach. A CLF system dynamically attempts to effect its environment, instead of just replying to external activities. This characteristic compliments the autonomy of every active IoT object. An IoT object listens for activities to happen in the environment and creates new reports based on the defined rules and their interpretation by its rules engine.

Listing 1: Integrating two information retrieval services using CLF

```
1 waitForNext @ sample (Motion ) <-> send (Motion ) @
  check_activity (Motion )
2 activity ( normal ) @ interval ( normal ) <-> #b
3 activity ( normal ) @ interval ( high ) <-> adjust_interval ( normal
  )
4 activity ( high ) <-> adjust_interval ( fast ) @ inform_neighbour
5 (Motion , Position , Interval )
6 received (Motion , position , Interval ) @ interval ( normal )
7 <-> sample (Motion ) @
8 send (Motion ) @ adjust_interval ( fast )
9 received (Motion , Position , Interval ) @ interval ( high ) <-> #b
```

When implementing the above described scenario, the coordination rules are located on and maintained by the individual IoT objects to enable several objects to collaborate autonomously, i.e., such rules apply to individual objects. The waitForNext primitive is a read-only token, which returns a value at the start of every sensing interval; it acts as a trigger to begin the search for an instance of the rule. The sample primitive stores the current motion readings from the motion sensor. The send primitive accepts the sensor sample and transmits it to the higher layer objects (e.g., cluster head or directly to the sink). The check_activity primitive examines the level of soldiers movement or activity within the object's sensing coverage area. If the level of activity, activity(normal), is within normal bounds and the transmission interval is also normal, interval(normal), then the object remains idle and continuous listening. Once the object's present level of activity is normal and the transmission interval is high, interval(high), the object sets the interval back to normal to save its energy, change_interval(normal). If the object detects a high activity within its area of coverage, then it sets its time-driven information retrieval service frequency to quick and sends a notification message to its neighbors informing them about the event in their region notify_neighbour(Motion; Location; Interval). The object includes in this notification message its sensor reading, position, and a new interval value. After an object receives a notification message, received(), it sets its periodic information retrieval service interval to quick and it sends its present sensor reading.

This application scenario demonstrates that coordination rules offer a practical and intuitive approach for integrating services from hybrid information retrieval classes. The coordination engine can manage the objects resources needed for composing a hybrid and distributed information retrieval service. What exactly is being coordinated, how the coordination is achieved, and what are the relevant grammar primitives that must be implemented and utilized, are all questions that we plan to address in the future.

VI. PERFORMANCE EVALUATION

The application scenario described in the previous section has been implemented on a small-scale testbed. The built testbed is used for testing and evaluating the efficiency of hybrid information retrieval service composition. The testbed is made up of six Sun SPOT (Sun Small Programmable Object Technology) nodes and two base stations acting like gateways to the Internet. Sun SPOT is a wireless sensor node for a WSN application prototyping, which was developed by Sun Microsystems. This node uses the IEEE 802.15.4 protocol for its communication, and different from other available hardware platforms, it runs the Squawk Java Virtual Machine. Sun SPOT processor board has ARM architecture 32 bit CPU with ARM920T core running at 180 MHz. It has 512 KB RAM and 4 MB flash memory. A 2.4 GHz IEEE 802.15.4 radio had an integrated antenna. The sensor board comes with an integrated a 3-axis accelerometer, light sensor and temperature. An abstract representation of the testbed is given in Figure 3.

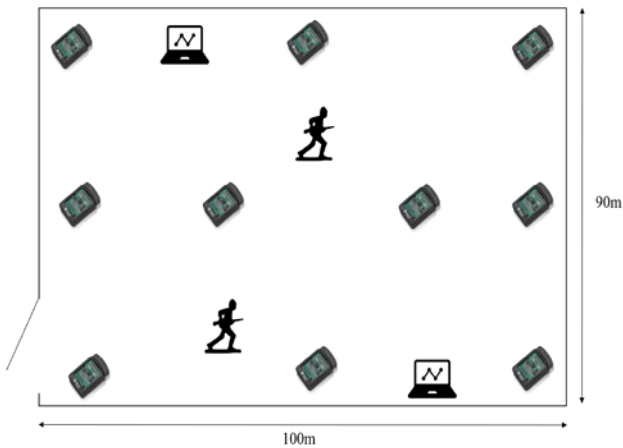


Fig. 3. Testbed layout showing Sun SPOT devices distributions and two laptops connecting the base stations. The running soldier icons determines the location of activity to be captured by the deployed nodes

Due to lack of resources to install motion sensors to detect the location of moving objects, the sensor devices were placed on students tests in a large lecture theater. To generate activity, the lights in the room were switched off and two students were asked to wander in the room using powerful LED torches. The light sensor readings were then used to indicate the presence of a target in the room.

The goal of the experiment was to compose a hybrid service on the run time. This service combines both periodic and event-driven information retrieval approaches. Initially, sensor nodes were operating in the periodic mode with large sensing intervals, i.e., every 3 seconds. When an object is detected, sensor nodes will switch the event driven service to continuously monitor the environment. The event driven service will check any sampled readings against preset thresholds, which were calculated based on the ambient light level in the room.

We compare our proposed approach to EDSOA [26] from the literature as they both offer similar services for the development of service-oriented IoT applications. EDSOA is

an information-centric session approach to describe service behavior working upon distributed events, called event session. We define two benchmarks service communication delay and sensing accuracy. Service communication delay is the communication time for two services running on different nodes before a composite service is built.

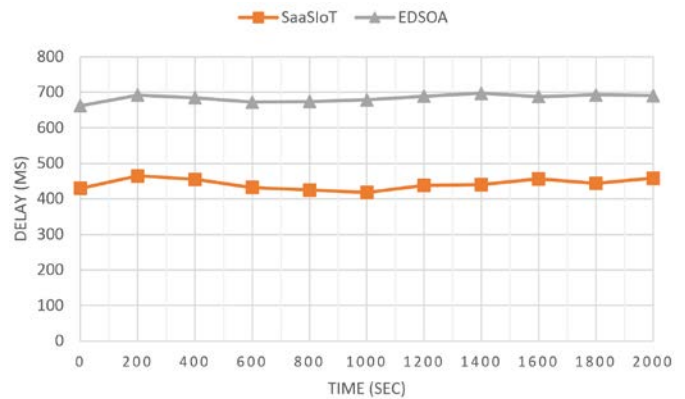


Fig. 4. Communication delay before service composition

Figure 4 shows the communication delay time comparison results. The delay time starts from when an object is detected by the first node up to the complete service composition. Compared to EDSOA, the communication delay before a service is composed has been significantly reduced. This is mainly due to the fact that the composition rules have been predefined. Even though these rules are static, their application is dynamic. The implementation of the coordination rules was executed in parallel on multiple devices. The devices that are in the same coordination state will be ready to collaborate with neighboring nodes without any delays.

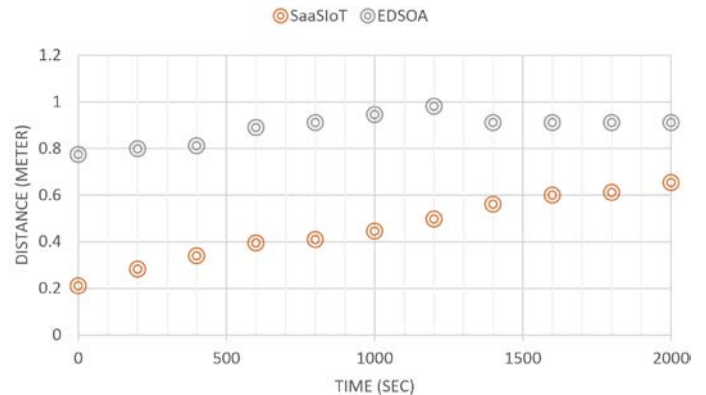


Fig. 5. Target detection accuracy

Figure 5 shows the detection accuracy rate. The coordination based approach has again performed significantly better than EDSOA due to the high levels of coordination between sensing objects. The ability of the composite information retrieval service to switch from one sensing mode to another and dynamically adjust the sensing interval has also contributed to the reduction of target detection localization delay and consequently increases the accuracy of the returned information.

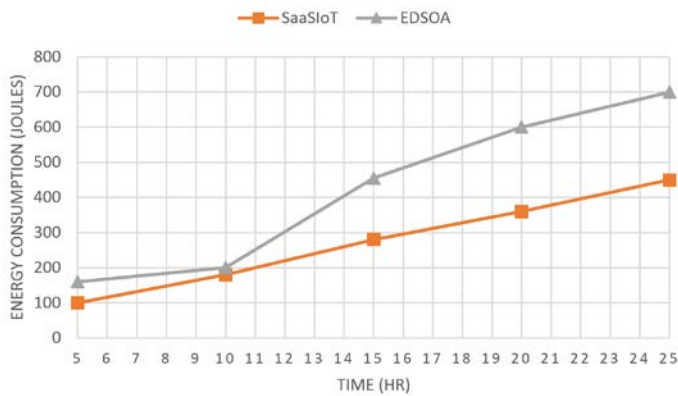


Fig. 6. Energy consumption in Joules

Figure 6 shows the cost of the service composition in Joules. The number of exchanged messages is significantly less than that of EDSOA due to the execution of the coordination rules independently on each device. The outcome of the coordination rules, which is based on the current environmental condition imposes that only the nodes that share knowledge about the target needs to communication and perform collaborative sensing. Due to the nature of the sensed environment, such nodes are normally physically located close to each other. This reduces the bridging distance between devices and results in smaller energy consumption.

VII. CONCLUSION

This paper attempted to address the challenge of utilizing heterogeneous sensor data to allow collaboration and information extraction from a group of objects at various levels of abstractions. In the presented approach, the logical grouping of objects is a crucial notion for service composition and coordination. Coordination languages can provide a hardware/software platform independent solution for the development of run-time hybrid and effective information retrieval services. We advocate adding a new level of abstraction to group objects with mutual mission restrictions. The proposed solution permits objects to collaborate and support network-level connectivity with a shared state of data and applications.

The contributions of this paper are three folds. First, we present a service stack of the proposed SOA collaborative platform and detail its individual components. Second, we present a practical approach to transforming the IoT object to a service. Finally, we demonstrated the applicability of such an approach using a scenario application implementation in the CLF coordination language. We have shown that the coordination engine can manage the objects resources needed for composing a hybrid and distributed information retrieval service. What exactly is being coordinated, how the coordination is achieved, and what are the relevant grammar primitives that must be implemented and utilized, are all questions that we plan to address in the future.

ACKNOWLEDGMENT

I am thankful for Al-Balqa' Applied University (BAU) for supporting me to do this research.

REFERENCES

- [1] Abuarqoub, A., M. Hammoudeh, and T. Alsoubi, An Overview of Information Extraction from Mobile Wireless Sensor Networks, in Internet of Things, Smart Spaces, and Next Generation Networking. 2012, Springer Berlin / Heidelberg. p. 95-106.
- [2] Alsoubi, T., et al., Information Extraction from Wireless Sensor Networks: System and Approaches. *Sensors and Transducers*, 2012. 14-2: p. 1-17.
- [3] Gracanin, D., et al., A service-centric model for wireless sensor networks. *IEEE J.Sel. A. Commun.*, 2006. 23(6): p. 1159-1166.
- [4] M. Hammoudeh, O. Aldabbas, S. Mount, S. Abuzour, M. Alfawair and S. Alratrout, "Algorithmic construction of optimal and load balanced clusters in Wireless Sensor Networks," *Systems Signals and Devices (SSD)*, 2010 7th International Multi-Conference on, Amman, 2010, pp. 1-5.
- [5] Lee, B.-D., Adaptive Data Dissemination Protocol for Wireless Sensor Networks, in *Security-Enriched Urban Computing and Smart Grid*. 2010. p. 188-195.
- [6] Lee, C.-H., C.-W. Chung, and S.-J. Chun, Effective processing of continuous group-by aggregate queries in sensor networks. *J. Syst. Softw.*, 2010. 83: p. 2627-2641.
- [7] Bhargavi, R., et al., Complex Event Processing for object tracking and intrusion detection in Wireless Sensor Networks, in *Control Automation Robotics Vision (ICARCV)*, 2010 11th International Conference on. 2010. p. 848 -853.
- [8] Bahrepour, M., Meratnia, N. and Havinga, P.J., 2009, July. Sensor fusion-based event detection in wireless sensor networks. In *Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous*, 2009. *MobiQuitous' 09*. 6th Annual International (pp. 1-8). IEEE.
- [9] Abreu, J.a. and J.L. Fiadeiro, A coordination model for service-oriented interactions, in *Proceedings of the 10th international conference on Coordination models and languages*. 2008, Springer-Verlag: Oslo, Norway. p. 1-16.
- [10] Hammoudeh, M.; Newman, R.; Dennett, C.; Mount, S.; Aldabbas, O. Map as a Service: A Framework for Visualising and Maximising Information Return from Multi-Modal Wireless Sensor Networks. *Sensors* 2015, 15, 22970-23003.
- [11] King, J.C., Atlas: a service-oriented sensor and actuator network platform to enable programmable pervasive computing spaces. 2007, University of Florida: Gainesville, FL, USA.
- [12] Delicato, F., et al., Exploiting Web Technologies to Build Autonomic Wireless Sensor Networks, in *Mobile and Wireless Communication Networks*. 2006, Springer Boston. p. 99-114.
- [13] Blumenthal, J. and D. Timmermann, Resource-Aware Service Architecture for Mobile Services in Wireless Sensor Networks, in *Wireless and Mobile Communications*, 2006. *ICWMC '06*. International Conference on. 2006. p. 34.
- [14] D. Georgakopoulos, P. P. Jayaraman, M. Zhang and R. Ranjan, "Discovery-Driven Service Oriented IoT Architecture," 2015 IEEE Conference on Collaboration and Internet Computing (CIC), Hangzhou, 2015, pp. 142-149.
- [15] V. B. C. Souza, X. Masip-Bruin, E. Marin-Tordera, W. Ramirez and S. Sánchez-López, "Towards the scalability of a service-oriented PCE architecture for IoT scenarios," *Networks and Optical Communications - (NOC)*, 2015 20th European Conference on, London, 2015, pp. 1-6.
- [16] Vanitha, V., V. Palanisamy, and K. Baskaran, Automatic Service Graph Generation for Service Composition in Wireless Sensor Networks. *Procedia Engineering*, 2012. 30(0): p. 591 - 597.
- [17] Hammoudeh, M., Newman, R., Dennett, C. and Mount, S. (2013), Interpolation techniques for building a continuous map from discrete wireless sensor network data. *Wirel. Commun. Mob. Comput.*, 13: 809–827. doi: 10.1002/wcm.1139
- [18] Newton, R., G. Morrisett, and M. Welsh, The regiment macroprogramming system, in *Proceedings of the 6th international conference on Information processing in sensor networks*. 2007: Cambridge, Massachusetts, USA. p. 489-498.
- [19] Mottola, L. and G.P. Picco, Programming wireless sensor networks with logical neighborhoods, in *Proceedings of the first international*

- conference on Integrated internet ad hoc and sensor networks. 2006, ACM: Nice, France.
- [20] Hammoudeh, M., R. Newman, and S. Mount, An Approach to Data Extraction and Visualisation for Wireless Sensor Networks, in Proceedings of the 2009 Eighth International Conference on Networks. 2009, IEEE Computer Society. p. 156-161.
- [21] Mohammad Hammoudeh, Robert Newman, Information extraction from sensor networks using the Watershed transform algorithm, Information Fusion, Volume 22, March 2015, Pages 39-49, ISSN 1566-2535, <http://dx.doi.org/10.1016/j.inffus.2013.07.001>.
- [22] Findler, R.B. and M. Felleisen, Contracts for higher-order functions, in Proceedings of the seventh ACM SIGPLAN international conference on Functional programming. 2002, ACM: Pittsburgh, PA, USA. p. 48-59.
- [23] Meyer, B., Applying "Design by Contract". Computer, 1992. 25(10): p. 40-51.
- [24] Mount, S., et al., CSP as a Domain-Specific Language Embedded in Python and Jython, in CPA'09. 2009. p. 293-309.
- [25] Andreoli, J.-M., S. Freeman, and R. Pareschi, The Coordination Language Facility: coordination of distributed objects. JOURNAL OF THEORY AND PRACTICE OF OBJECT SYSTEMS (TAPOS, 1996. 2: p. 1-18.
- [26] Yang Zhang, Li Duan, and Jun Liang Chen. 2014. Event-Driven SOA for IoT Services. In Proceedings of the 2014 IEEE International Conference on Services Computing (SCC '14). IEEE Computer Society, Washington, DC, USA, 629-636. DOI=<http://dx.doi.org/10.1109/SCC.2014.88>
- [27] M. Hammoudeh, J. Shuttleworth, R. Newman and S. Mount, "Experimental Applications of Hierarchical Mapping Services in Wireless Sensor Networks," Sensor Technologies and Applications, 2008. SENSORCOMM '08. Second International Conference on, Cap Esterel, 2008, pp. 36-43.

Fingerprint Gender Classification using Univariate Decision Tree (J48)

S. F. Abdullah

Optimization, Modelling, Analysis, Simulation and
Scheduling (OptiMASS) Research Group
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia

Z.A. Abas

Optimization, Modelling, Analysis, Simulation and
Scheduling (OptiMASS) Research Group
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia

A.F.N.A. Rahman

Optimization, Modelling, Analysis, Simulation and
Scheduling (OptiMASS) Research Group
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal, Melaka, Malaysia

W.H.M. Saad

Faculty of Electronic and Computer Engineering
Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal,
Melaka, Malaysia

Abstract—Data mining is the process of analyzing data from a different category. This data provide information and data mining will extracts a new knowledge from it and a new useful information is created. Decision tree learning is a method commonly used in data mining. The decision tree is a model of decision that looklike as a tree-like graph with nodes, branches and leaves. Each internal node denotes a test on an attribute and each branch represents the outcome of the test. The leaf node which is the last node will holds a class label. Decision tree classifies the instance and helps in making a prediction of the data used. This study focused on a J48 algorithm for classifying a gender by using fingerprint features. There are four types of features in the fingerprint that is used in this study, which is Ridge Count (RC), Ridge Density (RD), Ridge Thickness to Valley Thickness Ratio (RTVTR) and White Lines Count (WLC). Different cases have been determined to be executed with the J48 algorithm and a comparison of the knowledge gain from each test is shown. All the result of this experiment is running using Weka and the result achieve 96.28% for the classification rate.

Keywords—*fingerprint; gender classification; global features; Univariate Decision Tree; J48*

I. INTRODUCTION

A decision tree is a graph that uses a branching method to illustrate every possible outcome of the decision. A decision tree consists decision nodes and leaf nodes, where the decision node specifies a test over one attribute and a leaf node represent the class value [1]. A decision tree is a most powerful approach in knowledge discovery and data mining [2]. It is a non-parametric supervised learning method which is used to learn a classification function. It creates a model that predicts the value of the target variables by learning a simple decision rule from the data features.

Decision tree always be used with a complex bulk of data to enable a knowledge extraction in order to discover a useful pattern [2]. There are two approaches for decision tree [3] which is a univariate decision tree and multivariate decision

tree. The univariate decision tree is a decision node which considers only one feature that leads to the axis splits while the multivariate decision tree is a decision nodes that divide the input space into two widths an arbitrary hyperplane and leading to an oblique splits [4]. A J48 algorithm is an extension of an ID3 algorithm which is also from the univariate decision trees. For this study, the J48 algorithm has been used a proposed technique as it has more accuracy rate [5] compared to the available univariate decision tree.

Since 2006 until now, researchers keep finding the best classifier for gender classification problem. But until today there is no implementation of decision tree in gender classification based on the fingerprint. Badawi *et al.* [6] used three different types of classifier which are Neural Network (NN), Fuzzy C-Means (FCM) and Linear Discriminant Analysis (LDA) as a classifier for gender classification using the fingerprint. From his study, all three classifiers achieved above 80% of classification rate and the best classifier are NN with 88.5% of classification rate.

Verma *et al.* [7] used Support Vector Machine (SVM) as a classifier for fingerprint-based gender classification problem. SVM is used to separate the two classes of gender, which is male and female. From the study, SVM is able to get 88.00% of classification rate.

In the year of 2011, Arun *et al.* [8] used SVM to classify gender and they achieved 96.00% of classification rate using Radial Basis Function (RBF) kernel SVM. Early 2012, Ganasivam *et al.* [9] applied k-Nearest Neighbors (kNN) on the same problem and they achieved 88.28% of classification rate at k=1.

In the year of 2014, there are some researchers studies on gender classification problems to enhance and improve fingerprint-based gender classification problem. Gupta *et al.* [10] used the back propagation neural network as classifier to classify the gender and they achieved 92.67% of the classification rate. Agrawal *et al.* [11] used multi-SVM as a classifier to classify gender based fingerprint and they achieved

81.00% of classification rate which is lower than Verma *et al.* [7] and Arun *et al.* [8] even though they are applied the same classifier for the same problem.

Abdullah *et al.* [12][13] used several popular classifier for classification such as Multilayer Perceptron Neural Network (MLPNN), Support Vector Machine (SVM), Bayes Net and k-Nearest Neighbor (kNN) in classifying gender using the fingerprint features. They achieved above 95% of overall classification rate using 10-fold cross validation test. But in the study, there is a problem with MLPNN and kNN which is the popular overfitting problem. In order to overcome this problem, the number of features needs to reduce or needs to do the feature selection process before the classification part.

All the literature studies is shown in Table 1 below. From that, we can conclude that until now there is still a problem in the gender classification problem especially in terms of the accuracy rate. Thus, this study aims to see the performance of the J48 algorithm on fingerprint-based gender classification where J48 is commonly used in classification problem for the univariate decision trees. The performance of the J48 is compared with three different test cases, whereby each test case has a different number of fingerprint features selected.

TABLE I. PREVIOUS STUDIES ON FINGERPRINT BASED GENDER CLASSIFICATION

	Classifier	Accuracy
Badawi <i>et al.</i> [6]	Neural Network (NN)	80.39%
	Fuzzy C-Means (FCM)	86.50%
	Linear Discriminant Analysis (LDA)	88.50%
Verma <i>et al.</i> [7]	Support Vector Machine (SVM)	88.00%
Arun <i>et al.</i> [8]	Support Vector Machine (SVM)	96.00%
Gnanasivam <i>et al.</i> [9]	k-Nearest Neighbor (kNN)	88.28%
Gupta <i>et al.</i> [10]	Back Propagation Neural Network	91.45%
Agrawal <i>et al.</i> [11]	Support Vector Machine (SVM)	81.00%
Abdullah <i>et al.</i> [12]	Multilayer Perceptron Neural Network (MLPNN)	97.25%
Abdullah <i>et al.</i> [13]	Support Vector Machine (SVM)	96.62%
	Bayes Net	96.28%
	k-Nearest Neighbor (kNN)	95.27%
	Multilayer Perceptron Neural Network (MLPNN)	95.95%

The paper is organized as follows. Section II presents the methodology that has been done in this study, while the result analysis and discussion in Section III. Lastly, Section IV present the conclusion and future work.

II. METHODOLOGY

The sample of this study consist of four extracted features of 296 respondent which is Ridge Count (RC), Ridge Density (RD), Ridge Thickness to Valley Thickness Ratio (RTVTR) and White Lines Count (WLC). The database of the extracted fingerprint features are obtained from Abdullah *et al.* [14]. The process of classification is done using Weka programme with a 10-fold cross validation test. All features are arrange as shown

in Figure 1 and save as a Comma Delimited (CSV) file format.

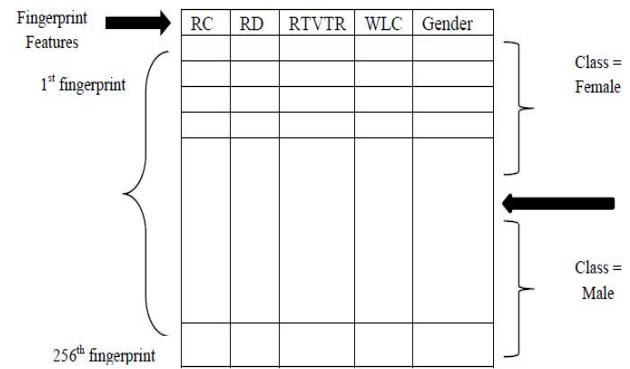


Fig. 1. The extracted features arrange in the database format

The four extracted features are save into four different files. The first file contain two types of fingerprint features which are Ridge Density (RD) and Ridge Thickness to Valley Thickness Ratio (RTVTR), the second file contains of three types of fingerprint features, which are Ridge Density (RD), Ridge Thickness to Valley Thickness Ratio (RTVTR) and White Lines Count (WLC). The third files contains of three types of fingerprint features, which are Ridge Density (RD), Ridge Thickness to Valley Thickness Ratio (RTVTR) and Ridge Count (RC) and the last file contain of all the features which are Ridge Count (RC), Ridge Density (RD), Ridge Thickness to Valley Thickness Ratio (RTVTR) and White Line Count (WLC). All these files are used to evaluate the performance of J48 algorithm in term of number of features involved in a test as shown in Figure 2. The result of this study is shown in a form of accuracy and decision tree.

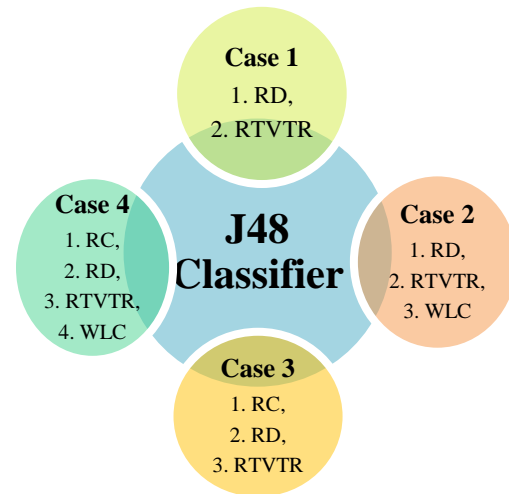


Fig. 2. Different number of features used in J48 Classifier Test Case

III. RESULT AND DISCUSSION

The result of each test case is given in Table II and the result is illustrated in a bar chart as shown in Figure 3. It can be seen that Test Case 3 gives a higher classification rate, which is

96.28% compared to Test Case 1, Test Case 2 and Test Case 3. The accuracy of Test Case 2 is 94.96%, which are the lowest classification rate for these 4 test cases. Each test case gives slightly different results in accuracy.

TABLE II. ACCURACY OF DIFFERENT TEST CASE

	Features Used	Accuracy
Case 1	RD & RTVTR	95.61%
Case 2	RD, RTVTR & WLC	94.96%
Case 3	RD, RTVTR & RC	95.61%
Case 4	RC, RD, RTVTR & WLC	96.28%

The accuracy of each case shown that there is slightly different of accuracy for each test case. As the higher number of features involved in a test case, the higher accuracy we get. But, there is a problem of Test Case 2, where 3 features involved in this test case give lower accuracy compared to the Test Case 1 which only involved two features. This is due to the additional features in Test Case 2, where White Lines Count (WLC) gives an impact to the classification rate. From this result, we can say that WLC are not reliable or suitable to be a feature for classifying gender of a person and this is proved by seeing the accuracy of the Test Case 3 which is Test Case 3 also involved 3 features which is Ridge Density (RD), Ridge Thickness to Valley Thickness Ratio (RTVTR) and Ridge Count (RC) gives a better accuracy compared to the Test Case 2. The other features like RD, RTVTR and RC is a good feature of this problem and this is supported by the T-test of each feature. t-Test is used to examine whether the fingerprint features of two classes which is male and female is statistically differ.

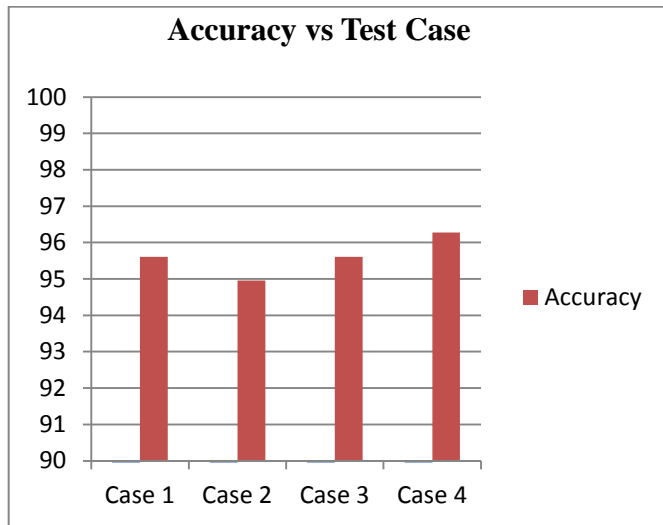


Fig. 3. Accuracy of different test case

Table III shows the result of the t-Test of the means of the four features which are RD, RTVTR, RC and WLC. It is shown that the female had a statistically significantly higher number of RD ($0.654 \pm 0.002 \text{ mm}^2$), RTVTR (0.811 ± 0.034) and RC ($16.34 \pm 1.242 \text{ per } 25\text{mm}^2$) compared to a male which lower numbers of RD ($0.470 \pm 0.002 \text{ mm}^2$), RTVTR (0.537 ± 0.008) and RC ($11.71 \pm 1.346 \text{ per } 25\text{mm}^2$). As we can see from Table III, the value of the variance of female for the White

Lines Count (WLC) is higher than the variance for male. We decided that the WLC feature is not to be include as a reliable feature for the gender classification in this work.

Table IV shows the number of respondents in term of correct classification, misclassification and the confusion matrix. For the Test Case 1, it is shown that 283 of 296 respondents are correctly classified as a male and as a female while another 13 of that are incorrectly classified. While for Test Case 2, it is shown that 285 respondents are correctly classified as a male and as a female. For Test case 3, 281 of 296 respondents are correctly classified as a male and as a female, while another 15 respondents are incorrectly classified as male and female. As we can see from the confusion matrix of test case 3, from 15 respondents who are incorrectly classified, nine of them are actually a female and six of them are males.

TABLE III. T-TEST OF THE MEANS OF THE FOUR FEATURES

Feature	Female		Male	
	Mean	Variance	Mean	Variance
Ridge Density (RD)	0.654	0.002	0.470	0.002
	P(T<=t) one-tail		8.3864E-102	
	t Critical one-tail		1.650255746	
	P(T<=t) two-tail		1.6773E-101	
	t Critical two-tail		1.968381923	
Ridge Thickness to Valley Thickness Ratio (RTVTR)	0.811	0.034	0.537	0.008
	P(T<=t) one-tail		3.64613E-41	
	t Critical one-tail		1.651564228	
	P(T<=t) two-tail		7.29225E-41	
	t Critical two-tail		1.970423195	
Ridge Count (RC)	16.34	1.242	11.71	1.346
	P(T<=t) one-tail		5.496E-106	
	t Critical one-tail		1.650161656	
	P(T<=t) two-tail		1.0992E-105	
	t Critical two-tail		1.968235174	
White Lines Count (WLC)	17.38	5.099	11.18	2.138
	P(T<=t) one-tail		7.36677E-83	
	t Critical one-tail		1.650559157	
	P(T<=t) two-tail		1.47335E-82	
	t Critical two-tail		1.968855173	

TABLE IV. NUMBER OF CORRECT CLASSIFICATION, MIS-CLASSIFICATION AND CONFUSION MATRIX FOR EACH DIFFERENT TEST CASE

	Correct Classification	Mis-Classification	Confusion Matrix									
Case 1	283	13	<table border="1"> <tr> <td></td> <td>M</td> <td>F</td> </tr> <tr> <td>M</td> <td>150</td> <td>6</td> </tr> <tr> <td>F</td> <td>7</td> <td>133</td> </tr> </table>		M	F	M	150	6	F	7	133
	M	F										
M	150	6										
F	7	133										
Case 2	285	11	<table border="1"> <tr> <td></td> <td>M</td> <td>F</td> </tr> <tr> <td>M</td> <td>151</td> <td>5</td> </tr> <tr> <td>F</td> <td>6</td> <td>134</td> </tr> </table>		M	F	M	151	5	F	6	134
	M	F										
M	151	5										
F	6	134										
Case 3	283	13	<table border="1"> <tr> <td></td> <td>M</td> <td>F</td> </tr> <tr> <td>M</td> <td>150</td> <td>6</td> </tr> <tr> <td>F</td> <td>7</td> <td>133</td> </tr> </table>		M	F	M	150	6	F	7	133
	M	F										
M	150	6										
F	7	133										
Case 4	281	15	<table border="1"> <tr> <td></td> <td>M</td> <td>F</td> </tr> <tr> <td>M</td> <td>147</td> <td>9</td> </tr> <tr> <td>F</td> <td>6</td> <td>134</td> </tr> </table>		M	F	M	147	9	F	6	134
	M	F										
M	147	9										
F	6	134										

Figure 4, Figure 5, Figure 6 and Figure 7 shows the tree visualization of the univariate decision tree generated by WEKA for each test cases. For Test Case 1, the total number of leaves in the generated tree is four and the size of the tree is seven.

For Test Case 2, the total number of leaves in the generated tree are nine and the size of the tree is 17. For Test Case 3, the total number of leaves in the generated tree are four and the size of the tree is 7 while for Test Case 4, the total number of leaves in the generated tree are seven and the size of the tree is 13. The number of leaves in the generated tree and size of tree from Test Case 2 is higher than Test Case 3, Test 4 and Test Case 1.

The decision tree has many benefits to classification process, where it can handle a variety input data which is nominal, numeric and text data. Decision tree also is able to handle effectively the missing values and had a high performance when tested with a small number of efforts. On the other hand, the J48 algorithm is still having a problem where it will slows down the process of generate a tree especially if the data sets used are large and not clean.

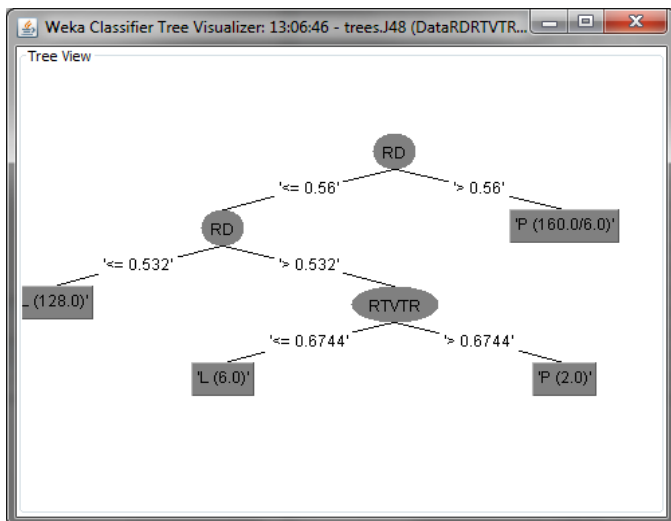


Fig. 4. Decision tree for Test Case 1

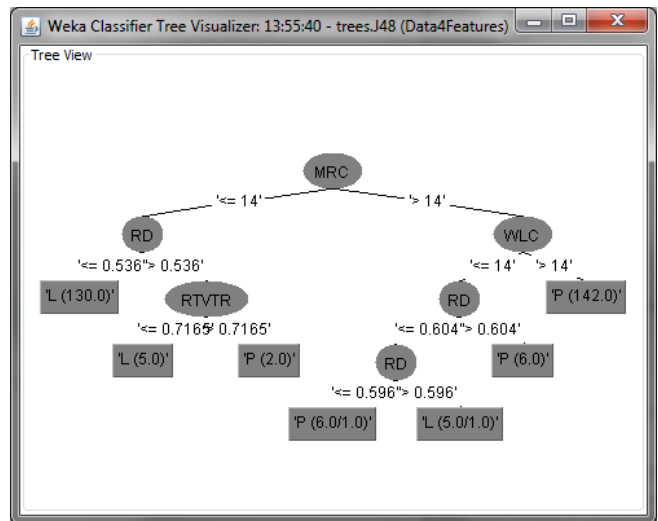


Fig. 5. Decision tree for Test Case 4

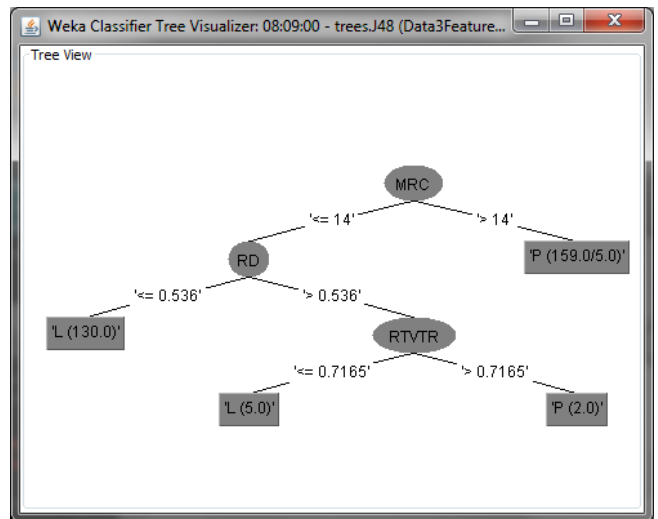


Fig. 6. Decision tree for Test Case 3

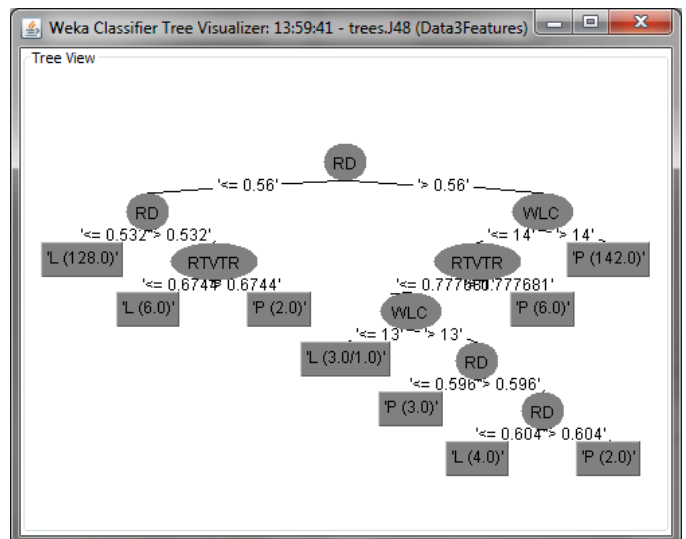


Fig. 7. Decision tree for Test Case 2

IV. CONCLUSION

In conclusion, this paper used J48 decision tree in fingerprint gender classification problem, and the accuracy of the approach is approximately 96.28% for the four fingerprint features used. Using WEKA tool, the decision tree is generated and we got the higher correctly classified male and female which is 285 from 296 respondents. In this study, we can state that we can use J48 decision tree as a classifier for fingerprint based gender classification.

ACKNOWLEDGMENT

We would like to thank Optimization, Modelling, Analysis, Simulation and Scheduling (OptiMASS) Research Lab for the supporting given throughout the research activities and the Centre of Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM) for the financial support.

REFERENCES

- [1] Sahu, Shailendra, and B. M. Mehtre. "Network intrusion detection system using j48 decision tree." In *Advances in Computing, Communications and Informatics (ICACCI)*, 2015 International Conference on, pp. 2023-2026. IEEE, 2015.
- [2] Bhargava, Neeraj, Girja Sharma, Ritu Bhargava, and Manish Mathuria. "Decision tree analysis on j48 algorithm for data mining." *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering* 3, no. 6 (2013).
- [3] Korting, Thales Sehn. "C4. 5 algorithm and multivariate decision trees." *Image Processing Division, National Institute for Space Research-INPE Sao Jose dos Campos-SP, Brazil* (2006).
- [4] Yıldız, Olcay Taner, and Ethem Alpaydın. "Comparing Univariate and Multivariate Decision Trees."
- [5] Jain, Yogendra Kumar. "Upendra: An efficient intrusion detection based on decision tree classifier using feature reduction." *International Journal of Scientific and Research Publication* 2, no. 1 (2012).
- [6] Badawi, Ahmed M., Mohamed Mahfouz, Rimon Tadross, and Richard Jantz. "Fingerprint-Based Gender Classification." In *IPCV*, pp. 41-46. 2006.
- [7] Verma, Manish, and Suneeta Agarwal. "Fingerprint based male-female classification." In *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08*, pp. 251-257. Springer Berlin Heidelberg, 2009.
- [8] Arun, K. S., and K. S. Sarath. "A machine learning approach for fingerprint based gender identification." In *Recent Advances in Intelligent Computational Systems (RAICS)*, 2011 IEEE, pp. 163-167. IEEE, 2011.
- [9] Gnanasivam, P., and Dr S. Muttan. "Estimation of age through fingerprints using wavelet transform and singular value decomposition." *International Journal of Biometrics and Bioinformatics (IJBB)* 6, no. 2 (2012): 58-67.
- [10] Gupta, Samta, and A. Prabhakar Rao. "Fingerprint based gender classification using discrete wavelet transform & artificial neural network." *International Journal of Computer Science and mobile computing* 3, no. 4 (2014): 1289-1296.
- [11] Agrawal, Heena, and Siddhartha Choubey. "Fingerprint Based Gender Classification using multi-class SVM."
- [12] Abdullah, S. F., A. F. N. A. Rahman, Z. A. Abas, and W. H. M. Saad. "Multilayer Perceptron Neural Network in Classifying Gender using Fingerprint Global Level Features." *Indian Journal of Science and Technology* 9, no. 9 (2016).
- [13] Abdullah, S. F., A. F. N. A. Rahman, Z. A. Abas, and W. H. M. Saad. "Support Vector Machine, Multilayer Perceptron Neural Network, Bayes Net and k-Nearest Neighbor in Classifying Gender using Fingerprint Features" *International Journal of Computer Science and Information Security (IJCSIS)*, 14 (7), 2016.
- [14] S. F. Abdullah, A.F.N.A. Rahman, Z.A. Abas and W.H.M. Saad, "Development of a Fingerprint Gender Classification Algorithm Using Fingerprint Global Features" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 7(6), 2016

Enhancing Wireless Sensor Network Security using Artificial Neural Network based Trust Model

Dr. Adwan Yasin¹

Dept. Engineering and Information Technology
Arab American University
Jenin, Palestine

Kefaya Sabaneh²

Dept. Engineering and Information Technology
Arab American University
Jenin, Palestine

Abstract—Wireless sensor network (WSN) is widely used in environmental conditions where the systems depend on sensing and monitoring approach. Water pollution monitoring system depends on a network of wireless sensing nodes which communicate together depending on a specific topological order. The nodes distributed in a harsh environment to detect the polluted zones within the WSN range based on the sensed data. WSN exposes several malicious attacks as a consequence of its presence in such open environment, so additional techniques are needed alongside with the existing cryptography approach. In this paper an enhanced trust model based on the use of radial base artificial neural network (RBANN) is presented to predict the future behavior of each node based on its weighted direct and indirect behaviors, in order to provide a comprehensive trust model that helps to detect and eliminate malicious nodes within the WSN. The proposed model considered the limited power, storage and processing capabilities of the system.

Keywords—Wireless sensor network; security; Artificial neural network; trust rate; malicious node; trust model; threat

I. INTRODUCTION

Wireless sensor network is a distributed system that contains a collection of autonomous spatially distributed nodes cooperating together to produce a globally useful information from its local raw sensed data. WSN introduced significant advantages upon traditional communication technologies in many fields such as healthcare applications which could be wearable or even implemented in the patient body, transportation, military operations and environmental conditions monitoring like fire and natural disaster detection.

WSN node contains a set of components; one sensor or more to sense the environmental conditions as the concentration of a specific chemical element in our case, a small processing unit, storage and a power supply (battery). Each of these components have to be used in a rational way since the resources of the node are limited and it is difficult or even impossible to feed or replace it [1] [2].

Wireless sensor network may contain tens, hundreds or even thousands of autonomous nodes equipped with sensors. It is essential to choose a suitable network topology that enables the communication between several nodes, and the transmission of sensed data. The main driving factor in selecting which topology should be used is the limited power supply within WSN nodes, and also the need to reduce the price as much as possible because hundreds of nodes should be connected and interact. It is possible to use the bus topology,

ring, star, mesh, tree or hybrid topology. A hybrid topology that combines star, mesh and ring topologies together is proposed to provide a reliable, fault tolerant and power efficient communication and data transmission [3].

Using WSN in water pollution monitoring requires the existence of sensing nodes in a harsh and changeable environment, that makes it exposed to several security threats and dangers. A sensing node could be damaged due to environmental changes and conditions, also it may be a target for an opponent party and simply be replaced or even modified in such a way that facilitates a passive or active attack by the opponent. As a result preserving the system security is an important and essential issue to prevent any unauthorized access to its components. Several techniques are used such as symmetric encryption, it depends on the use of a single key in both sides sender node and receiver one, but such methodology is not enough and we need an additional strategy that distinguish a malicious node even if it somehow obtained the key. Building trust between the different nodes is the intended approach [4].

Building trust between system nodes requires the use of a trust model to provide trust ratings for WSN sensor nodes depending on their performance and sensed data, higher the trust ratings for a specific node higher its effectiveness, while lower the trust rate means higher probability to remove it from the system especially when it becomes lower than a specific threshold [5].

Several trust schemes had been used to discover the malicious nodes in WSN, all of them are based on two essential characteristics in terms of WSN resources limitations; in one hand they are lightweight and need less power, processing and communications, in the other hand they are powerful and capable of managing trust between various heterogeneous nodes [6]. The proposed trust model aims to enhance the use of artificial neural network (ANN) in WSN using radial basis function benefiting from its simplicity in implementation to provide a comprehensive trust model that supports system security and rationalizes resources consumption.

The rest of the paper is ordered as the following; in section II a literature review that lists the state-of-the-art security approaches and trust models, while section III proposes a comprehensive architecture for the WSN system including system components and the network topology. In section IV securing the WSN is discussed, while section V focuses on

how to build trust between several nodes in WSN using ANN, VI includes a proposed enhanced trust model. Finally section VII concludes with the future work.

II. LITERATURE REVIEW

The raised popularity of WSN had been facing several security threats and permutations. Developing corresponding countermeasure mechanisms suffered from challenges represented by the sensors size, processing power and memory limitations. Data within a water pollution monitoring system should be protected from any unauthorized party since water is the basic resource of life for all countries, so security mechanisms are essential to perceive confidentiality, availability and integrity of the WSN components including hardware devices, software, networking equipments and collected data [7].

Attacks in WSN are categorized to two main approaches; either attacks against the employed security mechanism or attacks against the routing mechanism. An attack that aims to hack the security mechanism by exploiting its weaknesses depends on the mechanism characteristics while the last depends on hacking the routing algorithms within the network [8] [12].

Denial of service attack (DoS) prevents the normal use of the communication facilities within the network by exhausting its resources with extra transmitted packets; it aims to flood the network with useless data and may eventually disrupt the whole network [5]. Sybil attack is another threat in which a node claims numerous identities so it behalf and interact as a set of legitimate nodes, data integrity and resource utilization degrades and as a result network protocols may be disrupted [11]. In black whole attack a malicious node attempts to track and attract the traffic in the network, once the opponent can access, communicate and participate in the network the entire readings could be affected especially in the hierarchal network topologies where data is transmitted passing through several nodes. Hello flood attack is incorporated by a foreign adversary who can flood hello request to any legitimate node in the network and break the security mechanism, while in wormhole attack the attacker record the packets and forward to another location, one or more fake nodes are used with a route between them, once the malicious node starts its work a fake route is used to provide a path that is shorter than the original one, and as a result the data is tunneled within the undesirable route [9].

Basically WSN defensive line depends on the use of conventional key approaches for intrusion detection and prevention, symmetric encryption techniques help to hide the content of transmitted packets through the network such that no malicious node can make use of encrypted data even if it get the ciphered packets, and this is distinguished by its reduced overhead compared with public encryption algorithms. The network is still facing the previous types of attacks but although the available resources within the network are limited additional techniques are needed beside cryptography to ensure the system security [10].

Several Approaches including steganography, physical layer secure access, data aggregation schemes, multilayer

approaches and trust building are used integrally to provide a comprehensive defense line for the WSN. While cryptography aims to hide the content of transmitted packets steganography mechanism is capable of hiding the existence of the transmitted packets entirely, confusing the adversary who expects to get readings transmitted through the network [7]. Physical layer secure access strengthens the system security through frequency hopping in WSN, mainly by transmitting signals and quickly switching a carrier between several frequency channels. The effectiveness of this technique is lies in the efficient design for hopping order which is modified in less time compared to required time for discovery [9].

Multilayer defense approaches aims to guarantee WSN security in layers of protocols stack, providing a strong combination of malicious behavior prevention by employing a set of mechanisms at various layers within the OSI protocol. The main drawback of such approach is that the attack could be detected by several mechanisms causing in redundant detection and high power consumption [10].

Modeling trust in WSN is widely used for the early detection of malicious nodes and its subsequent effect prevention. Sensor nodes need to ensure the trust of the next node within the routing path to forward data packets, also the node needs to trust other neighbors to check anomalous readings. Several schemes had been proposed for trust modeling in WSN.

A. Trust management for resilient geographic routing (TM-RGR)

An algorithm is used to prevent attacks on geographic routing of data. The idea here is to reward a good behaving node and giving it additional confidence and trust raise every time it forwards a data packet successfully while punishing illegal node that lie about its location. Honest node remains longer time in the set of packet forwarding. After establishing a routing table for a specific node; it monitors the behavior of its next neighbor using snooping technique. TM-RGR is very simple and updating trust value does not take a lot of time, but in the other hand the accuracy is modest and the opportunity of false positives and false negatives is raised [13].

B. Hybrid Trust and Reputation Management (HTRM)

This scheme combines both behavior based approaches and certificate based approaches to update a node trust, behavioral based approaches depends on both direct and indirect behavioral information collected by the surrounding nodes. Trust of a node is calculated after gathering enough number of evidences from a certificate authority or any other trusted neighbor; in case where negative evidences are obtained the certificate is revoked immediately. As a result of the combination between certificates, direct and indirect behavior more power consumption is required for evaluating node trust which is not available within a single node [12].

C. Group Based Trust Management Scheme (GBTMS)

In this model instead evaluating the trust for a single sensor node, a light weight algorithm is used to evaluate the trust for a group of nodes within the WSN. A cluster head is capable of evaluating trust for sensor nodes within its cluster, and other cluster heads depending on both direct and indirect behaviors.

Memory consumption in GBTMS is minimized since trust of a group of nodes is evaluated and information is stored at the cluster head, but the amount of resources needed are more since the trust is calculated based on the previous behaviors [13].

D. Weighted Trust Algorithm (WTA)

WTA is used to detect the malicious nodes by observing its reported data to associate a weight for each node in the network. All sensing node are initialized with the same weight value. The a node weight is updated every cycle if it sends a report that differs from the reports of other sensing nodes. If a sensor node sends its report inconsistent with the final decision which is based on the reported data from others nodes its weight would be decreased and if it became lower than a specific threshold, the corresponding node will be identified as a malicious one. Weights are updated dynamically, but there is a high opportunity for false positive and false negative probabilities [14].

E. Behavior Trust based on Geometric Mean Approach (BTGMA)

BTGMA is a distributed trust scheme in which trust management is spread over the whole WSN; every node within the network is responsible for evaluating its trust based on direct and indirect behavior. Direct behavior is obtained by calculating the geometric mean of the quality of service (QoS) characteristics for a specific node, such as amount of consumed power, transmitted data rate, and reliability. Larger number of QoS characteristics employed larger the amount of consumed energy, and this is not consistent with the WSN limitations [13] [18].

F. Lightweight and Dependable Trust management Scheme (LDTS)

A light weight scheme is used for trust management in clustered WSN. Evaluating trust for a node is calculated depending on indirect evidences where indirect behavior is obtained using the feedback reported by cluster head node. LDTS improves system efficiency because it works even in cases where direct behavior is not accessible or insufficient, but the main drawback is its complete dependency on the cluster head, any unexpected damage in the CH disrupts the whole approach [8] [13].

G. Swarm intelligence based method.

Swarm intelligence approach is used to find the most reputable path in the network; nodes within this path are considered as trustworthy nodes and in result obtain a higher trust evaluation. Ant colony is a good example for swarm intelligence, searcher ants leave a pheromone in the path of food resource, where higher concentration of the pheromone higher attraction for collector ants to that path. Finding the shortest path between the source and destination node is computed easily using an algorithm such as Dijkstra's algorithm [8].

H. Artificial neural network method.

Neural networks have the ability to mimic the human brain behavior using a set of crude and simple approximations of the human neurons which is used to learn and generalize from

training data. ANN consists of two main phases; training stage and generalization. In training phase weights of the input patterns are learned until a specific error is reached or after a specific number of iterations, so that the network learn the decision boundaries from the training patterns. Generalization phase includes using untrained inputs to find the output using the trained ANN, i.e. this phase is essential to classify untrained data correctly. ANN has many features that encourage its usage in WSN such as its parallelism, efficiency and noise tolerance which is very important in such harsh environment [15].

The existing ANN trust models based on calculating the trust rate of each node depending on its direct and indirect behavior which is obtained from the current and previous readings of its neighbors either in a cluster based network or any other topological WSN. The obtained reading which is called the expected reading is then compared with the actual one and trust rate is updated according to the convergence between both. The main drawback in the state-of-the-art ANN based trust models is the method for calculating node behavior. Nodes readings are treated equally without concerning about the spatial and temporal dimensions of these readings [16].

III. SYSTEM ARCHITECTURE

Providing a comprehensive WSN architecture that satisfies the robustness, availability and reduced power consumption is a challenging role. The current WSN architectures try to guarantee rationalized data delivery using star topology which reduces the consumed power by reducing the nodes within the transmission route. Several routing techniques are employed to minimize the wasted power by controlling the route for the transmitted data from sensing node to sink node. Concentrating on rationalizing the data transmission power consumption indicates that it consumes the largest portion of the available resources [17].

A. WSN components

WSN for water pollution monitoring system consists of several nodes equipped with sensors to get the concentration of water components in the surrounding environment, these nodes are collaborating together and linked via a wireless data link to the main node (sink node) that aggregates the collected readings from individuals, preprocesses and sends it to a main station to be processed and stored as shown in fig.1

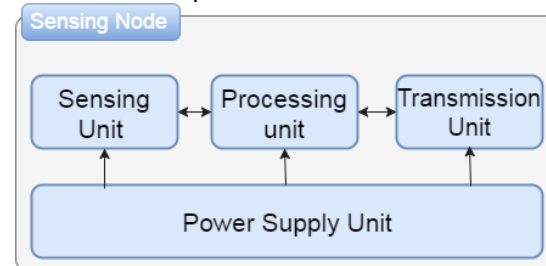


Fig. 1. WSN node components

microprocessor is the brain in each node, it is responsible for the basic processing of collected data via the sensing units, controlling the work of other node's components and managing the overall power consumption since each node has a

limited power resource, the main power consumption refers to the transmission of data which is the transceiver responsibility [17].

Water as a vital resource of life is polluted due to either chemical or natural changes. Natural water pollution includes the alteration of water natural properties due to increased salinity and temperature variation; so we need a sensor to measure water temperature and salt concentration. In the other hand chemical pollution implies water exposure to dangerous chemicals such as petroleum, arsenic and insecticides, these components change the properties of water and can affect the water potability. Sensing unit contains sensors to detect dangerous chemicals alongside with sensors that sense natural water pollution resources.

B. Networking topology

In WSN we proposed a cluster based WSN architecture that uses a hybrid network topology which combines star, mesh and ring topologies together. The three topologies are used to enhance the reliability of the system while taking in to account the available resources as following:

- a collection of the sensing nodes are connected to a central node cluster head(CH) that has an additional capabilities and resources among other nodes using star topology, as a result a set of clusters are obtained. Star topology has many advantages compares with other topologies including its scalability and power usage reduction; if a cluster head fail other nodes within the same cluster can reconfigure the topology and elect a new CH.
- Clusters heads are connected together as a mesh to provide a reliable communications, clusters heads had additional resources that facilitate the use of such topology. To access a specific cluster head the sink node needs an algorithm to detect the shortest path to that head to reduce consumed power as much as possible. All clusters heads are connected directly to the sink node which acts as a gateway to the outside world.
- Data collected through sensors in each node could be aggregated and sent to the base station by passing through the cluster head and sink node, if a node fail then the network will reconfigure itself around the other nodes, even if the radio link from a sensing node to its cluster head gone down due to interference for example then the access to that head would be done through an alternative ring connection, the ring path is chosen depending on the shortest path from the cluster head passing by the sensing node and returning to the head again. Fig.2 clarifies the overall architecture of WSN using hybrid network topology.

Cluster head aggregates the corresponding sensing nodes readings to be delivered to the base station via the sink node. Data delivery model decides when the collected data should be sent to the sink node, there are four main models for data delivery: continuous delivery, query driven, event driven and hybrid delivery models. Continuous delivery model sends the data periodically to the sink node while query driven models depend on the sink query for the data, event driven model

depends on sending data when an event occurs. The proposed WSN system uses a hybrid data delivery model that enhances effective monitoring and enables the responsible party to take decisions accurately.

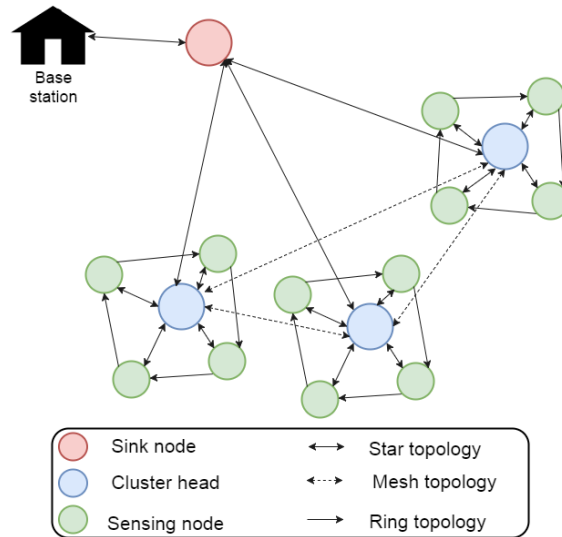


Fig. 2. System architecture

IV. SECURITY IN WIRELESS SENSOR NETWORK

Building Underground Water pollution monitoring system is a challenging task since several environmental conditions and changes may affect the WSN and cause in suspicious measurements as a result. For example a node could be disrupted due to environmental conditions or battery leakage, and that affects the availability of the system which is one of the main security goals for the WSN. A malicious node may enter the network and provide wrong data, change data or even transmit data about the validity and quality of the monitored water supply to another party; and this threatens the confidentiality and integrity of our system.

Generally, as any wireless sensor network built in a harsh environment the proposed system is exposed to security attacks such as DOS, wormhole, Sybil attack, hello flood and node-capturing attacks. The use of cryptographic algorithms alone cannot encounter and cope with the various types of attacks, so building trust between several nodes within the network is important to distinguish legitimate nodes from malicious ones [6].

V. BEHAVIOR PREDICTION IN WSN USING RBANN

Due to scalability, expandability and openness features of the WSN as a distributed system, additional new nodes can enter the system at different times; this exposes the network to several types of attacks and requires a strategy to distinguish legitimate nodes from foreign ones. Building trust is essential to assure the legitimacy of several nodes and protect the system so that no harmful node can masquerade or pretend to be a good one [9].

Different trust models have been proposed to provide an ultimate mechanism for detecting malicious nodes within WSN. All of the proposed trust models based on calculating

trust rate for every node within the system depending on its behavior which is captured by either direct or indirect fashion, depending on these rates the controller or director decide to consider a node as legitimate node and in result raise its trust rate or it could be a layer and punished by reducing its trust rate, if the rate decreased below a specific threshold the node may be discarded or monitored to be treated later [6].

ANN is widely used in real world approaches that build efficient systems to solve real world problems such as time series prediction. Employing ANN in Time series prediction applications is done by transferring the problem into a simple function that maps inputs to output using activations [16]. As shown in fig.3 predicting the reading of a sensor within the WSN is done depending on the previous n readings of the sensor as an input to the neural network, to accomplish the prediction based on both direct and indirect behaviors the network is expanded to include additional neighbors readings history.

Radial Base Artificial Neural Network uses three layers feed forward neural network; input layer, output layer and one hidden layer [19]. To predict the behavior of a node within the network the network works as following:

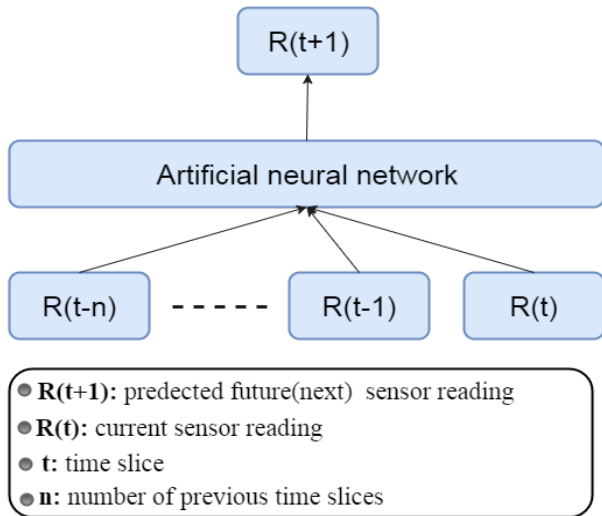


Fig. 3. Time Series Prediction using artificial neural network

- Input layer contains a set of readings (patterns) as input for nodes used in the direct behavioral prediction, but for indirect behavioral prediction the previous readings of other surrounding nodes are considered as a part of the input patterns.
- Single hidden layer that contains a set of radial bases functions (Gaussian functions) as an activation function to train the network from the available labeled readings. The training of ANN is essential to enable its ability to predict future readings of unlabeled patterns, the process can be described as preparing the network to be equipped for data it didn't see before and provide right predictions.
- Output layer implements a linear weighted sum function that calculates the predicted reading for a specific node.

Fig.4 shows how a RBFNN is used to predict the future reading of a sensing node based on its direct and indirect behavior. To find the activation function $\Phi(x)$ equation 1 is used [15]. The input to hidden basis function parameters $\{\mu, \sigma\}$ can be set using any number of unsupervised learning techniques.

$$\phi_j(x) = \exp\left(-\frac{\|x - \mu_j\|^2}{2\sigma_j^2}\right) \quad (1)$$

Where j is a hidden neuron.

The output at output layer is calculated using both $\Phi(x)$ and weights from hidden to output layer as in equation 2.

$$y_k(x) = \sum_{j=0}^M w_{kj} \phi_j(x) \quad (2)$$

Where x- is the input value, j- is the hidden neuron; k is the output neuron, W_{kj} the weight from neuron k in hidden layer to neuron j in output layer.

Weights are trained and updated so that the sum-square output error is minimized; when the error is minimized to a specific threshold training is stopped and the network becomes ready for the generalization phase which is essential to assure the ability of the network to generalize from trained data and predict a correct reading for untrained samples, as a result the RBANN is capable of providing an expected reading for every node in the WSN for water pollution monitoring system. [15].

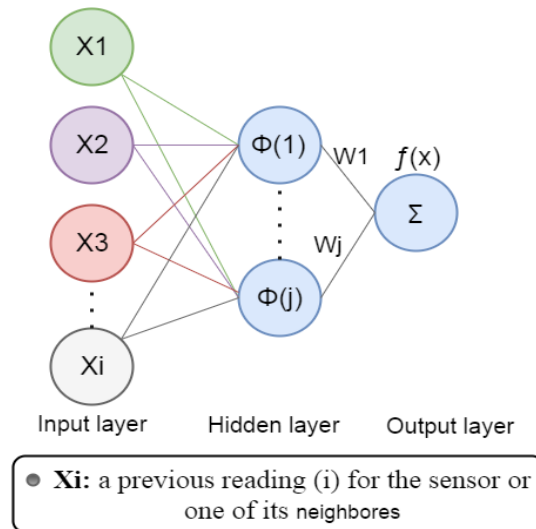


Fig. 4. Radial base neural network

VI. PROPOSED TRUST MODEL

The proposed approach is to enhance the use of neural network for malicious nodes detection in WSN taking in to account the power supply and computational restrictions in such systems. A RBANN is used to find the expected reading of a node using a set of direct and indirect behaviors, the output is compared with the real reading of that node and the trust rate is then determined according to the convergence or divergence between both readings. Two main sequential stages are included in the proposed trust model:

A. Clusters heads trust rating.

To assure that a cluster head is not a malicious node and prevent it from distorting and changing sensed readings captured by sensing nodes within the same cluster, RBANN is used to calculate the expected behavior for that cluster head. Inputs of the ANN training are the n previous readings of the intended cluster head, current reading and previous n readings of other surrounding cluster heads, other clusters heads are given weight that determine the ratio with which each one affect the intended cluster head depending on the distance between both; shorter the distance larger the contribution in determining the expected behavior.

B. Sensing nodes trust rating.

RBANN is used to calculate the expected output and reading of each sensing node within each cluster using the previous and current readings of the remaining sensing nodes in the same cluster in addition to the previous n readings of the intended node. Each sensing node within the cluster affect the intended one with a specific ratio depending on the distance between both, also previous n readings for a node varies in its contribution based on the temporal differences between them, older the reading less its contribution in calculating a specific node expected reading.

Cluster heads passes the actual readings of all nodes to base station through the sink node. ANN resides in the base station where we have unlimited power and computational capabilities, that helps to avoid resources limitations in WSN. The expected readings obtained from the ANN is compared with the actual ones in database, if both are convergent trust rate is raised, otherwise trust rate is minimized. Generally trust rating for each sensing node is maintained according to the comparison between both actual and expected behavior.

As shown in fig.5 the proposed algorithm is constituted by the following steps which are applied periodically every time a sink node asks for sensed data:

- 1) Basically when the WSN system is constructed all nodes are initialized with equally trust rate such as 1.
- 2) The expected trust rate of each cluster head CHi is calculated based on the previous n readings of that head, current reading and previous n readings of all other cluster heads selected depending on the distance between each one of them and the CHi. Each cluster head affects the trust rate of CHi with R ratio depending on the distance between both and even previous n readings of a specific head have different proportions in calculating the cluster head expected reading depending on its time, older reading has less effect.
- 3) RBANN is calculates the expected output, if it converges to the current actual reading for that head then it is considered as trusted head and trust is raised; otherwise trust is minimized with respect to the variance between actual and expected reading as shown in fig.6.

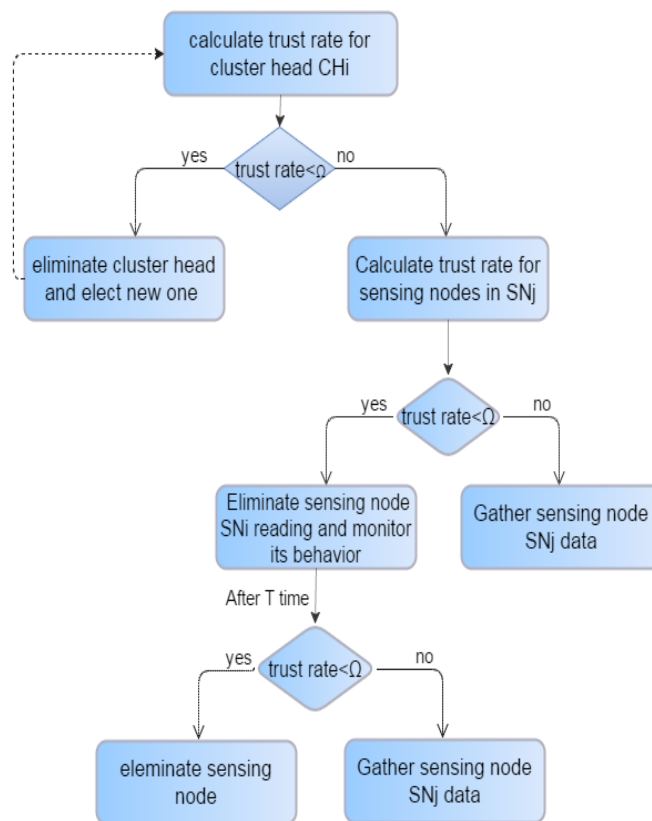


Fig. 5. Proposed algorithm for trust rate calculation

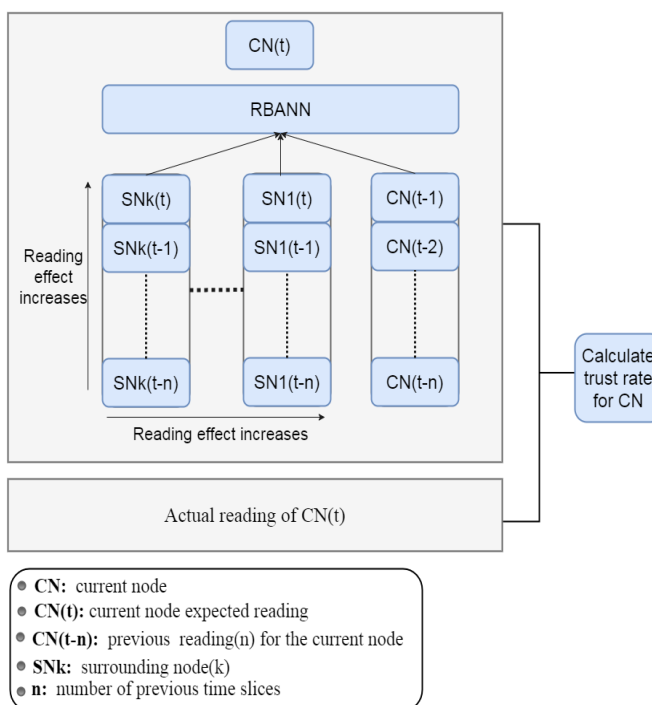


Fig. 6. WSN node trust rate calculation

4) If trust rate for a specific cluster head became less than a specific threshold Ω cluster head is eliminated and another new one from the same cluster is elected, the winner is the richer node which has abundant available resources compared to other cluster nodes.

5) After ensuring the trust of a cluster head, corresponding sensing nodes trust within the same cluster need to be checked. As in fig.6 trust rating for each sensing node is done by comparing its actual reading passed to the base station by a trusted cluster head and the corresponding expected one obtained using the ANN. The inputs to ANN are the n previous readings of the intended node, current and n previous readings for all sensing nodes within the same cluster. Each sensing node contributes in calculating the expected trust rate depending on its distance from the intended one, closer the distance larger the influence in calculations.

6) If the trust rate is raised, sensing node is considered a trusted one and its gathered data is considered as a valuable data, otherwise trust rate degrades and after it becomes smaller than Ω the node is considered a suspicious one and its reading is discarded for t of time while its behavior under monitoring, if the node behavior didn't improve then it is considered a malicious one and its collected data discarded, otherwise trust rating improves and became more than Ω so the node is reconsidered trustful and its readings are taken in to account.

Since cluster head has larger influence, permissions and effect on the system as a whole in more comprehensive fashion, a suspicious cluster head will be eliminated directly without any other considerations because it is capable of influencing and harming the entire system by either changing the collected data sent from sensing nodes within the same cluster or even by providing a misleading trust rate evaluation of other cluster heads.

Employing the proposed algorithm provides the administrator at the base station with a mean to track the trust of all network nodes and eliminate malicious or damaged ones. The implementation of the algorithm in two sequential hierarchical stages helps to reduce the consumed power for strangers' detection and also reduces the wasted efforts in case where a cluster head masquerades or forges nodes within the same cluster, exactly as any hierarchical arrangement the higher branch always affects all other sub-branches so credibility of a cluster head is checked to trust what it passes to the base station.

VII. CONCLUSION

WSN used for Water pollution monitoring system requires a powerful defense line against threats and changes in the surrounding, taking in to account resources challenges in such system. Enhanced ANN based trust model is proposed to overcome the weaknesses in the existing WSN trust models. By using a modified radial base ANN we improve the way in which ANN is used to predict the expected readings of network nodes, and calculate their trust rate based on spatial and temporal weighting of both sensing nodes, and clusters heads. In addition we adapted a dependable, fault tolerant hybrid architecture that combines mesh, star and ring topologies in

order to provide a comprehensive robust WSN that consumes less resources compared with the existing WSN. In the future work we aim to improve RBANN accuracy using additional inputs to represent the direct behavior for the node being evaluated depending on the geometric mean of its quality of service (QoS) characteristics.

REFERENCES

- [1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless sensor networks: a survey," Computer Networks, December 2011.
- [2] N. Rodriguez, S. Rossetto, "Distributed systems with wireless sensor networks," 2012.
- [3] Q. Mamun, "A qualitative comparison of different topologies for wireless sensor networks," 2012.
- [4] A. Devasena, B. Sowmya, "Wireless sensor network in disaster management," Indian journal of science and technology, July 2015.
- [5] A. K. Pathan, H. W. Lee, C. S. Hong, "Security in wireless sensor networks: Issues and challenges," ICACT, 2006.
- [6] M. Momani, S. Challa, "Survey of trust models in different network domains," 2010.
- [7] W. Stallings, "Cryptography and Network Security," sixth edition, 2013.
- [8] V. U. Rani, K. S. Sundaram, "Review of Trust Models in Wireless Sensor Networks," International scholarly and scientific research & innovation, 2014.
- [9] H. Rathore, H. Badarla, S. Jha, A. Gupta, "Novel approach for security in wireless sensor network using bio-inspirations," IEEE, 2014.
- [10] G. Kulkarni, R. Shelk, K. Gaikwad, V. Solanke, S. Gujar, P. Khatawkar, "Wireless sensor network security threats," IET, July 2015.
- [11] E. P. k Gilbert, B. Kaliaperumal, and E. B. Rajsingh, "Research issues in wireless sensor network applications: A survey," International journal of information and electronics engineering, September 2012.
- [12] H. A. N Jitender, S. Deogun and E. D. Manly, "Secure and energy aware routing against wormholes and sinkholes in wireless sensor network", IEEE, 2006.
- [13] R. W. Anwar, M. Bakhtiari, A. Zainal and K. Naseer Qureshi, "A survey of wireless sensor networks and routing techniques," Research Journal of Applied Sciences, Engineering and Technology, 2015.
- [14] V. Reshmi, M. Sajitha, "A Survey on trust management in wireless sensor networks," International journal of computer science & engineering technology, February 2014.
- [15] S. Haykin, "Neural networks and learning machines," third edition, 2009.
- [16] A. Doboli, "Discovery of malicious nodes in wireless sensor networks using neural predictors," WSEAS transactions on computer research, February 2007.
- [17] S. Sharma, D. Kumar and K. Kishore, "Wireless sensor networks- A review on topologies and node architecture," International journal of computer sciences and engineering open access, 2013.
- [18] F. Bao, I. R. Chen, M. Chang, and J. H. Cho, "Hierarchical trust management for wireless sensor networks and its applications to trust-based routing and intrusion detection," IEEE, JUNE 2012.
- [19] M. Awad, H. Pomares, I. Rojas, O. Salameh and M. Hamdon, "Prediction of time series using RBF neural networks: A new approach of clustering," The International Arab Journal of Information Technology, April 2009.

AUTHOR PROFILE



Adwan Yasin is an associate Professor, Former dean of Faculty of Engineering and Information Technology of the Arab American University of Jenin, Palestine. Previously he worked at Philadelphia and Zarka Private University, Jordan. He received his PhD degree from the National Technical University of Ukraine in 1996. His research interests include Computer Networks, Computer Architecture, Cryptography and Networks Security.

Kefaya Saba'neh is a Computer Science master student in the Arab American University of Jenin, she also received her bachelor degree in Multimedia Technology from the Arab American University of Jenin in 2010.

Security and Privacy Issues in Ehealthcare Systems: Towards Trusted Services

Isra'a Ahmed Zriqat
Computer Science Department
Applied Science Private University
Amman, Jordan

Ahmad Mousa Altamimi
Computer Science Department
Applied Science Private University
Amman, Jordan

Abstract—Recent years have witnessed a widespread availability of electronic healthcare data record (EHR) systems. Vast amounts of health data were generated in the process of treatment in medical centers such hospitals, clinics, or other institutions. To improve the quality of healthcare service, EHRs could be potentially shared by a variety of users. This results in significant privacy issues that should be addressed to make the use of EHR practical. In fact, despite the recent research in designing standards and regulations directives concerning security and privacy in EHR systems, it is still, however, not completely settled out the privacy challenges. In this paper, a systematic literature review was conducted concerning the privacy issues in electronic healthcare systems. More than 50 original articles were selected to study the existing security approaches and figure out the used security models. Also, a novel Context-aware Access Control Security Model (CARE) is proposed to capture the scenario of data interoperability and support the security fundamentals of healthcare systems along with the capability of providing fine-grained access control.

Keywords—*Electronic health records; Systematic review; Privacy; Security regulations; Interoperability*

I. INTRODUCTION

The widespread availability of ubiquitous medical wearable devices such as smart medical sensors and the using of medical management software systems led the revolution of collecting healthcare data. In this context, sensors and medical systems can be operated by very diverse organizations to continuously sensing patient data during the medical process. However, only authorized users such as medical staff should have access to the collected health data as it almost always contains confidential and sensitive data.

In fact, several pieces of regulations and standards have been proposed to protect individual privacy. One can consider, the HIPAA (Health Insurance Portability and Accountability Act of 1996) that provides data privacy for personal health care information, the European Data Protection Directive 95/46/EC, the GLBA (Gramm-LeachBliley Act, the Sarbanes-Oxley Act, and the EUs Safe Harbour Law [1]. These laws usually require strict security measures for sharing and exchanging health data, and failure to comply with them is strongly sanctioned, with severe penalties being imposed.

In this context, electronic healthcare systems (EHRs) employee such rules and thus were categorized as security critical systems [2]. These systems are differentiated in one important aspect to other systems: The balancing between confidentiality and availability. The tension between these goals is clear: while all the patient's data should be available to be shared and monitored to deliver professional healthcare services; for security reasons, part of the data may be considered confidential and must not be accessible. Clearly, reconciling between the pair goals should be achieved to provide the best possible care for patients.

Indeed, EHRs are real-time, patient-centered systems that make data available and managed by authorized providers in a digital format. In fact, EHR was built upon the standards of collecting data from patients and is composed of three main components: A set of intelligent physiological sensors with a personal server to gather the vital signs, a heterogeneous network, and a remote health care server. In EHRs, users may be a health data owner (i.e., patients) or a requester (i.e., doctors or pharmacists), servers, in turn could be local or cloud servers that store, process and analyze the gathered health data [3,4]. Networks, on the other hand, act as the bridge connecting between patients and the medical staff to support the transmitting and sharing of data [4]. Fig.1 illustrates the typical architecture of EHR system.

Although of many benefits provided by healthcare systems, nevertheless, there are vulnerable to a wide range of security threats because of their portability and design [10]. Specifically, threats were emerged at each level of the system, for instance: *At data collection level* [5-10], *At transmission level* [11-14], and *At storage level* [15-19]. These threats were described in Section III. In addition to the aforementioned threats, some patients worry while using healthcare systems applications. So, it is necessary to ensure patients feel fully confident to use the system and have their own privacy control over it [11]. To this end, in this paper, we conduct an in-depth survey study to analyze the healthcare system's security and privacy threats. Then, we propose a novel security model that captures the scenario of data interoperability and supports the security fundamental of EHR along with the capability of providing fine-grained access control [20].

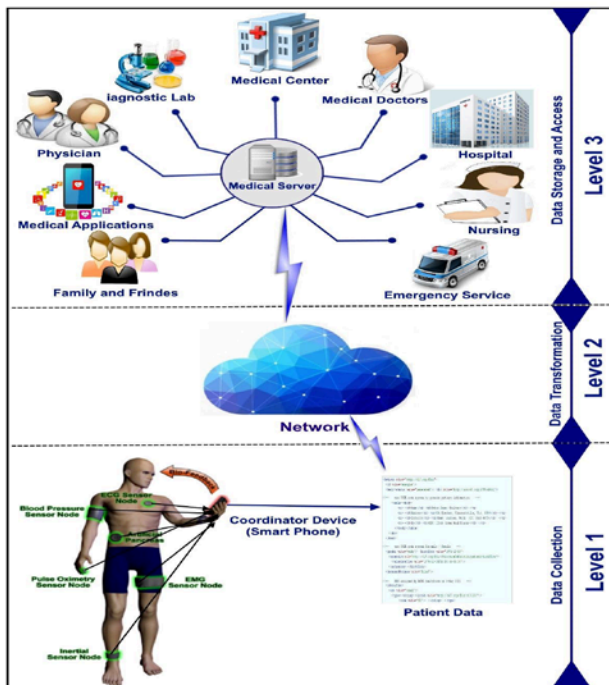


Fig. 1. The architecture of Healthcare Monitoring System

The remainder of the paper was organized as follows. In Section II, we discuss the privacy requirements of healthcare systems; its security attacks were then presented in detail in Section III. Section IV presents a set of exiting security models. The proposed model was discussed in Section V. Final conclusions, and the future work was offered in Section VI.

II. PRIVACY REQUIREMENTS IN HEALTHCARE SYSTEMS

Several general security and privacy requirements should be satisfied to provide the appropriate level of privacy in EHR system. Authors defined more than twenty security requirements that were found on surveys such as [3, 15-18, 21-26]. Due to space limitations, we list the most important requirements:

1) Access control is the ability to limit and control the access to resources by authorized users [3,21]. It makes use of three different security and privacy requirements: identification, authentication, and authorization. Identification is not an original security issue in itself, but its purpose is to identify users. Thus, it is used to affect the way a user can be authenticated [22,23,24]. Authentication, in turn, provides assurance that the requesting data access is authentic and valid [3] and has the identity claims before accessing [21]. It also ensures that the communication is with an authorized party on the other side [22]. Finally, the authorization process determines which part of data can be restricted to an external requester upon the security policy. It is important to mention that a proper access control mechanism should ensure patient privacy and also provide a good balance between availability and confidentiality [15, 23] security goals.

2) Availability is the property of a system and resource being accessible, usable and available upon demand by authorized users [15,18,21] anytime anywhere in the healthcare system [25]. Ensuring availability also involves preventing service disruptions due to hardware failures, power outages, and system upgrade [16,22].

3) Dependability guarantees easily retrievable of medical data at any time even if there are some threats caused by the network dynamic or failure node [18,26]. Usually in most medical cases, unable to retrieve accurate data is due to threats caused by the network dynamics, threaten the patient's life. Fault tolerance is a necessary requisite for dependability.

4) Flexibility is to enable unauthorized participant who is not on the permissible list to access specific data in an emergency case to save the patient's life. Inability or prevention the access rules may threaten a patient's life [18].

III. SECURITY ATTACKS IN HEALTHCARE SYSTEM

Healthcare systems are vulnerable to penetration by malicious attacks or intentionally from users for profit. This damages the effectiveness or deterioration the performance of healthcare systems [4,27,28]. Specifically, insulin pump sensors, hospital networks, or the personal health data can be hacked or stolen by malicious users [19, 26].

1) Attacks at data collection level

These attacks may cause several threats to data collection level such as altering information, dropping some important data, or resending data messages.

Jamming Attack: refers to interference attacker's radio signal with frequencies of the BAN (Body Area Networks). Resulting in isolating and preventing sensor node within the range of the attacker signals for giving or receiving any message among the affected nodes and other sender nodes as long as the jamming signal continues [5, 7].

Data Collision Attack: takes place when two or more nodes attempt to transmit simultaneously. Also, it refers to jamming attacks when a foe may strategically generate extra collisions by sending repeated messages on the channel [6, 7]. When the frame header is changed due to a collision, the error checking mechanism at the receiving end detects that as an error and rejects received data. Thus, a change in the data frame header is a threat to data availability in the BAN [5].

Data Flooding Attack: the attacker repeatedly broadcast many requests to the victim node for connection until using all the power of its resources reach a maximum limit, causing a flooding attack [8].

Desynchronization Attack: in this type of attack, the attacker's tampers messages between sensor nodes by copy it many times using a fake sequence number to one or both endpoints of an active connection, which leads the WBAN to an infinite cycle, resulting in causing the sensor nodes transmits messages again and wastes their energy [6, 8].

Spoofing Attack: where the attacker targets the routing information to perform several disruptions such as spoofing,

altering, or replay the routing information, leading to complicate the network by creating routing loops [9].

Selective Forwarding Attack: it takes place when the attacker malicious node in a data flow path forwards selected messages and drops the others. The damage becomes serious when these malicious nodes were located proximity to the base station [13].

Sybil Attacks: in Sybil, the attacker malicious node represents more than one identity in the network [6]. It has important effect in geographic routing protocols. Where the location information is required to be exchanged between the nodes and their neighbors to route the geographically addressed packets efficiently [6, 13]. Unfortunately, detecting Sybil attackers are not easily captured due to the unpredictable paths and high mobility they use [4, 27].

2) Attacks at transmission level

These attacks may cause several threats to transmission level such as spying, altering information, interrupting communication, sending extra signals to block the base station and networking traffic.

Eavesdropping of Patient's Medical Information: Monitoring system will record patient's health data from BANs to be transmitted to the healthcare providers. Unprincipled developers can easily build systems with the ability to spy on the patient's data through wireless technology. Thus the developer needs to apply controlling authority whenever they develop a system, which protects the patient's information against eavesdroppers and reduces the number of people who try to take and breach the patient's privacy [8, 11].

Man in the Middle Attacks: the attacker intercepts a communication between the end points and exchange messages between them. The communication is completely controlled by the attacker enable him being able to read, insert and modify the data in the intercepted communication [5, 12].

Data Tampering Attack: where a tampering attacker may damage and replace encrypted data by authorized network nodes [6, 13].

Scrambling Attacks: is a kind of jamming attack on radio frequency for short intervals of time during transmission of control or management information WiMAX frames to affect the normal operation of the network. It interrupts the communication that can prevent the patient's smartphone from sending data causing availability issue [5].

Signaling Attacks: Before patient's smartphone starts transmitting data, there is some preliminary signaling operation need to be performed with the serving base station. Signaling operations contain authentication, key management, registration, and IP-based connection establishment. The attacker can initiate a signaling attack on the serving base station by actuating extra state signals that block the base station. Thus, the excessive load on the base station results in DoS attacks, and the patient's smartphone cannot send data due to base station unavailability [5].

Unfairness in allocation: it lacks the network performance by interrupting the Medium Access Control (MAC) priority schemes [13, 14].

Message Modification Attack: In this type of attack, the attacker can capture the patient wireless channels and extract the patient medical data to be tampered later, which can mislead the involved users (doctor, nurse, family) [8].

Hello Flood Attack: these types of attacks are used to fool the network. Where the attacker sends a hello message with a high powered radio transmission to the network to convince all nodes to choose the attacker for routing their messages [6, 13].

Data Interception Attack: this type of attack can take place via interception the patient's information by the attacker during exchanging them between computers of healthcare system through hospital LAN [5].

Wormhole Attack: this type of attack known as a silent and severe type of attack because it copies the packet at one location and replays them at another location or within the same network without any changes in the content. It aimed to damage the network topology and traffic flow through creating a tunnel between the two attackers to be used for transmitting between them [10,13].

3) Attacks at storage level

These attacks may cause several threats to storage level such as modifying patient medical information or changing the configuration of system monitoring servers.

Inference Patient's Information: Attackers try to combine authorized information and combine them with other available data, which leads them to identify sensitive patient data such as diseases [8, 11, 17]. Thus, patient's data should be anonymous to cover their identities or data before publishing/posting the data [3].

Unauthorized access of Patient Medical Information: this type of attack can take place by unauthorized Individual without valid authentication, so patient's data will be accessed then it might cause problems such as damaging significant data [18]. Thus, it is necessary to protect patient privacy against breaching, capturing, and misusing by unauthorized users [11, 16].

Malware Attack is a malicious software program designed to perform harmful actions [19]. This type of attacks has the ability to infect and propagate to the whole hospital server that can cause unavailability and disruption. Whereas, Changing and updating in software configuration of patient monitoring servers making system configuration unstable, resulting in system malfunctioning and communication interruption [5, 12].

Social Engineering Attacks: in this type of attack, a third party attacker can gain access to the system by fooling either the patient or authorized user to access the information. Here, authorized users can also disclose patient's data to concerned parties such as Health Insurance Company for unethical personal intends [5, 12].

Removable Distribution Media Attack: In this type of attacks it is possible to theft or loss computer or data storage medium, such as a USB flash drives, can be used to steal information and to propagate viruses in a healthcare monitoring system [5].

Others issues: several hardware and software issues can cause an interruption in the healthcare system. Hackers may develop new techniques or discover new software vulnerabilities. It is possible also that the system can be exposed to various types of software attacks such as viruses, worms, Trojans, and spyware attacks [19].

IV. E-HEALTHCARE SECURITY MODELS

To improve the quality of healthcare delivery, patient's data could be shared across a variety of users, which may lead to privacy disclosure. So, e-Health systems need to be protected through convenient security models to ensure proper access controls [29,49,50,51]. In fact, encryption is the traditional solution used. Although it provides a simple access control, it is not applicable for complex EHR systems that require various access requirements. That is, keeping the e-Health data secured is a big challenge due to two main reasons: the significant computational overhead when encryption techniques were used, and the sensitivity of personal medical information from changing when modification techniques are employed [30]. In this section, a detailed description of a set of security models, along with their corresponding levels, are presented.

1) Security Models for Data Collection Level

O. G. Morchon and K. Wehrle in [31] present a modular access control system for pervasive healthcare applications. The system extends the traditional RBAC model for two main issues: Firstly, to assign and distribute access control policies to sensor nodes. Secondly, to store the current medical context (location, time, health information) that influences access control decisions upon patient's medical situation (critical, emergency or normal situation). The modular design makes the system's configuration more effective and simplifies the composition of policies to deploy safer and more secure medical sensor networks. However, when a critical or emergency case raised, the medical stuff can override the restrictions to access sensitive data that was restricted in normal condition. One of the limitations of this model is that there is no detection mechanism for unauthorized access when critical situations occur.

S. Amini et al. [32] examined a set of security protocols such as TinySec, MiniSec, LLSP, and RC4-based along with different ciphers algorithms (Skipjack, AES, and RC4) to proposed an approach to design a lightweight security model. To this end, authors combined different types of attacks (data loss, spoofing of sensors, and eavesdropping and replay) and applied the ciphers algorithms. They found that RC4 and Skipjack cipher algorithms are the most efficient to fulfill confidentiality regarding of RAM, ROM, and clock cycles per byte (CPB). Despite the advantages of such study, they did not consider other types of security threats.

H. A. Maw et al. [33] proposed an Adaptive Access Control model that provides fine-grained access control for

medical data in BSNs and WSNs. The model considers privilege overriding and behavior, so users might be able to override a denial of access when unexpected events occur. Here, there is no need for a human effort to pay pass authorizations and policies since users initialize their sessions in behavior trust model based on users, location, time, and action. However, the main limitation of this model is that there is no prevention or detection mechanism to check user's data access when the critical situation occurs.

Authors of [34] and [35] proposed a three-tier security framework based on pairwise key pre-distribution schema. The framework has two separate key pools: one for the mobile sink to access the network, and the second for pairwise key establishment between the sensors. To further improve the network resilience and reduce the damages caused by stationary access node replication attacks, they have strengthened the authentication technique between the sensor and the stationary access node in the proposed framework. However, in basic key predistribution schemes, an attacker can gain a number of keys by catching a small fraction of nodes, and hence, can gain control of the network by deploying a replicated mobile sink preloaded with some compromised keys.

S. N. Ramlil et al. [36] proposed a biometric-based security framework for data authentication within WBAN. In particular, signals like sender's Electrocardiogram (ECG) feature can be utilized as a key to ensuring that patients' data will not be mixed since each patient has his/her own specific biometrics, which results in reducing computational complexity and improving the efficiency over the using of cryptographic key distribution. Thus, it saves resources while convenient security measures are employed. The main limitation of this work is that the authentication process was based on the sensors themselves, which restricts the process with their limited resources.

M. Kun and L. Li. [37] proposed an efficient key management scheme for WSNs group-based key pre-distribution scheme. The proposed scheme consists of three phases which are initialization phase; share-key discover phase and path-key establishment phase. Here, every sensor node has a given security level (high to low), where a low-security level node cannot access the collected data for a higher-level security sensor. Thus, a compromised sensor node (e.g., with a low-security) cannot disclose the key information in the sensor node (with high-security). The analysis of their proposed scheme offers a stronger resilience against node capture attack.

2) Security Models for Data Transmission Level

A. Boonyarattaphan et al. [30] proposed a secure framework for authentication and data transmission using Encryption techniques for implementing two mechanisms: Data and Channel security. The channel security was provided by utilizing the SSL on the HTTP layer, while the data security is provided on the SOAP layer constructed above the HTTP. They emphasized that RBAC should be used along with multi-factor authentication to guarantee proper authorization and authentication. Depend on the roles of stakeholders and data sensitivity; communication was divided

into different layers where different authentication and encryption settings can be adapted. The only limitation here is that it is dealt only with the web-based eHealth services.

N. Kahani et al. [38] proposed a new and secure scheme that supports both secure authentication and scalable fine-grained data access control. The scheme is based on a zero-knowledge protocol to verify and maintain the anonymity of the user's identity. This approach uses combination of a system public key and a secret session key generated by Derive Unique Key Per-Transaction (DUKPT) scheme to establish secure communication between different interacting entities. The access control mechanism was implemented in two phases: the first one utilizes a static authorization method to determine the highest access rights of users. And, the second one grants the user the minimum access permissions on the required data according to the user's intention of access and the maximum rights determined by the first phase. To keep user's data confidential against malicious users and to decrease computational and communication overhead on data owners, data were stored in encrypted format in the cloud. However, by storing patient's health data in the cloud, patients lose the control over their data. Moreover, because of using encryption technique, it is difficult to achieve fine-grained access control to patient's data in a scalable and well-organized way.

Z. Guan et al. [39] considered the data security and privacy for cloud-integrated body sensor networks. They proposed a novel encryption outsourcing scheme named Mask-Certificate Attribute-Based Encryption (MC-ABE) by combining seven encryption algorithms. In this schema, data owner (patient) encrypts the outsourcing data to mask the row data before storing it securely in storage service provider (cloud servers). Furthermore, to achieve more effective access control, a unique authentication certificate is introduced for each user, which was verified before accessing data. Experimental results showed that the proposed scheme has less computation cost and storage cost compared with other common models. However, because of using encryption technique, it is difficult to achieve fine-grained access control, and still it requires some degree of computational overhead.

M. A. Simplicio et al. [40] proposed SecureHealth lightweight security framework based on very lightweight mechanisms such as (TLS/SSL) for securing the data exchanged with the server without needing an extra security layer. SecureHealth provides security services for both stored and transmitted data. Moreover, it includes many security features such as user authentication, data confidentiality, and the lack of connectivity. This framework depends on Even the SecureHealth was designed to prevent an outsider from illegally accessing or tampering with the system's data; it also gives managers the ability to identify misbehavior from insiders.

3) Security Models for Data Storage and Access Level

Lili Sun and Hua Wang [29] considered the notion of *purpose* to design a comprehensive usage access control model. Specifically, purpose notation was used for specifying privacy policies and giving the privilege to access private data. The proposed model consists of eight core components which

are, subject attributes, objects, object attributes, rights, authorizations, obligations, and conditions. Whereas, authorizations, obligations, and conditions are components of usage control decisions used to determine whether a subject is allowed to access an object. The existence of obligations and conditions helped in solving certain shortcomings that have been common in access controls. That being said, the main limitation of this model is that it represents only a first step for authorization model in purpose data with usage control.

M. Barna et al. [1] proposed a security scheme based on different privacy levels. In short, the access control process was done in the centralized infrastructures. Here, the attribute-based encryption (ABE) was used rather standard way; privileges were mapped into roles and roles into ABE access structures. The data then is moved to the cloud-based storage, which enables the e-Health care service providers to decrease the overall maintaining cost of data and allows data to be online anytime and anywhere. However, because the data was stored in a centralized server, it becomes like a bottleneck when data requests were issued from different users.

To solve the aforementioned problem, L. Guo et al. [41] took into account the distributed nature of eHealth system when designing a privacy-preserving authentication system. In this system, instead of letting centralized infrastructures take care of authentication, the two end users (patients and physicians) do the authentication process. In particular, users are allowed to authenticate each other without disclosing their attributes and identities, which solves the problem of maintaining privacy and variability of each user's attributes.

R. Gajanayake et al. [42] proposed a privacy oriented access control model for satisfying eHealth's requirements. The model was designed by combining three existing access control models (DAC, MAC, and RBAC) into a novel module that enables patients and healthcare professionals to determine and setting the access privileges. The module has been tested to demonstrate different scenarios of policy settings and data access. It proves that it can be used as a standalone security model to achieve HER requirements.

M. Barua et al. [43] Proposed a secure patient-centric personal health information schema for sharing and providing access control in cloud computing based on Proxy Re-encryption Protocol. The proposed schema has five main phases: transmitting patient's data to the Health-Service Provider, defining access policy, storing patient's data at cloud, validating data-access requester, and finally auditing the stored encrypted data. Their schema exploits attribute-based encryption to ensure patient-centric access control. The performance analysis shows that the proposed schema is extremely efficient to resist several possible attacks and malicious behaviors.

In the same vein, M. R. Kumar et al. [44] suggest a new patient-centric framework based on the same encryption technique (ABE). Here, the users were categorized into two main domains namely: public and personal domains to face the key management complexity. In the public domain, users utilize multi-authority ABE (MA-ABE) to improve the fine-grained security countermeasures. While, in the personal domain, an owner is permitted to access/encrypt the data

under his attributes. The limitations of this model is that integration ABE into large scale PHR system, required significant issues such as key management scalability, efficient on demand revocation, and lively policy updates which are nontrivial to resolve and remains up-to-date.

H. Zhu et al. [45] also proposed a secure and efficient personal scheme based on the attribute-based encryption (ABE) and re-encryption under the attribute group keys using RSA-Based proxy encryption. The proxy encryption technology is used to introduce an efficient privilege separation mechanism to ensure the validity of patients' data. Here, the write privilege keys were distributed to professional people and the read privilege keys to patients, so that the data is not only fully controlled by the patient to authorize access, but also have the great validity. As a result, the computational overhead was reduced, and the key escrow problem was solved by employing re-encryption under the attribute group keys. Thus, the health provider could be prevented from obtaining the read keys without multiple-authority ABE.

V. Sunagar and C. Biradar [46] proposed a secured framework based on advanced encryption standard (AES) algorithm to encrypt every patient's data according to the security policy. AES enables the users to maintain data in a secured cloud environment. Ultimately, the framework consists of three modules: PHR Owner/patient module, Data confidentiality module, and Cloud Server module, which provides a high level of security.

Finally, W. Liu et al. [47] proposed a generic framework that depends on hierarchical identity-based encryption (HIBE) schema and the role-based access control (RBAC). While the HIBE is used to encrypt patients' data before outsourcing them to the storage server, the RBAC facilitates forwarding users' privileges. The experimental results of this model show that it is a practical solution to keeping data secure and confidential. However, the framework does not provide accurate access control requirements, as in some specific situations, patients might not have access to their own sensitive data (e.g., psychotherapy notes) without proper authorization according to HIPAA regulations. Such approaches suffer from the well-known encryption drawbacks [48].

V. CARE SECURITY MODEL

The Context-aware Access Control Security Model (CARE) architecture is based on the scenario of data interoperability and supports the security fundamentals of healthcare systems along with the capability of providing fine-grained access control. Specifically, the CARE model could be located on the healthcare server, which serves as an access point for users' requests. Fig.2 depicts the architecture of CARE.

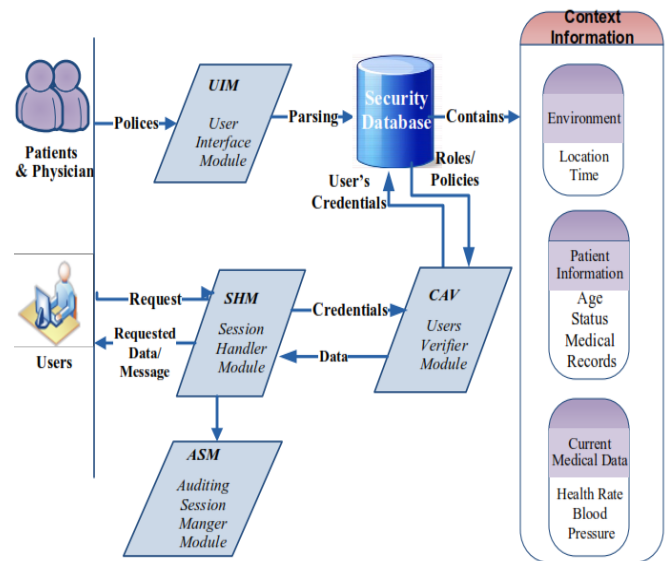


Fig. 2. CARE Model architecture

In CARE, policies were defined by using the User Interface Module (UIM), which could be a website or a mobile application. Patients and physician define the policies together to save patients privacy. All the defined policies are then parsed into its components (e.g., constraints) and stored in a centralized security database, which is represented using an Extended-RBAC model. The ERBAC consists of roles, permissions, and users. Roles were created for various job functions, with permissions, with permissions for specific operations. Users are assigned particular roles, and through those role assignments acquire permissions to perform certain operations. The consolidation of access control for many users into a single role entry allows for much easier management of the overall system and much more effective verification of security policies. Three different types of context-data were considered in this model. The Environmental context refers to location or time, the Personal information context regarding age, status, or medical record and finally, the current medical data such as Heart rate or blood pressure [31].

Upon a user's request data, a session was established between the requester and the server side by the Session Handler Module (SHM). The requester's credentials (e.g., digital certificate, or user-name/password) were then extracted to be verified against a list of valid user accounts stored in the security database. The established session may involve more than one message, and are secured since secured transmission protocols were employed in all communications. It is important to mention here that session's information were saved to be able to communicate later. To this end, the Auditing Session Manger (ASM) takes this responsibility and

states all the established sessions that could be used to retrieve multiple data for the subsequent access requests. This is opposed to stateless communication where it consists of independent requests (needs multiple authentications).

After establishing the session, the *Users Verifier Module (CAV)* verifies the requester credentials and then determines if the user is allowed to access the requested data or not. This is done by contacting the security database and retrieving the applicable policies and requester's assigned roles. CAV also classifies the request's cases as critical, emergency, or normal depending on the context-aware information and then adjusts the final access decision. In particular, when the patient's life is in danger the security settings are adapted by removing the need for user authentication to access the data.

VI. CONCLUSION AND FUTURE WORK

Patient's data should be kept securely in medical provider servers so that physicians can provide proper treatments. To ensure secure storage and access management, in this paper, we argue the security attacks in healthcare system along with the proposed security models that aim to prevent such attacks. Specifically, threats were categorized into three types depending on the its emerged level of the healthcare system, for instance: at data collection level; at transmission level; and at storage level. These attacks may cause several threats such as altering information, dropping some important data, interrupting communication, or sending extra signals to block the base station and increasing networking traffic.

After that, we briefly discussed a novel context-aware access control security model that supports the security fundamentals of healthcare systems and providing fine-grained access control. The model consists of multiple modules, each of which is in charge of taking a different type of task. This modular design aims at simple and efficient access control decision depending on the patient's situation and the requester's assigned roles.

ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman-Jordan, for the full financial support granted to cover the publication fee of this research article.

REFERENCES

- [1] Barua, M., et al. PEACE: An efficient and secure patient-centric access control scheme for eHealth care system. in Computer Communications Workshops (INFOCOM WKSHP), 2011 IEEE Conference on. 2011. IEEE.
- [2] Fernández-Alemán, J.L., et al., Security and privacy in electronic health records: A systematic literature review. *Journal of Biomedical Informatics*, 2013. 46(3): p. 541-562.
- [3] Zhang, K. and X.S. Shen, Security and Privacy for Mobile Healthcare Networks. 2015.
- [4] Shinde, S.S. And D. Patil, Review On Security And Privacy For Mobile Healthcare Networks: From A Quality Of Protection Perspective *International Journal of Engineering Research-Online Peer Reviewed International Journal* 2015. 3(6).
- [5] Habib, K., A. Torjusen, and W. Leister. Security analysis of a patient monitoring system for the Internet of Things in eHealth. in Proceedings of the International Conference on eHealth, Telemedicine, and Social Medicine (eTELEMED'15). 2015.

- [6] Saleem, S., S. Ullah, and K.S. Kwak, A study of IEEE 802.15. 4 security framework for wireless body area networks. *Sensors*, 2011. 11(2): p. 1383-1395.
- [7] CHELLI, K. Security Issues in Wireless Sensor Networks: Attacks and Countermeasures. in Proceedings of the World Congress on Engineering. 2015.
- [8] Kumar, P. and H.-J. Lee, Security issues in healthcare applications using wireless medical sensor networks: A survey. *Sensors*, 2011. 12(1): p. 55-91.
- [9] Saleem, S., S. Ullah, and H.S. Yoo, On the Security Issues in Wireless Body Area Networks. *JDCTA*, 2009. 3(3): p. 178-184.
- [10] Om, S. and M. Talib, Wireless Ad-hoc Network under Black-hole Attack. *International Journal of Digital Information and Wireless Communications (IJDIWC)*, 2011. 1(3): p. 591-596.
- [11] Ramli, R., N. Zakaria, and P. Sumari, Privacy issues in pervasive healthcare monitoring system: A review. *World Acad. Sci. Eng. Technol*, 2010. 72: p. 741-747.
- [12] Partala, J., et al. Security threats against the transmission chain of a medical health monitoring system. in e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on. 2013. IEEE.
- [13] Niksaz, P. and M. Branch, Wireless Body Area Networks: Attacks and Countermeasures.
- [14] Bonab, T.H. and M. Masdari, Security attacks in wireless body area networks: challenges and issues. *ACADEMIE ROYALE DES SCIENCES D OUTRE-MER BULLETIN DES SEANCES*, 2015. 4(4): p. 100-107.
- [15] Santos-Pereira, C., et al. A secure RBAC mobile agent access control model for healthcare institutions. in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013. IEEE.
- [16] Zhang, R. and L. Liu. Security models and requirements for healthcare application clouds. in 2010 IEEE 3rd International Conference on Cloud Computing. 2010. IEEE.
- [17] Drosatos, G., et al., Towards Privacy by Design in Personal e-Health Systems. 2016.
- [18] Fatema, N. and R. Brad, Security Requirements, Counterattacks and Projects in Healthcare Applications Using WSNs-A Review. arXiv preprint arXiv:1406.1795, 2014.
- [19] Wellington, K., Cyberattacks on Medical Devices and Hospital Networks: Legal Gaps and Regulatory Solutions. *Santa Clara High Tech. LJ*, 2013. 30: p. 139.
- [20] Yu, S., et al. Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing. in INFOCOM, 2010 Proceedings IEEE. 2010.
- [21] Zubaydi, F., et al. Security of mobile health (mHealth) systems. in Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on. 2015. IEEE.
- [22] Nagaty, K.A., Mobile Health Care on a Secured Hybrid Cloud.
- [23] Kotz, D. A threat taxonomy for mHealth privacy. in COMSNETS. 2011.
- [24] Mare, S., et al. Adapt-lite: Privacy-aware, secure, and efficient mhealth sensing. in Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. 2011. ACM.
- [25] Sun, J., et al., Security and Privacy for Mobile Healthcare (m-Health) Systems. 2011, Amsterdam, The Netherlands: Elsevier.
- [26] Wang, J., et al., A Research on Security and Privacy Issues for Patient Related Data in Medical Organization System. *International Journal of Security and Its Applications*, 2013. 7(4): p. 287-298.
- [27] Zhang, K., et al., Security and privacy for mobile healthcare networks: from a quality of protection perspective. *IEEE Wireless Communications*, 2015. 22(4): p. 104-112.
- [28] Zhang, K., et al., Sybil attacks and their defenses in the internet of things. *IEEE Internet of Things Journal*, 2014. 1(5): p. 372-383.
- [29] Sun, L. and H. Wang. A purpose based usage access control model for e-healthcare services. in Data and Knowledge Engineering (ICDKE), 2011 International Conference on. 2011. IEEE.

- [30] Boonyarattaphan, A., Y. Bai, and S. Chung. A security framework for e-health service authentication and e-health data transmission. in Communications and Information Technology, 2009. ISCIT 2009. 9th International Symposium on. 2009. IEEE.
- [31] Garcia-Morchon, O. and K. Wehrle. Efficient and context-aware access control for pervasive medical sensor networks. in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on. 2010. IEEE.
- [32] Amini, S., et al. Toward a security model for a body sensor platform. in Consumer Electronics (ICCE), 2011 IEEE International Conference on. 2011. IEEE.
- [33] Maw, H.A., H. Xiao, and B. Christianson. An adaptive access control model for medical data in wireless sensor networks. in e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on. 2013. IEEE.
- [34] Linciya, T. and K. Anandkumar, Enhanced Three Tier Security Architecture For Wsn Against Mobile Sink Replication Attacks Using Mutual Authentication Scheme. International Journal of Wireless & Mobile Networks, 2013. 5(2): p. 81.
- [35] Rasheed, A. and R.N. Mahapatra, The Three-Tier Security Scheme in Wireless Sensor Networks with Mobile Sinks. IEEE Transactions on Parallel and Distributed Systems, 2012. 23(5): p. 958-965.
- [36] Ramli, S.N., et al. A biometric-based security for data authentication in wireless body area network (wban). in Advanced Communication Technology (ICACT), 2013 15th International Conference on. 13. IEEE.
- [37] Mu, K. and L. Li, An efficient pairwise key predistribution scheme for wireless sensor networks. Journal of Networks, 2014. 9(2): p. 277-282.
- [38] Kahani, N., K. Elgazzar, and J.R. Cordy, Authentication and Access Control in e-Health Systems in the Cloud.
- [39] Guan, Z., T. Yang, and X. Du, Achieving secure and efficient data access control for cloud-integrated body sensor networks. International Journal of Distributed Sensor Networks, 2015. 2015: p. 142.
- [40] Simplicio, M.A., et al., SecourHealth: a delay-tolerant security framework for mobile health data collection. IEEE journal of biomedical and health informatics, 2015. 19(2): p. 761-772.
- [41] Guo, L., et al. Paas: A privacy-preserving attribute-based authentication system for ehealth networks. in Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on. 2012. IEEE.
- [42] Gajanayake, R., R. Iannella, and T. Sahama, Privacy oriented access control for electronic health records. electronic Journal of Health Informatics, 2014. 8(2): p. 15.
- [43] Barua, M., R. Lu, and X. Shen. SPS: Secure personal health information sharing with patient-centric access control in cloud computing. in 2013 IEEE Global Communications Conference (GLOBECOM). 2013. IEEE.
- [44] Kumar, M.R., M.D. Fathima, and M. Mahendran, Personal Health Data Storage Protection on Cloud Using MA-ABE. International Journal of Computer Applications, 2013. 75(8).
- [45] Zhu, H., et al. SPEMR: A new secure personal electronic medical record scheme with privilege separation. in 2014 IEEE International Conference on Communications Workshops (ICC). 2014. IEEE.
- [46] Sunagar, V. and C. Biradar, Securing Public Health Records in Cloud Computing Patient Centric and Fine Grained Data Access Control in Multi Owner Settings. 2014.
- [47] Liu, W., et al. Auditing and Revocation Enabled Role-Based Access Control over Outsourced Private EHRs. in High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICCESS), 2015 IEEE 17th International Conference on. 2015. IEEE.
- [48] Katz, J. and Y. Lindell, Introduction to modern cryptography. 2014: CRC press.
- [49] Altamimi, A., SecFHIR: A Security Specification Model for Fast Healthcare Interoperability Resources. International Journal of Advanced Computer Science and Applications(ijacsa), 7(6), 2016.
- [50] Sahama, T., Simpson, L., Lane, B., Security and Privacy in eHealth: Is it possible?. In e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on. 2013 pp. 249-253.
- [51] Leyla, N., MacCaul, W., A Personalized Access Control Framework for Workflow-Based Health Care Information. In International Conference on Business Process Management 2011. (pp. 273-284). Springer Berlin Heidelberg.

Estimation of Trajectory and Location for Mobile Sound Source

¹Mehmet Cem Catalbas, ¹Merve Yildirim, ¹Arif Gulden, ¹Hasan Kurum, and ²Simon Dobrišek

¹University of Firat Faculty of Electrical and Electronics Engineering, Elazig, Turkey

²University of Ljubljana Faculty of Electrical Engineering Ljubljana, Slovenia

Abstract—In this paper, we present an approach to estimate mobile sound source trajectory. An artificially created sound source signal is used in this work. The main aim of this paper is to estimate the mobile object trajectory via sound processing methods. The performance of generalized cross correlation techniques is compared with that of noise reduction filters for the success of trajectory estimation. The azimuth angle between the sound source and receiver is calculated during the whole movement. The parameter of Interaural Time Difference (ITD) is utilized for determining azimuth angle. The success of estimated delay is compared with different types of Generalized Cross Correlation (GCC) algorithms. In this study, an approach for sound localization and trajectory estimation on 2D space is proposed. Besides, different types of pre-filter method are tried for removing the noise and signal smoothing of recorded sound signals. Some basic parameters of sound localization process are also explained. Moreover, the calculation error of average azimuth angle is compared with different GCC and pre-filtered methods. To conclude, it is observed that estimation of location and trajectory information of a mobile object from a stereo sound recording is realized successfully.

Keywords—Sound processing; sound source localization; azimuth angle estimation; generalized cross-correlation; interaural time difference; interaural level difference

I. INTRODUCTION

The general definition of sound is pressure change versus time which is examined by microphone or microphone arrays. Also, these differences are used for sound recognition, localization, and classification, etc. The sound localization is implemented by using this information. Generally, Sound localization process is inspired from the human hearing system [1-2]. There are several areas of research for sound processing [3]. One of the most important areas is sound localization [4]. In parallel to technological improvement, sound localization is used in very different areas [5-6] such as security systems, cell phone, voice command application, and conference system, etc. The term of sound localization is determining the coordinates of sound source in 2D or 3D Space [7-8]. The 2D coordinate information is adequate for some of the basic applications [9-10]. Besides, the information of 3D coordinates is necessary for complex sound processing applications. In this study, we obtain 2D coordinates information for moving person via stereo sound recording [11]. A sound dataset which is created artificially is used for this work [12-14].

In this paper, Section II explains sound localization. In Section III, 2D sound localization via microphone pair is described. Estimations of sound delay and sound source

trajectory are carried out in Section IV and V, respectively. Finally, conclusions are given in the last section.

II. SOUND LOCALIZATION

The term of sound localization is determining of the coordinates of sound sources by using some of the signal processing methods. Generally, the users must have least three microphones or sensors for localization.

We propose a different approach for sound localization in this paper. In contrast to the other studies, the locations of sound sources are taken by only two microphones. Furthermore, two common terms for sound localization process which are ITD and Interaural Level Differences (ILD) are used [15-16]. The parameters of ITD and ILD are time and amplitude difference between signals received by microphones or sensors [17]. The parameter of ITD is used generally for determining azimuth angle [18-19]. ITD and ILD are illustrated in Fig. 1 and Fig. 2, respectively.

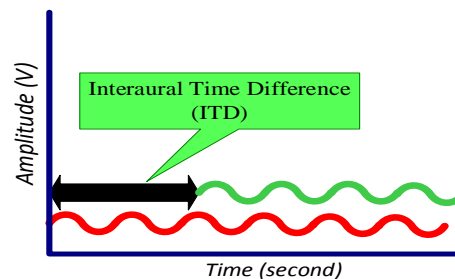


Fig. 1. Interaural Time Difference (ITD)

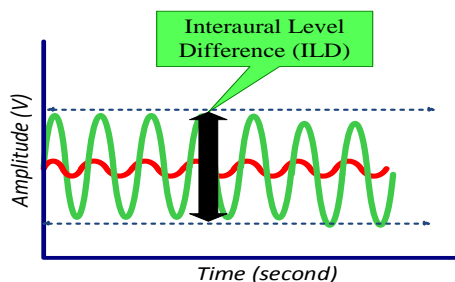


Fig. 2. Interaural Level Difference (ILD)

The parameter of ITD is much more reliable compared to that of ILD. The ITD is easily calculated by cross-correlation methods. The result of ITD is sample difference between received signals. The azimuth angle is also calculated by this parameter and sampling frequency is taken as $f_s = 44100$ Hz in this study. The calculation of azimuth angle is shown in Fig.

3. The speed of sound is very sensitive to environmental conditions such as especially ambient humidity and temperature [20].

The calculation of the speed of sound depended on ambient temperature and azimuth angle is shown in (1) and (2), respectively [21].

$$V_{sound} = 20.05\sqrt{273.15 + C_t} \quad (1)$$

$$\alpha = \arcsin\left(\frac{V_{sound}T_d}{d}\right) \quad (2)$$

where C_t , d , V_{sound} , and T_d are the ambient temperature ($^{\circ}\text{C}$), the distance between sound receivers (meter), speed of sound (m/sec), and time differences between received signals (sec), respectively.

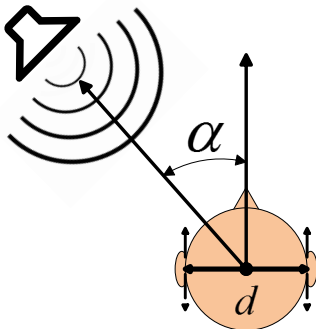


Fig. 3. Azimuth angle determination

The sound localization process is given in Fig. 4. The angles and distances between points are taken as $(\beta=\gamma=\delta=60^{\circ})$ and $(d_x = d_y = d_z)$, respectively. V_m is movement speed of the sound source.

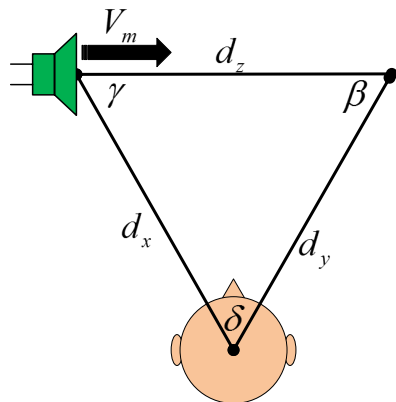


Fig. 4. Illustration of test environment

III. 2D SOUND LOCALIZATION VIA MICROPHONE PAIR

In this section, 2D localization via microphone pair is explained in detail. The sound localization problem is solved by three or more microphones and the researchers use Time Difference of Arrival (TDOA) between microphone pairs [22-23]. The advantage of this approach is that the user can determine x and y coordinates via only one microphone pair. However, the users have to obtain not only the parameter of ITD but also ILD parameter of sound signals [24]. In this

approach, the inverse square law of energy propagation is used for sound localization. The received sound signals are defined as s_1 and s_2 [21]. $n_i(t)$ represents to the white noise on environment.

The energy received from microphone is

$$E_i = \int_0^w x_i^2(t)dt \quad i=1,2,\dots,n \quad (3)$$

$$E_i = \frac{1}{d^2} \int_0^w s^2(t)dt + \int_0^w n_i^2(t)dt$$

where w , E_i , and d are the screen size of the observation, the energy of sound signal x_i , and the distance between sound source and receiver, respectively.

The relationship between the energies and distances can be obtained by using these equations

$$E_1 d_1^2 = E_2 d_2^2 + \eta \quad (4)$$

where $\eta = \int_0^w [d_1^2 n_1^2(t) - d_2^2 n_2^2(t)]dt$ is error term and (x_i, y_i) and (x_s, y_s) are the coordinates of an i -th microphone and the sound source, respectively.

$$d_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2} \quad (5)$$

Since the signals acquired by each microphone can be assumed as delayed replicas of the source signal, localizing a source is to estimate time delay estimation (TDE) between the signals of two microphones. τ_i is time delay and d_i is ignored in TDE model. When time delay is measured, hyperbolic equation can be satisfied by Cartesian coordinate.

$$\sqrt{(x_1 - x_s)^2 + (y_1 - y_s)^2} - \sqrt{(x_2 - x_s)^2 + (y_2 - y_s)^2} = c\tau_{12} \quad (6)$$

where c and τ_{12} are sound speed and TDOA of mic1 and mic2, respectively. By using (4) and (5), (7) can be obtained by

$$(x_1 - x_s)^2 + (y_1 - y_s)^2 = \frac{1}{\gamma^2} [(x_2 - x_s)^2 + (y_2 - y_s)^2] + \frac{\eta}{E_1} \quad (7)$$

where $\gamma = \sqrt{E_1/E_2}$. The noise term η/E_1 can be ignored in a high Signal to Noise Ratio (SNR) environment. According to this, (7) can be inserted into (6) and following equations are obtained as

$$(x_s - x_1)^2 + (y_s - y_1)^2 = \left(\frac{c\tau_{12}}{1-\gamma}\right)^2 \quad (8)$$

$$(x_s - x_2)^2 + (y_s - y_2)^2 = \left(\frac{c\tau_{12}\gamma}{1-\gamma}\right)^2 \quad (9)$$

The exact source position can be found by composing (8) and (9). The existence of solution is defined by

$$\left(\frac{c\tau_{12}}{1-\gamma}\right) |1-\gamma| \leq d \leq \left(\frac{c\tau_{12}}{1-\gamma}\right) (1+\gamma) \quad (10)$$

where d is the distance between two circle centers. When τ_{12} is greater than zero, it means that source reaches mic1 later than mic2 and γ is also less than one. Thus, τ_{12} and $(1-\gamma)$ are positive or negative. In case of $E_1 \neq E_2$, it can be determined by

$$c|\tau_{12}| \leq d \leq c|\tau_{12}| \frac{1+\gamma}{|1-\gamma|} \quad (11)$$

In the case of $E_1 = E_2$, it means that there will not be an intersection to determine source position. To solve this problem, (8) and (9) are used again as follows

$$(x_s - x_1)^2 + (y_s - y_1)^2 = \left(\frac{c\tau_{12}}{1-\gamma}\right)^2 = \tau_1^2 \quad (12)$$

$$(x_s - x_2)^2 + (y_s - y_2)^2 = \left(\frac{c\tau_{12}\gamma}{1-\gamma}\right)^2 = \tau_2^2 \quad (13)$$

According to this, the equations are obtained as follows

$$x_i x_s + y_i y_s = \frac{1}{2}(K_i - r_i^2 + R_s^2); \quad i = 1, 2 \quad (14)$$

where

$$K_i = x_i^2 + y_i^2; \quad (i = 1, 2) \\ R_s = \sqrt{x_s^2 + y_s^2} \quad (15)$$

Matrix form is defined by following equations

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} \begin{bmatrix} x_s \\ y_s \end{bmatrix} = \frac{1}{2} \left\{ R_s^2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} K_1 - r_1^2 \\ K_2 - r_2^2 \end{bmatrix} \right\} \quad (16)$$

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}^{-1} \frac{1}{2} \left\{ R_s^2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} K_1 - r_1^2 \\ K_2 - r_2^2 \end{bmatrix} \right\} \quad (17)$$

$$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (18)$$

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}^{-1} \begin{bmatrix} K_1 - r_1^2 \\ K_2 - r_2^2 \end{bmatrix} \quad (19)$$

$$x = \begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} b_1 & a_1 R_s^2 \\ b_2 & a_2 R_s^2 \end{bmatrix} \quad (20)$$

where R_s is the source coordinate.

Equation (21) is obtained by inserting (20) into (15)

$$(a_1^2 + a_2^2)R_s^4 + 2(a_1 b_1 + a_2 b_2)R_s^2 + b_1^2 + b_2^2 = 0 \quad (21)$$

The solution to R_s^2 is

$$R_s^2 = \frac{c_b \pm c_c}{c_a} \quad (22)$$

where

$$c_a = a_1^2 + a_2^2 \\ c_b = 1 - a_1 b_1 + a_2 b_2 \\ c_c = \sqrt{(1 - a_1 b_1 + a_2 b_2)^2 - (a_1^2 + a_2^2)(b_1^2 + b_2^2)} \quad (23)$$

Positive root means the square of the distance from the source to an origin. The final source coordinate can be found by using R_s calculation in (20).

IV. SOUND DELAY ESTIMATION

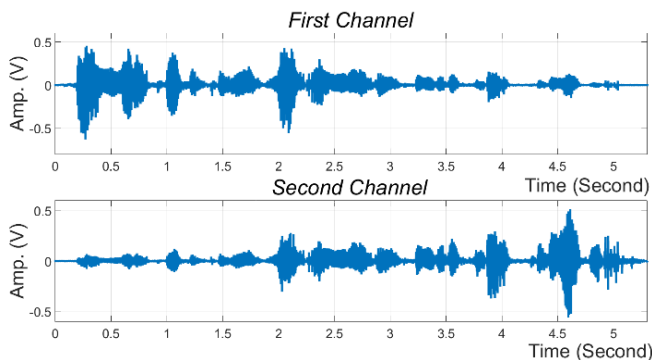


Fig. 5. Sound signals

In this section, trajectory estimation of sound sources and determination of the coordinates are explained. The dual-

microphone source location method is used in 2-D space. This method is realized by database created artificially. The ambient temperature is selected as $20\text{ }^\circ\text{C}$ for this work. Stereo sound recording is shown in Fig. 5.

The sliding window techniques are used for energy and delay estimation of sound channels. The width of sliding window and step size are selected as 1024 sample and 10, respectively. Delays between sound channels are calculated by GCC algorithms [25]. Two different types of GCC algorithms which are basic GCC and Generalized Cross Correlation with Phase Transform (GCC-PHAT) are used in this paper [26-27]. The GCC of the signals recorded by two microphones is given by

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \psi_{ij}(w) X_1(w) X_2(w) e^{jw\tau} dw \quad (24)$$

where $X_1(w)$ and $X_2(w)$ are the signals recorded by one and two microphones in Fourier domain and w is angular frequency. The weighting function $\psi_{ij}(w)$ is designed to optimize the given performance criteria. Many different types of weighting function are used in the literature. The most common one is Phase Transform (PHAT) which is defined in (25).

$$\psi_{ij}(w) = \frac{1}{|X_1(w)X_2^*(w)|} = \frac{1}{|X_1(w)||X_2(w)|} \quad (25)$$

Three different types of signal smoothing filter such as the moving average filter are utilized for increasing the success of delay estimation in this study. The moving average filter is shown in (26). The parameters including $x[i]$, $y[i]$, and M refer to the input signal, the output signal, and the number of points used in the moving average filter, respectively [28]. The weighted moving average filter is also given in (27).

$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} x[i+j] \quad (26)$$

$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} w_j x[i+j] \quad \sum_{i=1}^M w_i = 1 \quad (27)$$

V. ESTIMATION OF SOUND SOURCE TRAJECTORY

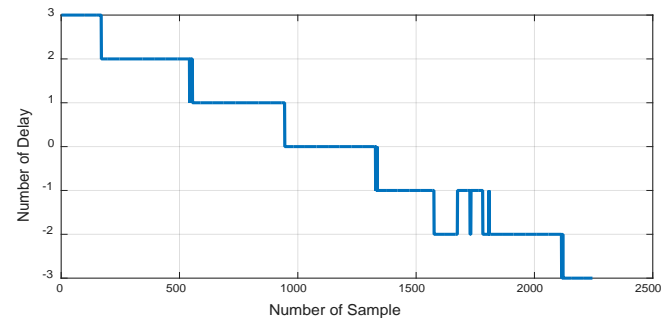


Fig. 6. Calculated delays

The estimation of sound trajectory is mentioned in this section. The implementation of our approach consists of several steps. Firstly, noise removal filters are applied to raw sound recordings. After that, the estimation of delays between

sound channels is realized by sliding windows. The estimated delays are converted to the angle by using (2). First order polynomial is fitted to determine the angle of linear movement of sound source versus time. The calculated delays and azimuth angle of the sound signal are shown in Fig. 6 and 7, respectively.

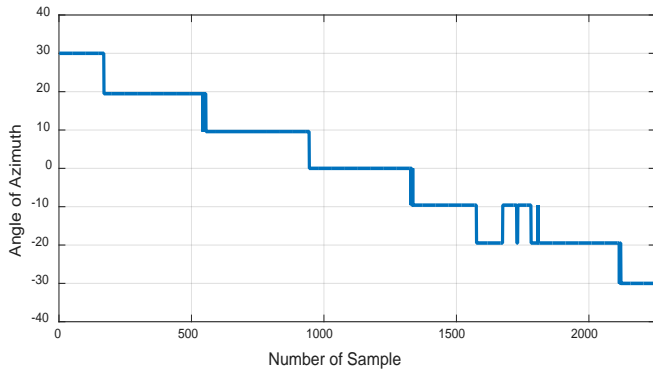


Fig. 7. Azimuth angle of movement

The first order polynomial interpolation result of azimuth angle is illustrated in Fig. 8.

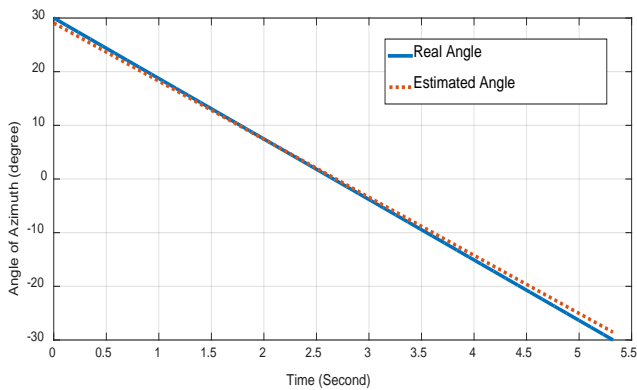


Fig. 8. Comparison of real and estimated azimuth angle

By using dual-microphone source location method, the estimation and curve fitting change of X and Y coordinate values versus time are shown in Fig. 9 and 10, respectively.

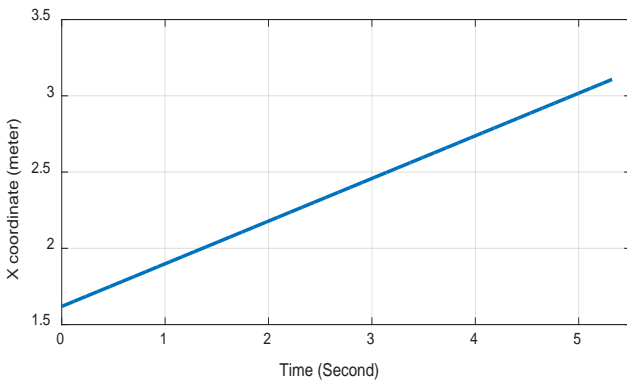


Fig. 9. X coordinate change versus time

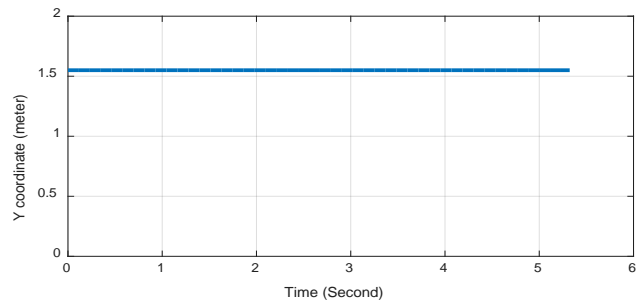


Fig. 10. Y coordinate change versus time

The 2D coordinate map of sound source motion is shown in Fig. 11. As seen in the figure, the estimated motion of sound source is from A to B by using 2D sound source localization approach.

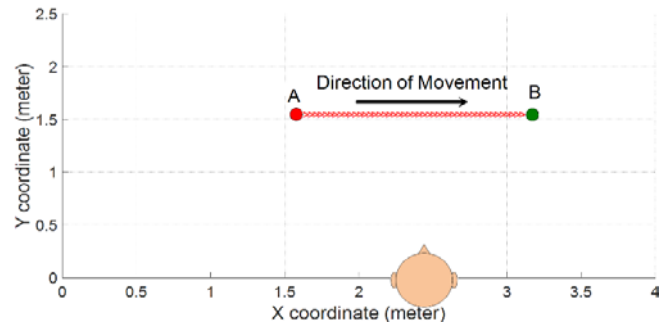


Fig. 11. Direction of motion on coordinate map

The mean calculation error of azimuth angle versus different algorithms is shown in Table I. The order of filters is selected as $N=5$ for the whole process.

TABLE I. CALCULATION OF MEAN ERROR OF THE AZIMUTH ANGLE

GCC Method/Filter	Moving Average	Moving Weighted Window	Median Filter
Unfiltered Correlation	0.053	0.0372	0.0257
PHAT Filter	0.1763	0.7568	0.0641

As shown in Table I, the best combinations of GCC method and pre-filter are generic GCC and Median filter. The mean difference between the estimated and values of the real coordinates is 8.544 cm.

VI. CONCLUSION

In this paper, we propose a different approach for sound source localization which is 2D sound localization microphone pair in contrast to common approaches. Some basic parameters of sound signal processing are explained and they give an idea about sound signal localization, azimuth angle determination and exact localization of sound sources. This method is related not only delay between sound channels but also the energy ratio between recording sound signals. The 2D sound localization using microphone pair approach is implemented for sound signal created artificially. The result of calculated

azimuth angle is compared with different GCC and noise removing filters methods. Mobile sound source trajectory depended on the time is also calculated. Furthermore, curve fitting method is applied for estimation of the sound source motion. The energy and delay parameter are also used for calculating the coordinates of the sound source. The optimal combination of methods is examined for sound source localization. The coordinates of mobile sound source are calculated very precisely and successfully depending on the time. It is seen that the azimuth angle for linear motion of sound sources is calculated in a minor error. In future work, real time sound localization application with similar approach will be performed by embedded systems such as Raspberry Pi.

REFERENCES

- [1] L. A. Jeffress, "A place theory of sound localization," *Journal of comparative and physiological psychology*, vol. 41, no. 1, 1948.
- [2] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual Review of Psychology*, vol. 42, no. 1, pp. 135-159, 1991.
- [3] S. K. Mitra and K. Yonghong, "Digital signal processing: a computer-based approach," *New York: McGraw-Hill*, vol. 2, 2006.
- [4] S. L. Gay and J. Benesty, "Acoustic signal processing for telecommunication," *Springer Science & Business Media*, vol. 551, 2012.
- [5] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems*, vol. 27, no. 4, pp. 199-209, 1999.
- [6] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition," *INTERSPEECH*, pp. 1013-1016, 2001.
- [7] Y. Liang, Z. Cui, S. Zhao, K. Rupnow, Y. Zhang, D. L. Jones, and D. Chen, "Real-time implementation and performance optimization of 3D sound localization on GPUs," *IEEE Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 832-835, 2012.
- [8] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3D sound localization system with miniature microphone array for virtual reality," *7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2012.
- [9] L. Battista, E. Schena, G. Schiavone, S. A. Sciuto, and S. Silvestri, "Calibration and uncertainty evaluation using monte carlo method of a simple 2D sound localization system," *IEEE Sensors Journal*, vol. 13, no. 9, pp. 3312-3318, 2013.
- [10] A. Deleforge and H. Radu "2D sound-source localization on the binaural manifold," *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1-6, 2012.
- [11] S. Aoki, T. Masayoshi, and T. Norihisa, "Sound localization of stereo reproduction with parametric loudspeakers," *Applied Acoustics*, vol. 73, no. 12, pp. 1289-1295, 2012.
- [12] B. C. J. Moore, "An introduction to the psychology of hearing," *Academic Press*, 5th edition, 2003.
- [13] N. Birbaumer and R. F. Schmidt, "Biologische psychologie," *Springer*, 4th edition, 1999.
- [14] Frithjof Hummes, Hendrik Buschmeier. 2005. Stereo Perception and Sound Localisation. [ONLINE] Available at: <http://buschmeier.org/bh/study/soundperception/>. [Accessed 22 June 2016].
- [15] R. H. Gifford, D. W. Grantham, S. W. Sheffield, T. J. Davis, R. Dwyer, and M. F. Dorman, "Localization and interaural time difference (ITD) thresholds for cochlear implant recipients with preserved acoustic hearing in the implanted ear," *Hearing research* 312, pp. 28-37, 2014.
- [16] M. Kyweriga, W. Stewart, C. Cahill, and M. Wehr, "Synaptic mechanisms underlying interaural level difference selectivity in rat auditory cortex," *Journal of Neurophysiology*, vol. 112, no. 10, pp. 2561-2571, 2014.
- [17] X. R. Xiong, F. Liang, H. Li, L. Mesik, K. Zhang, D. B. Polley, and L. I. Zhang, "Interaural level difference-dependent gain control and synaptic scaling underlying binaural computation," *Neuron* vol. 79, no. 4, pp. 738-753, 2013.
- [18] V. Benichoux, M. Stimberg, B. Fontaine, and R. Brette, "A unifying theory of ITD-based sound azimuth localization at the behavioral and neural levels," *BMC Neuroscience*, 2013.
- [19] C. P. Brown and O. R. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 476-488, 1998.
- [20] N. Bilaniuk and SK W. George, "Speed of sound in pure water as a function of temperature," *The Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1609-1612, 1993.
- [21] A. Pourmohammad and M. A. Seyed, "N-dimensional N-microphone sound source localization," *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1-19, 2013.
- [22] F. Gustafsson and G. Fredrik, "Positioning using time-difference of arrival measurements," *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'03*, vol. 6, 2003.
- [23] G. Mellen, M. Pachter, and J. Raquet, "Closed-form solution for determining emitter location using time difference of arrival measurements," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 3 pp. 1056-1058, 2003.
- [24] W. Cui, Z. Cao, J. Wei, "DUAL-Microphone source location method in 2-D space," *Toulouse in Proceedings of the ICASSP*, pp. 845-848, 2006.
- [25] GC Carter, "Time delay estimation for passive sonar signal processing," *IEEE T Acoust. S.*, pp. 462-470, 1981.
- [26] J. Velasco, M. J. Taghizadeh, A. Asaei, H. Bourlard, C. J. Martín-Arguedas, J. Macias-Guarasa, and D. Pizarro, "Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2669- 2673, 2015.
- [27] B. Kwon, P. Youngjin, and P. Youn-sik. "Analysis of the GCC-PHAT technique for multiple sources," *IEEE International Conference on Control Automation and Systems (ICCAS)*, 2010.
- [28] S. W. Smith, "The scientist and engineer's guide to digital signal processing," 1997.

Proposed Bilingual Model for Right to Left Language Applications

Farhan M Al Obisat
Computer and Information
Technology Dept.
Tafila Technical University
Tafila, Jordan

Zaid T Alhalhouli
Computer and Information
Technology Dept.
Tafila Technical University
Tafila, Jordan

Hazim S. AlRawashdeh
Department of Computer Science
Buraydah Private Colleges
Buraydah, Saudi Arabia

Abstract—Using right to left languages (RLL) in software programming requires switching the direction of many components in the interface. Preserving the original interface layout and only changing the language may result in different semantics or interpretations of the content. However, this aspect is often dismissing in the field. This research, therefore, proposes a Bilingual Model (BL) to check and correct the directions in social media applications. Moreover, test-driven development (TDD) For RLL, such as Arabic, is considered in the testing methodologies. Similarly, the bilingual analysis has to follow both the TDD and BL models.

Keywords—software; testing; languages; right to left; development; application; bilingual; social media

I. INTRODUCTION

Test-driven development (TDD) comprises the major component of the values of the Agile Manifesto and the agile development drives from Extreme Programming (XP). However, TDD is not original. In fact, TDD was mention in the NASA Project Mercury, which was launch in the 1960s [1]. Some encouraging properties are reported to be attainable with the use of TDD. Moreover, while it is often considered a testing method, TDD is also design and development method where in tests already written before the code to ensure an error-free code.

In TDD, tests are added to the code. Then, this is restructured to achieve better internal structure as soon as the test is successfully passed. Usually, this process is iterated several times until all functions are fully verified to be well implementing.

Any software development process encompasses the following main activities:

- 1) problem analysis (specification),
- 2) Software design,
- 3) Software implementation,
- 4) Software testing,
- 5) Software maintenance, and
- 6) Software operation.

The TDD consists of the following six basic steps:

- 1) Writing a test meant for a part of functionality,
- 2) Running the tests to check whether the output of the new test would fail,
- 3) Writing codes aims to pass the tests,

- 4) Running the test to check whether they pass or not,
- 5) Code rewriting, and
- 6) Running the tests to check if the rewriting did not alter the external behavior [2].

The first step is known as test writing, and it includes writing a code for the purpose of testing the function or functionality of the tasks. The second stage is performed to confirm whether the test is working correctly (this means that the test should not fail at this point as functionality has yet to be implemented). When the test passes at this phase, the test is incorrect and it needs rewriting and validation. The third step involves writing the code into short segments so that it can effectively pass the test. Finally, all the tests must be run to confirm whether any desired functionality has been implemented. The internal structure of the code should be further improved through rewriting when all tests pass [3].

At this point, researchers perform the TDD test first. To answer the question as to why this must be done first, one should consider the key benefits behind adopting this model. The advantages of the TDD test are listed below.

- 1) TDD test can capture the intent of the developer or domain expert (e.g., about RTL languages)
- 2) It allows thinking about program design.
- 3) It ensures that the tests are written (and real)
- 4) It provides a higher quality code and runs faster because it has fewer integration problems.
- 5) The Ping Pong Pair Programming-style TDD leads to better distribution of knowledge in the team and reduces the “truck factor” (worst case scenario).

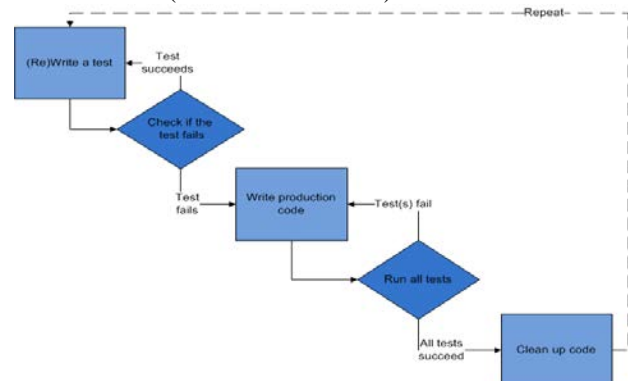


Fig. 1. Test-driven development cycle diagram [4]

TDDs are mostly coded for left to right languages with no consideration for other languages' directions, such as a right to left languages (RLL) or top-down languages, to name a few. In this work, researchers establish the many issues to be used when developing software for RLL applications.

As mentioned earlier, TDD is a software development procedure consisting of short development cycles that are repeated. At first, the developer creates an (initially failing) automated test case, which delineates a targeted improvement (or a new function). Then, the developer creates the minimal number of codes to pass that test. Finally, the new code is refactor to acceptable standards [4].

Software testing is an activity that assesses the features of a code and guarantees that the resulting code meets the essential output. Accurately finding all the errors (semantic or syntaxes) in the program is a difficult task. The choice of a correct approach at the right time will result in an efficient and effective software testing procedure [5].

“Software testing” refers to a process of the running code with the goal of finding errors. This method adapts test case designs into execution steps that are well-planned. This leads to the design and creation of successful software. The goal of software testing is to uncover the possible errors that may be present in the software. Thus, the main goal of a test case is to generate a set of tests with the highest likely hood of uncovering the errors. Also, software testing ensures that the computer code does what it must do. This is similar to a destructive process of identifying errors. The purposes of software testing may include reliability estimation, validation, quality assurance, or verification [6].

Software testing also assesses codes with the purpose of checking out errors within it, Software testing is a method that aims to evaluate a capability or attribute of a code or product and determine whether it satisfies quality requirements. Software testing is similarly employed to test the code for other factors use to assess software quality, such as usability, reliability, maintainability, integrity, capability, efficiency, security, portability, compatibility, and so on [7].

Software production includes developing codes according to assured requirements. Software testing is performed to validate and verify whether the code has been designed to satisfy these specifications [8]. Software testing for RLL apps has yet to be completely investigated by researchers in this field. Hence, this paper tries to establish a software testing approach for RT languages.

In the next sections the paper clarified the problem statement in details such as why bidirectional is important, and why the developers need to test their applications before pilot any developed software. Also, the research described test cases from software Facebook and Bocketcode. In the last section the Proposed Bilingual Model (BL Model) where discussed in detail.

II. PROBLEM STATEMENT

All the TDDs are mostly coded for a left to right languages with no consideration for other language directions, such as

RLL or top-down languages, to name a few. In this study, the work team will focus in TDD for RTL such as Arabic, Hebrew, Farsi, Urdu, and more. Right to left (RTL) text is supported in widespread consumer software. Often, this support is explicitly enabled. Thus, mixing RTL with the left to right (bidirectional) text is necessary.

There are many user interfaces (UI) points to consider in dealing with RTL languages. These components involve the following:

- Arrow direction
- Forms
- Text fields
- Dropdown fields (list/menu/jump menu)
- Scrollbars
- Data entry fields
- Checkbox fields
- Radio buttons
- Bulleted and numbered lists
- Buttons
- Labels
- Pocket Code: Bricks and formulas
- Facebook (direction of the post, i.e., arrow)

A. Why perform tests for RTL features?

Several reasons exist as to why tests must be executed for RTL features. These reasons are listed below.

- Developers often have little knowledge about how RTL languages are rendered.
- Later changes in the source code (i.e., refactoring) can result in layout problems in other languages, especially RTL, that are already solved in earlier versions.
- Tests are important in documenting coding decisions relevant to these other languages.
- Automatic tests allow quick locating problems and also hedging against reoccurring bugs (i.e., regression tests).

III. AN EXPERIENCE REPORT (POCKET CODE, FACEBOOK, AND SCRATCH)

Pocket Code is a software use to create applications and games especially for students dealing with school works through their smart phones, Figure 2 Snapshot from Pocket code with the bidirectional feature.

Interface for mathematical operation in Pocket Code where the places of x and y are in the left of the interface but should be in the right position as in Figure 3 Interface and direction of the text. User interface for Pocket Code with mathematical sin() function and the places of x and y are on the left of the interface but should be in the right position figure 4.

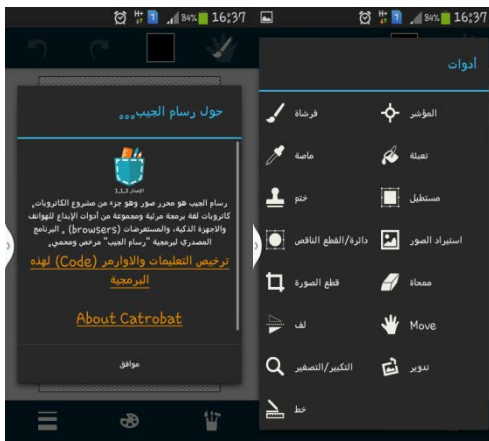


Fig. 2. Snapshot from Pocket code with bidirectional feature

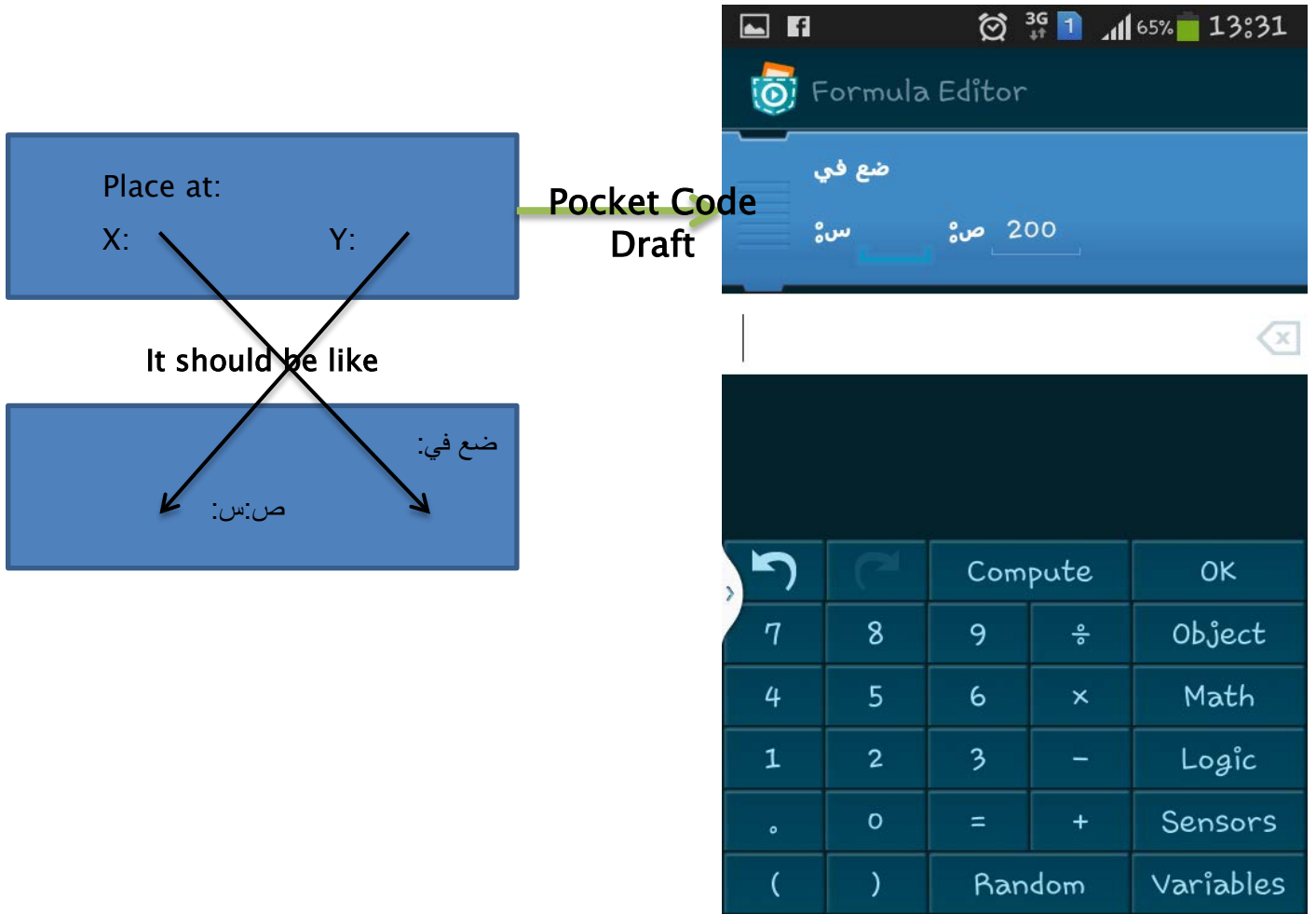


Fig. 3. Interface and direction of the text

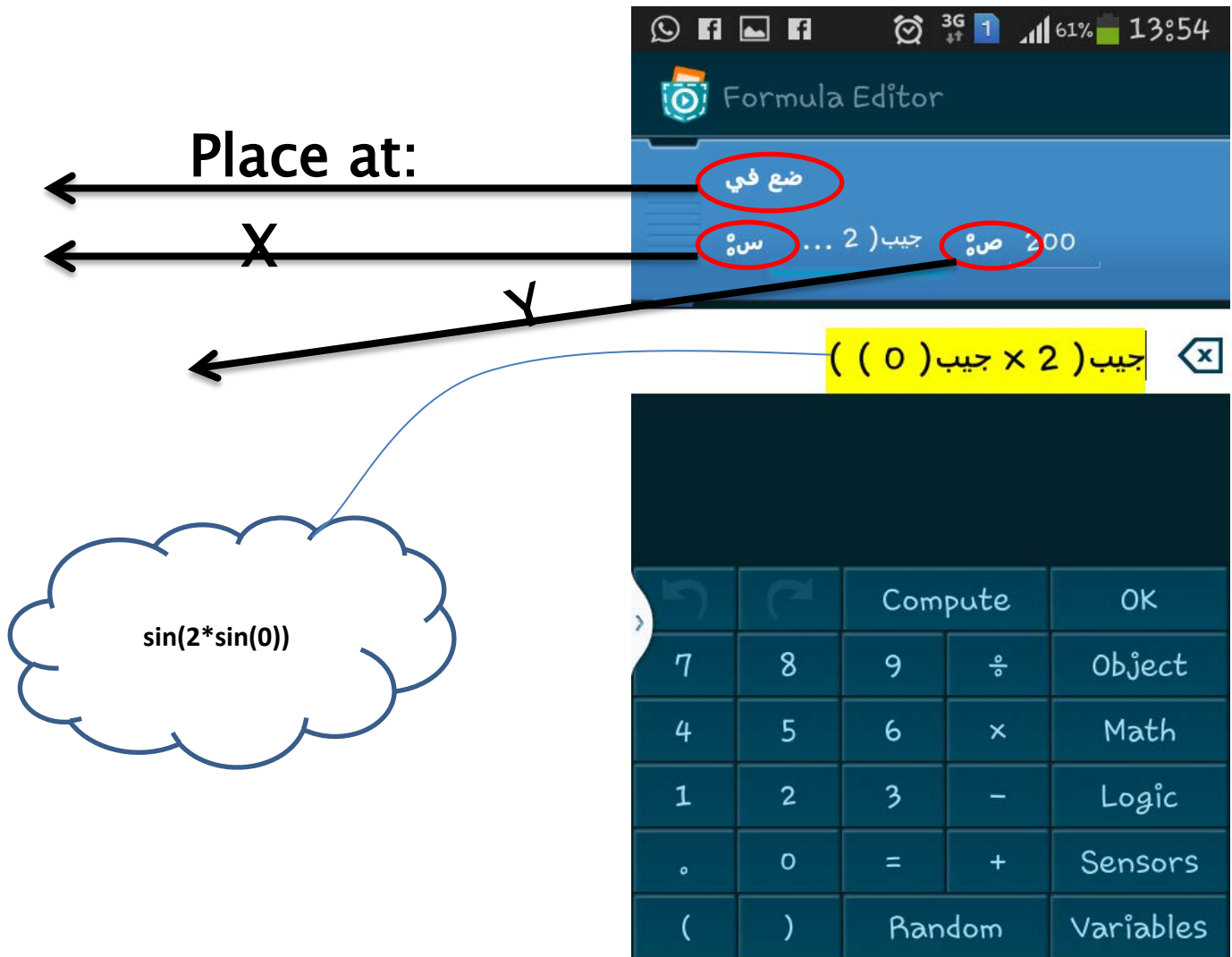


Fig. 4. Interface with mathematical operation

In Figure 5, Pocket Code accepts the equation as it appears in the snapshot from the Pocket Code where there is just one parenthesis opened, and the number of closed ones is three; the places of x and y are on the left of the interface but should be in the right position.

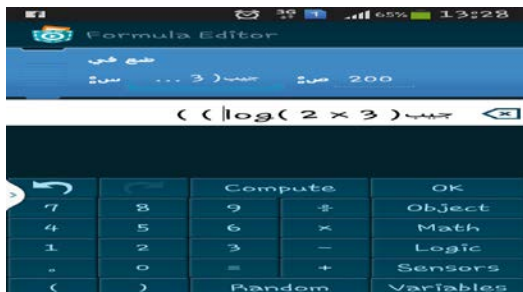


Fig. 5. Pocket Code equation

In figure 6, we can see that the direction of the arrow is wrong for RTL languages, and the position of the photo for the left snapshot should be on the right of the post.



Fig. 6. Snapshots from Facebook

IV. PROPOSED BILINGUAL MODEL (BL MODEL)

In the previous section, the study discussed the RTL problem, which is one of the strongest problems faced by researchers in this area. As shown in Figure 7, the problems appear in the directions of the profile image, sender and receiver information, and the publishing space for users. All these problems come from using different languages that also have different directions. Accordingly, researchers tried to construct a new model to solve this problem. Figure 8 clarifies the main components and steps or sequences of the processes that reduce the bidirectional problems in social media.



Fig. 7. Users posts in Social Media applications before applying the BL Model

The new model consists of nine phases as shown in Figure 8. The first and second phases are related to the language for both applications and user input so it is important to determine whether the languages used in this input, are compatible with the application language. In turn, this can help determine the correctness of the direction of the post. Accordingly, the test of the language compatibility will be conducted in the next phase. If the test succeeds, then no change is required, and the code is clean; otherwise, if the test fails, then there is a problem that must be investigated.

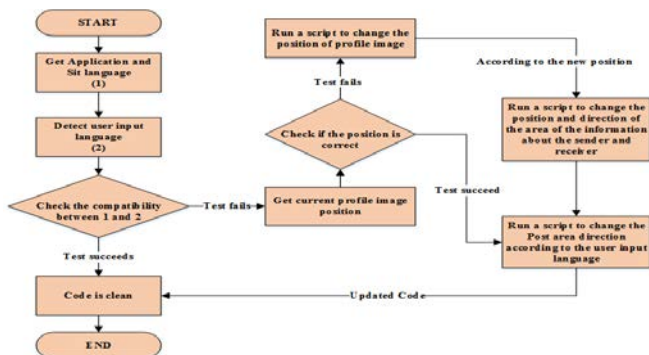


Fig. 8. Proposed Bilingual Model (BL) for social media applications

If a problem appears in the compatibility of the languages, the model must obtain the current location of the profile image and its details. In this stage, the model will retrieve the following information: Division width (DW), image width (IW), offset top (X), and offset left (Y) (see Figure 9). The model checks the validity of the position of the profile image

according to the retrieved information. In case the location of the image is incorrect, the model will apply two different scripts: the first points the new location and the second changes the direction of the post area according to the user input language. Also, the model will proceed to apply the second script directly if the test succeeds. In the end, the interfaces of social media applications are supposed to appear, as shown in Figure 9, 10, after applying the proposed BL Model.

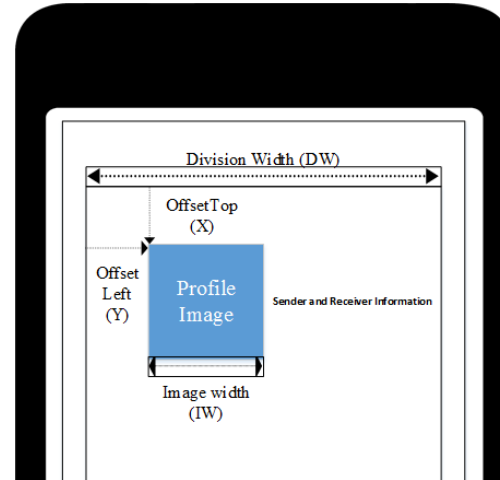


Fig. 9. Profile image details in social media applications



Fig. 10. User posts in social media applications after applying the BL Model

V. FUTURE WORKS

Researchers and software developers in general do not consider testing for RLL applications. This work established new issues concerning RLL applications, which should be considered when testing the newly developed applications. This research work also presents the proposed BL model, which requires more testing and validation. Our future goal is to examine these issues to develop additional techniques to test the RTL or bidirectional language applications.

VI. CONCLUSION

The goal of this research is to establish and consider TDD issues related to RLL apps. Researchers have also shown that there are many points that should be considered when

developing software for RTL or bidirectional languages. In this work, we discuss cases from real working software, such as Pocket Code and Facebook, in which research considered forms, text fields, drop-down fields (list/menu/jump menu), scrollbars, arrow direction, data entry fields, checkbox fields, radio buttons, bulleted and numbered lists, buttons, photo positions, labels, and places of all objects in the interface.

REFERENCES

- [1] http://projekter.aau.dk/projekter/files/204129305/Report_swd903e13_pdf (2014)
- [2] Shrivastava DP, Jain RC. Metrics for Test Case Design in Test Driven Development. International Journal of Computer Theory and Engineering. 2010 Dec 1;2(6):952.
- [3] Kumar S, Bansal S. Comparative study of test driven development with traditional techniques. IntJ Soft Comput Eng (IJSCE). 2013;3(1):2231-307. http://en.wikipedia.org/wiki/Test-driven_development.
- [4] Khan ME. Different forms of software testing techniques for finding errors. International Journal of Computer Science Issues. 2010;7(3):11-6.
- [5] Thakare S, Chavan S, Chawan PM. Software Testing Strategies and Techniques. International Journal of Emerging Technology and Advanced Engineering. 2012.
- [6] Sawant AA, Bari PH, Chawan PM. Software testing techniques and strategies. International Journal of Engineering Research and Applications (IJERA). 2012 May;2(3):980-6.
- [7] Batra S. Improving Quality using testing strategies. Journal of Global Research in Computer Science. 2011 Jul 7;2(6):113-7.

Between Transition from IPv4 and IPv6 Adaption: The Case of Jordanian Government

Iman Akour

Management Information Systems Department
University of Sharjah
Sharjah, UAE

Abstract—IPv6 is being the new replacement for its predecessor IPv4, IPv6 has been used by most Internet services and adopted by most internet architecture these days. Existing protocol IPv4 reveals critical issues such as approaching exhaustion of its address space, continuous growth of the internet and rising new technologies lead to increasing the complexity of the configuration, etc. To healing from the limitations of IPv4 Internet Engineering Task Force (IETF) developed the next generation IP called IPv6. Jordan, like many other countries, is endeavoring to adapt and transit from IPv4 to in an efficient way that will provide an excellent level of service as coveted by its citizens. In this study, the author tried to navigate IPv6 concept from the literature and review the thoughts, steps, and challenges that the Jordanian government pursued in transiting from IPv4 to IPv6.

Keywords—IP networks; IPv6 protocol; IPv6 road map; ipv6 transition; IPv6 adoption

I. INTRODUCTION

Since 1999, the IPv6 is a fact, it isn't a theory anymore. IPv6 forum has appeared as a result of joint efforts of more than 65 companies/organizations worldwide. The main objective of the IPv6 forum is to promote the IPv6 protocol and to educate the market on its benefits and advantages and to deploy its use worldwide. The appearance of IPv6 protocol coincided with the emergence of the document that outlined the IPv6 specification, but unfortunately, even with the exhaustion of IPv4 looming, the extent of IPv6 adoption remains low [1].

Now is the right time for deploying the IPv6 globally, because the IPv4 addresses are becoming rare. There are a lot of efforts around the world to increase broadband penetration; the greater number of smart phones and network-ready devices are entering the market, and the number of Internet users is growing steadily. It is necessary to supply larger number of global IP addresses than the IPv4 pool can to maintain the sustainable, long-term development of an extensive and open Internet.

Worldwide deployment of the IPv6 is vital to the continuous growth and stability of the Internet. Government, business, and technical fields have been cooperating and getting ready to adopt IPv6 since it fulfills the growth of addressing requirements.

Many countries in North America, Caribbean, Latin America, European Union, Middle East, Asia Pacific, and Africa have committed to IPv6 deployment within prescribed

timelines, and many of them have established Task Forces, bringing together stakeholders from the public and private sectors [2].

In this paper, the author's aim is to define IPV6, its stages, identifies its benefits, challenges, Jordan's position compared it other countries in their Internet Protocol usage, as well as the challenges facing the roll out of IPv6 in Jordan.

The author attempt to investigate the status of spreading and using IPv6 in Jordan from telecommunication perspective and customer's intention. Three main hypotheses are formulated and addressed based on three main factors that could have an impact on spreading and moving to IPv6 (i.e., educational qualification, gender, and place of residence).

The study is organized as follows: Section 2 illustrates the readiness of IPv6 in Jordan. Section 3 provides the preliminary investigation which was performed to survey the research area, hypotheses and sampling study. Result and discussion are presented in section 4. The study is concluded in section 5.

II. IPV6 READINESS IN JORDAN

There are many studies in the literature that offer valuable data on the IPv6 adoption process from various perspectives. Jakub Czyz et al. [3] divided the stakeholders into three types: Internet content providers, service providers, and content consumers. They mentioned that these three categories encapsulate the key perspectives should be considered to realistically assess deployment. They found that the regional adoption is not uniform compared to IPv4. Moreover, it was clear that over the last three years, the nature of IPv6 utilization—in terms of traffic, content, reliance on transition technology, and performance—has shifted dramatically from prior findings, indicating a maturing of the protocol into production mode.

One of the important issues must be explored is the migration from IPv4 to IPv6 in Jordan. IPv4 is installed in almost all the companies Internet infrastructures. Therefore, the transition process from IPv4 to IPv6 is challenging and expensive.

Even though several techniques were introduced to avoid the transition or even to postpone it like CIDR, dual stack, tunneling and NAT, the pool of IPv4 addresses is depleting and the only solution is to move towards IPv6 [4].

The collaborative applications on Internet protocol version 4 faced many problems. For instance, there were many

difficulties in MSN Messenger implementation to get it operates with NAT or other IPv6 alternative solutions. Due to this fact, IPv6 has been designed to support both real-time and multimedia applications. Another example of such applications is collaborative gaming which needs a good QoS, because they require the use of audio and video. Some of Microsoft Window versions support IPv6 with the intention of developing complex real-time multiplayer games, VoIP, and IP television [5].

Organizations have to decide whether is it the right time to change or not now. The change is associated with cost and time. If hardware infrastructure does not support the IPv6 then it will have to be replaced. Also, it is necessary to upgrade the software. So, such change requires a transition plan. However, this study looks at the issues that are relevant to a decision to be made towards IPv6 adoption in Jordan.

Figure 1 shows the cumulative IPv6 and assignments in Jordan. It can be seen how the number of new assignments is getting lower in comparison for 2013 and 2014.

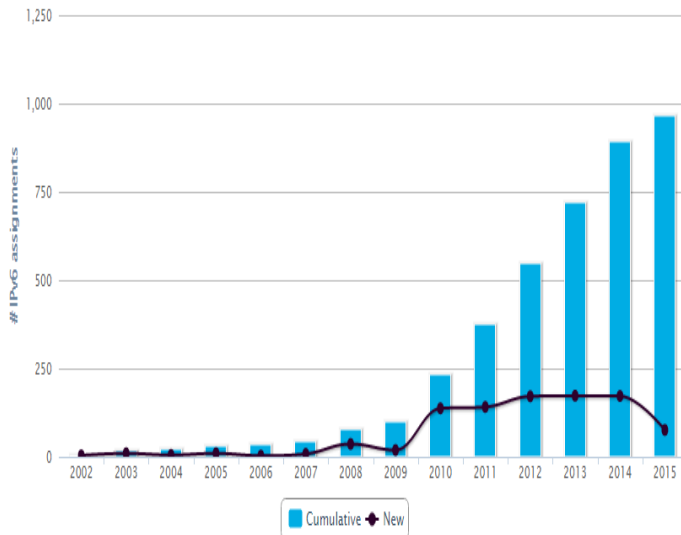


Fig. 1. IPv6 assignments in Jordan 2015 [6]

According to IPv6 test [9] report on August 2015, there is 25 internet service providers for IPv6 in Jordan. They expect a slow growth towards 50% IPv4 / 50% IPv6. The report shows the percentage of browsers that default to IPv6 vs. IPv4 when visiting the ipv6-test connection test. Based on the resulted percentage, hopefully, in the distant future, the IPv6 part will grow taller than the IPv4 one. Finally, the report reveals the percentage of browsers that default to IPv6 vs. IPv4 for users that have both v4 and v6 connectivity. Usually, a system will default to v6 when it's available, but in some cases with tunneled connections, v4 stays the default.

III. PRELIMINARY INVESTIGATION

To the best of my knowledge, this investigation never been carried out in Jordan. The study intended to explore the perception of Jordanians toward major issues related to IPv6 by exploring the attitudes of diverse group: IT and telecommunication universities student, telecommunication firms and computer centers employees and developers. The main objectives of the study are:

- 1) To address the level of the existing knowledge about IPv6 within the Jordanian society.
- 2) To explore whether Jordanians are aware of the benefits and the challenges of utilizing IPv6.
- 3) To navigate the impact of three main factors that could influence recognizing and moving to IPv6 (gender, educational qualification, and place of residence).

B. Research Hypotheses

For the purpose of addressing to which extends the knowledge, usage of the IPv6 exists in Jordan, three main Hypotheses are formulated.

- 1) There are statistically significant differences from the gender perspective in the level of knowledge and usage of IPv6 in the Hashemite Kingdom of Jordan.
- 2) There are statistically significant differences from educational qualification perspective in the level of knowledge and usage of IPv6 in the Hashemite Kingdom of Jordan.
- 3) There are statistically significant differences from the place of residence perspective in the level of knowledge and usage of IPv6 in the Hashemite Kingdom of Jordan.

C. Sampling description and Designed Survey

The population of this research study is Jordanian Internet specialist users and developers to control the bias of internet knowledgeable and not knowledgeable users. The study utilized a systematic random sampling of internet specialist users from three institutions: the Jordanian Ministry of Information and Communication Technology (MOICT), National Information Technology Center (NITC), Yarmouk University, Hashemite University, and Albalqa Applied University.

The purpose of using several institutes from different places is to control the main three factors the study is addressing (i.e., gender, educational qualification, and place of residence). The surveys distribution and collection were administered by graduate students at Yatmouk University. As a total of 320 surveys were distributed among the above mentioned institutions. The participation in this study was voluntary, a total number of complete surveys is 300.

Participants were asked to complete 3 pages survey on their experience and knowledge about IPv6. Likert 5-point scale was employed in this study to determine the participants answers of questions (i.e. Strongly agree, Agree, Neutral, Do not agree, and Strongly not agree). Table 1 summarizes the participants in the study. The participants are divided into three categories based on the three studied factors.

TABLE I. PARTICIPANT SUMMARIZATION

Factors	Category	Number
Gender	Male	48
	Female	52
	Total	100
Place of residence	City	54
	Village	46
	Total	100
Educational	Diploma	15
	Bachelor	45
	Master	40
	Total	100

As shown in Table 1 the percentage of male participants is 52%, while the percentage of female participants is 48%. The city participant's percentage was 8% higher than the village percentage. Finally, it seems most of the participants have bachelor degree.

IV. DATA ANALYSIS AND DISCUSSION

In order to judge the reliability of the used scales of the questionnaire, Cronbach's alpha methodology was employed [7] to calculate a coefficient of internal consistency. Reliability Coefficients for Current Situation of using the IPv6 in the Hashemite Kingdom of Jordan is $\alpha = 0.89 \geq 0.7$, which is a high and acceptable value for the purpose of this study.

To answer the main question of this study, to which extend IPv6 is known and used in Jordan; the arithmetic means and standard deviations were used for all questions that measure the level of using IPv6 in Jordan.

The study adopt the following thresholds to judge the degree of the estimate arithmetic average:

- Arithmetic average (less than 2.33) degrees estimate low.
- Arithmetic average (2.34-6.66) degrees estimate medium.
- Arithmetic average (less than 3.67) degrees estimate high.

Table 2 (in Appendix) summarizes the arithmetic mean of the vertebrae, which measure the level of use of IPv6 in the Hashemite Kingdom of Jordan which ranged between 2.70 and 3.80. The employed questionnaire in this study composed of 30 questions.

TABLE II. RESEARCH QUESTIONNAIRE

Number	Question	SMA	The standard deviation	class	Grade
1)	Do you know what is the IPv6	3.42	1.49	14	Medium
2)	Do you think we need to move to IPv6	3.70	0.96	3	High
3)	Do you think the IPv6 enables you to perform actions/functionalities not offered by IPv4?	3.61	1.01	7	Medium
4)	Do you think by moving to IPv6, IPv4 will be obsolete?	3.16	1.10	24	Medium
5)	Are you ready to start using IPv6?	3.48	0.95	10	Medium
6)	Is Kingdom of Jordan starts establishing the required infrastructure and architecture for launching the IPv6?	2.73	1.01	33	Medium
7)	Do you believe that Kingdom of Jordan has already IPv6?	2.89	2.89	31	Medium
8)	Do you prefer establishing a test-bed for IPv6?	3.29	3.29	22	Medium
9)	Do you have the necessary skill to develop applications/ sites work on IPv6?	3.39	0.94	18	Medium
10)	Do you believe that IPv6 will cost more that IPv4?	3.46	0.83	11	Medium
11)	Do you think Kingdom of Jordan does not provide IPv6 to ISPs	3.32	0.89	20	Medium
12)	Do you think ISPs play essential roles in not disseminating Ipv6 in Kingdom of Jordan	3.73	0.93	2	High
13)	Do you thing ISPs in Kingdom of Jordan do not move to IPv6, because the benefit that they gain from using IPv4	3.68	0.93	5	High
14)	Can you deploy and use websites on IPv6 without ISPs intervention?	3.07	1.04	28	Medium
15)	Do you believe that ISPs in Jordan already have IPv6 and it is working?	3.20	0.92	23	Medium
16)	Do you believe that ISPs in Jordan already have IPv6 but they don't activate it	3.38	0.97	19	Medium
17)	Do you know any institute/s use IPv6?	2.70	0.93	34	Medium
18)	Do you have the willingness to participate in any IPv6 workshops in Kingdom of Jordan?	3.80	1.15	11	High
19)	Can you conduct new services on IPv6 while you are still keep using IPv4	3.10	0.85	27	Medium
20)	Do you think the main reason behind not moving to IPv6 yet is the reluctance of Jordanian firms as it needs a lot of works	3.41	1.00	16	Medium
21)	Do you think the main reason behind not moving to IPv6 yet is the reluctance of Jordanian firms as it needs a lot cost	3.46	0.98	11	Medium
22)	Can you collaborate and communicate with the glob if we don't activate the IPv6 in Kingdom of Jordan?	3.15	1.02	25	Medium

23)	Do you thing IT sectors development in Jordan are in parallel with the development of IT sectors in other countries in term of using IPv6	2.84	1.04	32	Medium
24)	Do you think Ministry of IT and communication plays essential role in activating IPv6 in kingdom of Jordan	3.69	0.97	4	High
25)	Will you face any difficulties in moving from IPv4 to IPv6?	3.45	0.85	13	Medium
26)	Do you think using IPv6 will require new security policies and procedures?	3.66	0.82	6	Medium
27)	Do you prefer new services to be offered on IPv6?	3.41	0.88	16	Medium
28)	Do you think Jordan economic is behind not having IPv6 in all state of kingdom of Jordan	3.51	1.16	8	Medium
29)	Do you have any experience or even knowledge of IPv6 services?	3.30	0.96	21	Medium
30)	Do you think Telecommunication Sectors in Jordan is advanced in comparison with other technological countries?	2.99	1.04	29	Medium

Most notably, Question number (18), which states " Do you have the willingness to participate in any IPv6 workshops in Kingdom of Jordan?" got the highest arithmetic mean and high degree. Then, Question number (12) got the second place which achieves a mean (3.73) and high degree, which states: "Do you think ISPs play essential roles in not disseminating Ipv6 in Kingdom of Jordan?". The lowest arithmetic mean was for Question number (17), which states, "Are you aware of companies use Ipv6? ", the arithmetic mean was (2.70) with moderate degree.

In order to check the hypotheses, the average and standard deviations of using IPv6 in Jordan were used by taking into account the three main factors (gender, educational qualification, and place of residence,). The analysis of variance (ANOVA) was utilized in this study to compare the means of responses given by male and female, city, etc. respondents in order to explore differences [8].

TABLE III. AVERAGES AND STANDARD DEVIATIONS OF USING OF IPV6 UPON THREE STUDIED FACTORS

Variable	Category	SMA	The standard deviation
Gender	Male	3.39	0.53
	Female	3.26	0.42
Place of residence	City	3.27	0.49
	Village	3.49	0.38
education	Diploma	3.28	0.44
	Bachelor	3.32	0.55
	Master	3.36	0.35

As shown in table 3 and 4, there were no statistically significant differences at the level of significance ($\alpha \leq 0.05$) in the level of using the IPv6 in the Hashemite Kingdom of Jordan in corresponding with gender and educational qualification factors (i.e., F value doesn't reach to the level of statistical significance (0.05). Therefore, hypotheses 1 and 2 are rejected.

From table 3 and 4, there were statistically significant differences at the level of significance ($\alpha \leq 0.05$) in the level of using the IPv6 in the Hashemite Kingdom of Jordan in corresponding with place of residence factor. Village arithmetic mean was 3.49, while the arithmetic mean for city factor was 3.27. The value of (f) for Place of residence factor was 4.605 at the level of statistical significance (0.03), which is less than 0.05. Therefore hypothesis number 3 is accepted.

TABLE IV. STATISTICAL SIGNIFICANT COMPARISON IN TERMS OF THREE FACTORS

Contrast source	Sum of squares	Degrees of freedom	Average squares	F	Statistical significance
Gender	0.316	1	0.316	1.439	0.233
Place of residence	1.010	1	1.010	4.605	0.034
educational	0.373	2	0.186	0.850	0.431
The error	20.832	95	0.21		
Total debugger	22.393	99			

V. CONCLUSION AND FUTURE WORKS

The IPv6 as a protocol and technologies starts to march out across the world because of its exaggerated addressing space. Therefore, Internet addresses with IPv6 will meet the internet demand for the expected future, unlike the existing and exhausted IPv4 address space. As a result, the world has already start adopting IPv6 addresses, for instance Google and Facebook officially adopted IPv6 since 2012. Organizations worldwide are tirelessly planning for the migration. However, some organizations are slower than others to make the switch. Even though, many private and government bodies are moving to IPv6 and all major modern organizations have made the switch, still few challenges has to be addressed in order for governments to move forward towards IPv6 adoption. In this study, there is emphasizes on the need to migrate from IPv4 to IPv6 and investigate the factors that are slowing the migration process in Jordan. The conducted study reveals that Jordan people who live in villages are not very aware of the new internet protocol advantages compared to those lived in cities. This clearly indicates that the awareness programs for IPv6 among Jordanian people are targeted programs for certain sectors from the industry such as the telecommunication players who in turns keeping the granted IPv6 blocks until their allocated IPv4 addresses are sold out.

REFERENCES

- [1] Jari Arkko et al., "The Seven Stages of IPv6 Adoption", IETF 74 IPv6 PANEL, IETF Journal, June 2009, Volume 5, Issue 1, pp 14-17.
- [2] NRO Statistics: <https://www.nro.net/>, (accessed 10/05/2016)

- [3] Jakub Czyz et al., "Measuring IPv6 adoption", ACM SIGCOMM Computer Communication Review - SIGCOMM'14, Volume 44 Issue 4, October 2014, Pages 87-98, ACM New York, NY, USA
- [4] Kalwar, S.; Bohra, N.; Memon, A.A., "A survey of transition mechanisms from IPv4 to IPv6 — Simulated test bed and analysis," IEEE 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), pp.30-34, 3-5 Feb. 2015
- [5] Oxley, A., "Issues affecting the adoption of IPv6," IEEE 2014 International Conference on Computer and Information Sciences (ICCOINS), pp.1-6, 3-5 June 2014
- [6] APNIC, Access online on June 2016. Available <https://www.apnic.net/>.
- [7] Santos, J. R. (1999). Cronbach's Alpha: A tool for assessing the reliability of scales. *Journal of Extension [On-line]*, 37(2). Available at: <http://www.joe.org/joe/1999april/tt3.html>
- [8] Larson, M. G. (2008). Analysis of variance. *Circulation*, 117(1), 115-121.
- [9] <http://ipv6-test.com/>, Access online on June 2016.

A Machine Vision System for Quality Inspection of Pine Nuts

Ikramullah Khosa

Department of Electrical Engineering
COMSATS Institute of Information Technology
Lahore, Pakistan

Eros Pasero

Department of Electronics and Telecommunication
Politecnico di Torino
Torino, Italy

Abstract—Computers and artificial intelligence have penetrated in the food industry since last decade, for intellectual automatic processing and packaging in general, and in assisting for quality inspection of the food itself in particular. The food quality assessment task becomes more challenging when it is about harmless internal examination of the ingredient, and even more when its size is also minute. In this article, a method for automatic detection, extraction and classification of raw food item is presented using x-ray image data of pine nuts. Image processing techniques are employed in developing an efficient method for automatic detection and then extraction of individual ingredient, from the source x-ray image which comprises bunch of nuts in a single frame. For data representation, statistical texture analysis is carried out and attributes are calculated from each of the sample image on the global level as features. In addition co-occurrence matrices are computed from images with four different offsets, and hence more features are extracted by using them. To find fewer meaningful characteristics, all the calculated features are organized in several combinations and then tested. Seventy percent of image data is used for training and 15% each for cross-validation and test purposes. Binary classification is performed using two state-of-the-art non-linear classifiers: Artificial Neural Network (ANN) and Support Vector Machines (SVM). Performance is evaluated in terms of classification accuracy, specificity and sensitivity. ANN classifier showed 87.6% accuracy with correct recognition rate of healthy nuts and unhealthy nuts as 94% and 62% respectively. SVM classifier produced the similar accuracy achieving 86.3% specificity and 89.2% sensitivity rate. The results obtained are unique itself in terms of ingredient and promising relatively. It is also found that feature set size can be reduced up to 57% by compromising 3.5% accuracy, in combination with any of the tested classifiers.

Keywords—pine nuts; Image processing; neural networks; feature extraction; classification

I. INTRODUCTION AND BACKGROUND

In recent times, automatic inspection of product good as well as raw ingredients has gained more attention in food industry. Efforts have been made for non-destructive investigation of key ingredients in agriculture and food business. In this context, x-ray imaging has been a preferred technique which lets one examine the ingredient internally without causing any damage to the ingredient itself. It reveals the internal details which allow the presence of worm damage and other defects to be determined in a safe way [1-3]. In nuts selection, the key objective is to reduce the amount of nuts with navel orange worm damage passed to the consumer. For inspection of the ingredient, image processing is a potential

tool for unveiling the hidden damage present inside the ingredient, and to highlight the concealed facts. Work has been carried out in processing the images of food ingredients for identifying the damage present in it [1-4]. However, the efforts made to demonstrate the extraction of individual nutmeats - Regions of Interest (ROI) - are limited, in particular, where an ample image is captured of a large number of ingredients by the x-ray source. In a real time scenario, it is unlikely to activate x-ray source for each individual ingredient, instead, a batch of ingredients can be captured. We demonstrate development of an image processing method which is capable of identifying and extracting image samples of each individual ingredient that passes under the x-ray source, while discarding any external object simultaneously.

Afterward, converting image samples into significant features which hold discriminative properties of the target class is also a vital task. Extensive work has been carried out for classification of agriculture products as well as food ingredients by the use of several kinds of features extracted from their images. Keagy et al. [4] made use of statistical and histogram features for damage detection in pistachio nuts. Guyer et al. [5] used spectral imaging for defect detection in cherries. Park et al. [6] proposed content-based image classification using texture properties and diagonal moment. In extracting significant features for quality inspection of ingredients related to food and agriculture industry, more attention has paid towards texture analysis of images. Considering nuts as food ingredient for quality assessment and sorting, pistachio has been widely used in previous studies [7-11]. Hazelnuts and almonds are also studied; however, pine nuts are rarely reported. One of the reasons is its high cost around the globe which leads to its limited consumption in the food industry. In addition, its size is small which limits it to be graded efficiently and autonomously. We used pine nuts as raw food ingredient to develop a machine vision system for its automated quality inspection using x-ray imaging.

In a classification task, the success rate highly depends upon the suitable selection classifier as well. A classifier identifies objects as one of the target classes by using the features extracted from them. Many classification techniques have been used for quality assessment of goods in food and agriculture industry. Among them, Artificial Neural Network (ANN) has shown potential for resolving problems in estimating a mathematical relationship where some inputs and their corresponding target outputs are known [12-14]. Extensive work has been carried out employing this technique

by using several kinds of features including statistical, spectral, texture and color features [15-23]. Support Vector Machine (SVM) is another state of the art technique, used in binary classification problems [24-25]. It has the capability to separate the linear as well as non-linear data by estimating a hyper plane between classes. SVM classifier has been employed in several different applications [26-30].

In this work, the aim of the study is to propose a machine vision system which is initially capable of the extracting image of each unit ingredient (from source image which encompasses the large number of ingredient), and then classify it as healthy or unhealthy (damaged/diseased). We used raw pine nuts for Quality inspection. The technical goals of the study include developing a method for real-time extraction of unit ingredient (individual nut image) by processing the captured image by an x-ray scanner which is possibly mounted above the feeding belt containing non-overlapped nuts in a serial manner. Then estimation of fewer meaningful features to be extracted from each image sample which show a significant contribution to classification, and propose an appropriate classifier model to efficiently grade the sample ingredient as one of the target class. Figure 1 shows the in-line inspection scenario for quality classification of raw pine nuts.

The rest of the paper is scheduled as: Section 2 describes the image acquisition and the method for individual ingredient extraction. Section 3 includes the features choice and their combinations for classification. Classifiers' detail is presented in section 4. The parameters of performance evaluation and the results are discussed in sections 5 and 6 respectively.

II. MATERIAL AND METHODS

A. X-ray Imaging

Raw pine nuts (Unprocessed) are obtained as samples. For the x-ray imaging, we considered an x-ray image reader machine: FCR (Fuji Computed Radiography) PRIMA (Model: CR-IR 391RU) of FUJIFILM Corporation [31]. It is the Computed Radiography Machine primarily used for medical imaging. With the intension of developing an extraction

method (extraction of invidual nutmeats from the captured x-ray image) only, and due to the limited resources; this imaging solution is adopted on an experimental basis. The reading capability of the device is 10 pixels /mm and processing capacity is up to 29 IPs (Imaging Plates) per hour. The imaging plate of type ST-VI is used with dimension 35×35 cm. A 24 bit, 1760×1760 output image in JPEG (Joint Photographic Experts Group) format is generated using the FCR PRIMA Console workstation. A sample output image with pine raw nuts laid on the image plate is shown in Fig. 2. Later, each sample was carefully marked, manually internally examined and then labeled as healthy or unhealthy.

B. Pre-processing

The x-ray output image is a wider image containing a large number of raw pine nuts. In the Fig. 1, It is apparent that the background intensity is not constant rather there are vertical background stripes in the image with higher light intensity. The problem occurs due to crude environmental illumination conditions. We obtained such an image due to limited resources, however, it is assumed that for a practical scenario, non-overlapped nuts are laid on the feeding belt, and an X-ray source is mounted on top which captures the image of a batch of nuts (see Fig. 1). For pre-processing and independent nutmeat region extraction from the captured image, few useful image processing techniques are employed discussed in the following.

The obtained digital x-ray image (Fig. 1) is a 24 bit RGB image. It is converted to grayscale image by eradicating the hue and saturation information, while holding the luminance [32]. Considering the real-time setup with proper source as indicated in Fig. 1, we show the sample image with few ingredients as shown in Fig. 3a. Figure 3b represents the inverted intensity version. Next, to separate regions of interest from background, the grayscale image is converted to binary image with white color (pixel intensity~255) as the region and black (pixel intensity~0) as background pixel based on global thresholding, where the threshold was selected by trial and error for the current database (see Fig. 3c).

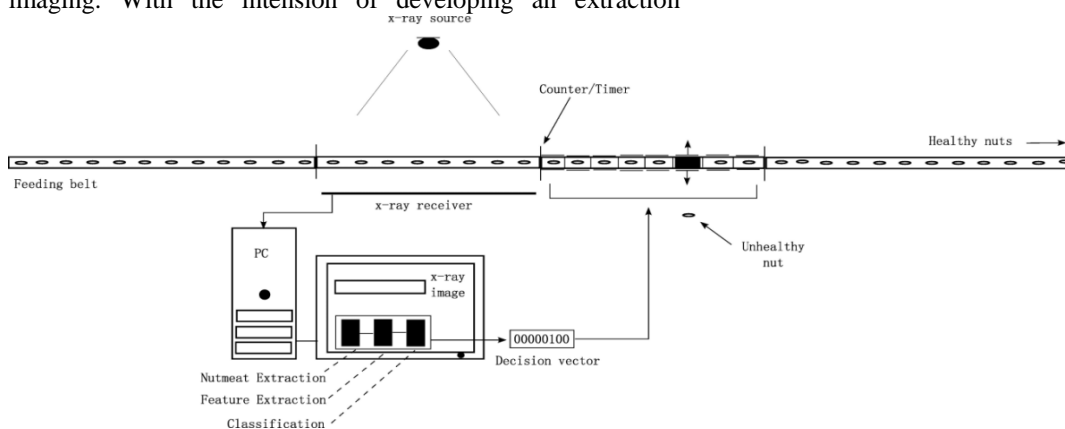


Fig. 1. Proposed schematic setup for selection of nuts with the aid of in-line quality assessment system

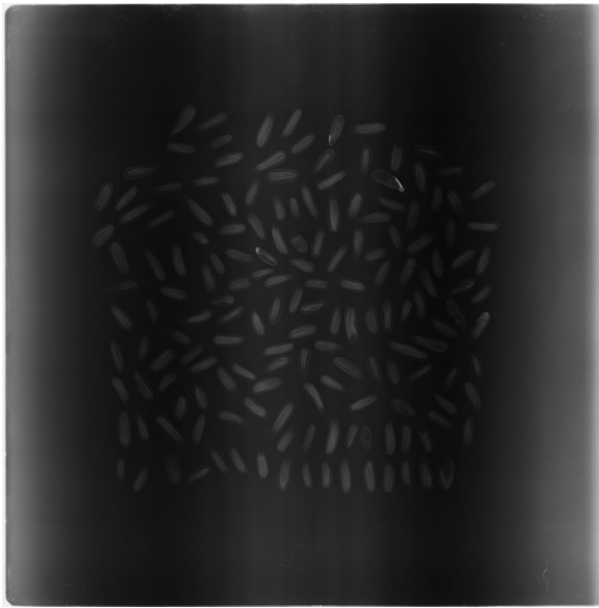


Fig. 2. Sample digital x-ray output image of an image plate containing pine nuts

For all the detected regions in Fig. 3c, the centroids, and the bounding boxes are calculated [32]. Bounding box contains the global coordinates of the top left corner of the region and the rectangular size of the region. These rectangular dimensions are estimated by using the binary image (fig. 3c) with the help of edge detection along the horizontal as well as the vertical axes. Centroids (global coordinates of central pixel of the region) are estimated for each region by calculating the mean of global coordinates of pixels belonging to the region along the horizontal and the vertical axes. Figure 3d shows all the detected regions with their corresponding centroids marked on

them. There are many tiny regions detected which are not our ROIs. To skip them, a median filter mask of size 5×5 is applied. These false regions act as salt and pepper noise, and so many of them were disappeared after applying the median filter. As an additional benefit of this filter, the boundaries of true regions became sharper as shown in Fig. 3e.

C. True Nutmeat Region Estimation and Extraction

Finally, to detect the correct nutmeat regions only, we approximated the size of a single average pine nut image. So an area-based threshold is used; $A = A_n \pm 15\%$, where A_n is the estimated normal nut image area. A region was marked as the true region if it satisfies the threshold criteria or otherwise discarded. The resultant regions with their corresponding centroids can be seen in Fig. 3f.

Figure 3g shows the true nutmeat regions with both corresponding centroids and the bounding boxes. Since the bounding boxes represent the global rectangular measurement of the region, each ingredient is extracted (cropped) from the original image (Fig. 3a) with the help of its corresponding bounding box information. Finally, each individually extracted pine nut sample is shown in Fig. 3 (at the bottom). It is worth mentioning here that each ingredient image is the part of original source x-ray image, and not of the processed image. The processing was done for detection of only real ingredient, and the efficient estimation of their size for successful cropping.

As mentioned earlier, the individual ingredients were manually inspected and labeled as binary label; 0 for healthy, and 1 for unhealthy. Figure 4 shows few image samples representing each of target categories.

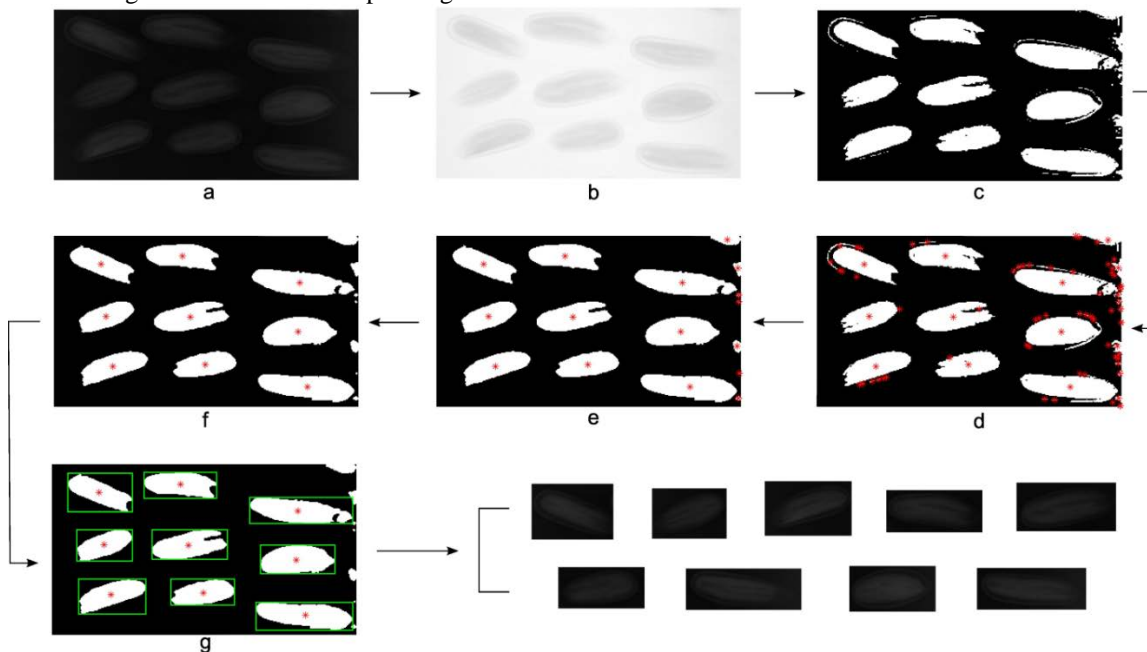


Fig. 3. Image processing steps: from captured collective image to individual nut sample image (a) A sample sub-image (b) negative transformed image (c) binary image after the region of interest based thresholding (d) region detection with marked centroids (e) Regions detection after applying a 5×5 median filter mask (f) Regions detection after applying area threshold (g) True nutmeat region extraction (by cropping) from source image by using respective bounding boxes

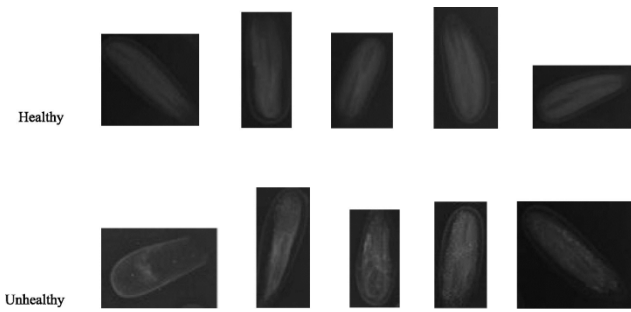


Fig. 4. X-ray image samples of Healthy and Unhealthy pine nuts

III. FEATURE SELECTION

Depending on the nature of task, attention has been paid in mining the meaningful and significant information from the images. The aim is to estimate and compute the properties which hold the discriminating characteristics of the image sample representing the target category.

Regarding quality inspection using images, texture analysis have been extensively exercised as discussed in the introduction section. For this quality assessment task, we chose the features providing the multilevel statistical and texture statistical properties of the sample images.

A. Global statistical features

We calculated a set of statistical features on the global level from each of the sample image. For an image I with highest pixel intensity N , these features are calculated by using the mathematical expressions summarized in Tab. 1. Minimum, maximum, and median are the first order features and represent the corresponding pixel intensities in the image. Mean represents the average pixel intensity and the standard deviation is the measure of average contrast. Variance is calculated as the square of standard deviation. These features are mathematically represented in Tab. 1.

B. Texture statistical features on global level

Next, four characteristics are calculated on global level which represents the texture statistical analysis of image presented in Tab. 2. Smoothness represents the measure of relative softness of the intensity in a region, ranging between 0 and 1. A constant intensity image corresponds to zero smoothness. Third moment determines the skewness of histogram of the image. Uniformity is the measure opposite to smoothness; hence a constant intensity image corresponds to maximum uniformity. Entropy is the statistical measure of randomness. Mathematical expressions for calculation of these features are represented in Tab. 1.

C. Texture statistical features from co-occurrence matrices

In addition to global level characteristics, we extracted features from Gray-Level Co-occurrence Matrices (GLCMs). A Co-occurrence matrix is largely used to measure the texture of an image [33]. The size of this matrix depends on the number of gray levels present in the image. The elements in the GLCM depend on the position operator, which is described by a vector containing direction and distance parameters (also called offset). We calculated four GLCMs from each of sample images using four position operators shown in Fig. 5.

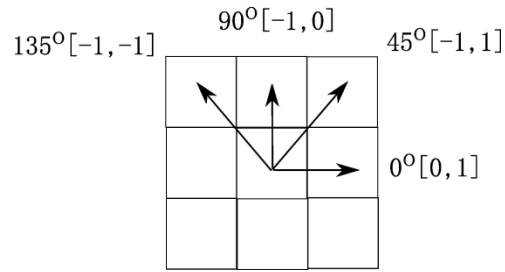


Fig. 5. Position operator to calculate Gray Level Co-occurrence Matrix (GLCM) with angles and offsets

A GLCM for an image I of size $m \times n$ is calculated as follows

$$C_{\Delta x, \Delta y}(i, j) = \sum_{p=1}^n \sum_{q=1}^m \begin{cases} 1, & \text{if } I(p, q) = i \text{ and } I(p + \Delta x, q + \Delta y) = j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where Δx and Δy represent the horizontal and vertical distances.

From each ingredient image sample, four GLCMs are calculated. Next using each GLCM, following four features: contrast, correlation, energy, and homogeneity, are calculated respectively as;

$$c = \sum_{i, j} (i - j)^2 c_{ij} \quad (2)$$

TABLE I. STATISTICAL AND TEXTURE STATISTICAL FEATURE EXTRACTED FROM X-RAY IMAGE SAMPLES ON GLOBAL LEVEL

Ser No	Feature	Expression
1	Minimum	$\min(I)$
2	Maximum	$\max(I)$
3	Median	$\text{med}(I)$
4	Mean	$\mu = \sum_{i=0}^{N-1} a_i p(a_i)$
5	Standard Deviation	$\sigma = \sqrt{\sum_{i=0}^{N-1} (a_i - \mu)^2 p(a_i)}$
6	Variance	$\text{Var} = \sigma^2$
7	Smoothness	$s = 1 - \frac{1}{1 + \sigma^2}$
8	Third Moment	$M_3 = \sum_{i=0}^{N-1} (a_i - \mu)^3 p(a_i)$
9	Uniformity	$u = \sum_{i=0}^{N-1} p^2(a_i)$
10	Entropy	$H = - \sum_{i=0}^{N-1} p(u_i) \log_2 p(u_i)$

$$C = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)c_{ij}}{\sigma_i \sigma_j} \quad (3)$$

$$e = \sum_{i,j} c_{ij}^2 \quad (4)$$

$$h = \sum_{i,j} \frac{c_{ij}}{1 + |i - j|} \quad (5)$$

Where i and j represent pixel intensities and c_{ij} is their count of co-occurrences according to the specified position operator.

D. Features Organization

Concretely, from each image sample, we calculated global statistical features, and statistical texture features on global level as well as from co-occurrence matrices. To assess the significance of different features in classifying the unseen data, we organized these features into six different combinations. These combinations of features are shown in Tab. 2.

Figure 6 shows the flow chart of the entire system including image processing, feature extraction, and classification phases.

IV. CLASSIFIER CHOICE

As discussed earlier in the introduction section, the selection of appropriate classifier for a particular task is a vital part. ANN has become a state of the art choice as nonlinear classifier in recent times due to available improved computing and parallel processing capabilities. We opted two state-of-the-art non-linear classifiers: ANN and SVM for this classification task. The specifications and application details for each of the classifiers are described in the following subsections.

A. Artificial Neural Network

Artificial neural networks are the computing systems, composed of large number of highly inter-connected units (called neurons) that emulate the structure and operation of biological nervous system. There are many types and architectures of neural networks, fundamentally depending on their learning mechanisms. Multilayer Perceptrons (MLPs), also called Multilayer Feed Forward Neural network (MFNN) has an architecture comprised of an input layer, one or more hidden layers and an output layer. Typically, a MFNN with one hidden layer is sufficient to map any kind of linear or non-linear approximation. An example of three-layer neural network architecture is shown in Fig. 7. An MLP operates in two phases: learning and recall. For the learning of MLP,

TABLE II. STATISTICAL AND TEXTURE STATISTICAL FEATURE EXTRACTED FROM X-RAY IMAGE SAMPLES ON GLOBAL LEVEL

	Properties	Feature Set 1	Feature Set 2	Feature Set 3	Feature Set 4	Feature Set 5	Feature Set 6
Global Statistical Features	Minimum			×		×	
	Maximum			×		×	
	Median			×		×	
	Variance			×		×	
	Mean	×		×	×	×	
	Standard Deviation	×		×	×	×	
Global Texture Statistical Features	Smoothness	×			×	×	×
	Third Moment	×			×	×	×
	Uniformity	×			×	×	×
	Entropy	×			×	×	×
Textures Statistical Features from GLCMs	Features from GLCM (calculated at 0°)		×	×	×	×	×
	Features from GLCM (calculated at 45°)		×	×	×	×	×
	Features from GLCM (calculated at 90°)		×	×	×	×	
	Features from GLCM (calculated at 135°)		×	×	×	×	
	Total Features	06	16	22	22	26	12
	Feature Set Characteristics	Global texture statistical features	GLCM texture features	Global statistical and GLCM texture features	All Texture statistical features	Global and GLCM Statistical & texture features	Fewer Texture statistical features

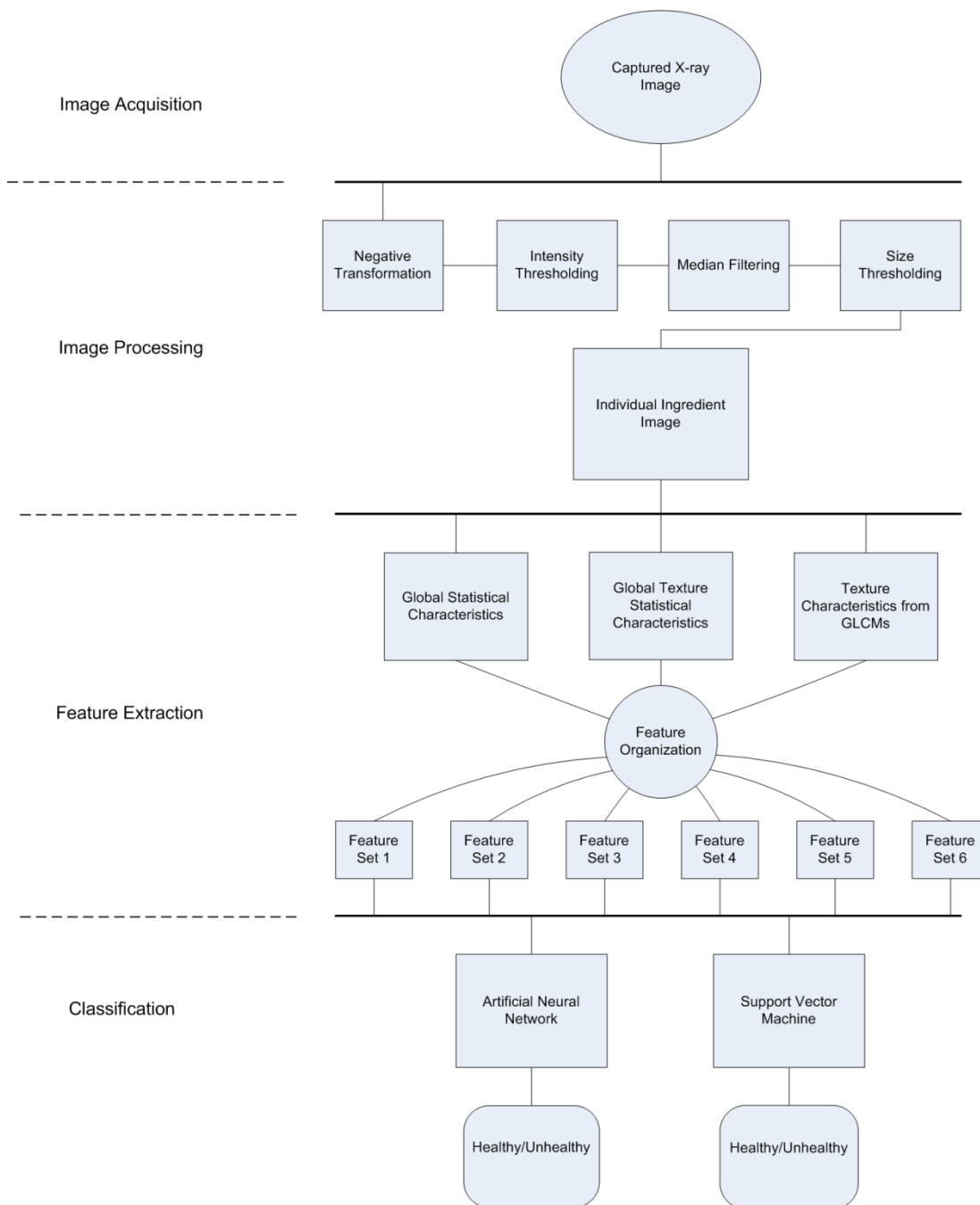


Fig. 6. Flow chart of proposed quality inspection system

special training algorithms have developed based on the learning rules similar to learning mechanisms of biological systems. By using the training data having inputs and corresponding targets, the weights of the classifier associated with the inputs are updated iteratively by the learning algorithm to approximate the target behavior. Back Propagation (BP) algorithm is typically used to update the weights, by minimizing the error function [12]. There are many types of BP algorithm available, and it is desired to select one which best fit the data. We employed the Levenberg-Marquardt (LM) algorithm for the learning [34]. It was designed to come up to second order training speed without computing the

Hessian matrix. The Hessian matrix can be calculated as $H=J^T J$, and the gradient can be computed as $G=J^T e$. Where J is the Jacobian matrix, which contains the first derivatives of network error, and e is the network errors vector.

The iterative update in the weights incorporated by the LM algorithm is calculated as

$$w_{j+1} = w_j - \frac{J^T e}{J^T J + \alpha I} \quad (6)$$

Where w represents the network weights, α is the learning parameter and I is the identity matrix. A large value of α

corresponds to smaller step size in gradient descent approximation and vice versa. It was fixed as 0.9.

For the network training, the above procedure is followed. We selected an MLP with one hidden layer. To estimate the optimized number of hidden layer neurons, we used the cross-validation data. For each of organized feature set as described in Tab. 2, the size of hidden layer is varied, and network performance is repeatedly observed against cross-validation error. The network configuration with the best cross-validation outcome is selected for classification of test data. Table 3 represents the estimated MLP architectures for different feature sets. The number of epochs for training was limited to 100. It was approximated after performing several training sessions.

B. Support Vector Machines

Support Vector Machines (SVM) is a widely used technique which involves supervised learning for binary classification. In this technique, a learning algorithm estimates a plane which separates the data between different classes. SVM have been employed both for linear and non-linear classification problems. It is worth repeating basic concepts of SVM classifier here. For the training data $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ with x_i as feature vector of i th sample and y_i as the corresponding target class, a linear SVM hyper plane fulfills the following conditions;

$$\begin{aligned} x_i \cdot w + b &\geq 1 \text{ for } y_i = +1 \\ x_i \cdot w + b &\leq -1 \text{ for } y_i = -1 \end{aligned} \tag{7}$$

Where w represents the weight vector associated with x_i and b represents the bias value.

In the case of two classes which are linearly non-separable, a suitable function (kernel) is used to transform the input feature space X into another feature space L ($L = f\{X\}$), where it is possible to separate the classes linearly. Figure 8a shows linearly separable data with the hyper planes separating the classes with different margins. Non-separable data can be mapped by a mapping function to higher feature space, and can be separated by a linear hyper plane as shown in Fig. 8b [35].

For the purpose of visualization, we applied Principal Component Analysis (PCA) [36]. PCA is primarily used for data representation in a lower dimension. The first principal component holds maximum variance among features. The second principal component holds the second highest variance and so on. We used first two principal components to produce the features representing data samples (holds ~70 % of the variance of original data). It was observed that the data is not separable in original feature space (the data plots can be seen in figures referred in results section). To classify the data with SVM, we transformed the input feature space to another feature space using Gaussian kernel function given as:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{8}$$

Where σ is the scaling factor in kernel function

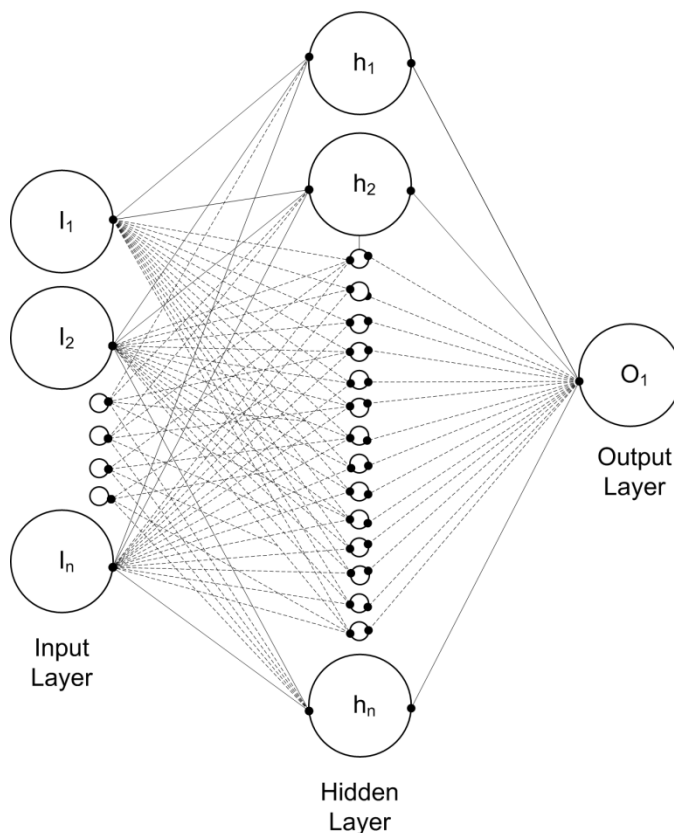


Fig. 7. A typical three layered artificial neural network architecture

Training data is used to train the SVM classifier. After training, the classifier is optimized by using the cross validation data. Two variables were selected to optimize the classifier's performance on cross-validation data: scaling factor of the kernel function " σ ", and " C " to control the soft margin between classes and the hyper plane. We geometrically varied the values for these parameters such that each value for σ is tested in combination with each value of C . Following is the batch represents the options to select the value of these parameters;

$$\text{Batch} = \{.01 .02 .05 .08 .1 .2 .4 .6 .8 1 1.2 1.5 1.8 2 5 10 15 20 30 40 50 70 80 100\} \tag{9}$$

The classifier is optimized by using the cross-validation data, in the following three ways independently:

TABLE III. OPTIMIZED ARTIFICIAL NEURAL NETWORK ARCHITECTURES ESTIMATED FOR DIFFERENT SET OF FEATURES

Artificial Neural Network Architectures	
Feature set	Input layer - Hidden layer - Output layer
1	6-9-1
2	16-12-1
3	22-15-1
4	22-15-1
5	26-20-1
6	12-14-1

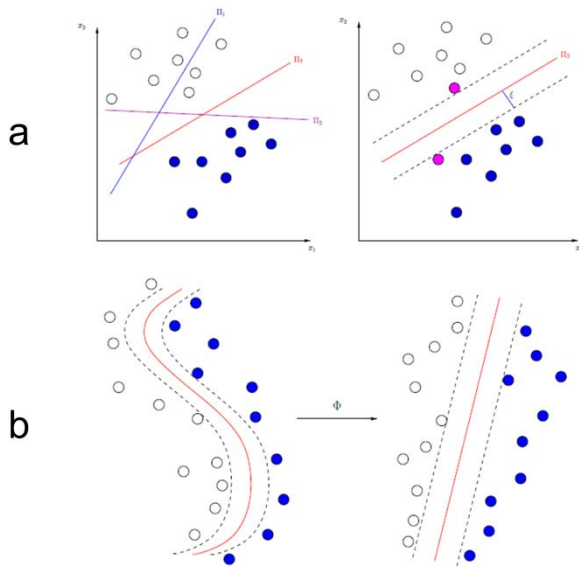


Fig. 8. (a) Left: Example of a linear discriminant analysis using SVM algorithm. Three hyperplanes are shown: π_1 does not separate the two classes, π_2 separates the two classes with a small margin and π_3 corresponds to the best separating hyperplane. Right: Illustration of canonical hyperplanes (black dashed lines), the support vectors in magenta color and the distance between support vectors and best separating hyperplane ξ . (b) Left: Samples which are not linearly separable. Right: Non-linear data mapped to another feature space using the function Φ where linear separation is achievable

- Achieving overall best classification accuracy regardless of individual class accuracy
- Achieving the maximum sensitivity regardless of specificity or classification accuracy
- Achieving best trade-off between specificity and sensitivity while prioritizing the sensitivity

Later with each of the optimized classifier's model, test data is classified and results are recorded. The parameters accuracy, sensitivity and specificity are defined in the following section.

V. PERFORMANCE EVALUATION

Since this is a binary classification task: the data is labeled as one of two categories: image sample of healthy nut (referred as Negative example) and image sample of unhealthy nut (referred as Positive examples). 77 percent of the database is composed of negative examples (631 samples), while the rest 23% (187 samples) of images belong to positive examples. To be used with the classifier, the data of each class is divided as: 70% for training purpose, 15% for cross validation and 15% for test purpose. Training data is used to train the classifier, cross validation data is dedicated to optimize the classifier, and results are calculated on test (unseen) data. To ensure the generalized performance of classifiers, they were trained using randomly selected data. For each set of input features, we rotated the data five times in all three divisions: training, validation and test data. Finally an average of five outcomes of test data is calculated and presented as classification results.

The performance is evaluated regarding parameters defined as;

Accuracy: the proportion of true results (both true positives and true negatives) in the total samples.

Specificity: (or true negative ratio) probability of ingredient recognized as healthy, given that the ingredient was healthy

Sensitivity: (or true positive ratio) probability of ingredient recognized as unhealthy, given that the ingredient was unhealthy

These are calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (10)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (11)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

Where

TP is equal to total number of true positives: the sample is an Unhealthy nut AND the classifier correctly classifies it as Unhealthy

FN is the total number of false negatives: the sample is an Unhealthy nut BUT the classifier incorrectly classifies it as Healthy nut

TN is the total number of true negatives: the sample is a Healthy nut AND the classifier correctly classifies it as Healthy nut

FP is the total number of false positives: the sample is a Healthy nut BUT the classifier incorrectly classifies it as Unhealthy nut

VI. RESULTS AND DISCUSSION

The two classifiers after being optimized by using the cross validation data were used to classify the test data. Their performance is evaluated and discussed in the following section.

A. Artificial Neural Network

Initially, ANN classifier is used for performance evaluation by using each of the organized set of features. At first, global texture statistical features (set 1) are fed to ANN, and results are produced. On average, the network demonstrated 81.8% classification accuracy, with 94.26% correct recognition of healthy nuts. However, the sensitivity rate was less than 50%, which is insufficient. Next, feature set 2 is fed at input layer of its corresponding estimated ANN architecture. The network showed a slight improvement in accuracy, less than 1%. On the contrary, recognition of unhealthy nuts is improved by 13.4%, and so the accuracy was compromised by comparatively lower specificity rate of 89.9%. Texture features from GLCMs (set 2) showed improved performance in recognizing unhealthy nuts, which is desired, but at the same time the size of this feature set - 2.6 times larger than set 1 - is noticeable. Next, the combination of global statistical features and texture features from GLCMs (set 3) is used in order to seek improved network

performance. The network with these features showed enhanced performance than previous outcomes. It showed 59.8% and 92% specificity and sensitivity rate respectively. Hence this combination of features has increased the classification accuracy - 83.7% - but still unsuccessful in achieving a significantly good recognition rate for bad nuts. Feature set 4, which primarily projects the texture statistical characteristics of images and sounds a comprehensively superior choice of features, is then used with ANN. As a result, overall 84.5% accuracy is achieved and 93.28% healthy nuts were correctly identified. Unexpectedly, the sensitivity rate is even lower than the last outcome (since texture statistical features (set 1) seem having more significant information than simple statistical properties (set 3)). Next, we selected feature set 5 (contains all kinds of features extracted from test data) and fed to its corresponding estimated (trained and optimized) ANN architecture. The classification accuracy produced by the network using all the features was 87.6%, with 94.46% correct recognition of healthy nuts. 62% of unhealthy nuts were correctly recognized which was the highest percentage of its kind so far. Hence using feature set 5, the network outperformed all previous outcomes on each evaluation parameter.

The key task in such a binary classification problems is to achieve the good recognition rate for unhealthy nuts while a minor rejection rate of healthy nuts is acceptable. It can be noticed by and large that taking into account the enlarging in feature set on size and information, the progress in classification accuracy is relatively not promising. Keeping this in mind, we tried to estimate a smaller feature set which yet is capable of providing the comparable results as obtained earlier with improved computational efficiency. It can be fairly expected that a small number of significant features may reflect comparable classification results. For this purpose, we selected eight texture statistical features from GLCMs calculated at an angle of 0° and 45°, and, also, four on the global level. These are combined and formed as the feature set 6, which is then fed to the network of its corresponding estimated architecture (see Tab. 3), and results are calculated. The network demonstrated the classification accuracy of 83.9%, as 3.74% less than reported using feature set 5. 90.7 percent specificity rate is

achieved which is 3.72% Less than that of obtained by set 5. On the contrary, the true positive rate (sensitivity) is closer, with a difference of 1.32%.

In general, ANN classifier showed good recognition rate for healthy nuts, but the overall accuracy is poor due to meager performance of the network toward identifying the unhealthy nuts. It is observed that the classifier showed comparable results by feature sets 5 and 6. Hence a trade is perceivable between accuracy and the computational pace. Consequently by using feature set 6 (having less than 50% of information as compared to set 5), the computational efficiency can doubled in terms of feature extraction, and as a whole resulting in more computationally efficient network with small architecture. In this case, less than 2% sensitivity, 3.7% of specificity rate and overall 3.7% accuracy is compromised. Region Operative Characteristic (ROC) curves plotted for test data and total data with different choice of feature sets are shown in Fig. 9.

B. Support Vector Machine

Next we present the performance evaluation of SVM classifier by using different features (as given in Tab. 2) on the test data. As mentioned in section 4.2, the classifier was optimized on cross-validation data in three ways: for best classification accuracy, for best sensitivity, and for best trade-off between Specificity and sensitivity. Note that the SVM classifier was trained using training data, optimized on validation data with the appropriate choice of σ (the scaling factor in Gaussian radial basis function kernel) and C (estimate of the soft margin between the classes), and later tested by using the test data. During optimization process using cross-validation data, it was first optimized for overall accuracy so as to estimate the General performance of the classifier. Secondly, it was optimized to achieve highest sensitivity rate (considering the fact that recognition rate for the unhealthy nut is most important). Finally, it was optimized to achieve maximum sensitivity rate while maintaining a possibly fair specificity rate. These three-way optimizations were achieved by varying the parameters σ and C . The vector having choices of each of these parameters is given in (9). Individually, we employed the feature combinations (sets) given in Tab. 2, and calculated the results by using the test data as well as total data.

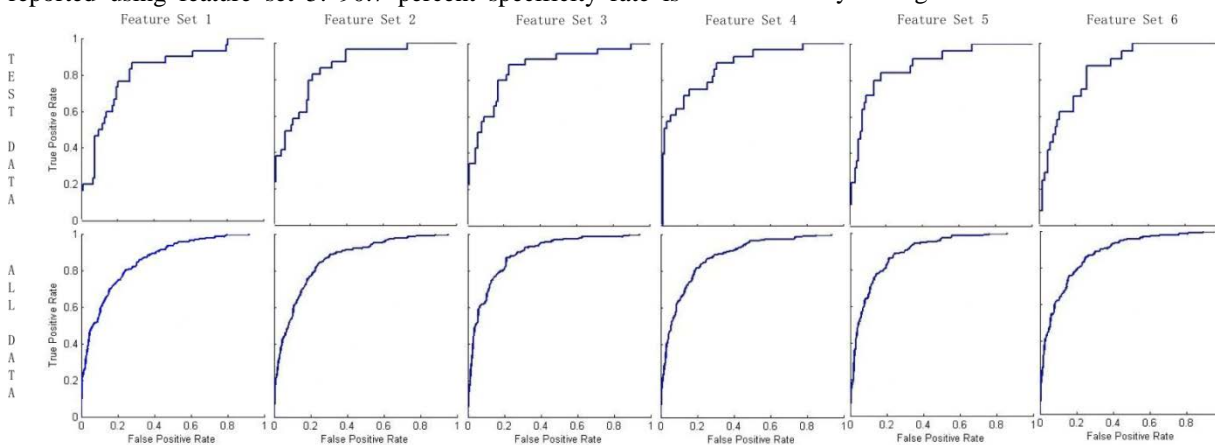


Fig. 9. Receiver Operative Characteristic (ROC) curves produced by artificial neural network classifier for test data as well as total data using different feature sets

TABLE IV. CLASSIFICATION RESULTS OF TEST DATA PRODUCED BY DIFFERENT MODELS OF SUPPORT VECTOR MACHINE CLASSIFIER, ESTIMATED AND OPTIMIZED USING CROSS-VALIDATION DATA FOR ACHIEVING (A) BEST ACCURACY (B) HIGHEST SENSITIVITY RATE (C) BEST TRADE-OFF BETWEEN SPECIFICITY AND SENSITIVITY

Feature Set	Support Vector Machines								
	A			B			C		
	Acc. (%)	Spec. (%)	Sens. (%)	Acc. (%)	Spec. (%)	Sens. (%)	Acc. (%)	Spec. (%)	Sen. (%)
1	75.6	83.7	58.7	45.5	32.6	89.3	77.4	80.3	71.4
2	78.2	85.3	63.2	50.4	36.8	96.4	78.8	81.1	73.8
3	82.5	87.1	72.6	53.6	40	100	81.9	82.3	80.1
4	85.1	89.6	75.5	56.1	43.1	100	83.7	84.2	82.1
5	88.9	94.3	77.6	54.5	41	100	87	86.3	89.3
6	84.1	88.8	74.1	57.7	45.3	100	83.7	83.2	85.7

As stated in section 4.2, to visualize the data separation performed by the classifier hyper plane, we applied PCA to each of feature sets. From each of feature sets, after applying PCA, first two principal components were used as features to represent each example. These features were then used with their corresponding optimized SVM classifier model. The hyperplanes separating the classes, estimated by the classifier optimized for overall best accuracy, are shown in Fig. 10 for different choice of features sets. Similarly, Fig. 11 represents the data separation with hyperplanes created by the classifier optimized for achieving maximum sensitivity rate. The classifier optimized for achieving best trade-off between recognition rate for healthy and unhealthy nuts, separated the data by the hyperplanes shown in Fig. 12 for different feature sets. The results produced by SVM classifier for the test data are summarized in Tab. 4. Considering all-purpose performance, the classifier produced best results with the choice of feature set 5. The classification accuracy equal to 88.9% is achieved having 94.3% and 77.6% correct recognition rate for healthy and unhealthy nuts respectively. On the contrary, if the focus is dedicated towards achieving highest sensitivity rate i.e. by using the optimized classifier for achieving maximum sensitivity, 100% unhealthy nuts can be identified. However, with this model more than 50% of healthy nuts are discarded, which are a huge figure and not a choice. Hence, it is necessary to optimize the classifier fair enough for both the target classes, while being generous towards achieving sensitivity rate. So the best result of SVM classifier considering both the true negative and the true positive rate is achieved when it is optimized for estimating best trade-off between specificity and sensitivity. Using feature set 5, 89.28% Unhealthy nuts were correctly identified while 86.3% healthy nuts were correctly spotted. Figure 13 presents the test data results calculated by three-way optimized classifier.

Comparing the two classifiers regarding individual evaluation parameters, ANN showed good performance in recognizing healthy nuts while SVM demonstrated the true positive ratio relatively better than that of produced by ANN. The finest performance was achieved with the choice of all

extracted features (set 5) for both the classifiers. The relative improvement in accuracy was higher in the SVM case while choosing different features sets. The feature set 6 was estimated containing fewer significant features, to estimate a computationally efficient classifier. SVM proved to be the better choice as a classifier with this set showing 25% higher accuracy for recognizing unhealthy nuts, although the overall accuracy is similar for both classifiers. The comparative results for ANN and SVM with the best trade-off model is given in Tab. 5.

As discussed in the introduction section, pine nuts are rarely reported as the ingredient for quality inspection task. In contrast, pistachio is generally used in studies where nuts are used as ingredients for the quality assessment task. A comparison of results from studies where nuts are inspected as the ingredient is presented in Tab. 6. The table also presents the results of our previous study with pine nuts where a similar but much smaller database of x-ray images was used [37].

VII. CONCLUSION

In this article, binary classification of pine nuts using x-ray Images were presented. X-ray images were obtained by using a commercial x-ray machine on an experimental basis, and later each nut image was labeled by careful manual inspection. Image processing techniques were employed to develop a method which is capable to individually identify and extract the nuts when captured while moving on feeding belt. For features, statistical texture properties were calculated from image samples on the global level as well from co-occurrence matrices. Features were organized in different combinations to estimate their effectiveness for classification. Two state of the art non-linear classifier: ANN and SVM were used for classification. Classifiers were trained on training data, and optimized using cross validation data. The results were calculated on test data in different scenarios with different variants of features. On the whole, SVM performed better by achieving higher recognition rate for unhealthy nuts, while showing similar level of overall accuracy as demonstrated by ANN classifier.

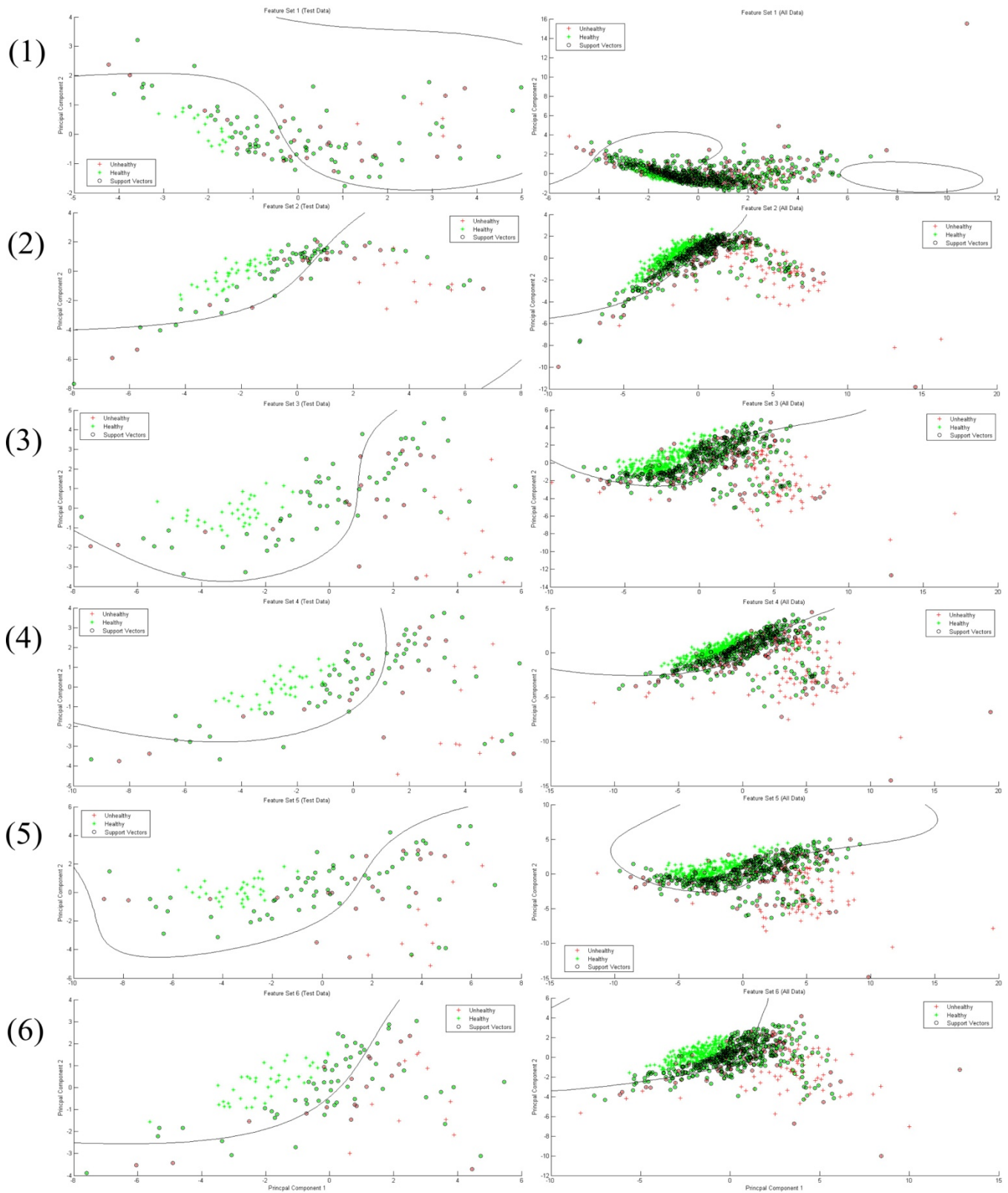


Fig. 10. Classification results of test set data as well as total data by using different feature sets, produced by Support Vector Machine (SVM) model, (trained using training data and) optimized by achieving highest accuracy on cross validation data, $C = 2$, $\sigma = 2$

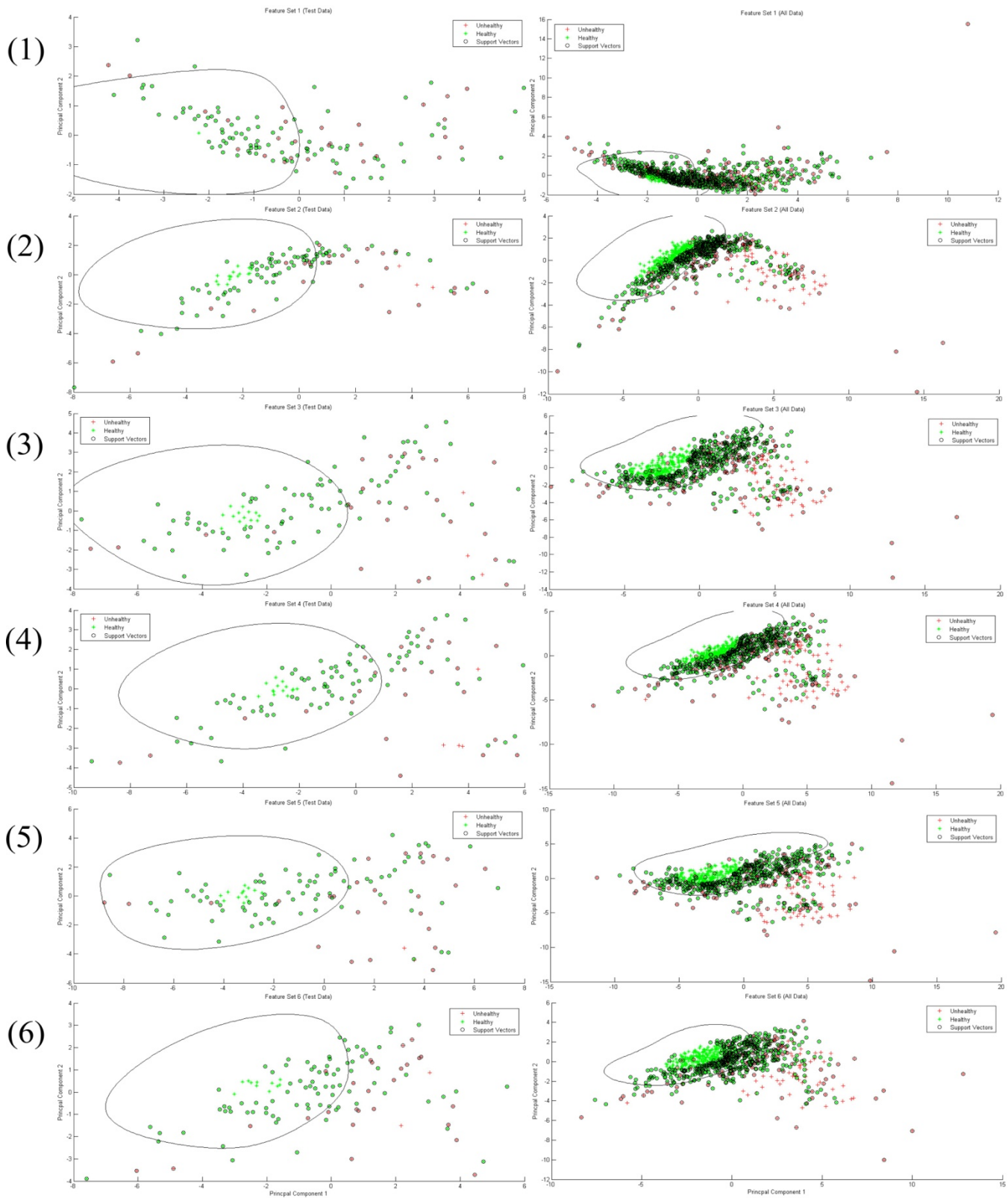


Fig. 11. Classification results of test set data as well as total data by using different feature sets, produced by Support Vector Machine model (trained using training data and) optimized by achieving highest sensitivity rate on cross validation data, $C = 0.1$, $\sigma = 1$

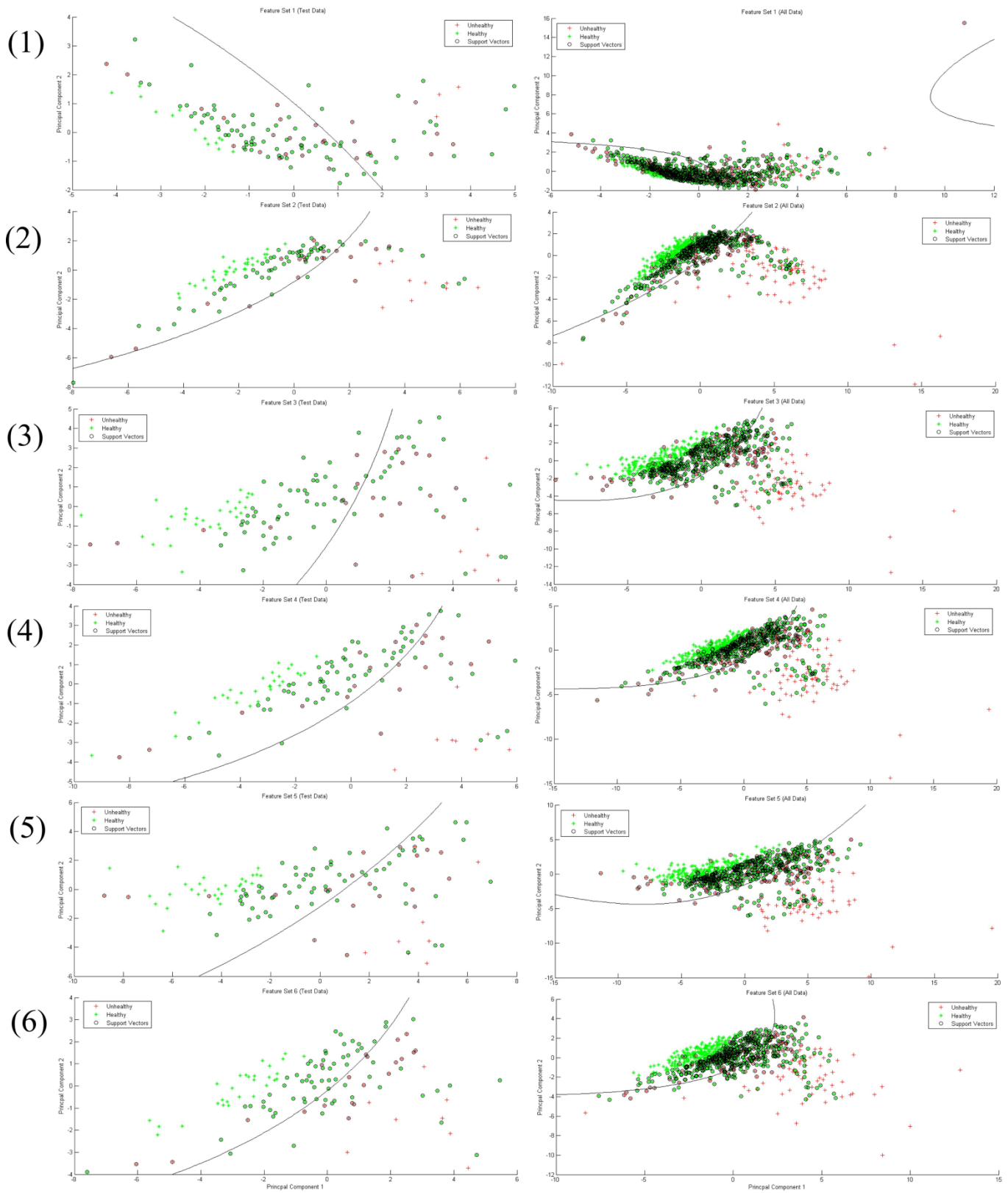


Fig. 12. Classification results of test set data as well as total data by using different feature sets, produced by Support Vector Machine model (trained using training data and) optimized by achieving best trade-off between specificity and sensitivity rate on cross validation data, $C = 25$, $\sigma = 10$

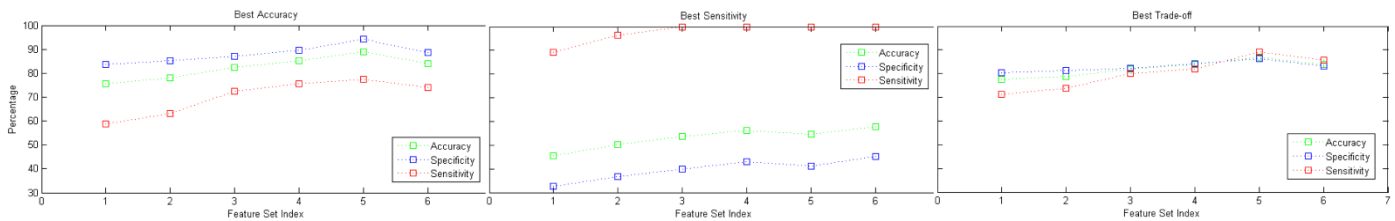


Fig. 13. Comparative test data results produced by SVM models optimized on cross validation data for best accuracy, best sensitivity, and best trade-off.

TABLE V. TEST SET RESULTS OBTAINED USING ARTIFICIAL NEURAL NETWORK AND SUPPORT VECTOR MACHINE (WITH BEST TRADE-OFF OPTIMIZATION)

Feature Set	Artificial Neural Network			Support Vector Machine		
	Accuracy (%)	Specificity (%)	Sensitivity (%)	Accuracy (%)	Specificity (%)	Sensitivity (%)
1	81.8	94.3	42.2	77.4	80.3	71.4
2	82.1	89.9	55.6	78.8	81.2	73.8
3	83.8	92.2	59.8	81.9	82.3	80.1
4	84.5	93.3	51.4	83.7	84.2	82.1
5	87.6	94.5	62.2	87	86.3	89.3
6	83.9	90.7	60.8	83.7	83.2	85.7

TABLE VI. RELATIVE COMPARISON OF RESULTS WITH OTHER INGREDIENT STUDIES

Study	Ingredient	Features	Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)
[4]	Pistachio nuts	Statistical features from Histogram	--	89	99	50
[7]	Pistachio nuts	Histograms and global statistical features	Neural network	88	85.4	90.4
[8]	Pistachio nuts	Statistical features	Distance measure	86	92	80
[37]	Pine nuts (smaller database < 100 samples)	Histogram and Texture statistical features	Logistic regression	98.3	99.1	97.6
Our results	Pine nuts (818 samples)	Texture and statistical features	ANN	87.6	94.5	62.2
			SVM	87	86.3	89.3

REFERENCES

[1] D. S. Narvankar, C. B. Singh, D. S. Jayas, N. G. D White "Assessment of soft x-ray imaging for detection of fungal infection in wheat" Biosys Eng. Vol. 103(1), pp. 49-56, 2009.

[2] R. P. Haff, N. Toyofuku "X-ray detection of defects and contaminants in the food industry" Sens. & Instrumen. Food Qual. Vol. 2, pp. 262-273, 2008.

[3] D. Mery et al. "Automated fish bone detection using x-ray imaging" Jour. Food. Eng. Vol. 105(3), pp.485-492, 2011.

[4] P. M. Keagy et al. "Expanded image database of pistachio x-ray images and classification by conventional methods" Proc. SPIE. 2907, pp.196-204, 1996.

[5] D. Guyer, X. Yang "Use of genetic artificial neural networks and spectral imaging for defect detection on cherries" Comp & Elec. in Agr. Vol. 29, pp.179-194, 2000.

[6] S. B. Park, J. W. Lee, S. K. Kim "Content-based image classification using neural network" Pat Rec. Let. Vol. 25, pp.287-300, 2004..

[7] D. A. Casasent, M. A. Sipe, T. F. Schatzki, P. M. Keagy, L. C. Lee "Neural net classification of x-ray pistachio nut data" Food Science and Technology. Vol. 31(2), pp.122-128, 1998.

[8] T. C. Pearson "Machine vision system for automated detection of stained pistachio nuts" Food Science and Technology, Vol. 29(3), pp.203-209, 1996.

[9] A. Ghazanfari, J. Irudayaraj, A. Kusalik "Grading pistachio nuts using a neural network approach" Trans of ASAE, Vol. 39(6), pp. 2319-2324, 1996.

[10] A. Ghazanfari, J. Irudayaraj, A. Kusalik, M. Romaniuk "A Machine vision grading of pistachio nuts using fourier descriptors" J Agri. Eng. Res. Vol. 68(3), pp.247-252, 1997.

[11] M. Omid, A. Mahmoudi, M. H. Omid, "An intelligent system for sorting pistachio nut varieties" Exp. Sys. App. Vol. 36(9), pp.11528-11535, 2009.

[12] S. Haykin, Neural network: a comprehensive foundation, Prentice Hall, 1999.

- [13] I. A. Basheer, H. Hajmeer "Artificial neural networks: fundamentals, computing, design and applications" J. Microbiological Methods, Vol. 43, pp.3-31, 2000.
- [14] L. Fausett "Fundamentals of neural networks: architectures, algorithms and applications" Eaglewood Cliffs, NJ, USA: Prentice Hall, 1994.
- [15] M. H. Nadian, S. Rafiee, M. Aghbashlo, S. Hosseinpour, S. S. Mohtasebi, "Continuous real-time monitoring and neural network modeling of apple slices color changes during hot air drying" Food and Bioproducts Processing, Vol. 94, pp.263-274, 2015.
- [16] A. Mueen, M. S. Baba, R. Zainuddin "Multilevel feature extraction and x-ray image classification" J. App. Sci. Vol. 7(8), pp.1224-1229, 2007.
- [17] E. Borràs et al. "Data fusion methodologies for food and beverage authentication and quality assessment – A review" Analytica Chimica Acta, Vol. 891, pp.1-14, 2015.
- [18] S. Kumar, V. Kumar, R. K. Sharma "Sugarcane yield forecasting using artificial neural network models" Int. J. Art. Int. & App. Vol. 6(5), 2015.
- [19] H. Pu, D. Sun, J. Ma, J. Cheng "Classification of fresh and frozen-thawed pork muscles using visible and near infrared hyperspectral imaging and textural analysis" Meat Science, Vol. 99, pp.81-88, 2015.
- [20] I. Kavdir "Discrimination of sunflower, weed and soil by artificial neural networks" Comp & Elec. in Agr. Vol. 44(2), pp.153-160, 2004.
- [21] D. Moshou et al. "Automatic detection of yellow rust in wheat using reflectance measurements and neural networks" Comp & Elec. in Agr. Vol. 44(3), pp.173-188, 2004.
- [22] P. M. Granitto, P. F. Verdes, H. A. Ceccatto, "Large-scale investigation of weed seed identification by machine vision", Comp & Elec. in Agr. Vol. 47, pp.15-24, 2005.
- [23] K. Y. Huang "Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features", Comp & Elec. in Agr. Vol. 57 (1), pp. 3-11, 2007.
- [24] C. Cortes, V. N. Vapnik, "Support vector networks", J Mach Learn Res, Vol. 20(3), pp.273-297, 1995.
- [25] I. Steinwart, A. Christmann. Support vector Machines, Springer, 2008.
- [26] X. Petros, P. M. Pardalos, and T.B. Trafalis "Support Vector Machines." Robust Data Mining. Springer, New York, pp.35-48, 2013.
- [27] X. Kong, X. Liu, R. Shi, K. Y. Lee "Wind speed prediction using reduced support vector machines with feature selection." Neurocomputing, Vol. 169, pp. 449-456, 2015.
- [28] L. J. Kao, T. S. Lee, C. J. Lu "A multi-stage control chart pattern recognition scheme based on independent component analysis and support vector machine" Journal of Intelligent Manufacturing, Vol. 27(3), pp. 653-664, 2016.
- [29] Tian, Yingjie, et al. "Nonparallel support vector machines for pattern classification." IEEE transactions on cybernetics, Vol. 44(7), pp.1067-1079, 2014.
- [30] A. T. Azar, S. Ahmed El "Performance analysis of support vector machines classifiers in breast cancer mammography recognition" Neural Computing and Applications, Vol. 24(5), pp.1163-1177, 2014.
- [31] www.fujifilm.com
- [32] Gonzalez RC, Woods RE, Eddins SL. Digital image processing using matlab. New Jersey, USA: Pearson Education, Inc. 2008.
- [33] R. M. Haralick, K. Shanmugam, I. H. Dinstein "Texture features for image classification" IEEE Trans on Sys, Man and Cyb, Vol. 3(6), pp.610-621, 1973.
- [34] J. J. More "The levenberg-marquardt algorithm: implementation and theory" J Numerical Analysis, Vol. 630, pp.105-116, 1978.
- [35] M. Colangeli, F. Rugiano, E. Pasero "Pattern recognition at different scales: a statistical perspective" Chaos, Solitons & Fractals, Vol. 64, pp.48-66, 2014.
- [36] I. Jolliffe. Principal component analysis. Wiley, 2005.
- [37] I. Khosa, E. Pasero, "Pine nuts selection using x-ray images and logistic regression" Proc of World Sym on Comp App & Res, 2014.

Predicting CO₂ Emissions from Farm Inputs in Wheat Production using Artificial Neural Networks and Linear Regression Models

“Case study in Canterbury, New Zealand”

Majeed Safa

Department of Land Management and Systems, Lincoln University, Christchurch, New Zealand

Peter Nuthall

Department of Land Management and Systems, Lincoln University, Christchurch, New Zealand

Mohammadali Nejat

Department of Agricultural of Sciences, Payame Noor University, Saveh, Iran

Bruce Greig

Department of Land Management and Systems, Lincoln University, Christchurch, New Zealand

Abstract—Two models have been developed for simulating CO₂ emissions from wheat farms: (1) an artificial neural network (ANN) model; and (2) a multiple linear regression model (MLR). Data were collected from 40 wheat farms in the Canterbury region of New Zealand. Investigation of more than 140 various factors enabled the selection of eight factors to be employed as the independent variables for final the ANN model. The results showed the final ANN developed can forecast CO₂ emissions from wheat production areas under different conditions (proportion of wheat cultivated land on the farm, numbers of irrigation applications and numbers of cows), the condition of machinery (tractor power index (hp/ha) and age of fertilizer spreader) and N, P and insecticide inputs on the farms with an accuracy of ±11% (± 113 kg CO₂/ha). The total CO₂ emissions from farm inputs were estimated as 1032 kg CO₂/ha for wheat production. On average, fertilizer use of 52% and fuel use of around 20% have the highest CO₂ emissions for wheat cultivation. The results confirmed the ANN model forecast CO₂ emissions much better than MLR model.

Keywords—Artificial neural networks; modelling; CO₂ emissions; wheat cultivation

I. INTRODUCTION

Around the world wheat is used as one of the main food sources to provide a large proportion of the calories and protein needed by human beings [1]. The world wheat production forecasted for 2020 varied depending on the prediction method used: 746 Mt [2], 840 Mt [3] and 1050 Mt [4]. To meet the target 2020 wheat production, the current average wheat yield of 2.7 t/ha needed to be increased by 40% [5]. The three options available to lift wheat production to meet the 2020 target include: expansion of cultivated land, intensification of cultivated land and increases in production per ha [6].

The use of plant genetics, new pest control methods, and more efficient fertilizers have increased farm production over the last 30 years [7]. At a global level, it would be too difficult to find additional areas for agriculture as most cultivable area is already under use. Intensification of the area currently

cultivated involves adopting more rigorous farm operation systems and the application of more chemical inputs (pesticides, fertilizers and fuel). It was expected that the newly-developed seed varieties would have improved yields over the last few decades; but, in many areas, due to the use of traditional farming methods by farmers and other technical limitations, yields are still lower than the desired production [4, 8].

Overall, New Zealand agriculture is dominated by high farm inputs [9, 10]. Agricultural production is a victim of, and contributor to, global warming [10-13]. CO₂ contributes significantly to greenhouse gases (GHG) [14]. The link among production, energy consumption and CO₂ emissions in agricultural activities is well understood [10, 15-18]. CO₂ is emitted during different farming activities, such as land use changes, application of fertilizers and pesticides, the ignition of fossil fuels and plant waste, decay of organic matter and microorganisms in the soil [12, 19, 20]. GHGs could change the current environmental conditions that have uncontrolled impacts on agricultural production. To monitor CO₂ emission reduction targets, the effects of direct and indirect factors on CO₂ emissions should be investigated.

MLR models have been used widely in agricultural projects more than other prediction techniques [21, 22]. A simple model with a high r^2 can be developed through the use of sufficient numbers of samples and independent variables. Input variables are always maintained in the best model if the actual and predicted data are correlated with a p value of 0.05 [23]. In the first step, corroboration between CO₂ emission and each input variable is verified with simple MLR using r^2 as the decision criterion. A MLR model is then established to predict CO₂ emissions as:

$$Y = a_0 + a_1V_1 + a_2V_2 + \dots + a_nV_n + \epsilon \quad [1]$$

where a_0 - a_n = coefficients of regression, V_1 - V_n = the input variables and ϵ = the error. The linear model represents the links between the independent (input) variables and the dependent (output) variable.

Artificial neural networks (ANN) have been used recently for investigating the connections between input and output parameters [24, 25]. Based on an analysis of the already-entered data, neural networks can find a link among the input and outputs, as well as the controlled and uncontrolled parameters [26]. To develop an effective ANN model, the number and accuracy of the data sampled are key issues, as ANNs require enough data to develop suitable connections, ANNs cannot develop the correct connections by themselves. ANN models are simple applications that can predict or classify different data to give with robust results. ANNs can estimate nonlinear input-output applications with high accuracy, so can play a vital role in simulating complex systems [27].

The feed-forward multi-layered perception (MLP) paradigm is the most common ANN structure used in modelling studies. The feed-forward MLP paradigm consists of independent variables, hidden layers and an output layer trained by the back propagation (BP) learning method. MLPs trained by BP are capable of modelling any function, so they are widely used for prediction models [28-30]. The neurons associated with the first hidden layer analysis, the weighted independent variables, use a transfer function to lead to the results. The most commonly used transfer functions include: logistic, linear, sine, Gaussian and hyperbolic-tangent. The results from the first hidden layer are then directed to the second hidden layer via weighted connections. Summation of the weighted inputs is processed by the neurons in the hidden layer using their transfer functions. The neuron outputs associated with the output layer are termed called the predicted output [22, 25].

The mean square error (MSE) between the predicted results and the measured data is minimized by adjusting the weights. The following relation is used for estimating the mean square error for a basic network having one output

$$MSE = \frac{1}{2N} \sum_i^N (t_i - z_i)^2$$

neuron:

where z_i = predicted outputs associated with the i th training pattern, t_i = the actual outputs associated with the i th training pattern and N = the sample size of the training patterns [25]. Furthermore, the root mean square error (RMSE) is used to show the errors in the units of the actual and predicted data.

Models with a minimum of input variables are preferable for problem solving. Therefore, data reduction is useful if the number of input variables is high and the available sample size is limited [25].

II. METHOD

The experiments were conducted on irrigated and dry land arable farms totalling 35,300 ha in Canterbury, New Zealand.

Canterbury is the dominant wheat production region in the country and shares almost 90% of wheat cultivation farms and wheat yield in New Zealand [31]. CO₂ emissions from wheat farms were investigated by considering different energy sources such as: fertilizers, pesticides, electricity, fuel and machinery. The following relationship was used to calculate total CO₂ emissions (E).

$$E = \Sigma(A_i C_i) \quad [2]$$

where Σ = summation, A_i = input factor and C_i = the CO₂ emission conversion coefficient for each factor.

Different conversion coefficients were used to convert farm inputs into CO₂ emissions. Selecting accurate conversion coefficients was the key point of this study. Apart from farm inputs, the impact of around 140 factors comprising both technical and social aspects such as the farmers' social status, the properties of tractors and equipment, farm conditions and yield, were investigated.

Except for fuel burning, where carbon dioxide was released directly, CO₂ was also released indirectly from farming activities. The use of most inputs associated with agriculture were converted into energy coefficients to obtain kg CO₂/MJ. Three different sources of data collection were included: a survey, a literature review and field measurements. This study was based on an analysis called the 'cradle-to-gate analysis', which meant that the transport and waste disposal components of the products' life cycles were not involved after they left the farm gate.

A limited number of independent variables were selected to ensure a practical model. The input variables were reduced by applying pre-processing based on correlation analysis, followed by principle component analysis (PCA). Analysis of various variables associated with the components of PCA led to the identification of a cumulative variance with eigenvalues greater than 1. Around 140 input variables were applied in the final ANN model. The analysis consisted of two steps. In the first step (pre-processing), input variables, which had no little correlation with each other but had a significant impact on CO₂ emissions, were selected. In the second step, 16 variables that demonstrated high links with CO₂ production were selected, and included: area of wheat cultivation (ha), percentage of wheat cultivation area on farms, number of cows, annual rainfall, age of farmers, educational background of farmers, irrigation frequency, capacity of tractors (hp), farm size (ha), inputs such as N, P, fungicides, and insecticides, age of fertilizer spreader, number of plough passes, number of sprayer passes, and age of sprayer. The PCA process was guided to select eight independent variables to be applied as independent variables in the ANN model. The eight independent variables selected included plough passage numbers, the proportion of wheat area on farms, irrigation frequency, number of cows, age of fertilizer spreader and farm inputs, nitrogen input (kg), insecticide input (kg), phosphate input (kg), age of sprayer and tractor power index (hp/ha). The selected eight independent variables had a threshold cumulative variance of around 72.3%.

Data were collected from 40 wheat farms. For training purposes 30 farms were selected randomly and the remaining 10 farms were used for model validation.

A limited number of hidden neurons were enough to describe simple nonlinear problems. In contrast, to solve the very nonlinear problems associated with large amounts of input variables large numbers of neurons were essential to predict an output variable with a low margin of error. Currently, neuron numbers were selected based mostly on trial

and error rather than science [32]. For the purpose of this study, the different aspects of the ANN model were optimised by using a genetic algorithm-based optimisation to determine a satisfactory model structure. A number of trials led to the selection of a modular neural network with two hidden layers containing two sub-networks (Figure 1).

For function approximation, the optimised model was trained. In the training process, the weight change followed by subsequent batch processing was controlled by the learning rate. A training process with a higher learning rate would be quicker; but the weights may oscillate around the lowest level of error, but never reach it [25]. Subsequently, this study used a learning rate of 0.01 (low). The learning method adopted (Quick Prop) was very fast in reducing flaws when finding promising results. Quick Prop adjusted weights by indirectly using the second derivative of error. In each trial of Quick Prop, weights were revised using following relationship:

$$w_{m+1} = w_m + \Delta w_m$$

$$\Delta w_m = \frac{d_m}{d_{m-1} - d_m} \Delta w_{m-1}$$

$$d_m = \sum_{n=1}^N \left[\frac{\partial E}{\partial w_m} \right]_n$$

where Δw_m = the existing weight, d_m = the average derivative of the error for the current epoch (batch) m ; and $\partial E/\partial w_m$ = the current error gradient for a particular input vector [25].

This study examined different functions, that included the logistic sigmoid, hyperbolic tangent (tanh), sine, Gaussian and linear functions. To propose the final model, the hyperbolic tangent function was selected for the input layer and the first hidden layer; the logistic function was applied for the second hidden layer; and the linear function was selected for the output layer (Figure 3). These functions can be written as:

$$L(u) = [1 + e^{-u}]^{-1}$$

$$\tanh(u) = \frac{1 + e^{-u}}{1 - e^{-u}}$$

where $\tanh(u)$ = the hyperbolic tangent function, $L(u)$ = the logistic function, and u = weighted sum of inputs into a neuron [25].

III. RESULTS AND DISCUSSION

A. CO₂ Emissions

The study revealed that an average of 1032 kg CO₂/ha, was released from each wheat cultivation farm. To achieve a wider perspective, farm inputs were divided into five categories: electricity, fertilizers, agrichemicals, machinery and fuel. As shown in Table 1, fertilizer (mostly nitrogen) was ranked highest on the farms studied, with 52% of total CO₂ emissions.

TABLE I. TOTAL CO₂ PRODUCTION FROM DIFFERENT AGRICULTURAL INPUTS APPLIED ON WHEAT CULTIVATION FARM (KG CO₂/HA)

	Fertilizer	Agrichemicals	Electricity	Machinery	Fuel	Total
Total	539 (52%)	55 (5%)	86 (8%)	149 (14%)	103 (20%)	1032

B. Model Development

MLR and ANN models were developed for forecasting CO₂ emissions from the wheat cultivation farms.

1) Multiple Linear Regression Model

MLR demonstrates the linear relationships between the input and output variables. In this study the MLR model was compared with an ANN for which data from 25% randomly-selected samples were used for model validation and the remaining 75% data were used for model training. The MLR model developed was used to estimate, the validation data. The MLR model was able to be fitted to the CO₂ emission data and accounted for around 35% and 70% of the variance in the validation and training data, respectively. Figures 1 and 2 demonstrate the relationships between the forecasted and measured CO₂ emissions, respectively, for the training and validation data. The MSE and RMSE estimated for validation data were 6977 and 84 kg CO₂/ha, respectively.

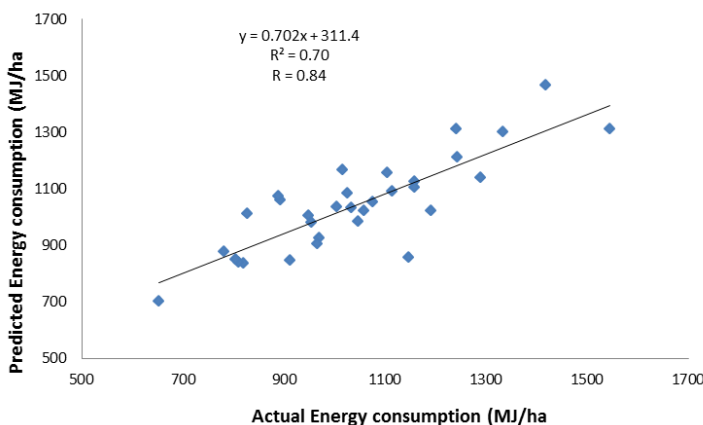


Fig. 1. Relationships between the field measurements and model-predicted CO₂ emissions (training) based on the MLR model

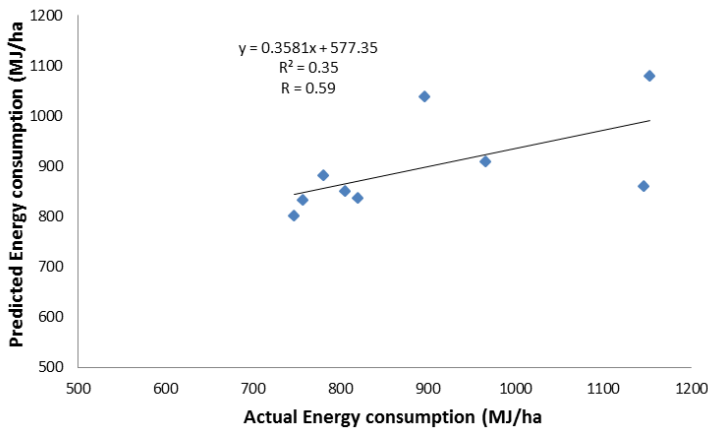


Fig. 2. Relationships between the field measurements and model-predicted CO₂ emissions (validation) based on the MLR model

2) Artificial Neural Network Model

After trialling different neuron activation functions, learning algorithms and network structures, a modular network with two hidden layers was developed (Figure 3). After the input layer, the modular network was divided into

two parts. There were 20 neurons in the input layer with two and 12 neurons, respectively, for the first and second parts of the modular network. The final output (CO₂ emissions) was produced by combining the results in the output layer.

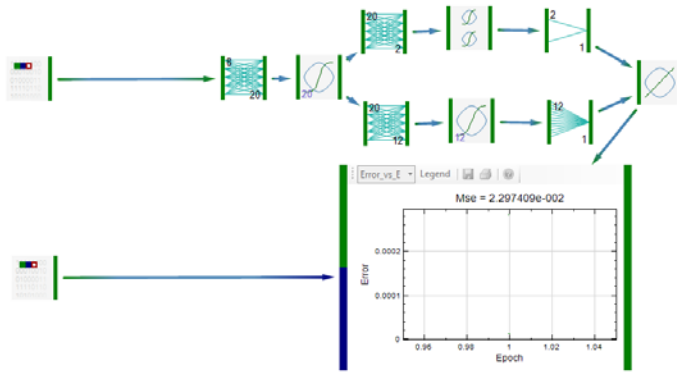


Fig. 3. Structure of the modular network and the number of neurons in each layer

Trials (1500) led to the production of the most satisfactory ANN model with a scaled MSE of 2.3×10^{-2} (with inputs and outputs ranged between -1 and +1). Compared to MLR, the ANN model predicted CO₂ emissions effectively and accounted for almost 90% of the variance (Figure 5) in the

validation data. Figures 4 and 5 show the relationship between the actual and predicted data for the training and validation of the ANN model. The r^2 was estimated at 0.82 and 0.89, respectively, for training and validation of the ANN model.

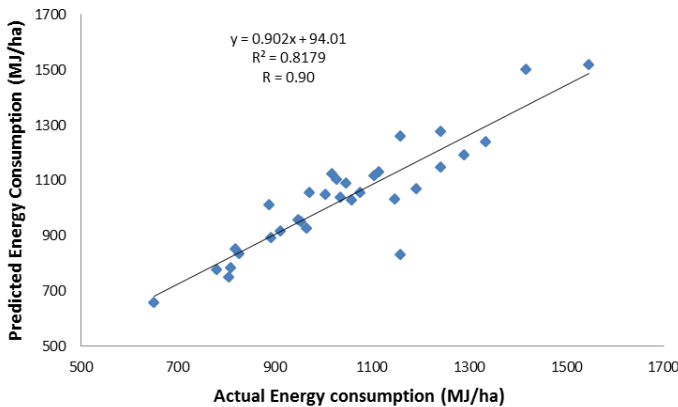


Fig. 4. Relationships between the field measurements and model-predicted CO₂ emissions (training) based on the ANN model

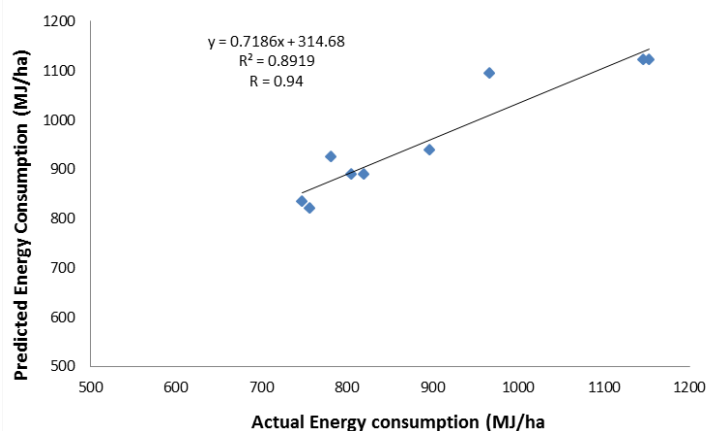


Fig. 5. Relationships between the field measurements and model-predicted CO₂ emissions (validation) based on the ANN model

Figures 6 and 7 illustrate the ANN predictions for the training and validation data, respectively. The four lines in each picture represent the desired output, network output, and the high and low boundaries of the confidence intervals. The region within which the correct answer was within the 95%

confidence level, as indicated by the grey area. As Figure 7 shows, the final model can predict CO₂ emissions up to ±113 kg CO₂/ha within the 95% confidence level. The results indicated the chance that the predicted errors would be more than ±113 kg CO₂/ha was only 5%.

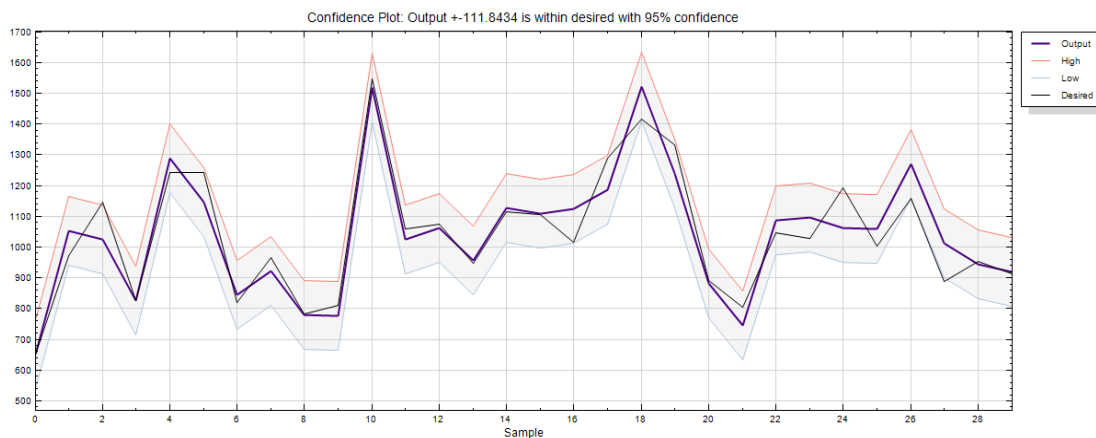


Fig. 6. Predicted, observed and 95% confidence interval for CO₂ emissions based on the ANN model (training data)

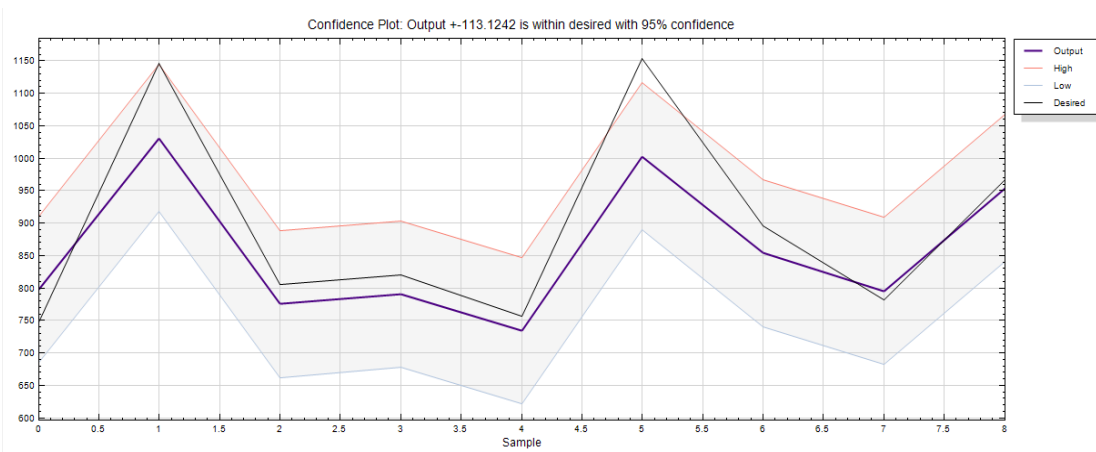


Fig. 7. Predicted, observed and 95% confidence interval for CO₂ emissions based on the ANN model (validation data)

For the training and validation data, the link between the measured and forecasted CO₂ emissions for the ANN model was much greater compared to that of the MLR model.

Compared to MLR, the ANN model had a noticeably lower RMSE for the validation data (Table 2).

TABLE II. MSE AND RMSE OF THE MLR AND ANN MODELS

	Linear		ANN	
	Training	Validation	Training	Validation
MSE	5635	6977	3641	3307
RMSE	75	83	60	57

There were a number of uncontrolled parameters that could affect CO₂ emissions in agricultural farms, and they made the results of these experiments quite interesting. The proposed model can predict CO₂ emissions in wheat farms within tolerable margins of error. There were some fixed independent parameters in the proposed model that could not be changed: such as farm conditions. Some of the independent variables, such as the farmer's education, would also influence the CO₂ emission indirectly. This calls for future studies to investigate the detailed links between the input parameters and CO₂ emissions from agricultural farms.

The ANN model can predict CO₂ emissions from farm inputs. The model can help farmers identify the factors with more potential to minimise CO₂ emissions on their farms. In addition, scientists and decision makers can evaluate CO₂ emissions in Canterbury.

IV. CONCLUSIONS

In this study an ANN model was developed to forecast CO₂ emissions from farm inputs using direct and indirect factors to predict CO₂ emissions from wheat farms. The proposed ANN model could forecast CO₂ emissions from farm inputs depending on each farm's conditions, such as the proportion of wheat cultivation area on farms, frequency of irrigation and number of cows, the condition of machinery (tractor power index (hp/ha) and age of fertilizer spreader) and inputs on the farm (N, P and insecticide use) in Canterbury agricultural farms with a margin of error of ±11% (±113 kg CO₂/ha). As there were numbers of uncontrolled factors in agricultural production, the size of error was acceptable. In addition, the results showed that the ANN model using heterogeneous data can better forecast CO₂ emissions than the MLR model (Table 2). Using dissimilar inputs, such as farm conditions and social factors, would help the relevant agencies view the problem from various angles.

The finding from these experiments indicated the capability of an ANN model for forecasting CO₂ emissions from agricultural inputs by adopting indirect factors. This improved model can support decision makers by providing information on predicted CO₂ emissions from a wide range of farm products. Analysis of the results made it clear that it was not possible to change some input parameters in the short term. However, for the scientific community and decision makers, the model would provide useful information to judge the best directions for CO₂ emission reductions in the future.

Testing the results for at least five years with larger sample sizes would lead a more accurate model for forecasting the trend of CO₂ emissions in agricultural farms under various situations. The outcomes of this research can be recognised as a first effort to propose methods appropriate for estimating CO₂ emissions by considering geographical, social and technical parameters together. This proposed approach can be

replicated to other farm production systems and cropping areas.

REFERENCES

- [1] Breiman A, Graur D. Wheat Evaluation. Israel J Plant Sci. 1995; 43: 58-95.
- [2] OECD-FAO. OECD-FAO Agricultural Outlook. Paris: ECD Publishing, OECD & FAO,; 2011.
- [3] Kronstad WE. Agricultural development and wheat breeding in the 20th century., Developments in Plant Breeding vol. 6. Kluwer Academic Publishers, Dordrecht, The Netherlands, p. 1-10.
- [4] Rosegrant MW, Agcaolli-Sombilla A, Perez N. Global Food Projections to 2020: implications for investment 2020 Vision Discussion papers. 1995;5.
- [5] Kole C. Cereals and millets. Berlin ; New York: Springer, 2006.
- [6] Vlek PLG, Rodríguez-Kuhl G, Sommer R. Energy Use and CO₂ Production in Tropical Agriculture and Means and Strategies for Reduction or Mitigation Environment, Development and Sustainability. 2004;6.
- [7] Pimentel D, Pimentel M. Food, energy, and society. 3rd ed. Boca Raton, FL: CRC Press, 2008.
- [8] Ozkan B, Akcaoz H, Fert C. Energy input-output analysis in Turkish agriculture. Renewable Energy. 2004;29(1):39-51.
- [9] Wells C. Total energy indicators of agricultural sustainability : dairy farming case study. Wellington [N.Z.]: Ministry of Agriculture and Forestry; 2001. p. vii, 81 p.
- [10] Safa M, Samarasinghe S. CO₂ emissions from farm inputs "Case study of wheat production in Canterbury, New Zealand". Environmental Pollution. 2012;171(0):126-32.
- [11] Sauerbeck DR. CO₂ emissions and C sequestration by agriculture – perspectives and limitations Nutrient Cycling in Agroecosystems. 2001;60:253-66.
- [12] Hillier J, Walter C, Malin D, Garcia-Suarez T, Mila-i-Canals L, Smith P. A farm-focused calculator for emissions from crop and livestock production. Environmental Modelling & Software. 2011;26(9):1070-8.
- [13] Rajaeifar MA, Ghobadian B, Safa M, Heidari MD. Energy life-cycle assessment and CO₂ emissions analysis of soybean-based biodiesel: a case study. Journal of Cleaner Production. 2014;66(0):233-41.
- [14] Janardhan V, Fesmire B. Energy explained. Lanham, Md.: Rowman & Littlefield, 2011.
- [15] USDA. U.S. Agriculture and Forestry Greenhouse Gas Inventory: 1990-2005. Washington, D.C: United States Department of Agriculture (USDA); 2008.
- [16] Stout BA. Handbook of energy for world agriculture. London; New York: Elsevier Science Pub. Co., 1990.
- [17] Snyder CS, Bruulsema TW, Jensen TL, Fixen PE. Review of greenhouse gas emissions from crop production systems and fertilizer management effects. Agriculture, Ecosystems & Environment. 2009;133(3-4):247-66.
- [18] Safa M, Samarasinghe S. Determination and modelling of energy consumption in wheat production using neural networks: "A case study in Canterbury province, New Zealand". Energy. 2011;36(8):5140-7.
- [19] Smith KA, Conen F. Impacts of land management on fluxes of trace greenhouse gases. Soil Use and Management. 2004;20:255-63.
- [20] Jeffrey S A. Effects of atmospheric CO₂ concentration on wheat yield: review of results from experiments using various approaches to control CO₂ concentration. Field Crops Research. 2001;73(1):1-34.
- [21] Colwell JD. Estimating fertilizer requirements : a quantitative approach. Wallingford: CAB International, 1994.
- [22] Safa M, Samarasinghe S, Nejat M. Prediction of Wheat Production Using Artificial Neural Networks and Investigating Indirect Factors Affecting It: Case Study in Canterbury Province, New Zealand. Journal of Agricultural Science and Technology. 2015.
- [23] Alvarez R. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. European Journal of Agronomy. 2009;30(2):70-7.

- [24] Parten C, Hartson C, Maren A. Handbook of neural computing applications: San Diego, CA (USA); Academic Press Inc, 1990.
- [25] Samarasinghe S. Neural networks for applied sciences and engineering : from fundamentals to complex pattern recognition. Boca Raton, FL: Auerbach, 2007.
- [26] Kalogirou SA. Applications of artificial neural-networks for energy systems. Applied Energy. 2000;67(1-2):17-35.
- [27] Hagan M, Demuth H, Beale M. Neural network design: Boston, USA: PWS Publishing Company, 2002.
- [28] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Elsevier Science Ltd Oxford, UK, UK 1989;2(5).
- [29] Heinzow T, Tol RSJ. Prediction of crop yields across four climate zones in Germany: an artificial neural network approach. Centre for Marine and Climate Research, Hamburg University, Hamburg. 2003.
- [30] Jebaraj S, Iniyar S. A review of energy models. Renewable and Sustainable Energy Reviews. 2006;10(4):281-311.
- [31] Statistics New Zealand. Agricultural Production Statistics (Final): June 2010. Wellington: Statistics New Zealand; 2010.
- [32] Kermanshahi B, Iwamiya H. Up to year 2020 load forecasting using neural nets. International Journal of Electrical Power & Energy Systems. 2002;24(9):789-97

E-Learning for Secondary and Higher Education Sectors: A Survey

Sadia Ashraf

Bahria University Islamabad,
Pakistan

Tamim Ahmed Khan

Bahria University Islamabad,
Pakistan

Inayat ur Rehman

COMSATS Institute of Information
Technology Islamabad, Pakistan

Abstract—Electronic learning (e-learning) has gained reasonable acceptance from educational institutions at all levels. There are various studies conducted by researchers considering different aspects of e-learning to investigate how we can benefit in imparting quality education. However, there is a requirement to find out how researchers consider different sectors of secondary and higher education (HE) sectors. In this paper, we carefully select published research article of past six years and study how the research was conducted and which research methods are applied to attain results. We also investigate how case studies are presented for evaluating results. We finally present our findings from conducting this study of e-learning research at secondary and the higher education levels.

Keywords—Distributed learning environments; elementary education; improving classroom teaching; intelligent tutoring systems; interactive learning environments; media in education; post-secondary education; secondary education; simulations

I. INTRODUCTION

E-learning is a learner-centered instructional strategy which provides students with the opportunity for an in-depth investigation of a given topic. With the advent of information technology and its growing use in education sector, a transformation is seen in traditional and conventional teaching methods that are used in schools. There is an increase in the competency requirement within pedagogical professions such that the modern-day teachers require improving their skill level through effective use of e-learning initiatives. The use of e-learning has a two-fold impact on the students' learning, i.e., we are able to provide a uniform system of education and, secondly, the students learning pattern can be recorded. E-learning is successfully used for augmenting students' learning in education sector. It has been noticed that e-learning is more effective in teaching subjects such as Mathematics, Science and English and impact of e-learning is more evident where tools specific to teachers' everyday use were utilized Cox and Britain (2003). In order to conduct an in-depth study of the impact of e-learning initiatives, it is important to classify different studies and published research into intermediate and HE levels. Our interest is in the classification of selected research papers according to different education levels.

The paper is organized as follows. We present our evaluation criteria in Section 1 and our finding from secondary and HE levels in Section 2 and Section 3. Finally, we conclude

our findings in Section 5.

II. OUR EVALUATION CRITERIA

We study related research papers of last six years, i.e., 2007 onwards and classify them according to the education sector forming the research focus. We consider two sectors of education, i.e. secondary and higher education. The secondary sector covers students of Grade-6 to Grade-10 where HE sector consists of college and university levels. We develop a classification so that selected papers may be distinguished on the basis of their main research focus to classify research presented in reviewed papers. We present this classification in Table 1 where each classification is given a tag or name and a short description to explain its meaning.

TABLE I. CLASSIFICATION OF MAIN FOCUS

S/N	Classification	Description
1.	Adaptability	Possible uses of e-learning methods for imparting quality education
2.	Tools	Tools to support e-learning
3.	Challenges	Challenges related to implementation of e-learning
4.	Impact Analysis	Analysis of impact after implementation of e-learning
5.	Environment	Present environment related experience (computer-based learning/Mobile based learning)

We present our findings according to our evaluation criteria based on the paper, main focus and effect of their focus area. Finally, we discuss selected research papers to provide the reader an overview of research in each educational sector. We do not include papers that do not fit in our evaluation criteria. Examples of such papers include research by Journell (2010) which is based on investigating perception of e-learning from students and teachers perspective, research published by Kalinga (2008) who developed an interactive Learning Management System (LMS) to support teachers and students for sharing learning materials across different Tanzanian schools. We also do not consider game based learning e.g., Krunoslav Bedi (2011), Nalin Warnajith et al. (2012) etc.

III. SECONDARY LEVEL

We devote this section to papers related to secondary level of education. We present our findings considering our evaluation scheme as shown in Table 1.

A. Adaptability

Kiilu Redempta (2012) examined Kenyan schools to find out if schools were ready and if they showed a positive attitude towards the adoption of e-learning. He conducted a desktop review supported by Current Situation Analysis (CSA). He concludes that less than 10% of Kenyan schools offer computer science as a subject in the curriculum, The current educational practices neglect e-readiness. Mildred A. Ayere (2010) presented a case study of New Partnership for Africa's Development (NEPAD) schools of Kenya. These schools were developed to integrate ICT and they are compared with the ones that are not implementing NEPAD. We call them Non-NEPAD schools. Twelve schools participated in the research where six NEPAD and remaining Non-NEPAD schools are selected. The researchers select NEPAD schools through saturated sampling whereas they select Non-NEPAD schools through random sampling. The research is a combination of exploratory approach using descriptive survey and ex-post-facto design with questionnaire as the main data collection instrument along with interviews involving students, principals and heads of departments. The authors use descriptive and inferential statistics for data analysis and do a comparison of both types of schools concluding that teachers should be fully facilitated and their roles should be strengthened in schools which offer ICT education. Glushkova (2012) used Sharable Content Object Reference Model (SCORM) standard templates and modified them to create e-lessons. The underlying approach in SCORM is a combination of stereotypical and overlay models which not only benefited students in terms of providing better guidance and distance learning but also helped special students.

Neyland (2011) investigated factors associated with integration of online learning in Sydney high schools. The research is carried out by conducting interviews and questionnaires in New South Wales secondary schools. The authors conclude that school support and the micro factors such as teacher capabilities are important.

B. Tools

Hsien-Sheng Hsiao, Lin, and Lin (2012) proposed a self-regulated web quest learning method for Chinese secondary schools with the aim of examining correlation between students' self-regulated behavior and their achievements. The proposed web quest learning was based on self-regulated learning assisted functions and traditional web quest learning.

The experiment was conducted on sixth grade students and results were analyzed by observing their self-regulated behavior as well as system records collected during in the learning process. The results showed that the presented system assisted students in learning and helped teachers in monitoring students' performance. Wei-wei Feng (2010) targeted Chinese secondary schools in Hong Kong and analyzed benefits of e-learning in teachers' training and emphasized importance of e-portfolio as a learning and assessment tool for teachers training. According to the authors, use of e-learning could make in-service training more easy and learning process could be made more transparent and innovative by applying e-portfolio.

C. Challenges

Micheuz (2007) focused on Austrian secondary academic schools to investigate how far e-learning technologies are established at this educational level. The research was carried out with the help of online questionnaires and the authors evaluated whether e-learning helped in learning and concluded that it is still an unanswerable question since it faces many obstacles like budget, teachers' adoption, etc.

D. Impact Analysis

Huan-Ming Chuang (2008) examined the impact of Knowledge Sharing Blog (KSB) among three groups of secondary students. Authors divided students in groups such that one of the groups used e-learning with KSB, second used e-learning without KSB and the third group comprised of those who studied in the traditional classroom environment. The results concluded that students of the first group showed significant improvement in their learning as compared to the other two groups.

E. Environment

Chiu-Pin Lin et al. (2010) studied learning of social studies and involved 6th-grade students to find out the impact of collaborative concept learning by having a comparative analysis of two different learning environments. In the first environment, each student used a computer system for learning purpose while in the second scenario; multiple students shared a single computer system. Analysis through questionnaires and interviews showed that students belonging to the first scenario demonstrated better performance as compared to the second one. We summarize our findings for e-learning research at secondary education level in Table 2.

TABLE II. E-LEARNING AT SECONDARY EDUCATION LEVEL

Paper	Main Focus	Effect
Kiilu Redempta (2012)	e-learning adoption	Current educational practices often neglect e-readiness
Mildred A. Ayere (2010)	e-learning in schools	e-learning implementation improved teaching and learning
Glushkova (2012)	Use of Sharable Content Object Reference Model and creation of e-lessons	better guidance and distance learning
Neyland (2011)	factors associated with integration of online learning	better educational outcomes achieved through online learning
Hsien-Sheng Hsiao, Lin, and Lin (2012)	self-regulated web quest learning method, correlation between students self-regulated behavior & their achievements	improved students' learning and assisted teachers in monitoring students performance
Weiwei Feng (2010)	benefits of e-learning in teachers' training and e-portfolio as a learning and assessment tool for teachers training	e-portfolio made learning process more transparent and innovative
Micheuz (2007)	establishment of e-learning technologies	availability of subjects' courses hasn't been achieved yet
Huan Ming Chuang (2008)	impact of Knowledge-sharing Blog (KSB)	improvement in students' learning
Chiu-Pin Lin et al. (2010)	impact of collaborative concept learning and comparative analysis of two different learning environments	better performance of students working in collaborative concept learning environment

IV. SECONDARY EDUCATION LEVEL

We classified the research papers of HE level on the same criteria used for secondary level of education. Here, we did not consider papers such as Nicole Wagner (2008) since they do not fit our classification and the purpose of this contribution. Nicole Wagner (2008) discussed stakeholder's point of view and discussed their role in successful implementation of e-learning in HE. Other such examples are Nikolaos Tselios and Papadopoulou (2011), Ksenija Klasnic and Seljan (2010), Oystein Sorebo, Gulli, and Kritiansen (2009), etc. The following is a careful selection of papers as per our classification discussed in Section 2.

A. Adaptability

Bradford S.Bell (2013) inspected the use of e-learning at post-secondary education in three different aspects. They investigated e-learning effectiveness in comparison with other traditional teaching methods; listed key features influencing e-learning effectiveness and discussed obstacles in adoption of e-learning. Meta-analysis, study and research review were used as research tools. While resolving first issue, the researchers found that e-learning is as effective as other delivery methods when used in similar instructional conditions.

Fageeh (2011) conducted in-depth-interviews along with literature review to find the attitude of undergraduate students of a Saudi Arabian university towards adoption of e-learning. A similar type of study was conducted by Nikolaos Tselios and Papadopoulou (2011). Shintaro Okazaki (2012) proposed the use of Technology Acceptance Model (TAM) to find out effects of gender on adoption of e-learning in Brazil. The study was carried out with the help of questionnaires from three Brazilian Universities. The results concluded that male students showed more flexibility towards e-learning adoption as compared to the female students. Liaw (2008) examined reasons that despite popularity of e-learning at university level, some students are uncomfortable while accepting it. The author claims that such problems can be handled effectively by individuals since using an e-learning system depends on self-efficiency and flexibility of the learner.

Adnan Riaz (2011) explored factors of successful acceptance of e-learning among university students. Teachers

and e-learning tools were found as the main factors. Ndume, F.N.Tilya, and H.Twaakyondo (2008) designed a tool for helping disabled students of Tanzanian institutions. Documentary review, structured questionnaires and interviews were used as data collection tools. Another feature of this research was analysis of challenges in e-learning acceptance which included management support, methodology, technology, resource accessibility and availability, etc. Toshie Ninomiya et al. (2007) developed a learning management system named WebClass RAPSODY for university students. Its purpose was to support personalized adaptability by improving LMS in lectures at university level. The learning mode of this system was able to monitor and analyze learners learning status and unit for contents to search and analyze contents status. After learning a content, this system indicated next suitable content, with data mining of learners status and contents status by genetic algorithm (GA). This function was able to support learner to sustain e-learning with understanding of contents and highly-motivation to learning. Mubarak M Alkharang (2013) explored challenges and barriers which influenced acceptance of e-learning in Kuwaitian HE institutions. Semi-structured interviews were used to gather data and it was found out that there are obstacles such as technological and language barriers, lack of management awareness and support in implementing e-learning in Kuwait HE system.

B. Tools

Colin Beard (2007) introduced media based material (files on CD-ROM) to provide e-learning support to post graduate students. The students were provided with a model that was a combination of film and text. The feedback about the model was received from the students in the form of questionnaires, interviews and reflective writing. The authors found that this concept was highly appreciated by distant learners and on-campus students which indicated that proposed model could be used as a diagnostic tool for design of learning experiences. Ann Heirdsfield, Tambyah, and Beutel (2011) designed an online survey to get response from faculty, staff and students about use of an online learning environment (Blackboard). The aim of this study was to help teaching staff in getting students perceptions and experience about online learning.

Schiaffino, Garcia, and Amandi (2008) designed a tool (e-Teacher) to assist e-learners. This tool created students' profiles by observing their behavior during the course. Student's profiles were reflection of their performance which showed learning style. Bayesian networks were used to automatically detect the students actions. Student was able to improve his performance by getting a continuous feedback on his learning behavior. Chin-Yeh Wang et al. (2010) built a humorous learning system to develop the interest of students in learning. This system provided an interactive learning environment so that students may not feel bored during lectures. After testing this software on different college students, it was concluded that students learning process could be made interesting by interacting with them through empathic activities. Regueras et al. (2009) presented a case study of the effects of competitive learning on the satisfaction and academic achievements of telecommunication students and a tool, QUEST, for active and competitive learning was used in an undergraduate course named Communication Networks. The data was collected through survey and tool was analyzed by using T-Test for students outcomes and results showed that overall students were satisfied with QUEST tool.

Ivana Simonova (2013) found out if better results of increased knowledge could be achieved by tailoring the ICT supported process of instruction to students individual learning style. The query was resolved by designing an e-application (plug-in) and considering a case study of the University of Hradec Kralove, Czech Republic. Assessment questionnaire was filled by 105 students of the university which clearly showed that 93% of them were fully satisfied with provided application. K. Koistinen (2009) studied various social media using concepts of virtual and mirror worlds. He evaluated them based on different parameters such as usefulness of media in teaching photogrammetry, pedagogical aspects etc. The author presented a case study of how Helsinki University of Technology (TKK) utilized various e-learning methods in eTKK project including administrative course information system, study and teaching portal, learning management system (LMS), etc. He proposed how to use new e-learning tools like Google Earth, Geocarching, innovation exercise, etc. The purpose of this research is to encourage innovative trials with new e-learning tools.

C. Challenges

Milan Puvaca (2010) studied Croatian institutions to investigate possible challenges faced in the adoption of e-learning. Suggestions are made to implement e-learning in different universities at common ground. Obstacles in implementing e-learning are identified such as resistance provided by faculty or students depending on their abilities and perception and the conservative nature of the organization to avoid change are discussed. Ndume, F.N. Tilya, and H. Twaakyondo (2008) designed a tool for helping physically challenged students of Tanzanian institutions. The researchers used documentary review, structured questionnaires and interviews as data collection tools. Analysis of challenges in e-learning acceptance was a feature of this research.

A. S. Sife and Sanga (2007) presented examples from Tanzanian institutions to discuss challenges faced by developing countries in HE while implementing new teaching

and learning technologies. Some of important issues found included lack of administrative and technical support, inadequate funds, staff development, etc. Mubarak M Alkharang (2013) explored those challenges and barriers which influenced acceptance of e-learning in Kuwaitian higher educational institutions. Semi-structured interviews were used to gather data and it was found out that there are obstacles such as technological and language barriers, lack of management awareness and support in implementing e-learning in Kuwait HE system. Paredes J. et al. (2008) covered 11 Spanish universities to analyze the problems, uses and effects of using platforms in European Space for Higher Education (ESHE). Data gathering was done through questionnaires and the findings revealed that there was a lack of institutional culture. It was also suggested that uses of distance learning platforms can be improved with student participation, mentoring, assessments, etc.

D. Impact Analysis

Olojo Oludare Jethro (2012) examined effects and benefits of e-learning in HE and discussed use of technology more efficiently. The research gathered data by performing empirical study from 1996 to 2008. He concluded that the National Assessment of Educational Progress in mathematics students who were using computers at home more frequently was more useful in at higher level in mathematics. The authors concluded that e-learning tools are more effective when computers and fast internet connection, improved software and reliable electricity is available.

Rodgers (2008) explained the role of e-learning in improving students' grades. The author used a fusion of e-learning methods and lectures. The research showed that students could excel in HE with the help of e-learning methods. Islam (2013) proposed a model to examine usefulness and role played by e-learning in improving students academic performance. Data collection was done by testing the model on university students and was analyzed with the help of Partial Least Squares (PLS). The results revealed that proper utilization of e-learning could be predicted by students' perceived academic performance. Ahmad Al-Adwan (2012) discovered factors having the effect on the implementation of e-learning in Jordanian HE. The researchers conducted study in two universities of Jordan involving staff and students. The researcher used Questionnaires and focus groups for data collection. Results suggested that e-learning facilities improve technological skills of faculty and students.

Shopova (2011) explained the role of e-learning in European HE. The author presented a case study of Bulgarian university and concluded that successful learning process and high quality results could only be achieved with the proper utilization of e-learning at HE. Michael Zastrocky (2008) conducted a survey considering timescale of ten years (1999 - 2008) to find maturity level of e-learning at HE. Fan Liu et al. (2010) presented a hypothesis to show whether an extension of TAM could be helpful in predicting that users will adopt online learning. The extension comprised of different variables like online course design, user-interface design, perceived interaction, etc. Questionnaires were used as a data collection tool and Structural Equation Modeling (SEM) was applied for

data analysis. This research has added new variables to the already existing TAM.

Gurmeet Singh (2009) chose the University of South Pacific as a case study to find out how far e-learning has been successful in improving the quality of learning at HE. The authors concluded that e-learning approaches are capable of improving learning process quality however the developing countries are unable to recognize key areas of e-learning due to which these are suffering in terms of development. Haverila (2011) carried out a study on a Finnish university to investigate the impact of e-learning experience on students perceived learning outcomes. Results recommended that some prior knowledge of e-learning experience such as collaborative and situated learning experience, construction of knowledge experience, etc. can help students in improving their learning ability. Lai (2011) explained how digital technologies can support teaching and learning practices. He surveyed literature to find the changing needs and expectations of today's students. The author also emphasized that students must be aware of their learning characteristics and they should be prepared for future as innovative knowledge creators by using formal and informal strategies.

Jamal F. Kakbra (2013) focused on Kurdistan (Iraq) to determine the effect e-learning and analyzed if it should be redesigned and implemented through MOODLE which is an LMS. There was lack of LMS due to which students and teachers were facing difficulties in the university under consideration. Sorebo and Sorebo (2009) conducted a study to investigate level of satisfaction experienced by Norwegian university teachers in using e-learning methodologies on the basis of their expectations, perceived usefulness and perceived competence. It was concluded that teachers' perceived usefulness seemed to be the best indicator of their satisfaction. Paul Lam (2011) research was based on analysis of issues related to perception of undergraduate students towards e-learning usage in teaching and learning and the affect of their previous e-learning knowledge on their perception of the value of e-learning. Cradduck (2012) discussed future of e-learning and interdependency between internet and e-learning in Australian institutions. It was concluded that e-learning cannot be utilized effectively without having required skills and proper internet access. Laura Asandului (2008) used mixed mode research methods to examine level of distribution of e-learning in Romanian HE. It was concluded that male students spent more time on computer as compared to the females. While comparing e-learning with traditional methods, students appreciated e-learning methods in terms of updating content, efficiency, and amount of knowledge while intelligibility was rated low.

Ainhoa Alvarez et al (2009) proved that inclusion of a recommendation system based on courses to study and contents of each course, the students can be provided

assistance in strengthening their own study process. The recommendation system allowed flow of information between online and off-line environments. Regueras et al. (2009) presented a case study analyzing effects of competitive learning on satisfaction and academic achievements of telecommunication students and a tool, QUEST, for active and competitive learning was used in an undergraduate course named Communication Networks. The data was collected through survey and tool was analyzed by using T-Test for students' outcomes and results showed that overall students were satisfied with QUEST tool. Xiaofei Chen (2010) presented characteristics, functions and investment benefits of e-Learning technology. Some important functions of e-Learning were found to be improving teaching and learning model, and the quality of teaching, enhancing college teaching efficiency, etc. Ivana Simonova (2013) designed an e-application (plug-in) and considering a case study of the University of Hradec Kralove, Czech Republic. Assessment questionnaire was filled by 105 students of the university which clearly showed that 93% of them were fully satisfied with provided application.

E. Environment

Graham Attwell (2007) research was based on Personal Learning Environments (PLEs) to investigate why PLEs are helpful in learning and how these can be developed with different services such as ubiquitous computing and development of social software. It was concluded that PLEs are a new approach which not only support learners to develop and share their ideas but this approach also bridges the gap between educational institutions and the outside worlds. YUE Jun, Yanqing, and Zetian (2009) offered a semantic retrieval approach that was based on semantic layer, semantic similarity and semantic mapping. The keywords could be analyzed by using semantic layer, semantic similarity is calculated to catch the users retrieval intention and proper semantic mapping from keywords to concepts in knowledge base was realized. Rafaela Lunardi Comarella, Silveira, and Catapan (2012) designed educational Linux for Brazilian high schools students. The benefit of using educational Linux was to provide Virtual Learning Environment (VLE) platform and availability of online practical activities on a virtual lab. The author presented findings on the basis of usage experiences of students. Donghuai Gao, Ning, and Zhang (2011) identified the issues of current e-learning environments [ELE] which were found in non-excellent environment structure, fragile support service and improper promotion policies. The authors discussed a four-layered ELE comprising of information infrastructure, application platform, information resource and support service to overcome these issues. In order to conduct the experiment, a system was designed for providing comprehensive information service to the teachers as well as the students in Fourth Military Medical University.

TABLE III. E-LEARNING AT HIGHER EDUCATION LEVEL-ADAPTABILITY

Paper	Main Focus	Effect
Bradford S. Bell (2013)	Comparison of e-learning effectiveness and features influencing e-learning	obstacles in the adoption of e-learning
Fageeh (2011)	Attitude of under graduate students in a university and adoption of e-learning	students' acceptance for e-learning initiatives
Shintaro Okazaki (2012)	Use of TAM and effects of gender on adoption of e-learning in Brazil	male students showed more flexibility towards e-learning
Liaw (2008)	students' (dis)comfort for e-learning acceptance	learners' behavior positively affected by perceived satisfaction / usefulness
Adnan Riaz (2011)	factors of successful acceptance of e-learning	students' commitment towards e-learning and characteristics as well as use of technology and resources
Ndume, F.N.Tilya, and H.Twaakyondo (2008)	tool for helping disabled students of Tanzanian institutions	challenges in e-learning acceptance
Toshio Ninomiya et al. (2007)	development of a LMS WebClass RHAPSODY and support personalized adaptability	well-understanding of contents, highly-motivation to learning
Mubarak M Alkharang (2013)	challenges and barriers which influenced acceptance of e-learning	Kuwaitian higher educational Institutions

TABLE IV. E-LEARNING AT HIGHER EDUCATION LEVEL-TOOLS

Paper	Main Focus	Effect
Colin Beard (2007)	media based material (files on CDROM) to support e-learning	Model used as diagnostic tool for the design of learning experiences
Ann Heirdsfield, Tambyah, and Beutel (2011)	use of an online learning environment (Blackboard)	help in teaching staff in getting students perceptions and experience about online learning
Schiaffino, Garcia, and Amandi (2008)	designing of tool (eTeacher)	improve students' performance
Chin-Yeh Wang et al. (2010)	Develop interest of students in learning	learning process made interesting by interaction through empathic activities
Regueras et al. (2009)	effects of competitive learning on satisfaction and academic achievements	overall students were satisfied with QUEST tool
Ivana Simonova (2013)	Tailoring ICT supported process of instruction to students individual learning style	93% students fully satisfied with provided application
K.Koistinen (2009)	Usefulness of media in teaching photo-grammetry, pedagogical aspects	Encourage innovative trials with new e-learning tools

TABLE V. E-LEARNING AT HIGHER EDUCATION LEVEL-CHALLENGES

Paper	Main Focus	Effect
Bradford S. Bell (2013)	comparison of e-learning effectiveness with other traditional teaching methods	key features which influence the e-learning effectiveness and the obstacles in the adoption of e-learning
Milan Puvaca (2010)	possible challenges faced in adoption of e-learning and suggestions to implement e-learning in different universities at common ground	Facilitation to teachers in achieving easier communication with students
Ndume, F. N. Tilya, and H. Twaakyondo (2008)	tool for helping disabled students of Tanzanian institutions	helped in developing trust
A. S. Sife and Sanga (2007)	Challenges faced by developing countries in HE while implementing new teaching and learning technologies	ICT adoption in teaching and learning can be effected by pedagogical, technical and cost issues
Mubarak M Alkharang (2013)	barriers which influenced acceptance of e-learning in Kuwaitian higher educational institutions	different hardware and software used by various departments result in e-learning adoption & implementation difficulties
Paredes J. et al. (2008)	problems, uses and effects of using platforms in European Space for Higher Education (ESHE)	uses of distance learning platforms show improvements

TABLE VI. HIGHER EDUCATION LEVEL-IMPACT ANALYSIS

Paper	Main Focus	Effect
Olojo Oludare Jethro (2012)	effects and benefits of e-learning in HE and efficient use of technology	students using computers at home performed better in mathematics
Rodgers (2008)	role of e-learning in improving students' grades	students could excel in HE with e-learning methods
Islam (2013)	Usefulness of e-learning in improving students academic performance	students academic performance predicted through e-learning tools
Ahmad Al-Adwan (2012)	Factors having effect on implementation of e-learning in Jordanian HE	students' lack of interest in e-learning and showing resistance in adoption
Shopova (2011)	role of e-learning in European HE successful learning process	Quality and proper utilization of e-learning at HE
Michael Zastrocky (2008)	Maturity level of e-learning at HE in ten years (1999 - 2008)	drastic increase in implementation rate of e-learning
Fan Liu et al. (2010)	How TAM helpful in predicting users' adaptation of online learning	added new variables to the already existing TAM
Gurmeet Singh (2009)	Use e-learning in improving the quality of learning at HE	e-learning approaches capable of improving learning process quality
Haverila (2011)	Impact of e-learning students perceived learning outcomes	Prior knowledge of e-learning experience helpful in learning
Lai (2011)	How digital technologies can support teaching and learning practices	help students in being prepared for future as innovative knowledge creators
Jamal F. Kakbra (2013)	Effect of ICT and e-learning through MOODLE	Positive response of teachers and students towards ICT and e-learning methodologies
Sorebo and Sorebo (2009)	Norwegian university teachers satisfaction in using e-learning methodologies	achieved teachers perceived usefulness as the best indicator of satisfaction
Paul Lam (2011)	Undergraduate students perception towards e-learning	students with previous experience of e-learning were more adamant in adoption
Dr. Laura Asandului (2008)	E-learning in Romanian HE and comparison of e-learning with traditional methods	students appreciation for e-learning methods
Ainhoa Alvarez et al. (2009)	recommendation system based on courses to study and contents of each course	allowed the flow of information between online and off-line environments
Regueras et al. (2009)	Satisfaction and academic achievements of telecom students and QUEST tool	overall students were satisfied with QUEST tool
Xiaofei Chen (2010)	Characteristic and functions of e-Learning technology	improving teaching and learning model and quality of teaching
Ivana Simonova (2013)	Increased knowledge by tailoring process of instruction to students individual learning	93% students fully satisfied with provided application

TABLE VII. E-LEARNING AT HIGHER EDUCATION LEVEL-ENVIRONMENT

Paper	Main Focus	Effect
Toshie Ninomiya et al. (2007)	Development of learning management system named for university students and support of a personalized adaptability by improving LMS in lectures at university level	learners get support to sustain e-learning with well understanding of contents and highly-motivation to learning
Graham Attwell (2007)	investigation of reasons why PLEs are helpful in learning and how PLEs can be developed with different services	learners develop and share their ideas and bridge gaps between educational institutions and the outside worlds
YUE Jun, Yanqing, and Zetian (2009)	semantic retrieval approach that was based on semantic layer, semantic similarity and semantic mapping	helped in analyzing retrieving intention
Rafaela Lunardi Comarella et al (2012)	Educational Linux for Brazilian high schools students and usage experiences of students	students' progress can be monitored
Donghuai Gao et al (2011)	issues of current ELE and design of four layers model	Benefits to students and teachers using comprehensive information system

V. ANALYSIS AND DISCUSSION

Our research assisted in getting a real picture of what has been done so far in the educational sectors with respect to e-learning. This paper clearly shows that a lot more work has been done in higher education as compared to the secondary education. While particularly discussing the secondary section, more attention has been paid to the adoption of e-learning.

Some researchers focused on the development of e-learning tools, few have emphasized on its impact and challenges faced in its adoption, and a small number of researchers have worked on the designing of suitable environment in support of e-learning. Following graphs, shown in Figure 1 provide statistical picture of research performed on e-learning in various fields of secondary and higher level during the course of this study.

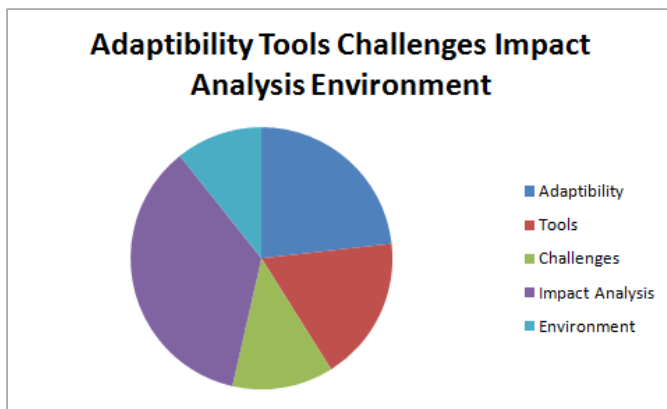


Fig. 1. Comparative analysis

On the other side, we find e-learning at higher education to be of great importance from the researchers' point of view and much work has been done at this level. In contrast to secondary education, we found most of the research on impact analysis of e-learning at the higher level. Other areas of interest are its adoption and designing of various e-learning tools; some researchers examined the challenges in its adoption, and few have paid attention to the development of e-learning supporting environment.

At the higher level, we conclude that use of various e-learning tools and methodologies plays a significant role to make the learning process more effective. However, it is important to dig out all the related aspects in order to get the maximum benefits of e-learning. We also conclude that most of the emphasis is on the learning process from students' perspective. It is pertinent to mention here that teachers' role is of pivotal importance as students and teachers share equal role in the successful learning process.

VI. CONCLUSION

Concluding our findings at the secondary and the higher levels of education, we found a mixed kind of research where most of the work has been performed on the adoption of e-learning. Different tools are developed and some of the authors have focused the challenges faced in implementing e-learning. Being the highest education level, this field has almost covered all e-learning aspects in terms of research with the main contribution in the development of tools and the impact of e-learning on the higher institutions and the stake holders. Not only this, the researchers have also explored the challenges and investigated the rate of adoption of e-learning and its related experience. The researchers have also discussed risks and shared proposals related to designing e-schools, e-colleges and e-universities. According to our study, important part of the theories and proposal of models related research was found generic in nature like e-learning usage, its role in effective learning, maturity level of e-learning in HE system, etc. Novel and innovative learning environments and systems have also been developed either for the HE level or secondary level.

REFERENCES

[1] S. Sife, E.T. Lwoga, and C. Sanga. 2007. "New technologies for teaching and learning: Challenges for higher learning institutions in developing countries." *International Journal of Education and*

Development using Information and Communication Technology 3 (2): 57–67.

[2] Adnan Riaz, Mubarak Hussain, Adeel Riaz. 2011. "Students' Acceptance and Commitment to E-Learning: Evidence from Pakistan." *Journal of Educational and Social Research* 1 (5): 21–30.

[3] Ahmad Al-adwan, Jo Smedley. 2012. "Implementing e-learning in the Jordanian Higher Education System: Factors affecting impact." *International Journal of Education & Development using Information & Communication Technology* 8(1):121–135.

[4] Ainhoa Alvarez, Smara Ruiz, Maite Martin Isabel Fernandez-Castro, and Maite Urretavizcaya. 2009. "Focusing on Personal Organization to Enhance Overall e-learning." 2009 Ninth IEEE International Conference on Advanced Learning Technologies 337–339.

[5] Ali M. Aseere, David E. Millard, Enrico H. Gerding. 2010. "A Voting-Based Agent System for Course Selection in E-Learning." 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 303–310.

[6] Ann Heirdsfield, Susan Walker, Mallihai Tambyah, and Denise Beutel. 2011. "Blackboard As An Online Learning Environment: What Do Teacher Education Students And Staff Think?." *Australian Journal of Teacher Education* 36 (7). Bradford S. Bell, Jessica E. Federman. 2013. "E-learning in Postsecondary Education." 23 (1): 165–185.

[7] Chin-Yeh Wang, Shu-Yu Ke, Hui-Chun Chuang, He-Yun Tseng, and Gwo-Dong Chen. 2010. "E-learning system design with humor and empathy interaction by virtual human to improve students' learning."

[8] Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia Pacific Society for Computers in Education 615–622.

[9] Chiu-Pin Lin, Lung-Hsiang Wong, Yinjuan Shao, Jitti Niramitranon, and Chih-Jong Tong. 2010. "1:1 Learning Technology to Support Collaborative Concept Mapping: A Case Study of Social Studies Lesson in Elementary School." The 6th IEEE International Conference on Wireless, Mobile, and Ubiquitous Technologies in Education 3–10.

[10] Colin Beard, Richard McCarter, John P. Wilson. 2007. "Towards a Theory of e-Learning: Experiential e-Learning." *Journal of Hospitality, Leisure, Sport and Tourism Education* 6 (2): 3–15.

[11] Cox, Margaret J, and Great Britain. 2003. *ICT and Attainment: A Review of the Research Literature; a Report to the DfES, DfES*.

[12] Craddock, Lucy. 2012. "The future of Australian e-Learning: Its all about access." *e-Journal of Business Education & Scholarship of Teaching* 6 (2): 1–11.

[13] Donghuai Gao, Xiajuan Shen, Yuwen Ning, and Ying Zhang. 2011. "Construction and Application of E-Learning Environment in Higher Schools." *Communication Software and Networks (ICCSN)* 636–640.

[14] Dr. Laura Asandului, Dr. Ciprian Ceobanu. 2008. "E-LEARNING IN ROMANIAN HIGHER EDUCATION: A study case." *Turkish Online Journal of Distance Education-TOJDE* 9 (3): 162–175.

[15] Dr.P.Nagarajan1, Dr. G. Wiselin Jiji2 .2010. "Online Educational System (e-learning)." *International Journal of u- and e-Service, Science and Technology* 3 (4): 37–48.

[16] Fageeh, Abdulaziz Ibraheem. 2011. "EFL students' readiness for e-learning: factors influencing e-learners' acceptance of the Blackboard in a Saudi university." *JALT CALL Journal* 7 (1): 19–42.

[17] Fan Liu, Meng Chang Chen, Yeali S. Sun, David Wible, and Chin-Hwa Kuo. 2010. "Extending the TAM model to explore the factors that affect Intention to Use an Online Learning Community." *Elsevier, Computers & Education* 54 600–610.

[18] Glushkova, Todorka. 2012. "Adaptive Model for E-Learning in Secondary School." <http://www.intechopen.com/download/get/type/pdfs/id/31951> 3–22.

[19] Graham Attwell, Pontydysgu. 2007. "Personal Learning Environments - the future of eLearning." *eLearning Papers*, www.elearningpapers.eu, ISSN 1887-1542 2 (1).

[20] Gurmeet Singh, Rafia Naz, R D Pathak. 2009. "e-Learning and Educational Service Delivery- A case study of the University of the South Pacific (USP)." *Conference Proceedings* .

- [22] Haverila, Matti. 2011. "Prior E-learning Experience and Perceived Learning Outcomes in an Undergraduate E-learning Course." *MERLOT Journal of Online Learning and Teaching* 7 (2): 206–218.
- [23] Hrmo, R., L. Kristofiakova, and D. Kucerka. 2012. "Developing the information competencies via e-learning and assessing the qualities of e-learning text." 1–4.
- [24] Hsien-Sheng Hsiao, Chung-Chieh Tsai, Chien-Yu Lin, and Chih-Cheng Lin. 2012. "Implementing a self-regulated Web Quest learning system for Chinese Elementary schools." *Australasian Journal of Educational Technology* 28 (2): 315–340.
- [25] Huan-Ming Chuang, Chia-Cheng Shen. 2008. "A Study on the Applications of Knowledge-sharing Blog Concepts to the Teaching in Elementary School." *International Conference on Cyber worlds 2008* 65–70.
- [26] Islam, A.K.M. Najmul. 2013. "Conceptualizing Perceived Usefulness in E-learning context and investigating its role in improving students' academic performance." *Proceedings of the 21st European Conference on Information Systems*.
- [27] Ivana Simonova, Pavel Kriz, Petra Poulouva. 2013. "E-application for Individualized e-Learning." 2013 *International Conference on Interactive Collaborative Learning (ICL)* 709–713.
- [28] Jamal F. Kakbra, Haval M. Sidqi. 2013. "Measuring the Impact of ICT and E-learning on Higher Education System With Redesigning and Adapting MOODLE System in Kurdistan Region Government, KRG-Iraq." *Proceedings of the 2nd e-learning Regional Conference State of Kuwait, 25-27 March 2013-01-13 Paper Code. No. eRC-125*.
- [29] Jan Skalka, Peter Svec, Martin Drlik. 2012. "E-learning Courses Quality Evaluation Framework as Part of Quality Assurance in Higher Education." *Interactive Collaborative Learning (ICL)*, 2012.
- [30] Journell, Wayne. 2010. "Perceptions of e-learning in secondary education: a viable alternative to classroom instruction or a way to bypass engaged learning?." *Educational Media International* 47 (1): 69–81.
- [31] Kalinga, Ellen Ambakisye. 2008. "Development of an Interactive E-learning management system (e-LMS) for Tanzanian secondary schools." *Blekinge Institute of Technology Licentiate Dissertation Series No 2008:03, ISSN 1650-2140, ISBN 978-91-7295-134-1*.
- [32] Kiilu Redempta, Muema Elizabeth. 2012. "An E-Learning Approach to Secondary School Education": E-Readiness Implications in Kenya." *Journal of Education and Practice* 3 (16): 142–148.
- [33] K.Koistinen. 2009. "NEW E-LEARNING TOOLS AND THEIR USEFULNESS IN TEACHING PHOTOGRAMMETRY." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science, XXXVIII6/W7*.
- [34] Krunoslav Bedi, Ana Coric, Nikolina Zajdela Hrustek. 2011. "Teaching vs. 3D gaming in secondary school." *MIPRO*, 1325-1330.
- [35] Ksenija Klasnic, Jadranka Lasic-Lazic, and Sanja Seljan. 2010. "Quality Metrics of an Integrated E-Learning System - students' perspective." *Safeullah Soomro (Ed.), ISBN: 978-953307-092-6, InTech, Available from: <http://www.intechopen.com/books/e-learning-experiences-and-future/quality-metrics-of-an-integrated-e-learning-system-students-perspective> 71–94.*
- [36] Kumiko Morimura, Shinji Suzuki, Jorg O. Entzinger. 2011. "The "SNOWBALLS" Web-Based e-learning System and its Development." *Professional Communication Conference (IPCC)*.
- [37] Lai, Kwok-Wing. 2011. "Digital technology and the culture of teaching and learning in higher education." *Australasian Journal of Educational Technology* 27 (8): 1263–1275.
- [38] Liaw, Shu-Sheng. 2008. "Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the Blackboard system." *Science Direct, www.sciencedirect.com, Computers & Education* 51 864–873.
- [39] Michael Zastrocky, Jan-Martin Lowendahl, Marti Harris. 2008. "E-Learning for Higher Education: Are We Reaching Maturity?." *Gartner, 2008, ID Number: G00156361*.
- [40] Micheuz, Peter. 2007. "E-Learning in (Austrian) Schools Some Empirical Findings, Obstacles, Theses & Visions" *Conference ICL*
- [41] Milan Puvaca, Ivica Zdrilic, Dinko Roso. 2010. "Challenges and Opportunities of Introducing e-Learning System." *Croatian Society for Information and Communication Technology, Electronics and Microelectronics* 1074–1079.
- [42] Mildred A. Ayere, J. O. Agak, F. Y. Odera. 2010. "E-learning in secondary Schools in Kenya: A Case of the NEPAD E-schools." *Educational Research and Reviews* 5 (5): 218–223.
- [43] Mubarak M Alkharang, George Ghinea. 2013. "E-learning in Higher Educational Institutions in Kuwait: Experiences and Challenges." *International Journal of Advanced Computer Science and Applications* 4(4).
- [44] Nalin Warnajith, Gamunu Dassanayake, DDGL Dahanayaka, Hideyuki Tonooka, Atsushi Minato, Satoru Ozawa, and Meepagalage Piyath Malaka Perera. 2012. "Prototype of E-Learning Management System for Secondary School in Sri Lanka." *IEEE, Information Technology Based Higher Education and Training (ITHET)*.
- [45] Ndume, Vitalis, F.N.Tilya, and H.Twaakyondo. 2008. "Challenges of adaptive e-learning at higher learning institutions: A case study in Tanzania." *International Journal of Computing and ICT Research* 2 (1): 47–59.
- [46] Neyland, Edwina. 2011. "Integrating online learning in NSW secondary schools: Three schools' perspectives on ICT adoption." *Australasian Journal of Educational Technology* 27 (1): 152–173.

Design of a Prediction System for Hydrate Formation in Gas Pipelines using Wireless Sensor Network

Ahmed Raed Moukhtar
Department of Electrical
Power Engineering
Helwan University
Helwan,
Egypt

Alaa M. Hamdy
Department of Electronic,
Communication, and
Computer Engineering
Helwan University
Helwan, Egypt

Sameh A. Salem
Department of Electronic,
Communication, and
Computer Engineering
Helwan University
Helwan, Egypt

Abstract—Before the evolution of the Wireless Sensor Networks (WSN) technology, many production wells in the oil and gas industry were suffering from the gas hydration formation process, as most of them are remotely located away from the host location. By taking the advantage of the WSN technology, it is possible now to monitor and predict the critical conditions at which hydration will form by using any computerized model. In fact, most of the developed models are based on two well-known hand calculation methods which are the Specific gravity and K-Factor methods. In this research, the proposed work is divided into two phases; first, the development of a three prediction models using the Neural Network algorithm (ANN) based on the specific gravity charts, the K-Factor method and the production rates of the flowing gas mixture in the process pipelines. While in the second phase, two WSN prototype models are designed and implemented using National Instruments WSN hardware devices. Power analysis is carried out on the designed prototypes and regression models are developed to give a relation between the sensing nodes (SN) consumed current, Node-to-Gateway distance and the operating link quality. The prototypes controller is interfaced with a GSM module and connected to a web server to be monitored via mobile and internet networks.

Keywords—WSN; Sensing Node; K-Factor; ANN; Link Quality Indicator; Hydrate Formation Temperature; Received Signal Strength Indicator

I. INTRODUCTION

WSN technology has proven a great potential in the field of data monitoring systems. Their simple device design and enhanced energy consumption gave them the opportunity to contribute in a wide range of applications such as the military, Environmental, Health and industrial applications [1]. In the field of process monitoring, various parameters such as the operating pressure, temperature, flow and specific gravity are collected and transferred wirelessly to the host location for operation and management [2]. Wireless sensor networks are mainly consists of three main elements which are the sensing elements, sensor nodes & sink node. The sink node can be either the gateway or the router node, while the sensor node are usually connected to a single or a group of sensors from which data is collected and transmitted to the gateway sink node. They are commonly structured of a microcontroller, energy source and a wireless transceiver [3]. One of the base

WSN applications in the oil and gas industry is monitoring the gas hydration formation process for remote production wells.

The Gas hydration is a well-known problem specifically in the oil and gas industry which costs millions of dollars due to production losses. They are ice-like crystalline solid compounds formed from water and molecular non-polar or slightly polar molecules (usually gasses) under low temperature [4]. At a defined temperatures & pressures, the production and transmission pipelines are blocked with what looked like to be ice.

The key for gas hydration prevention, is by using prediction systems that can give a random estimation of how far is the current operating conditions from the critical hydration values so that operator could take the necessary actions to shift the operating conditions away from the critical predicted values.

Most of the developed models depend on two methods; the first method is the specific gas gravity method [5], while the second method is the K-Factor method [6]. In fact, both of them are hand calculation methods that depend on some interpolations from data charts.

It is intended in this research to design a WSN prototype system capable of predicting the critical temperatures at which hydration will form based on the Artificial Neural Network algorithm. The research work developed three prediction models for determining the hydrate formation temperature.

For the input parameters, the first ANN model is based on the operating gas specific gravity values, the second model is based on the mole fraction of the flowing gas mixture components, while the input parameters for the third ANN model is the the Gas, Oil and Water production rates. The developed models are trained with data records taken from a live petroleum company history logs.

On the other side, two WSN prototypes are designed and implemented with different hardware devices arrangements and power analysis is carried out for their wirelessly operated sensing nodes. The research work in this phase aims to develop mathematical correlations to emphasis the relation between the sensing node consumed current, Node-to-Gateway distance and the operating link quality value. They can be used for real time estimation of the remaining

conserved battery power in the WSN nodes and also for the power efficient node localization.

The designed prototypes are interfaced with a GSM module to be able to send the predicted data wirelessly through mobile networks via SMS. In addition to the GSM module, the collected data is also uploaded to a web server by means of a designed web program, so that data can be accessed through internet networks.

II. RELATED WORK

A. Wireless Sensor Networks

The increasing demand for the automated monitoring systems make WSN technology a target of research for many Researchers during the last decade. Although WSN is considered as a promising technology, there still many open research contributions needed to overcome the current challenges so that it can fulfill the industrial needs.

One of the challenges that WSN technology faces is energy conservation of battery-powered nodes. This topic has attracted many researchers during the last years to develop new techniques that could prolong the battery life time. Some developed techniques went through designing an adaptive control system that controls the transmission power related to the operating link quality to save power loss during high link quality conditions.

In [7], "M. Tahir et al" proposed an Energy-efficient Adaptive Scheme for Transmission in WSNs. The scheme uses an open loop for link quality estimation and compensation due to temperature variations. While in [20], "Jang-Ping Sheu et al." proposed a distributed transmission power control algorithm that is based on investigating the impact of utilizing different transmission power on the link quality values. The Received Signal Strength Indicator (RSSI) and the Link Quality Indicator (LQI) parameters are used to determine the appropriate transmission power.

"Shan Lin et al." [8] and "Yong Fu et al." [9] also have designed an Adaptive Transmission Power Control (ATPC) algorithm using a feedback-based transmission power control algorithm that gives an indication of the operating link quality values over the time to estimate the amount of the transmission power needed.

"C. Behrens et al." [10] addressed the energy conservation topic from the different point of view. They studied the effect of temperature on batteries life time and developed a computationally linear model which is capable of calculating the residual energy as a function of temperature directly on the sensing node.

In [11], Eric Alberto de Mello Fagotto et al." used the "Received Signal Strength Indicator" to develop a mathematical model that can predict the power depletion at the sensing nodes. The model should be able to monitor the charge consumption process, giving the possibility for predicting the batteries behavior and choosing the right time to replace them.

B. Gas Hydrate Formation prediction Models

Prediction of the critical conditions at which hydration is formed requires a quite understanding of how hydration is formed, and finding out the relationship between the hydrate formers such as the Methane and carbon dioxide from one side and the pressure and temperature measurements from the other side.

Several studies have been carried out on the measurement and prediction of hydrate formation temperature (HFT) for various gas mixtures. These studies can be classified as follows:

a) Hand Calculation Methods

K-VALUE METHOD

The K-Value method was developed by "Wilcox et al" in 1941 [6], it utilizes the vapor-solid equilibrium constants for prediction. The hydrate forming conditions are predicted from empirically estimated Vapor-solid equilibrium constants given by:

$$K = \frac{y_i}{x_i} \quad (1)$$

Where, y_i is the mole fraction of the i^{th} hydrocarbon component in the gas phase on a water-free basis and x_i is the mole fraction of the same component in the solid phase on a water-free basis. For any given pressure, the value of each gas component x_i can be interpolated at any given y_i value. The hydrate formation conditions should satisfy the equation:

$$\sum_{i=1}^n \frac{y_i}{x_i} = 1 \quad (2)$$

Using this method requires applying the following steps:

1. First a random value for the critical temperature should be assumed.
2. Using the assumed temperature with the given current pressure and y_i value, x_i can be interpolated for each gas composition.
3. $K = \frac{y_i}{x_i}$
4. is calculated for each gas composition.
5. $\sum_{i=1}^n \frac{y_i}{x_i} = 1$
6. is calculated, if the summation result:
 - < 1 or > 1 , all the previous steps should be repeated again by assuming new value for the critical temperature T_c .
 - $= 1$ then the assumed temperature is the exact critical value at which hydration will starts to form.

In fact, this method is not commonly used due to the large number of interpolations needed to figure out the exact critical temperature at each given pressure values, which as a result impacts negatively on the generated error.

GAS GRAVITY METHOD

The gas gravity method was developed by "Katz" [5]. The chart shown in Figure 1 is a plot of the pressure, temperature

and the specific gravity of the flowing gas mixture. For any gas gravity in addition to the current operating pressure, T_c can be calculated. In fact, as this method is not considered to be accurate, yet it is frequently used in the industry as it gives a random estimation of the critical temperatures with the minimum input data.

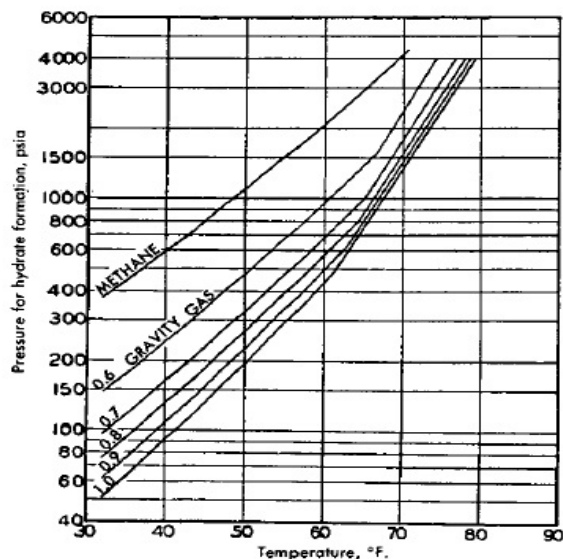


Fig. 1. Gas-gravity chart (Association, 1972)

TABLE I. THEORY AND TRAINING PARAMETERS OF THE DEVELOPED MODELS

No	Theory background	Training parameters
Model 1	Specific gravity method	Specific gas gravity and the operating pressure
Model 2	K-Factor method	Gas Compositions and the operating pressure
Model 3	N/A	Production rates (Oil, Gas, Water) and the operating pressure

b) Neural Network models

“Ehsan Khamehchi et al” [12] used 356 data records to train an ANN with a back propagation learning algorithm that is capable of determining the HFT at various system conditions, the training data used were gathered from the gas specific gravity chart and valid for the ranges between: 31.95 - 78.8 F, 50.98 - 3874.1 psi and 0.6-1 for the temperature, pressure and the specific gravity respectively. While “Amir Heydari et al” [13] developed a Neural Network model for estimation of temperature for gas hydrate formation using a training data of 167 input records ranging between 32-74 F, 50-4200 psia and 0.554-1 for temperature, pressure and specific gravity, respectively. Results in “Ehsan Khamehchi et al” [12] and Amir Heydari et al” [13] research work shows that that the proposed ANN models can be used successfully for the prediction of hydrate formation in natural gas within

their mentioned boundaries.

In [14] presented by “Jalal Foroozesh et al”, an investigation was formed for the relationship between growth rate of methane hydrate with temperature and pressure using Artificial Neural Network and Adaptive Neuro-Fuzzy Inference System (ANFIS). The results have shown that ANFIS is a more potential tool in predication relationship of kinetics of hydrate formation with temperature and pressure in comparison of ANN. The training data used in the mentioned paper was gathered from the experimental data of Chang Feng [15] and Chang Yu [16].

c) Data Regression models

Sharareh Ameripour [17] developed two correlations for calculating the hydrate-formation pressure or temperature for single components or gas mixtures. They are based on over 1,100 published data points of gas-hydrate formation temperatures and pressures and are valid for many hydrate formers such as methane, ethane, propane, carbon dioxide and hydrogen sulfide. Statistical Analysis Software (SAS) was used to find the best correlations among the input variables. They are applicable to temperatures up to 90°F and pressures up to 12,000 psi. The results have shown excellent agreement with the experimental data.

In [18], “A. Bahadori” used the Katz gas-gravity chart for developing a new correlation between pressure, temperature & molecular weights. It proves reliability for pressures between 1200 to 40000 kPa and temperatures between 265 K and 298 K, as well as the gas molecular weight within the range 16 to 29. While “Javanmardi et al.” [19] employed approaches for predicting the hydrate equilibrium based on the vdWP (Van der Waals and Platteeuw) hydrate equation. They cover the hydrate formers: ethane, carbon dioxide, xenon, and nitrogen.

III. AIM OF THE RESEARCH

Most of the developed hand calculation methods used for determining the hydrate formation temperature gives inaccurate results as they depend on some kind of interpolation. In addition to this, real-time protection for the Oil and Gas production wells from the hydrate formation process requires an effective online monitoring system based on an accurate computerized model which hand calculation methods cannot provide.

The first phase in this research work is minimizing the hand calculation error due to the interpolations by developing an accurate three computerized prediction models based on the neural network algorithm to accurately predict the critical hydrate temperature at which hydration will form. The intellectual contribution in this phase is developing a new computerized prediction model that is based on the Well production quantities of Oil, Gas and Water. Also two other models are developed based on the gas specific gravity and the K-Factor charts.

While the aim of the second phase, is developing a mathematical correlation between the sensing nodes consumed current, Node-to-Gateway distance and the operating link quality for modeling the remaining conserved power in the battery-operated nodes and for their better localization.

Experiment will run in this phase on two designed WSN prototypes implemented with National Instruments hardware devices. All the designed WSN prototypes uses ANN model 1.

IV. EXPERIMENTAL WORK

All the proposed ANN models in this research work are trained by the "trainlm." training function where the "trainlm" function updates the weights and biases according to the Levenberg-Marquardt optimization. During the training process, the ANN model starts to divide the input data records into three subsets according to the set division function. There are four types of them used in the ANN models which are the Dividerand, divideblock, divideint and the divideind function. For the proposed model the Dividerand default function is used to divide the data randomly into the Training, Validation and Testing subsets.

The training data set is used for training and updating the weight and biases of the proposed network, while the validation set is then used to tune up the developed network by comparing the model results to the validation data set, when the validation error is computed, the network is retrained by adjusting its weights again with the training data set.

Certain parameters are used to monitor the learning process of the developed network such as the min_grad, max_fail, mu, mu_dec, mu_inc and the mu_max parameters. For every successful step (validation error decreased) the mu parameter is multiplied by the mu_dec factor, while in case the validation error increases, the mu parameter is multiplied with the mu_inc factor. The training will stop when the validation error is kept constant for a certain number of steps determined by the max_fail parameter which means no further improvement in the network training process. Also the training stops when the mu parameter reaches either it's minimum or maximum values according to the min_grad and mu_max parameters respectively.

PHASE 1

A. ANN prediction model 1

- Architecture

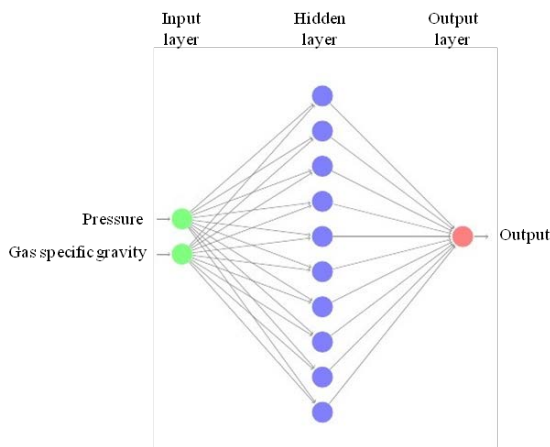


Fig. 2. ANN Model 1 architecture

- Regression curve

In this method, the regression curves in Figure 1 shows that the correlation coefficient is 1, 0.99998 and 1 for the training, validation and test sets respectively with a total average of 0.99997.

- Error

According to the model test results in Table II, the maximum error of 0.161088% was found between the ANN model output and the experimental test data.

TABLE II. MODEL 1 OUTPUT DATA RESULTS

Record	Experimental data	ANN model	Error %
1	62.312	62.308	0.005
2	62.221	62.251	0.049
3	60.554	60.527	0.042
4	49.465	49.483	0.036
5	62.613	62.625	0.020
6	61.621	61.621	5.17E-05
7	54.430	54.471	0.075
8	58.331	58.316	0.024
9	56.834	56.830	0.006
10	38.524	38.586	0.161

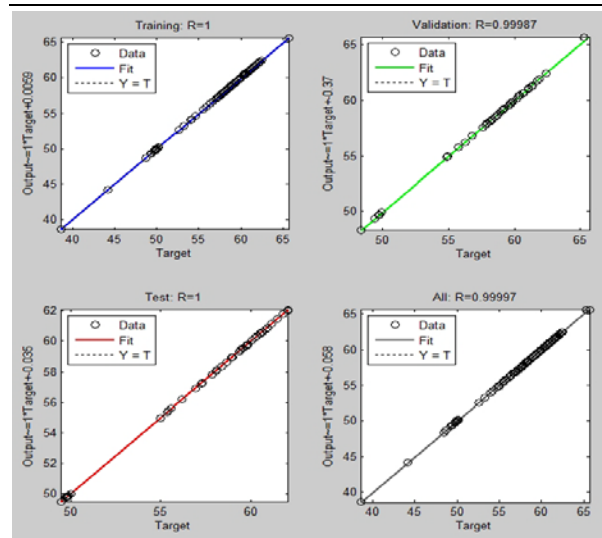


Fig. 3. Regression curves for Model 1

B. ANN prediction model 2

- Architecture

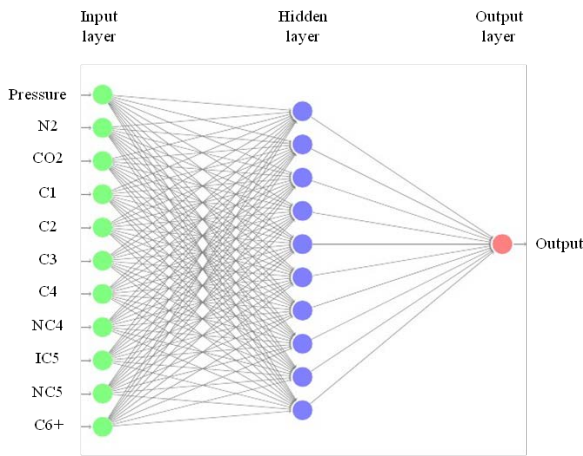


Fig. 4. ANN Model 1 architecture

- Regression curve

In this method, the regression curves shows that the correlation coefficient is 0.99713, 0.99225 and 0.96628 for the training, validation and test sets respectively with a total average of 0.99279.

- Error

According to the model test results in Table III, it is shown that a maximum error of 0.10684% was found between the ANN model output and the experimental test data.

TABLE III. MODEL 2 OUTPUT DATA RESULTS

Record	Experimental data	ANN model	Error %
1	62.312	62.311	0.045
2	60.693	60.694	0.001
3	60.171	60.171	0.106
4	57.823	57.823	0.001
5	61.494	61.493	0.003
6	60.434	60.434	0.002
7	57.731	57.730	0.004
8	49.897	49.896	0.041
9	57.185	57.184	0.003
10	56.911	56.910	0.059

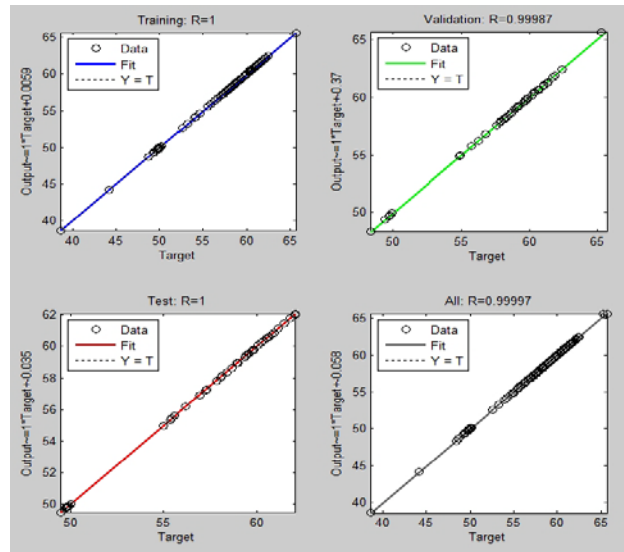


Fig. 5. Regression curves for Model 2

C. ANN prediction model 3

- Architecture

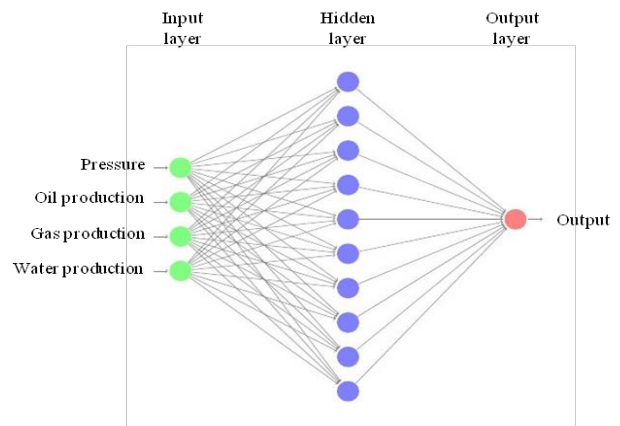


Fig. 6. ANN Model 1 architecture

- Regression curve

In the proposed model, the correlation coefficient for the training, validation and test curves shows high linearity between the model output and the target data. By plotting the

regression curves as shown in Figure 7, the correlation coefficient value was 0.99713, 0.99225 and 0.96628 for the training, validation and test sets respectively with a total average of 0.99279.

• Error

A maximum error of 1.8 % was found between the ANN output data and the experimental test data. The output data results of the ANN model are shown in Table IV.

TABLE IV. MODEL 3 OUTPUT DATA RESULTS

Record	Experimental test data	ANN model	Error %
1	62.312	62.280	0.050
2	61.140	60.963	0.288
3	59.324	59.774	0.758
4	59.666	59.968	0.506
5	57.920	58.144	0.387
6	49.947	49.938	0.017
7	60.980	61.195	0.352
8	58.521	58.971	0.770
9	38.649	39.038	1.007
10	54.927	53.935	1.804

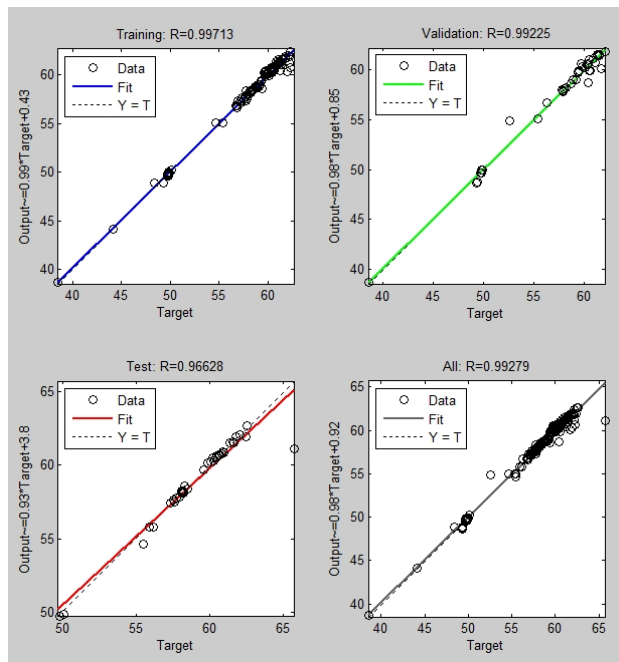


Fig. 7. Regression curves for Model 3

PHASE 2

A. Prototype Model 1

a) Hardware

All the hardware devices used in this prototype model can be used in the industrial environments except for the Sensors

and the TP-Link router. Table V shows the hardware devices used in this prototype.

TABLE V. PROTOTYPE MODEL 1 HARDWARE DEVICES

Name	Description
Wireless sensor node 1	NI WSN 3212
Wireless sensor node 2	NI WSN 3202
GSM Module	SEA 9703
Controller	Compact Rio 9075
Pressure sensor	Adjustable knobs for simulation
Specific gravity sensor	Adjustable knobs for simulation
Temperature sensor	Nonindustrial J type thermocouple
Gateway	NI 9791
Internet communication	TP-link router MR 3020 with Vodafone internet card

b) Distance Vs. Link quality analysis for both Nodes

The ambient temperature of the experiment is 20 Celsius to find out the effect of distance variation between the sensing nodes and the Gateway sink node on the link quality measurements. As shown in Figure 8, various line of sight test points as are selected at different distance points captured by Google Earth. The experiment results are shown in Table VI.

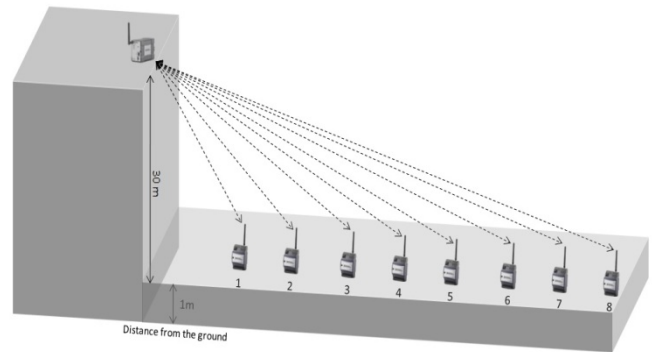


Fig. 8. Localization of node test points

TABLE VI. DISTANCE VS. LINK QUALITY RESULTS

POINT No.	DISTANCE (m)	LINK QUALITY
1	23	78
2	46	65
3	94	57
4	138	45
5	173	36
6	202	27
7	240	22
8	278	17

From the given results mentioned Table VI, Microsoft data regression tool developed a mathematical correlation Equation 3 between the Node-to-Gateway distance and the link quality values for better localization of the sensor nodes based on the current consumption rates.

$$L = 2.7745 \times 10^{-12} * d^6 - 3.3372 \times 10^{-9} * d^5 + 1.5452 \times 10^{-6} * d^4 - 0.000345 * d^3 + 0.038303 * d^2 - 2.17667 * d + 111.550 \quad (3)$$

c) Power analysis

The power consumption analysis in this research work concerning the battery powered sensor nodes uses the NI power quality tool kit, while the power calculations for the solar powered hardware devices refers to the manufacturer’s datasheet.

For the sensor node power analysis, the sensor node-link quality indicator is used to monitor the transmission efficiency between the sensor node and the gateway node, while the NI 9227 current input module of the NI power quality tool kit is used to measure the consumed current from the sensor nodes side at every monitored link quality value. Measurements are carried out at full battery voltage with one sample interval per second.

All the consumed current calculations are measured in terms of the Root Mean Square value (RMS) by the LabVIEW RMS function in the electrical power suite software kit, while the graphical representation used for simulation is programmed by the LabVIEW graphical functions.

NODE 1

- Current consumption vs. Link quality

This experiment was done for each node to investigate the relation between the sensing node consumed current at different link quality values. The link quality points selected for this node were taken randomly at values of 96, 87, 64, 44, 30, 23 and 11. The minimum and maximum current consumptions were found 13.31 RMS at link quality of 96 and 16.69 RMS at 11.

TABLE VII. POWER ANALYSIS FOR NI WSN 3202 AT DIFFERENT LINK QUALITY VALUES

Link quality	Voltage (v)	Current RMS (mA)	Power (w)
96	6.2	13.31	0.833
87	6.2	13.81	0.845
64	6.2	13.878	0.860
44	6.2	14.41	0.868
30	6.2	15.86	1.002
23	6.2	16.11	1.012
11	6.2	16.69	1.034

A chart representation for the generated results in Table VII is shown in Figure 9.

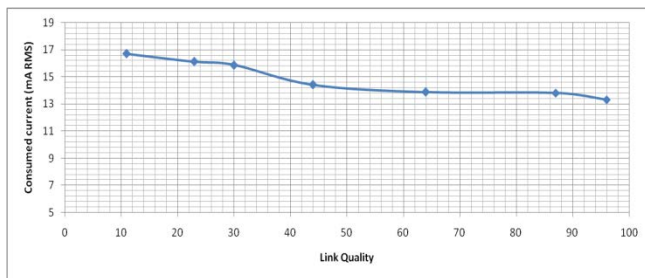


Fig. 9. NI Node 3202 consumed current at different link quality values

- Regression Equation

The regression Equation 4 is developed based on the generated results in Table VII from the node power analysis. Microsoft Excel regression tool is used to develop the regression equation with a total number of 7 input data points.

Where I is the node consumed current while x_L is the link quality value.

Regression results show a polynomial Equation 5-3 of the fourth-degree that gives a quite relation between the consumed current and different link quality values. The equation is valid at link quality values ranging between 11 and 96.

$$I = -6.663 * 10^{-7}x_L^4 + 0.000142x_L^3 - 0.009454x_L^2 + 0.17204x_L + 15.7496 \quad (4)$$

NODE 2

- Current consumption vs. Link quality

The randomly selected quality link points for this node are at values of 99, 95, 87, 67, 54, 41, 31, 22, 17 and 11. Results shown in Table VIII shows that the maximum current consumption was 10.98 mA at a link quality of 98 while, the minimum current consumption was 13.926 mA at link quality 11.

TABLE VIII. POWER ANALYSIS FOR NI WSN 3202 AT DIFFERENT LINK QUALITY VALUES

Link quality	Voltage (v)	Average Current RMS (ma)	Power (w)
99	6.3	10.98	0.691
95	6.3	11.054	0.696
87	6.3	11.07	0.697
67	6.3	11.246	0.708
57	6.3	11.12	0.700
41	6.3	11.057	0.696
31	6.3	13.167	0.829
22	6.3	13.873	0.874
17	6.3	13.896	0.875
11	6.3	13.926	0.877

A chart representation for the generated results in Table VIII is shown in Figure 10.

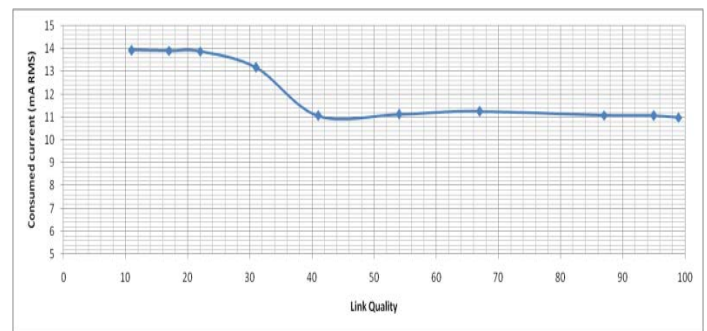


Fig. 10. NI Node 3212 consumed current at different link quality values

- Regression Equation

The regression Equation 5 is developed based on the generated results in Table VIII from the node power analysis, where (I) is the node consumed current while (x) is the link quality value.

$$I = 2.69 * 10^{-8}x_L^5 - 8.14 * 10^{-6}x_L^4 + 0.00091x_L^3 - 0.04477x_L^2 + 0.861x_L + 8.654 \quad (5)$$

B. Prototype Model 2

This prototype is designed to test the indoor hardware devices operation with the developed ANN model. A mobile workstation laptop is used instead of using the CompactRio controller, while the EFCOM GSM module is used instead of the SEA 9707 GSM communication module.

a) Software

In this prototype, the mobile workstation laptop is used as the host controller to interface with the gateway and the EFCOM GSM modules through the Ethernet and serial ports respectively. As shown in Figure 11 and Figure 12, the designed LabVIEW program is almost the same as what was in prototype 1 except for using the MATLAB script function in loop1. The main advantage of using the MATLAB script function is giving the opportunity for the LabVIEW software to execute the previously developed ANN model 1 through the MATLAB software directly.

In the second loop, the EFCOM GSM module is initialized and programmed by the AT commands using the LabVIEW block functions. At program execution, the module is initialized and enters into a while loop waiting for a Ring signal from the host cell phone. Once the module receives the ring signal, it cancels the host phone call and starts to collect the predicted data from the first loop to send it directly to the preconfigured host number. The time taken between receiving the ring signal and sending the predicted data via SMS was found to be with an average value of 10 seconds

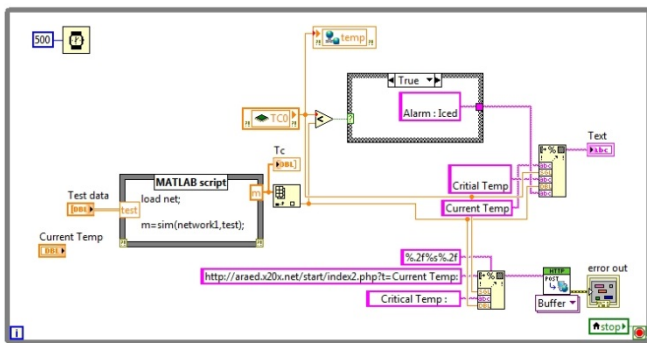


Fig. 11. Loop 1 LabVIEW program

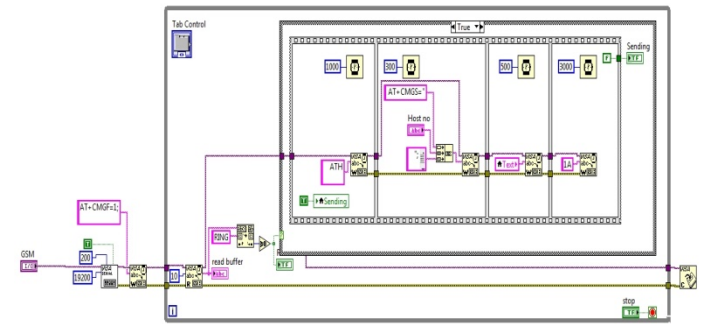


Fig. 12. Loop 2 LabVIEW program

b) Hardware

Most of the hardware devices used in this prototype model shown in Table IX are for indoor applications only except the WSN nodes and the NI 9791 Gateway.

TABLE IX. PROTOTYPE MODEL 2 HARDWARE DEVICES

Name	Description
Wireless sensor node 1	NI WSN 3212
Wireless sensor node 2	NI WSN 3202
GSM Module	EFCOM module
Controller	Dell M4800 Mobile Workstation Laptop
Pressure sensor	Adjustable knobs for simulation
Specific gravity sensor	Adjustable knobs for simulation
Temperature sensor	Non industrial J type thermocouple
Gateway	NI 9791

V. CONCLUSION AND FUTURE WORK

Designing wireless monitoring systems capable of predicting the hydrate formation temperatures is highly concerned by many industrial domains. The remotely located production wells, in the Oil and Gas industry make the wireless monitoring systems the ideal solution for the HFT prediction. The designing process in this research work is divided into two phases. In the first phase, three Artificial Neural Network models have been developed for Hydrate formation prediction with different input parameters. The developed models have been trained and tested with an input data of more than 700 data records collected from the history logs of a working Petroleum Company. The test results for all the developed models have shown correlation coefficient values nearly equals to one which indicates the high linearity between the models output and the target data records. ANN

model 1 test results confirmed the observations presented in the previous work giving the best results with a maximum correlation coefficient (R) of 0.9999 while ANN model 2 is 0.99721. A third model was proposed in this research work based on new input parameters that are not commonly known in HFT prediction systems. The model has shown acceptable results with a correlation coefficient of 0.9921. All the developed models are valid to use in computerized systems within the mentioned training data ranges.

While in the second one, two WSN prototypes have been designed and implemented with National Instruments hardware devices. Different analyses were carried out on the wireless nodes and two mathematical correlations have been developed. The developed correlations gave a quite understanding of the relation between the WS consumed current, Node-to-Gateway distance and the link quality parameters.

WSN is considered as one of the promising technologies in the field of monitoring and control, yet there still some barriers that prevent their fuller adoption in various industrial applications. Nodes power efficiency is considered as one of the existing challenges that are concerned by many researchers during the last decades. The remotely located nodes in most cases are powered up by internal batteries that have a specified life time. Different techniques have been developed for battery power conservation such as: the duty cycle based operation and the automatic transmission power control, while other techniques are used for maximizing the amount of the conserved power such as Energy harvesting from wind, mechanical vibration, temperature variation and the solar power sources. For the future work, it is intended to explore the opportunities and challenges of using the Energy Harvesting techniques combined with the rechargeable batteries and super capacitors for the energy storage. Developing the harvesting techniques enables the wireless sensor nodes to last potentially more than what it typically lasts by normal batteries.

REFERENCES

- [1] Zhao Gang Wireless Sensor Networks for Industrial Process Monitoring and Control: A Survey [Journal] // Network Protocols and Algorithms. - Davis CA : [s.n.], 2011. - Vol. 3.
- [2] Chiara Buratt Andrea Conti, Davide Dardari and Roberto Verdone An Overview on Wireless Sensor Networks Technology and Evolution [Journal] // Sensors Journal. - Newport : [s.n.], August 31, 2009. - 9 : Vol. 9. - pp. 6869-6896.
- [3] Neha Singh Prof.Rajeshwar Lal, Vinita Mathur Wireless Sensor Networks: Architecture, Protocols, Simulator Tool [Journal]. - [s.l.] : International Journal of Advanced Research in Computer Science and Software Engineering, May 2012. - 5 : Vol. 2.
- [4] Mohammad Javad Jalalnejhad Mohammad Ranjbar, Amir Sarafi and Hossein Nezamabadi-Pour Comparison of intelligent systems, artificial neural networks and neural fuzzy model for prediction of gas hydrate formation rate [Journal] // International journal for science and engineering. - 2014. - Vol. 7. - pp. 35-40.
- [5] Katz D. L. Prediction of conditions for hydrate formation in natural gases [Journal] // Trans. A.I.M.E.. - 1945. - Vol. 160. - pp. 140-149.
- [6] Willard I. Wilcox D. B. Carson, D. L. Katz Natural Gas Hydrates [Journal] // Industrial & Engineering Chemistry. - 1941. - 5 : Vol. 33. - pp. 662-665.
- [7] M. Tahir N. Javaid, A. Iqbal, Z. A. Khan, N. Alrajeh On Adaptive Energy Efficient Transmission in WSNs [Journal] // International Journal of Distributed Sensor Networks. - 2013. - Vol. 2013. - p. 10 pages.
- [8] Shan Lin Jingbin Zhang, Gang Zhou, Lin Gu, John A. Stankovic, and Tian He ATPC: adaptive transmission power control for wireless sensor networks [Journal]. - New York : In SenSys '06: Proceedings of the 4th international conference on Embedded networked sensor systems, 2006. - pp. 223-236.
- [9] Mo Sha Yong Fu, Gregory Hackmann and Chenyang Lu Practical Control of Transmission Power for Wireless Sensor Networks [Journal]. - Austin : The 20th IEEE International Conference on Network Protocols, October 2012. - pp. 1-10.
- [10] Behrens C., Bischoff, O., Lueders, M., Laur Energy-efficient topology control for wireless sensor networks using online battery monitoring [Journal] // Advances in radio science. - Deutschland : [s.n.], 2007. - Vol. 5. - pp. 1-4.
- [11] Inacio Henrique Yano Vitor ChavesDe Oliveira, Eric Alberto de Mello Fagotto, Alexandre De Assis Mota and Lia Toledo Moreira Mota Predicting battery charge depletion in wireless sensor networks using received signal strength indicator [Journal] // Journal of Computer Science. - Brazil : [s.n.], 2013. - Vol. 7. - pp. 821-826. - 1549-3636.
- [12] Ehsan Khamehchi Ebrahim Shamohammadi and Seyed Hamidreza Yousefi Predicting the Hydrate Formation Temperature by a New Correlation and Neural Network [Journal] // Gas Processing Journal. - Iran : [s.n.], 2013. - Vol. 1. - pp. 41-50.
- [13] Amir Heydari Keivan Shayesteh, Ladan Kamalzadeh Prediction of hydrate formation temperature for natural gas using Artificial Neural Network [Journal] // Electronic scientific journal "Oil and Gas Business". - Ardebil : [s.n.], 2006. - 2.
- [14] Jalal Foroozesh Abbas Khosravani, Adel Mohsenzadeh, Ali Haghghat Mesbahi Application of Artificial Intelligence (AI) Modeling in Kinetics of Methane Hydrate Growth [Journal] // American Journal of Analytical Chemistry. - 2013. - Vol. 4. - pp. 616-622.
- [15] C. Ma G. Chen and T. Guo Kinetics of Hydrate Formation Using Gas Bubble Suspended in Water [Journal] // SCIENCE CHINA Chemistry (Science in China Series B: Chemistry). - 2002. - Vol. 45. - pp. 208-215. - 1674-7291.
- [16] C. Y. Sun G. J. Chen, C. F. Ma, Q. Huang, H. Luo and Q. P. Li The Growth Kinetics of Hydrate Film on the Surface of Gas Bubble Suspended in Water or Aqueous Surfactant Solution [Journal] // Journal of Crystal Growth. - 2007. - 2 : Vol. 306. - pp. 491-499.
- [17] Ameripour Sharareh Prediction of gas-hydrate formation conditions in production and surface facilities [Journal]. - Texas, USA : Texas A&M University, August, 2005. - p. 79.
- [18] Alireza Bahadori Hari. B. Vuthaluru A novel correlation for estimation of hydrate forming condition of natural gases [Journal] // Journal of Natural Gas Chemistry. - Perth, Western Australia : [s.n.], 2009. - 4 : Vol. 18. - pp. 453-457.
- [19] Javanmardi Jafar, Partoon Behzad and Sabzi Fatemeh Prediction of hydrate formation conditions based on the vdWP-type models at high pressures [Journal] // The Canadian Journal of Chemical Engineering. - 2011. - 2 : Vol. 89. - p. 254.
- [20] Jang-Ping Sheu Kun-Ving Hsieh and Yao-Kun Cheng Distributed Transmission Power Control Algorithm for Wireless Sensor Networks [Journal] // Journal Of Information Science And Engineering. - Taiwan : [s.n.], 2009. - Vol. 25. - pp. 1447-1463.

The Role of Image Enhancement in Citrus Canker Disease Detection

K. Padmavathi

Research Scholar
Research and Development Centre
Bharathiar University, Coimbatore, India

Dr. K. Thangadurai

Assistant Professor and Head,
P.G & Research Department of Computer Science
Government Arts College (Autonomous)
Karur, India

Abstract—Digital image processing is employed in numerous areas of biology to identify and analyse problems. This approach aims to use image processing techniques for citrus canker disease detection through leaf inspection. Citrus canker is a severe bacterium-based citrus plant disease. The symptoms of citrus canker disease typically occur in the leaves, branches, fruits and thorns. The leaf images show the health status of the plant and facilitate the observation and detection of the disease level at an early stage. The leaf image analysis is an essential step for the detection of numerous plant diseases.

The proposed approach consists of two stages to improve the clarity and quality of leaf images. The primary stage uses Recursively Separated Weighted Histogram Equalization (RSWHE), which improves the contrast level. The second stage removes the unwanted noise using a Median filter. This proposed approach uses these methods to improve the clarity of the images and implements these methods in lemon citrus canker disease detection.

Keywords—Lemon tree; Citrus Canker; Recursively Separated Weighted Histogram Equalization; Median Filter; Image Enhancement; Disease detection

I. INTRODUCTION

In India, various citrus plants are cultivated in every state. However, the geographic areas of Andhra Pradesh, Karnataka, Punjab, Assam and TamilNadu are the leading citrus-growing states. The foremost important species and varieties of citrus grown in India are the loose-jacket oranges or santras, sweet oranges or tight jacket oranges, Sathgudl oranges, musambi oranges, acid limes or Kaghzi limes, and lemon. Several species of citrus might be affected by diseases like fungi, bacteria and viruses. In recent years, the foremost severe and customary disease is citrus canker, and it is considered a serious problem where lemon is mature on a large-scale. Citrus canker is seen on leaves, twigs, branches, fruit stalks, fruits and thorns and may lead to tree death and loss of yield. Citrus canker appears as yellowish spots or halos on leaves that gradually enlarge to 2 – 4 mm dark brown pustules. Canker on the fruit does not have the yellow halo seen on leaves. This proposed approach considers lemon leaves for detection of citrus canker disease and uses two major techniques to obtain clear, high quality images. The high quality images will facilitate easy disease identification. Within the past decade, various histogram equalisation and filtering techniques have been used to increase image quality in several applications such as eye and plant diseases. The

proposed approach collects or captures unhealthy leaf images from outside and uses RSWHE and a Median filter to get high quality images. These high quality images are necessary for leaf disease detection. The following figures show various stages of citrus canker on the leaves.



Fig. 1. Initial stage of Citrus Canker



Fig. 2. Maturity stage of Citrus Canker

II. EVALUATION OF PROPOSED SYSTEM

Researchers have implemented various techniques to enhance the images utilised in plant disease detection. A.Camargo and J.S.Smith[2009]_[11] utilised histogram equalization to distribute the intensities of an image and improve the image quality and visual appearance of the plant leaf images. Anjali Naik[2010]_[2] used histogram equalisation in dental disease detection to obtain high contrast images and better views of bone structure in radiographic images. T.Jintasuttisak and S. Intajag [2014]_[31] developed a technique called Rayleigh Contrast Limited Adaptive Histogram Equalization. Using this technique, the image contrast and overall appearance of images used to discover vision-related diseases was improved. Mary Kim and Min Gyo Chung

[2013]_[4] proposed another method, Recursively Separated and Weighted Histogram Equalization (RSWHE), to enhance the image contrast and brightness preservation. RSWHE offered higher leads in comparison to earlier methods in all aspects. With RSWHE, the input histogram was segmented into more than two sub-histograms recursively using mean or median values and a weighting process was applied to modify each sub-histogram and perform histogram equalization independently. Omprakash Patel, Yogendra.P and S.Maravi and Sanjeev Sharma[2013]_[5] reviewed many extensions of histogram equalization. Their analysis found that RSWHE produced less mean brightness error and a high peak signal to noise ratio. This method offers higher brightness preservation and contrast enhancement that conjointly offers scalable brightness preservation owing to its recursive nature. Samuel Oporto-Díaz, Rolando, Terashima-Marín [2005]_[6] proposed a method, sequential difference of Gaussian (DoG) filters, to detect microcalcification clusters in mammograms. DoG removes unwanted noise and classifies regions. G. Kale Vaishanw[2014]_[9] highlighted the uses of varied filtering techniques like Gaussian, Laplacian and Median filters. He emphasised the importance of image enhancement in X-ray lung images. He maintained that the filtering technique selection is based on both the application and process image enhancement. M.A. Shaikh and Dr. S.B.Sayyad[2014]_[10] used colour image enhancement in the agricultural domain. They implemented histogram equalization for contrast enhancement and compared linear (Wiener) filters and nonlinear(Median) filters. They observed that the Median filter was better at noise reduction and removed the blurred effect of the image. J.M. Durge, Prof. N.P. Bobade and Dr N.N. Mhala[2015]_[8] proposed to detect early stage lung cancer. They used the Median filter to enhance the image and suppress noise and other fluctuations of lung images.

Based on our literature review, we carried out our image enhancement using two different methods: histogram and filtering. Our proposed approach uses histogram equalization for contrast enhancement and filtering for unwanted noise removal.

III. PROPOSED APPROACH

The proposed methods play an important role in citrus canker disease detection by improving image quality with greater clarity. Here, we have developed two strategies to enhance the contrast and quality of the lemon leaf images: RSWHE increases the contrast of the image, and the Median filter removes the noise.

A. Mathematical Model

1) Recursively separated and Weighted Histogram Equalization (RSWHE)

RSWHE increases the contrast of an image in the following three stages:

- a. Histogram segmentation: Consider the image X and calculate the histogram H(X). H(X) is then divided into the number of sub-histograms.
- b. Histogram weighting: Change the sub-histograms using normalized power law.

- c. Histogram equalization: Equalize the weighted sub-histograms independently over the changed sub-histograms.

a) **Histogram Segmentation:** This divides the histogram H(X) using the recursion level r and creates 2r sub-histograms based on the mean value. Consider a sub-histogram $H^t(X)$ over a range $[X_l, X_u]$ at a recursion level t ($0 \leq t < r$). The sub-histogram $H^t(X)$ mean value (i.e., X_m^t) is computed using the formula: $X_m^t = \sum_{k=1}^u k \cdot p(k) / \sum_{k=1}^u p(k)$.

$H^t(X)$ divides into $H^{t+1}_L(X)$ and $H^{t+1}_U(X)$ based on $[X_m^t]$. Here, $H^{t+1}_L(X)$ and $H^{t+1}_U(X)$ over $[X_l, X_m^t]$ and $[X_m^t + 1, X_u]$, respectively.

b) **Histogram Weighting:** The histogram segmentation creates 2r sub-histograms $H_i^r(X)$ ($0 < i \leq 2^r - 1$). For r = 2, the histogram weighting changes the Probability Density Function p(k) of $H_i^r(X)$ as follows:

- i) Calculate the maximum and minimum probability (p_{max} and p_{min}) using the equations: $p_{max} = \max_{0 < k < L} p(x)$ and $p_{min} = \min_{0 < k < L} p(x)$.
- ii) Calculate a cumulative probability α_i using $\alpha_i = \sum_{k=1}^u p(k)$ for $H_i^r(X)$.
- iii) Change the p(k) into weighted p(k) using the following formula:

$$p_w(k) = p_{max} (p(k) - p_{min} / p_{max} - p_{min})^{\alpha_i} + \beta \quad (l \leq k \leq u)$$

Here, β adjusts the brightness and contrast of the image and $\beta \geq 0$. The $p_w(k)$ is normalized using $p_{wn}(k) = p_w(k) / \sum_{k=0}^{L-1} p(k)$ and forwarded to the next module.

c) **Histogram Equalization:** The $p_{wn}(k)$ contains 2^r curves and bounds by the range $[X_l, X_u]$ of $H_i^r(X)$. This module equalizes all 2^r sub-histograms and combines all resultant images. It provides high quality images as its output.

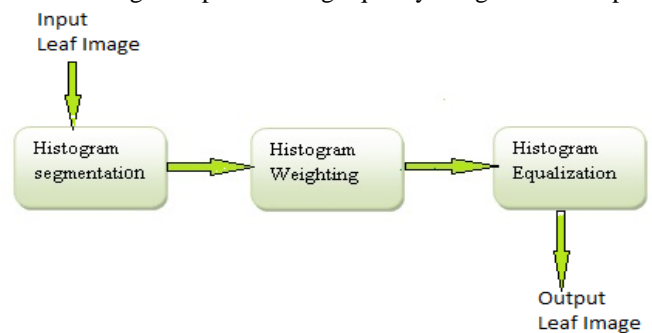


Fig. 3. A model of RSWHE

2) Median Filter

We can get the higher end in conserving, removing noise and isolating noise spikes using the Median filter. The median m could be a group of numbers in this group if half the numbers are less than m and half are greater. The median m could be a midpoint pixel value that is drawn from the neighbourhood sorted distribution values. It does not produce a new, unrealistic pixel value. The median operator arranges the values within the pixel element neighbourhood at each

pixel element location. This reduces edge blurring and loss of image detail.

Algorithm:

Step 1: Select a 3×3 size two-dimensional centre window $p(x, y)$ from the input image.

Step 2: Rank the pixel values $p(x, y)$ within the selected window in ascending order and find the median, maximum and minimum pixel values (P_m, P_{max} and P_{min}).

Step 3: If the pixel value $p(x, y)$ has the limit $P_{min} < P(x, y) < P_{max}$, $p(x, y)$ is taken into account as uncorrupted. Otherwise, $p(x, y)$ is taken into account as corrupted.

Step 4: The corrupted $p(x, y)$ has two categories:

Category 1: If $P_{min} < P_m < P_{max}$ and $0 < P_m < 255$, replace the corrupted $p(x, y)$ with P_m .

Category 2: P_{min} considered a noisy pixel. Here, calculate the difference between each adjacent pixel across the ranked values, find the maximum difference and mark it as the next processed pixel.

Step 5: Apply steps 1 to 4 to the entire image until the process is complete.

Neighborhood Values:

126,127,133,115,119, 135,118,120,150

Max: 135

Min: 115

TABLE I. MEDIAN FILTER-NEIGHBORHOOD VALUES SELECTION

130	140	123	125	126
134	126	120	118	122
124	127	150	135	118
119	133	115	119	123
120	110	116	111	130

IV. EXPERIMENTS AND RESULTS

Image enhancement methods have been applied to various lemon citrus canker diseased leaf images using MATLAB. RSWHE has been used to increase the brightness of images, while the Median filter has been used for de-noising corrupt images.

Here, we have taken sample citrus canker diseased leaves images and applied RSWHE and Median filter and got high quality citrus canker diseased images.



Fig. 4. Original image

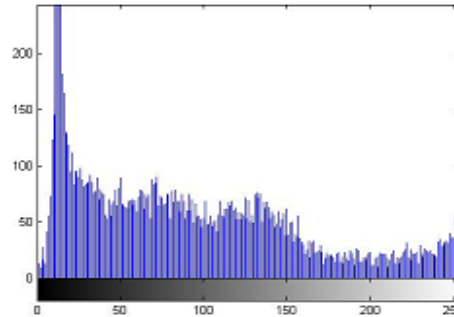


Fig. 5. Original Image Histogram



Fig. 6. Equalized Image

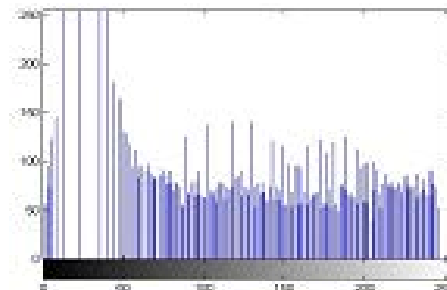


Fig. 7. Equalized Image Histogram

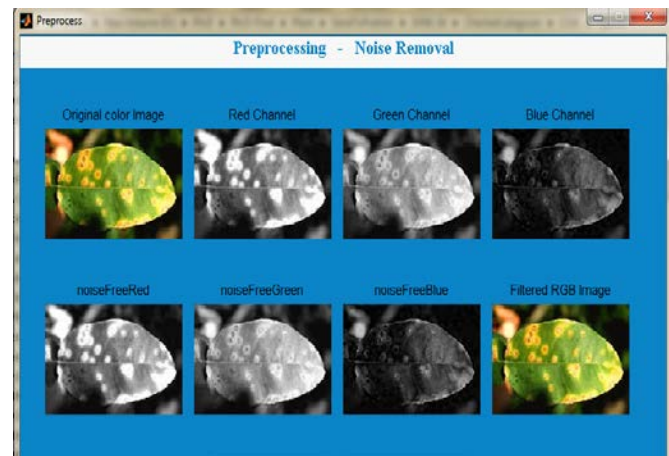


Fig. 8. Filtering process

V. CONCLUSION

Compared to the predominant strategies, our proposed approach has produced better image enhancement. This approach applies RSWHE and a Median filter to enhance

citrus canker diseased leaf images. Using this method, we tend to get clear, high quality leaf images for further processing. This approach highlights the results of the enhancement process in citrus canker disease detection, showing that the visual illustration of diseased leaf images is improved.

In the future, enhanced citrus canker images will be taken and applied segmentation method which is used to extract the diseased portion. They may be utilised for citrus canker disease detection and the identification of disease level.

REFERENCES

- [1] Camargo A, Smith J S. An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosyst Eng* 2009; 102: 9-21.
- [2] Anjali Naik, Tikhe S v, Bhide S D. Histogram Equalization for Class-Identification of Dental Disease Using Digital Radiography. *Communications in Computer and Information Science* 2010; 70: 144-151.
- [3] Jintasuttisak T, Intajag S. Color retinal image enhancement by Rayleigh contrast-limited adaptive histogram equalization. In: *Control, Automation and Systems (ICCAS) 2014 14th International Conference* ; 22-25 October 2014; Seoul: IEEE. pp. 692–697.
- [4] Kim M, Chung MG. Recursively Separated and Weighted Histogram Equalization for brightness Preservation and Contrast Enhancement. *IEEE T Consum Electr* 2008; 54: 1389-1397.
- [5] Omprakash Patel, Yogendra P. S. Maravi and Sanjeev Sharma. A Comparative Study of Histogram Equalization Based Image Enhancement Techniques for Brightness Preservation and Contrast Enhancement. *Signal & Image Processing: An International Journal (SIPIJ)* 2013; 4: 11-25.
- [6] Oporto-Díaz S, Hernández-Cisneros R, Terashima-Marín H. Detection of Microcalcification Clusters in Mammograms Using a Difference of Optimized Gaussian Filters. In: Kamel M, Campilho A, editors. *Image Analysis and Recognition*. Toronto, Canada: Springer Berlin Heidelberg Publisher, 2005; pp: 998-1005.
- [7] Swati R. Dixit, Dr.Amol Y. Deshmukh. Design Strategies for Various Edge Detection Techniques for Disease Detection. *International Journal of Scientific Engineering and Applied Science* 2015; 1: 498-505.
- [8] Durge J M, Bobade N P, Dr.Mhala N N. Image Enhancement to Detect The Lung Cancer at Early Stage Using Median Filter. *International Journal of Science, Engineering and Technology* 2015; 3: 936-939.
- [9] Kale Vaishanw G. X-Ray Lung Image Enhancement by Spatial filtering. In: *VISHWATECH 2014 Two days National Conference*; 21-22 February 2014; Maharastra, INDIA: *International Journal of Innovative Research in Science, Engineering and Technology*. Pp. 330-336.
- [10] Shaikh M A, Sayyad S B. Color Image Enhancement Filtering Techniques for Agricultural Domain Using MATLAB. In: *VIII International Symposium on Operational Remote Sensing Applications: Opportunities, Progress and Challenges*; 9-2 December 2014; Hyderabad, INDIA.
- [11] Gavhale KR, Gawande U, Hajari K O. Unhealthy region of citrus leaf detection using image processing techniques. In: *IEEE 2014 International Conference for Convergence of Technology*; 6-8 April 2014; Pune, INDIA: IEEE. Pp. 1–6.
- [12] Kim D G, Burks T F, Schumann A W, Zekri M, Zhao X, Qin J. Detection of Citrus Greening Using Microscopic Imaging. *Agricultural Engineering International: the CIGR Ejournal* 2009; XI: 1-17.
- [13] Chen E, Chung P, Chen C, Tsai H, Chang C. An Automatic Diagnostic System for CT Liver Image Classification. *IEEE T Bio-Med Eng* 1998; 45: 783-794.
- [14] Narvekar PR, Kumbhar M M, Patil S N. Grape Leaf Diseases Detection & Analysis using SGDM Matrix Method. *International Journal of Innovative Research in Computer and Communication Engineering* 2014; 2: 3365-3372.
- [15] Gwan D, Burks TF, Qin J, Bulanon D M. Classification of grapefruit peel diseases using color texture feature analysis. *Int J Agric & Biol Eng* 2009; 2: 41-50.
- [16] Kumar V, Bansal H. Performance Evaluation of Contrast Enhancement Techniques for Digital Images. *International Journal of Computer Science and Technology* 2011; 2: 23-27.
- [17] Huynh-The T, Le B, Lee S, Le-Tien T, Yoon Y. Using weighted dynamic range for histogram equalization to improve the image contrast. *Int J Image Video Process* 2014; 44.
- [18] Rani S, Kumar M. Contrast Enhancement using Improved Adaptive Gamma Correction with Weighting Distribution Technique. *International Journal of Computer Applications* 2015; 101: 47-53.
- [19] Sonia, Goel M, Goel R. Comparative Study of Image Enhancement Using Histogram Equalization Based Processing Techniques. In: *ITCSE 2015 National Conference on Innovative Trends in Computer Science Engineering*; 4 April 2015; Bahal, INDIA: IJRRRA. pp. 172-174.
- [20] Kotkar VA, Sanjay S. Gharde. Review of Various Image Contrast Enhancement Techniques. *International Journal of Innovative Research in Science, Engineering and Technology* 2013; 2: 2786-2793.
- [21] Chen SD, Ramli A R. Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation. *IEEE T Consum Electr* 2003; 49: 1301–1309.
- [22] Sim K S, Tso C P, Tan Y Y. Recursive sub-image histogram equalization applied to gray scale images. *Pattern Recogn Lett* 2007; 28: 1209–1221.
- [23] Gonzalez R, Woods R. *Digital Image Processing*. 3rd ed. Delhi, India: Pearson education, 2010.
- [24] Gonzalez R, Woods R, Eddins S. *Digital Image Processing using MATLAB*. 3rd ed. Delhi, India: Pearson education, 2012.

Analysis of Compensation Network in a Correlated-based Channel using Angle of Arrivals

Affum Emmanuel Ampoma

Centre for RFIC and System Technology,
School of Communication and Information Engineering,
University of Electronics Science and Technology of China,
Chengdu 611731, P.R. China

Paul Oswald Kwasi Anane

School of Information and Communication Engineering,
University of Electronics Science and Technology of China,
Chengdu 611731, P.R. China

Obour Agyekum Kwame O.-B

School of Information and Communication Engineering,
University of Electronics Science and Technology of China,
Chengdu 611731, P.R. China

Maxwell Oppong Afriyie

School of Information and Communication Engineering,
University of Electronics Science and Technology of China,
Chengdu 611731, P.R. China

Abstract—We explore combined effect of spatial correlation and mutual coupling matrix, and its subsequent effects on performance of multiple input multiple output (MIMO) systems After the decoupling process. We will also look at a correlation based stochastic channel model with the linear antenna arrays as the signal source. For the purpose of understanding, it is assumed that fading is correlated at both transmitter and receiver sides, in spite of the fact that the decoupling network enhances isolation between Receiving antenna array. In this paper, we model the transmit the antenna array in CST Microwave Studio, as a uniform linear Array with monopoles as antenna elements. On the receiving side, the scattering parameters of the coupled and decoupled monopole Array are measured in an anechoic chamber. The theoretical analysis and simulation results show the joint dependency of the system capacity on an angle of arrival (AoA) and antenna element spacing, with enhanced system performance at reduced AoAs with Increased antenna element separation. Consequently, essential benefits of MIMO system performance can be achieved with an efficient decoupling network while boosting the signal sources by adding further antenna elements.

Keywords—Angle of arrival (AoA); channel correlation; decoupling network; mutual coupling; MIMO

I. INTRODUCTION

One of the challenges of MIMO antenna design is the task of enhancing the isolation between ports nearly located within restricted space in the mobile handset. This is because of the way the array elements have to be contained in a reduced volume, which brings about substantial pattern/spatial correlation and Strong mutual coupling effect between the elements. It is the common conclusion that mutual coupling influences the performance of antenna arrays, as the increase in correlation Restricts the channel capacity. Moreover, if mutual coupling is solid, a massive portion of the power fed into one port will be coupled to the other port rather than radiating to free space; consequently diminishing the signal-to-noise ratio, radiation Sufficiency and channel capacity. Some of works have investigated the effect of mutual coupling on the performance of communication system [1]-[10]. For

this reason, building a successful decoupling technique to balance the performance degradation in MIMO antennas by mutual coupling effects has attracted the attention of the academic society recently.

In [11] researchers separates decoupling strategies into four classes: 1) *Eigen-mode Decomposition Scheme*: Its guideline is to diagonalize the scattering matrix of a compact array using 90° and or 180° [12]-[16]. 2) *The Inserted Component Scheme*: It works on the concept of inserting a section of transmission-line between the coupled antenna ports [17]-[21]. 3) *Artificial Structure Decoupling Scheme*: This method uses sub-wavelength EM structures such as electromagnetic band gap (EBG) structure [22], defected ground structures (DGS) [23], and magnetic meta-materials [24], [25]. 4) *Coupled Resonator Decoupling Scheme*: This method was proposed for the first time in 2014, and has the concept of decoupling pair of coupled elements using coupled resonators [11] and [26]-[29].

Various works have examined the impact of spatial correlation on the performance of communication systems by means of experimentation [30], [31], modeling [32], [33] and theoretical analysis [34]-[40]. Numerous works have characterized the effect of channel correlation in the performance and capacity of the wireless channels by mathematical analysis [34], [35], further focusing on linear receivers [36], [37], [40] using mainly random matrix theory [41], [42], [43]. In [33] and [44]-[48] precoder designs Specifically tailored for correlated channels are derived. An interesting channel model with transmitting correlation, based on the angular spread of the transmits antenna elements emission, was shown by the authors of [33], [49].

The above discussions focus on studying the influence of antenna separation on a set number of antennas and effects on The communication performance. The purpose of compensation or decoupling network, however, is to enhance the isolation between ports jointly located within restricted space in mobile handset. It is, therefore, reasonable to investigate the combined effect of spatial correlation and

mutual coupling matrix on correlation-based stochastic channel coefficients, and subsequent effects on system capacity and transmit diversity After the decoupling process. To recognize user equipment (UE) channels and enhance channel estimation, the correlation channel model introduces angle of arrivals (AoAs).

For this reason, we investigate the system performance of a user equipment (UE) channels at different orientations for the uniform linear antenna arrays at diverse antenna separations, while increasing the signal sources by adding the further antenna elements. For the purpose of demonstrating the effectiveness of the research, we utilized a prototype of two-element compensation network with insertion losses between input and Output ports better than 11dB. Using the spatial correlation model [33] and incorporating the effects of mutual coupling matrix before and after the decoupling process, the results reveal different system performances for the user equipment (UE) channels at reduced AoAs and antenna physical separation while adding more elements.

The paper is organized as follows: Section II presents the formulation of operating matrix and the design of the compensation network. Part III focuses on the system model. Analytical results and discussions are presented in Section IV. Finally, we give concluding remarks in Section V.

II. OPERATING MATRIX AND DESIGN OF COMPENSATION NETWORK

The Operating pattern for two-element receiving array for the remittance network is expressed as [54]

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{Z_{12}}{Z_L} \\ -\frac{Z_{21}}{Z_L} & 1 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} V_1 - \frac{Z_{12}}{Z_L} V_2 \\ V_2 - \frac{Z_{21}}{Z_L} V_1 \end{bmatrix} \quad (1)$$

Where V_1 and V_2 are the coupled voltages and the inputs to the network from the monopole terminals and the output voltages are U_1 and U_2 , also known as the compensation voltages. The compensation network is designed using a power divider of unequal power-dividing ratio with no active circuit elements to minimize extra circuit noise and two rat-race couplers. The power divider has three transmission lines (Z_a, Z_b and Z_c), each having impedance of $\sqrt{2}Z_o$, where Z_o is the system's characteristic impedance, but unequal electrical Lengths ϕ, Ψ and θ [55]. The electrical length ϕ can be defined as $\phi = \cos^{-1}(|Z_M/Z_L|)$, whereas Ψ and θ are 90° and $(90^\circ + \phi)$ respectively and Z_M representing the mutual impedance. We fabricate the circuit by using the substrate FR4 with dielectric constant 4.8 at operating frequency of 2.4 GHz as shown in Figure 1. The measured insertion losses between input and output ports of the decoupling network are shown in Figure 2.

III. SYSTEM MODEL

In this paper, we examine the theoretical performance of MIMO system in the correlation-based stochastic channel Models with the decoupling network. It has been accounted for in [50] that the correlation-based stochastic channel models could be applied to cases that the user equipment (UE) with multiple antennas work at millimeter wave, in any case, we experiment the performance of the MIMO systems at 2.4 GHz. For the purpose of understanding, it is assumed fading is correlated at both transmitter and receiver sides, in spite of the fact that, the decoupling network enhances isolation between antenna arrays at the user equipment (UE). When the linear the antenna array is assumed, the steering matrix R_k is expressed as [50], [51]

$$R_k = \frac{1}{D_k} \left[a(\theta_{k,1}), a(\theta_{k,2}), \dots, a(\theta_{k,D_k}) \right] \quad (2)$$

$$a(\theta_{k,i}) = [1, e^{j2\pi d/\lambda \sin \theta_{k,i}}, \dots, e^{j2\pi d(N-1)/\lambda \sin \theta_{k,i}}]^T \quad (3)$$

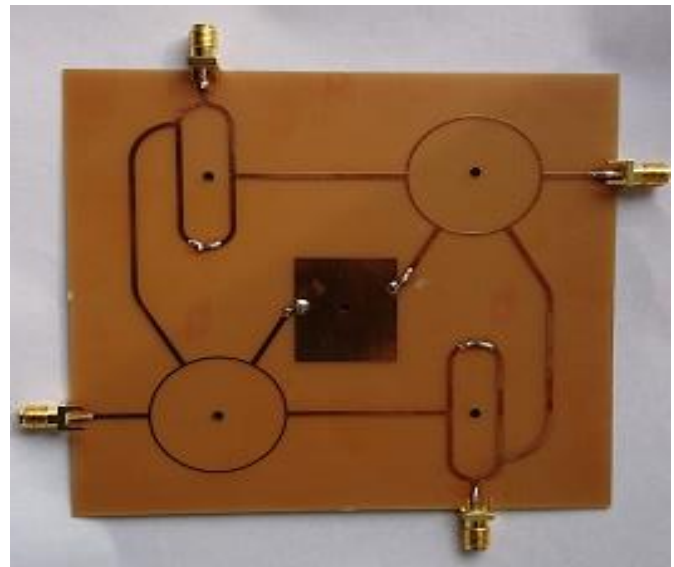


Fig. 1. Photograph of the fabricated inserted decoupling network

Where d is the distance between the adjacent antennas, λ is the carrier wavelength and N the number of elements. Incorporating mutual coupling the channel vector UE can be written as follows

$$h_k = ZR_k v_k, k = 1, 2, \dots, k \quad (4)$$

Where, $Z \in \mathfrak{M}^{N \times N}$ represents the mutual coupling matrix, $R_k \in \mathfrak{M}^{N \times D_k}$ denotes the steering matrix containing D_k steering vectors of the receiver array, $v_k \sim \mathfrak{M} N(0, I_{D_k})$. The mutual coupling matrix is defined as [51], [52]

$$Z = (Z_A + Z_L)(\Gamma + Z_L I)^{-1} \quad (5)$$

with

$$\Gamma = \begin{bmatrix} Z_A & Z_m & 0 & \dots & 0 \\ Z_m & Z_A & Z_m & \dots & 0 \\ 0 & Z_m & Z_A & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & Z_m & Z_A \end{bmatrix} \quad (6)$$

Where, Z_A, Z_L, Z_m are the antenna impedance, load impedance and mutual impedance respectively. For theoretical approximation of bilateral coupling (4) we assume that Z_A and Z_L have a value of 50Ω each. Z_m is calculated using the EMF method [7] using d .

A. S-parameter Based Mutual Coupling

Regarding scattering parameters the mutual coupling the Matrix in (5) of the linear array is expressed as [53]

$$Z_t = (I + S_t)(I - S_t)^{-1} * Z_o \quad (7)$$

Where Z_o represent the reference antenna impedance, and

$S_t \in \mathfrak{M}^{N \times N}$ is the S-parameter matrix of the antenna array. The voltage and current on the m -the antenna element are given as

$$v_m = \sqrt{Z_o}(a_m + b_m) \text{ and } i_m = \frac{1}{\sqrt{Z_o}}(a_m - b_m) \quad (8)$$

Where vectors a and b , are the complex envelopes of the inward-propagating and outward-propagating waves from the antenna elements respectively. In this paper, we model the transmit antenna array in CST Microwave Studio, as a uniform linear array with monopoles as antenna elements. For the receiving antenna array, the scattering parameters of the coupled and decoupled monopole array are measured in an anechoic chamber.

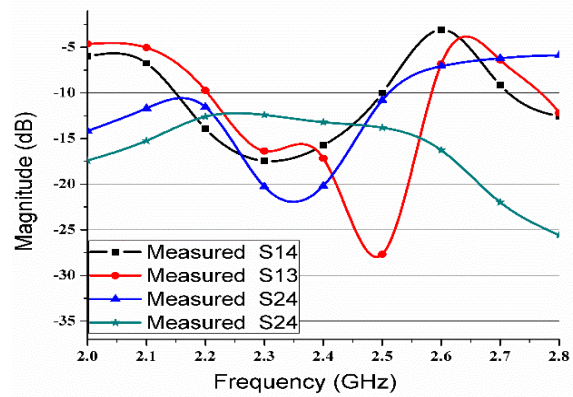


Fig. 2. Measured insertion losses between input and output ports of the Decoupling Network

IV. NUMERICAL RESULTS

This section illustrates the analytical performance and results according to the modeling of mutual coupling matrix at the Receiving side. The simulation assumes a uniform linear array With monopoles as antenna elements at the source. The compact antenna array consists of two parallel monopoles operating at 2.4 GHz and placed on a metallic ground. The monopoles have the length of 30 mm, a radius of 0.5 mm and fixed element separation of $\lambda/8$. The S-parameters for the receiving array are determined in two different conditions. Firstly, the scattering parameters of the coupled monopole array are measured in order to determine the coupled voltages (V_1 and V_2) and coupling matrix. Secondly, the monopole antennas in the array are connected to the decoupling network through equivalent length coaxial links, and scattering parameters of the output ports of the decoupling network are measured to determine the coupling matrix for the compensated voltages (U_1 and U_2).

The simulated channel is according to the channel model (1)-(6) and the angle of arrival is between 0° and 360° . For the purpose of demonstrating the effectiveness of the decoupling network, there are three different types of voltages listed in Table I. The last row in Table I is a ratio of the voltage obtained with monopole B to the voltage achieved with monopole A. It can be seen that the ratio of the compensated

voltage is very close to the uncoupled voltages, demonstrating that the compensated voltages have successfully taken off from the coupling effect.

A. Fixed Number of Antennas- Traditional View

We investigate the effect of antenna spacing on performance according to coupling matrix model at the receiving side. For reasons of reference, first, we analyze the performance of the uncorrelated channel when the receiving antenna array are closely spaced with a separation of $\lambda/8$. We do this for an angle of arrival at the receiving end as 360° for SNR of 10 dB, to illustrate the behavior of system performance regarding dependency between antenna spacing and average capacity. Figure 3 reveals that capacity of the decoupled receiving array is statistically better than that of the coupled receiving array. In order words, the results demonstrate the promising potentials of an efficient decoupling scheme in a correlated-based stochastic channel. However, system performance is restricted to antenna element spacing. With the decoupling network at the receiving end, therefore, results show the performance benefits that can be achieved by increasing the separation between antenna elements.

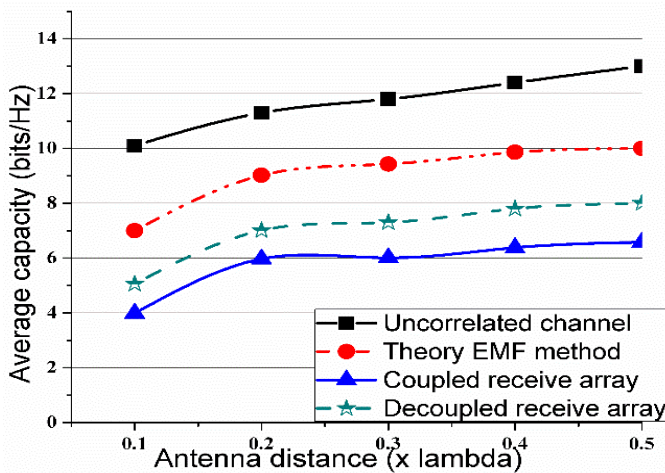


Fig. 3. Average capacity vs. antenna spacing d with SNR=10 dB, AoA= 360°

B. Effects of AoA on System Performance

We now move forward to investigate the behavior of channel capacity with different AoAs at various antenna Elements separations at the transmitter. In Figure 4 we compare the performances for 2×2 systems for AoAs of 60° and 360° at element separation of at $dt = \lambda/8, \lambda/4, \lambda/2$. We repeat the process with increasing number of transmit antennas for AoAs Of 90° and 120° in figure 5 for 4×2 MIMO system. The graphs reveal that system performance at reduced AoAs with the increase antenna element separation outperformed that with increased AoA and reduced antenna separation. The results also demonstrate the dependency of channel capacity on an antenna separation. We note that reducing the distance between antenna elements affect system performance, however, the larger number of antenna elements

at the signal source improves system performance. It is interesting to observe in Figure 5 that even though the channel is modeled with spatial correlation and mutual coupling at both ends, the performance for AoA of 60° at $\lambda/2$ a uncorrelated Mutual coupling at SNR=10 dB. The result demonstrates the advantage of decoupling system in a correlated channel with reduced AoA

C. Transmit Diversity

For larger performance gain of almost 20 dB at error rate of 10^{-2} figure 6 indicates no match between decoupled and coupled receive arrays for AoA = 360° . However, a close observation in Figure 7 reveals an improvement in the BPSK performance. In spite of the fact that there was little change in the QPSK performance, we take note of that after an increase of 10 dB, the performance for coupled array with both modulations remains Nearly the same. In general, our analysis indicates a close match between QPSK and BPSK diversity performances under both conditions. We include theoretical performance based on the EMF method which outperformed the simulation results.

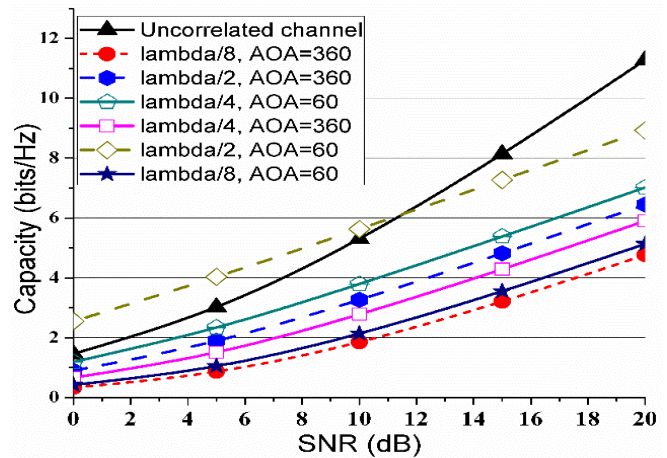


Fig. 4. Capacity vs. SNR for decoupling array for 2×2 MIMO at different AoAs (degrees) and antenna separation

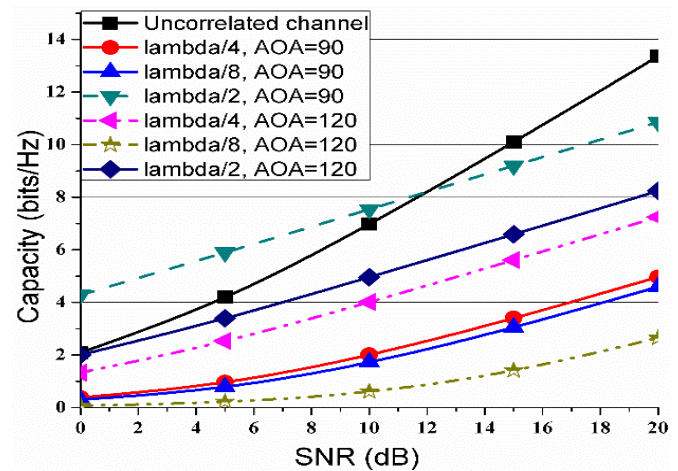


Fig. 5. Capacity vs. SNR for decoupling array for 4×2 MIMO at different AoAs and antenna element separation

V. CONCLUSION

In this paper, we have analytically explored the performance of decoupling network in the correlation-based stochastic model With uniform linear array at the transmitter. Performance evaluation indicates the joint dependency of system capacity on the angle, angle of arrival and antenna separation. Results show that an important benefit of system performance can be achieved for reduced AoA with increase antenna element separation at the transmitter. Our analysis demonstrates the promising potentials when mutual coupling matrix is modeled with efficient Decoupling network for a compact antenna array. Further work can be carried out towards investigating the performance the behavior of rectangular and circular arrays with decoupling network for MIMO spatially correlated transmitters.

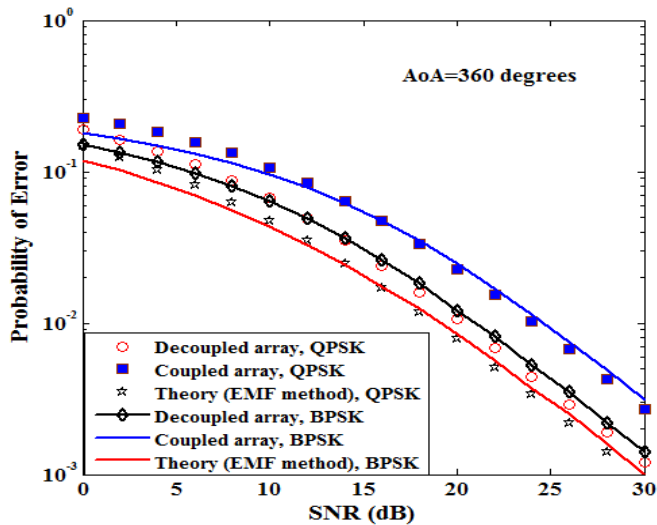


Fig. 6. Transmit diversity for decoupled and coupled array in correlated based stochastic model with linear array

TABLE I. DIFFERENT MEASURED VOLTAGES

		Uncoupled Voltages (reference)	Coupled voltages	Compensated voltages
Monopole A	mag (mV)	16.64	12.4	11.55
	angle (°)	-160.64	-166.67	34.967
Monopole B	mag (mV)	16.54	15.42	12.30
	angle (°)	-139.56	-141.46	55.16
B/A	mag (mV)	0.9939	1.2199	1.065
	angle (°)	21.08	25.208	20.193

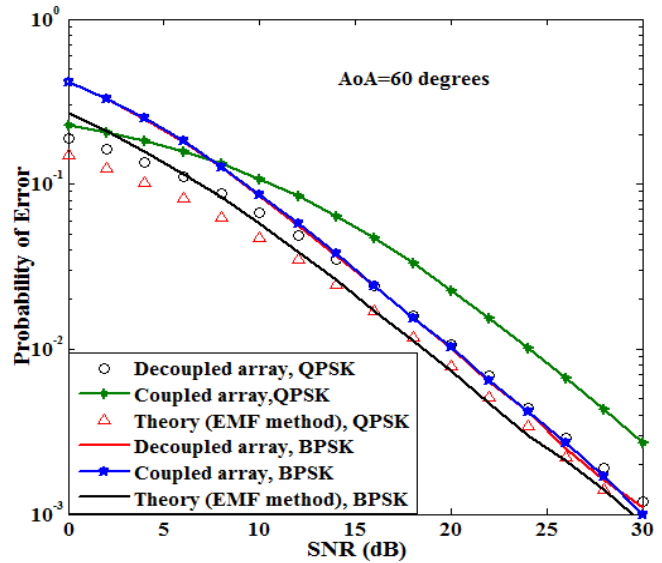


Fig. 7. Transmit diversity for decoupled and coupled array in correlated-based stochastic model with the linear array

ACKNOWLEDGMENT

The authors are grateful to all the members of Center for RFIC and System Technology and Multidimensional Signal Processing Laboratory (MSPL), School of Communication and Information Engineering, University of Electronic Science and Technology of China for relevant advice and discussion to this

REFERENCES

- [1] I. J. Gupta and A. A. Ksienski, "Effect of Mutual Coupling on the Performance of Adaptive Arrays," IEEE Transactions on Antennas and Propagation, AP-31, 5, September 1983, pp. 785-791.
- [2] T. Svantesson. "The Effects of Mutual Coupling Using a Linear Array of Thin Dipoles of Finite Length". In Proc. 8th IEEE SSAP, pages 232-235, Portland, September 1998.
- [3] T. Svantesson and A. Ranheim, "Mutual coupling effects on the capacity of multi-element antenna systems," in Proc. IEEE In Conf.: Acoustics, Speech, and Signal Processing, vol. 4, Salt Lake City, UT, May 7-11, 2001, pp. 2485-2488.
- [4] Griffith, K.A.; Gupta, I.J., 'Effect of mutual Coupling on the performance of GPS AJ Antennas' EEE/ION, Pages: 871 – 877, 2008.
- [5] Hao Yuan, Hirasawa, K. and Yimin Zhang, "The mutual coupling and diffraction effects on the performance of a CMA adaptive array" IEEE Transactions on Vehicular Technology, volume: 47, Issue: 3, Pages: 728 – 736, 1998.
- [6] S. Lu, H. T. Hui, M. E. Bialkowski, X. Liu, H. S. Lui, and N. V. Shuley, "Effects of antenna mutual coupling on the performance of MIMO systems," in Proc 29th Symposium on Information Theory, pp. 2945–2948.
- [7] C. A. Balanis, Antenna Theory: Analysis and Design, 3rd Edition.
- [8] A. A. Abouda and S. G. Haggman, "Effect of mutual coupling on capacity of MIMO wireless channels in High SNR," Progress in Electromagnetics Research, vol. 65, p. 27–40, 2006.
- [9] B. Clerckx, C. Craeye, D. Vanhoenacker-Janvier, and C. Oestges, "Impact of antenna coupling on 2 x 2 MIMO communications," IEEE Trans. Veh. Technol., vol. 56, no. 3, pp. 1009–1018, May 2007,
- [10] T. Ratnarajah and A. Manikas, "An H-infinity approach to mitigate the effects of array uncertainties on the MUSIC Algorithm," IEEE Signal Process. Lett., vol. 5, no. 7, pp. 185–188, Jul. 1998.

- [11] L. Zhao, L. K. Yeung, and K.-L. Wu, "A coupled resonator decoupling network for two-element compact antenna arrays in mobile terminals," *IEEE Trans. Antennas Propag.*, vol. 62, no. 5, pp. 2767–2776, May 2014.
- [12] J. C. Coetzee and Y. Yu, "Port decoupling for small arrays by means of an eigenmode feed network," *IEEE Trans. Antennas Propag.*, vol. 6, pp. 1587–1593, Jun. 2008.
- [13] S. Zuo, Y.-Z. Yin, Y. Zhang, W.-J. Wu, and J.-J. Xie, "Eigenmode decoupling for MIMO loop-antenna based on 180 coupler," *Progr. Electromagn. Res. Lett.*, vol. 26, pp. 11–20, 2011.
- [14] S. K. Chaudhury, H. J. Chaloupka, and A. Ziroff, "Multiport antenna Systems for MIMO and Diversity," *EUCAP, Barcelona, Spain*, Apr. 2010.
- [15] C. Volmer, J. Weber, R. Stephan, K. Blau, and M. A. Hein, "An eigen analysis of compact antenna arrays and its application to port decoupling," *IEEE Trans. Antennas Propag.*, vol. 56, no. 2, pp. 360–370, Feb. 2008.
- [16] L. K. Yeung and Y. E. Wang, "Mode-based beamforming arrays for miniaturized platforms," *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 1, pp. 45–52, Jan. 2009.
- [17] J. B. Andersen and H. H. Rasmussen, "Decoupling and descattering networks for antennas," *IEEE Trans. Antennas Propag.*, vol. 24, no. 6, pp. 841–846, Nov. 1976.
- [18] S. Chang, Y.-S. Wang, and S.-J. Chung, "A decoupling technique for increasing the port isolation between strongly coupled antennas," *IEEE Trans. Antennas Propag.*, vol. 56, no. 12, pp. 3650–3658, Dec. 2008.
- [19] C.-Y. Lui, Y.-S. Wang, and S.-J. Chung, "Two nearby dual-band antennas with high port isolation," presented at the *IEEE Int. Symp. Antennas Propag.*, San Diego, CA, USA, Jul. 2008.
- [20] A. Diallo, C. Luxey, P. L. Thuc, R. Staraj, and G. Kossiavas, "Study and reduction of the mutual coupling between two mobile phone PIFAs operating in the DCS1800 and UMTS bands," *IEEE Trans. Antennas Propag.*, vol. 54, no. 11, pp. 3063–3073, Nov. 2006.
- [21] C. Luxey, "Design of multi-antenna systems for UMTS mobile phones," in *Proc. Loughborough Antennas Propag. Conf.*, pp. 57–64, Nov. 2009.
- [22] F. Yang and Y. R. Samii, "Microstrip antennas integrated with electromagnetic band-gap EBG structures: A low mutual coupling design for array applications," *IEEE Trans. Antennas Propag.*, vol. 51, no. 10, pp. 2936–2946, Oct. 2003.
- [23] C. Y. Chiu, C. H. Cheng, R. D. Murch, and C. R. Rowell, "Reduction of mutual coupling between closely-packed antenna element," *IEEE Trans. Antennas Propag.*, vol. 55, no. 6, pp. 1732–1738, Jun. 2007.
- [24] M. M. Bait-Suwailam, M. S. Boybay, and O. M. Ramahi, "Electromagnetic coupling reduction in high-profile monopole antennas using single-negative magnetic metamaterials for MIMO applications," *IEEE Trans. Antennas Propag.*, vol. 58, no. 9, pp. 2894–2902, Sep. 2010.
- [25] B. K. Lau and J. B. Andersen, "Simple and efficient decoupling of compact arrays with parasitic scatterers," *IEEE Trans. Antennas Propag.*, vol. 60, no. 2, pp. 464–472, Feb. 2012.
- [26] L. Zhao, L. K. Yeung, and K.-L. Wu, "A coupled resonator decoupling network for two-element compact antenna arrays in mobile terminals," *IEEE Trans. Antennas Propag.*, vol. 62, no. 5, pp. 2767–2776, May 2014.
- [27] L. Zhao and K.-L. Wu, "A broadband coupled resonator decoupling network for a three-element compact array," in *Proc. IEEE MTT-S Int. Microw. Symp.* pp. 1–3, Jun. 2013.
- [28] L. Zhao, L. K. Yeung, and K. L. Wu, "A novel second-order decoupling network for two-element compact antenna arrays," *Proc. Asia-Pacific Microwave Conf.*, 2012.
- [29] K. Qian, L. Zhao, and Ke-Li Wu, "An LTCC Coupled Resonator Decoupling Network for Two Antennas" *IEEE Transactions on Antennas and Propagation*, vol. 63, No. 7, July 2015.
- [30] M. T. Ivrlac, W. Utschick, and J. A. Nossek, "Fading correlations in wireless MIMO communication systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 819–828, June 2003.
- [31] D. Piazza, N. J. Kirsch, A. Forenza, R. W. Heath, and K. R. Dandekar, "Design and evaluation of a reconfigurable antenna array for MIMO systems," *IEEE Trans. Antennas Propag.*, vol. 56, no. 3, pp. 869–881, Mar. 2008.
- [32] D. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.
- [33] C. Wang and R. D. Murch, "Adaptive downlink multi-user MIMO wireless systems for correlated channels with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2435–2446, Sep. 2006.
- [34] A. M. Tulino, A. Lozano, and S. Verdu, "Impact of antenna correlation on the capacity of multiantenna channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2491–2509, July 2005.
- [35] G. Alfano, A. M. Tulino, A. Lozano, and S. Verdu, "Capacity of MIMO channels with one-sided correlation," in *Proc. IEEE Conf. Spread Spectrum Techniques and Applications*, pp. 515–519, 2004
- [36] S. Jin, M. R. McKay, C. Zhong, and K. K. Wong, "Ergodic capacity analysis of amplify-and-forward MIMO dual-hop systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2204–2224, May 2010.
- [37] S. Jin, M. R. McKay, X. Gao, and I. B. Collings, "MIMO multichannel beamforming: SER and outage using new eigenvalue distributions of complex noncentral Wishart matrices," *IEEE Trans. Commun.*, vol. 56, no. 3, pp. 424–434, Mar. 2008.
- [38] M. Kiessling and J. Speidel, "Analytical performance of MIMO zeroforcing receivers in correlated Rayleigh fading environments," in *Proc. Conf. Signal Processing Advances in Wireless Communications*, pp. 383–387.
- [39] H. Liu, Y. Song, and R. C. Qiu, "The impact of fading correlation on the error performance of MIMO systems over Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2014–2019, Sept. 2005.
- [40] M. Matthaiou, C. Zhong, and T. Ratnarajah, "Novel generic bounds on the sum rate of MIMO ZF receivers," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4341–4353, Sep. 2011
- [41] A. M. Tulino and S. Verdu, "Random matrix theory and wireless communications," *Foundations and Trends in Commun. and Inf. Theory*, vol. 1, no. 1, pp. 1–182, 2004.
- [42] T. Ratnarajah and R. Vaillancourt, "Quadratic forms on complex random matrices and multiple-antenna systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2979–2984, Aug. 2005.
- [43] T. Ratnarajah, "Topics in complex random matrices and information theory," Ph.D. thesis, Univ. Ottawa, Ottawa, Canada, 2003.
- [44] A. Alexiou and M. Qaddi, "Robust linear precoding to compensate for antenna correlation in orthogonal space-time block coded systems," in *Proc. 2004 Sensor Array and Multichannel Signal Processing Workshop*, pp. 701–705.
- [45] H. R. Bahrami and T. Le-Ngoc, "Precoder design based on the channel correlation matrices," *IEEE Trans Wireless Commun.*, vol. 5, no. 12, pp. 3579–3587, Dec. 2006.
- [46] J. Akhtar and D. Gesbert, "Spatial multiplexing over correlated MIMO channels with a closed-form precoder," *IEEE Trans Wireless Commun.*, vol. 4, no. 5, pp. 2400–2409, Sep. 2005.
- [47] S. Zhou and G. B. Giannakis, "Optimal transmitter eigen-beamforming and space-time block coding based on channel correlations," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1673–1690, July 2003.
- [48] S. A. Jafar, S. Vishwanath, and A. Goldsmith, "Channel capacity and beamforming for multiple transmit and receive antennas with covariance feedback," in *Proc. IEEE International Conference on Communications*, vol. 7, pp. 2266–2270, 2001
- [49] T. Ratnarajah and A. Manikas, "An H-infinity approach to mitigate the effects of array uncertainties on the MUSIC Algorithm," *IEEE Signal Process. Lett.*, vol. 5, no. 7, pp. 185–188, Jul. 1998.
- [50] Kan Zheng, Suling Ou and Xuefeng Yin "Massive MIMO Channel Models: A Survey" *International Journal of Antennas and Propagation* Vol. 2014, Article ID 848071, June 2014.
- [51] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Large-scale MIMO transmitters in fixed physical spaces: the effect of transmit correlation and mutual coupling," *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2794–2804, 2013.
- [52] B. Clerckx, C. Craeye, D. Vanhoenacker-Janvier, and C. Oestges, "Impact of antenna coupling on 2×2 MIMO communications," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 3, pp. 1009–1018, 2007.
- [53] C. Masouros, J. Chen, K. Tong, M. Sellathurai, and T. Ratnarajah, "Towards massive-MIMO transmitters: on the effects of deploying

- increasing antennas in fixed physical space," in Proceedings of the Future Network and Mobile Summit, pp. 1–10, 2013.
- [54] H. T. Hui, "A Practical Approach to Compensate for the Mutual Coupling Effect in an Adaptive Dipole Array," *IEEE Transactions on Antennas and Propagation*, AP-52, pp. 1262-1269, May 2004.
- [55] Y. Yu and H.T. Hui "Design of a Mutual Coupling Compensation Network for a Small Receiving Monopole Array" *IEEE Trans. On Micro. Theory and Techniques*, vol. 59, no. 9, September 2011. J. C. Coetzee and Y. Yu, "Port decoupling for small arrays by means of an eigenmode feed network," *IEEE Trans. Antennas Propag.*, vol. 6, pp. 1587–1593, Jun. 2008.

Differential Evolution based SHEPWM for Seven-Level Inverter with Non-Equal DC Sources

Fayçal CHABNI, Rachid TALEB, M'hamed HELAIMI

Electrical Engineering Department, Hassiba Benbouali University, Chlef, Algeria
Laboratoire Génie Electrique et Energies Renouvelables (LGEER)

Abstract—This paper presents the application of differential evolution algorithm to obtain optimal switching angles for a single-phase seven-level to improve AC voltage quality. The proposed inverter in this article is composed of two H-bridge cells with non-equal DC voltage sources in order to generate multiple voltage levels. Selective harmonic elimination pulse width modulation (SHPWM) strategy is used to improve the AC output voltage waveform generated by the proposed inverter. The differential evolution (DE) optimization algorithm is used to solve non-linear transcendental equations necessary for the (SHPWM). Computational results obtained from computer simulations presented a good agreement with the theoretical predictions. A laboratory prototype based on STM32F407 microcontroller was built in order to validate the simulation results. The experimental results show the effectiveness of the proposed modulation method.

Keywords—selective harmonic elimination; multi-level inverters; differential evolution; cascade H-bridge inverters; optimization

I. INTRODUCTION

The direct current to alternative current (DC/AC) multi-level conversion systems have been drawing a lot of attention in the last few years especially for high voltage and renewable energy applications. The usage of power converters in high power applications have led to the development of various families and architectures such as Neutral Point Clamped (NPC)[1] Diode-clamped [2], and cascade multilevel inverters (CMLIs)[3].

The cascade multilevel inverters (CMLIs) have received a lot of attention due to their modular structure and simplicity of control; they offer a lot of advantages such as low switching losses [4], better electromagnetic compatibility [5] and low voltage stress on the switching devices [6]. The architecture can be formed by associating several individual H-bridge cells in series, and by adding more cells, the output voltage waveform become close to a sinusoidal waveform.

Several modulation strategies have been proposed and studied for the control of multilevel inverters such as Sinusoidal Pulse width modulation (SPWM)[7] and space vector pulse width modulation (SVPWM)[8]. A more efficient method called selective harmonic elimination pulse width modulation (SHE-PWM) is also used; the method offers a lot of advantages such as operating the inverters switching devices at a low frequency which extends the lifetime of the switching devices. The main disadvantage of this method is that a set of

non-linear equations must be solved to obtain the optimal switching angles to apply this strategy.

Multiple computational methods have been used to calculate the optimal switching angles such as Newton-Raphson (N-R)[9], this method depends on initial guess of the angle values in such a way that they are sufficiently close to the global minimum(desired solution). And if the chosen initial values are far from the global minimum, non-convergence can occur.

Selecting a good initial angle, especially for a large number of switching angles can be very difficult. Another approach is to use optimization algorithms such as genetic algorithm (GA) [10], firefly algorithm (FFA) [11] and particle swarm optimization (PSO)[12]. The main advantage of these methods is that they are free from the requirement of good initial guess.

The differential evolution (DE) is one of the most powerful optimization algorithms. Since its introduction in 1997 [13], the algorithm has drawn the attention of many scientists over the world, resulting in multiple variants derived from the original basic algorithm, with improved performance. The DE is a simple yet powerful algorithm; it is composed of three main operations mutation, crossover and selection [14]. The algorithm uses the difference of solution vectors to create new candidate solutions using the above-mentioned operators. This work investigates the use of (DE) as an optimization tool to implement the (SHEPWM) for a seven level inverter.

In [15] the author proposed the application of differential evolution for selective harmonic elimination in a three phase two level inverter with multiple switching angles in quarter of period to eliminate multiple low order harmonics, the main disadvantage of this work is that the proposed inverter cannot be used in high power application due to voltage stress exercised on the switching devices. This paper presents a simple and fast optimal solution of harmonic elimination of a seven level inverter with non-equal DC sources using the differential evolution algorithm. The algorithm is used to solve a system of non-linear equations that describes the waveform of the output voltage in order to obtain the optimal switching angles, to improve the output voltage quality.

This paper is organized as follows: the next section explains briefly the structure of the proposed multilevel inverter and its control, the third section covers the application of the differential evolution algorithm for the selective harmonic elimination, this section details the procedures to

obtain the optimal switching angles and the formulation of the objective function, the fourth section presents the simulation results obtained from the mathematical model of the system and the optimization method. The effectiveness of the selective harmonic elimination using DE is verified using a small scale laboratory seven level inverter based on STM32F407 Microcontroller unit, the section also presents and discusses the hardware implementation and the experimental results in details. The conclusion is presented in the last section.

II. TOPOLOGY

As mentioned before there are three main families of multilevel inverters. And those families are the diode clamped inverters, flying capacitor inverters and the cascade multilevel inverters. The multilevel cascaded configuration is a popular choice in high power applications, in addition to advantages mentioned before, this configuration does not require a large number of components and does not need clamping diodes or balancing capacitors [16-17], the modular structure of this topology allows easier maintenance.

Cascaded Multilevel Inverter topology rely on a simple principle based on the summation of voltages generated by each individual cell (H-bridge) to obtain a staircase output voltage waveform. Fig.2 illustrates the voltage waveform of a seven level inverter in a quarter of period. Fig.1 demonstrates the proposed single phase seven level asymmetrical inverter. It is formed by two H-bridges connected in series each bridge is powered by electrically isolated power supplies to generate the desired waveform.

Each H-bridge module is connected to its respective isolated DC source; each module can generate three voltage levels +V which is the positive voltage of the DC source ,0V and -V which is the negative voltage of the DC source , and as it can be observed in Fig.1 in order to obtain seven levels at the output of the inverter, the DC voltage source connected to the lower cell has to be twice the value of the DC source connected to the upper cell ($V_{dc2}=2 \times V_{dc1}$).

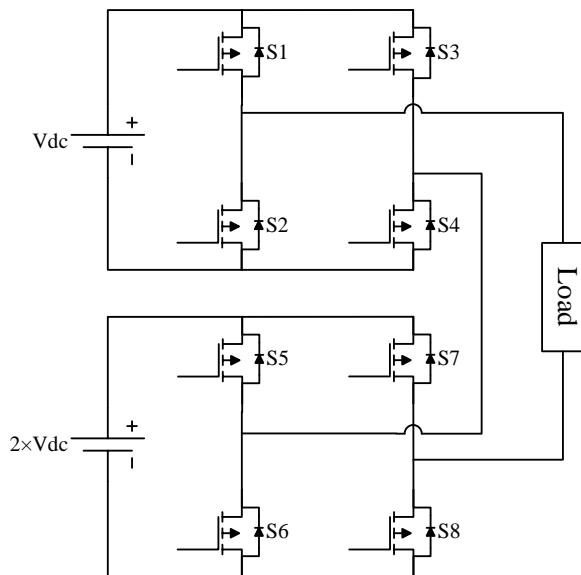


Fig. 1. Structure of the proposed multilevel inverter

The valid switching states for all possible combination of input voltage sources are given in Table.1. It can be seen that in order to generate a voltage level at least four switching devices have to be switched on.

TABLE I. SWITCHING STATES OF SEMICONDUCTOR DEVICES FOR 7-LEVEL INVERTER

Voltage levels (P.U)	Switches state							
	S1	S2	S3	S4	S5	S6	S7	S8
3	on	Off	off	on	on	Off	off	on
2	off	On	off	on	on	Off	off	on
1	on	Off	off	on	off	On	off	on
0	off	On	off	on	off	On	off	on
-1	off	On	on	off	off	On	off	on
-2	off	On	off	on	off	On	on	off
-3	off	On	on	off	off	On	on	off

III. SELECTIVE HARMONIC ELIMINATION USING DIFFERENTIAL EVOLUTION

The number of voltage levels that can be generated by CMLIs is generally presented by $2P+1$ where P represents the number of voltage levels or switching angles in a quarter waveform of the signal, and $P-1$ is the number of undesired harmonics that can be eliminated from the generated waveform. In a seven level inverter, the number of voltage levels in quarter waveform is three which means the number of harmonics that can be eliminated is two.

In order to eliminate the undesired harmonics, the switching angles θ_1 , θ_2 and θ_3 represented in Fig.2 must be computed.

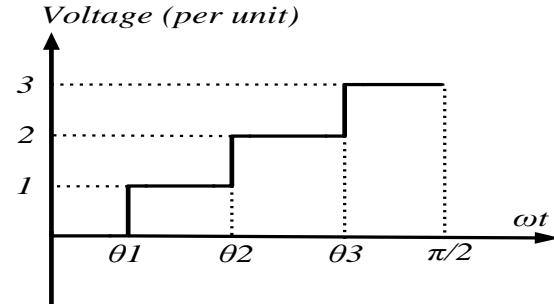


Fig. 2. quarter waveform of a seven-level inverter

For the staircase output voltage waveform of multilevel inverter as shown in Fig.2 there are 3 voltage levels (in quarter waveform) and 2 undesired harmonics.

To control the peak value of the output voltage to be V_1 and eliminate the 3rd and 5th harmonics the resulting equations and since the voltage waveform has quarter and half wave symmetry characteristics, the Fourier series expansion is given as:

$$V(\omega t) = \sum_{n=1,3,5,\dots}^{\infty} \left[\frac{4V_{dc}}{n\pi} \sum_{i=1}^p \cos(n\theta_i) \right] \sin(n\omega t) \quad (1)$$

Where n is rank of harmonics, $n = 1,3,5, \dots$, and $p = (N-1)/2$ is the number of switching angles per quarter waveform., and θ_i is the i^{th} switching angle, and N is the number of voltage levels per half waveform. The optimal switching angles θ_1 , θ_2

and θ_3 can be determined by solving the following system of non-linear equations:

$$\begin{cases} H_1 = \cos(\theta_1) + \cos(\theta_2) + \cos(\theta_3) = M \\ H_3 = \cos(3\theta_1) + \cos(3\theta_2) + \cos(3\theta_3) = 0 \\ H_5 = \cos(5\theta_1) + \cos(5\theta_2) + \cos(5\theta_3) = 0 \end{cases} \quad (2)$$

Where $M = ((N - 1)/2)r/4$, r is the modulation index.

The obtained solutions must satisfy the following constraint:

$$0 < \theta_1 < \dots < \theta_p < \pi/2 \quad (3)$$

An objective function is necessary to perform the optimization operation, the function must be chosen in such way that allows the elimination of low order harmonics while maintaining the amplitude of the fundamental component at a desired value Therefore the objective function is defined as:

$$F(\theta_1 \theta_2 \dots \theta_p) = \left(\sum_{n=1}^p \cos(\theta_n) - M \right)^2 + \left(\sum_{n=1}^p \cos(3\theta_n) \right)^2 + \left(\sum_{n=1}^p \cos(5\theta_n) \right)^2 \quad (4)$$

The optimal switching angles are obtained by minimizing Eq (4) subject to the constraint Eq(3).The main problem is the non-linearity of the transcendental set of Eq(2), the differential algorithm is used to overcome this problem.

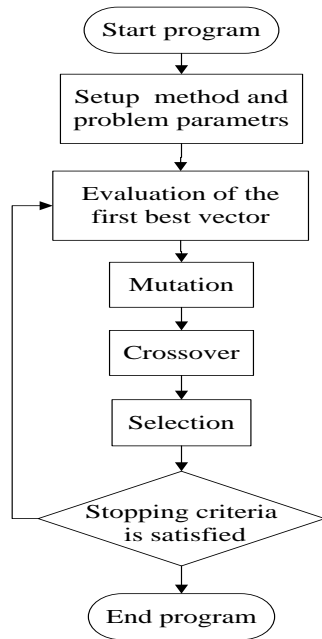


Fig. 3. Flowchart of DE algorithm

The differential evolution algorithm (DE) is an optimization method is composed of three main steps initialization, mutation and crossover. The general structure of a DE program is shown in Fig.3. The algorithm perturbs the population of vectors by employing the mutation, whereas its diversity is controlled by the cross-over process [18].

In the case of SHPWM, differential evolution algorithm is used as an optimization tool to perform a random search for the global minima, which is forcing the objective function (4) towards an allowable error value.

The optimization process starts by initializing the necessary parameters of the algorithm, such as the population size (NP), crossover probability (CP), upper and lower bounds (θ_{min} and θ_{max}) and the maximum number of iterations. It should be noted that the boundaries must satisfy equation (3). The next step is to randomly generate an initial population of switching angles in this process the algorithm creates

$$\theta_{ij}^{(0)} = \theta_{min\ ij} + rand_i(\theta_{max\ j} - \theta_{min\ j}) \quad (5)$$

With $i=1,2,\dots,NP$ and $j=1,2,\dots,N$

Where $\theta_{ij}^{(0)}$ is the initial population, i presents the population size in this study $NP=50$, j is the number of decision variables which represents the number of switching angles, in case of a seven level inverter $N=3$. After the initialization process, the generated population is evaluated, the evaluation of the fitness of each individual is carried out by using (4).

The mutation process creates a mutant v_{ij} vector based on the initial population; this process is described by the following expression

$$v_{ij} = X_{r1} + F(X_{r2} - X_{r3}) \quad (6)$$

X_{r1} , X_{r2} and X_{r3} are vectors randomly sampled from the generated population, $X_r = [\theta_{i1}, \theta_{i2}, \dots, \theta_{iN}]$, the indices $r1$, $r2$ and $r3$ are integers randomly chosen from the range $[1\ NP]$, they are also chosen to be different from the index i , the parameter F is the mutation constant which controls the amplification of the differential variation $(X_{r2} - X_{r3})$, the value of this parameter is randomly generated from the range $[0\ 1]$, it should be noted that multiple mutation methods were reported in[19].

To improve the diversity of the population, the crossover operation comes into play, after generating the mutant vector v_{ij} through mutation, this operation assures the production of fitter individuals, the result of this process is a vector u obtained by mixing the components of v_{ij} and X_i the process can be expressed as:

$$u = \begin{cases} v_{ij} & \text{if } rand \leq CP \text{ or } j = j_{rand} \\ X_i & \text{otherwise} \end{cases} \quad (7)$$

Where $rand$ is a random number in the range of $[0\ 1]$, CP is the crossover probability constant, it controls the diversity of the population and it has a value between 0 and 1 [20], j_{rand} is randomly chosen index. Once the crossover process is completed, the selection process comes into play to decide whether the u_i or X_i vector survives for the next generation, this process is carried out to choose the fittest individual. The selection process can be expressed mathematically as:

$$X_i^{G+1} = \begin{cases} u_i^{G+1} & \text{if } f(u_i^{G+1}) < f(X_i^G) \\ X_i^G & \text{otherwise} \end{cases} \quad (8)$$

Where $f(X)$ is the objective function to be minimized, and G is the generation count. Once the selection operation is completed, the algorithm loop is repeated until the stopping criteria is satisfied, in this study the DE algorithm is limited by maximum number of iterations $Nitr=1000$.

IV. SIMULATION RESULTS

In order to prove the theoretical predictions and to test the effectiveness of the proposed algorithm, the control method and the proposed inverter were developed and simulated using MATLAB/SIMULINK scientific programming environment; the optimization program was executed on a computer with Intel(R) Core(TM) i3 CPU@ 2.13GHz Processor and 4GB of RAM, the optimization algorithm takes 1274.463 seconds to complete the computation process.

To verify the effectiveness of the proposed method, total harmonic distortion (THD) is used as a performance indicator to evaluate the quality of output AC voltage waveform generated from the multilevel inverter, the THD is defined as the total amount of harmonics related to the fundamental, it can be calculated using the following formula:

$$THD\% = \frac{\sqrt{\sum_{n=3}^{19} H_n^2}}{H_1} \times 100 \quad (9)$$

The differential evolution algorithm is used to find the switching angles for each value of modulation index r ; the total harmonic distortion is computed also for each r , Fig.4 illustrates optimal switching angles (in degrees) versus modulation index r with $r \in [0.2 \ 0.95]$, the angles are computed with a fine step-size of 0.01, and it can be seen that in some ranges of the modulation index, the obtained solutions exceeded the 90 degrees limit, those solutions are not going to be taken in consideration. Fig.5 shows the variation of the total harmonic distortion versus the modulation index, these results are obtained by using equation (9) and (2).

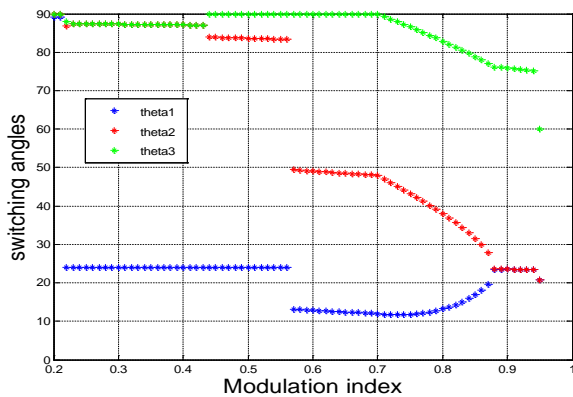


Fig. 4. Switching angles versus modulation index

To confirm the validity of the proposed algorithm, angles extracted from the obtained switching angles were applied to a mathematical model of a seven-level inverter. The fundamental frequency used in this simulation is 50Hz, the input voltages of the first bridge (upper cell) and the second bridge (lower cell) are respectively $V_{dc1}=25V$, $V_{dc2}=50V$ the switching angles to be applied (in degrees) are: $\theta_1= 16.87^\circ$, $\theta_2= 31.57^\circ$ and $\theta_3= 78.82^\circ$ which correspond to the modulation index $r= 0.85$.

Fig.6 and Fig.7 show the voltage waveforms generated respectively by the upper and the lower cell. From those two figures it can be seen that each bridge is responsible of generating three voltage levels. The summation of two voltages will generate the desired seven-level staircase waveform.

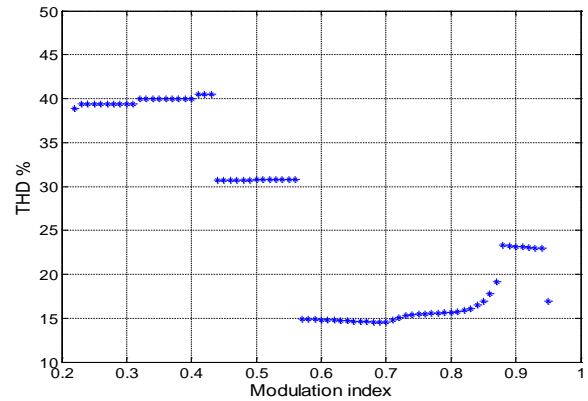


Fig. 5. THD versus modulation index

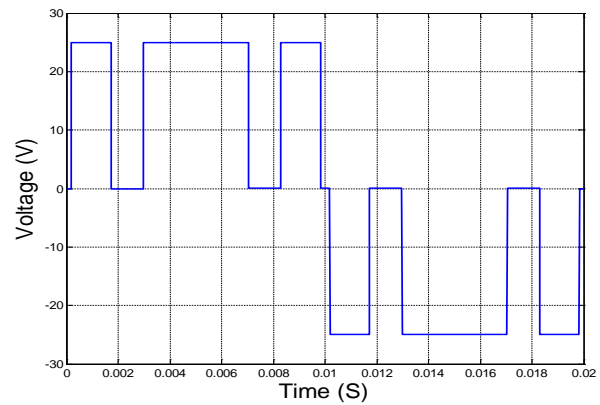


Fig. 6. Output voltage of the upper cell

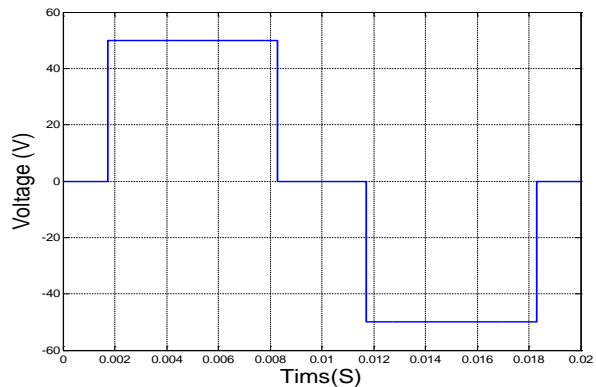


Fig. 7. Output voltage of the lower cell

Fig.8 shows the output voltage obtained from the multilevel inverter for $r=0.85$. Fig.9 shows its spectra of the output voltage. As expected, the selected harmonics (3rd and 5th) are successfully eliminated, the total harmonic distortion $THD=16.97\%$. Fig.10 demonstrates the gating signals for the semiconductor switches from S1 to S8.

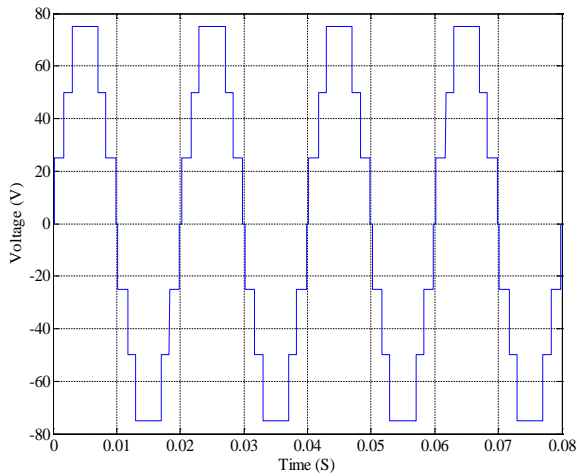


Fig. 8. Output voltage generated by the inverter

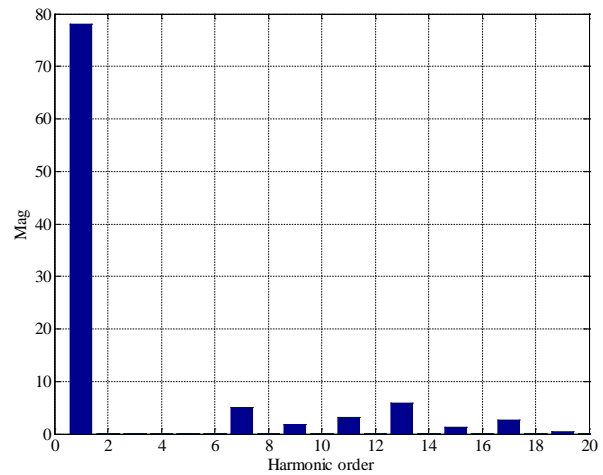


Fig. 9. FFT of 7-level inverter voltage output

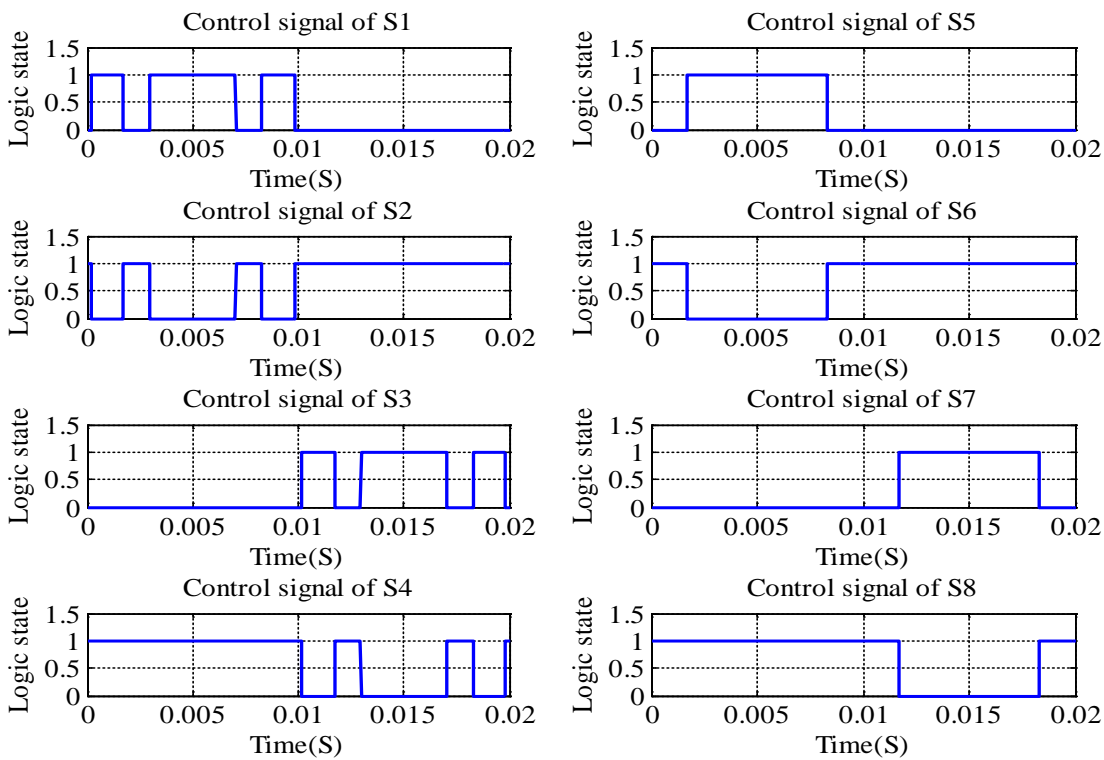


Fig. 10. Fig. Gating signals

V. EXPERIMENTAL RESULTS

The proposed method was validated by building a small scale laboratory prototype, IRF640(200V,18A) MOSFETs were used as switching devices SDS1000 oscilloscope 100MHz 500Ms/s was used to capture the voltage waveforms, an STM32F407 microcontroller was used to generate control signals for the switching devices, the FFT analysis was performed by computer connected to the oscilloscope through USB.

Fig.11 presents the block diagram of the laboratory prototype of the seven level inverter that is implemented as

mentioned before with eight IRF 640 Metal Oxide Semiconductor Field Effect Transistors (MOSFET), it should be noted that those switching devices are also equipped with freewheeling diodes. TLP250 photocouplers are used to provide electrical isolation between the MCU and the power circuits, and also to provide proper and conditioned gate signals to the MOSFETs. The switching angles are calculated using differential evolution algorithm by a computer, once the switching angles are obtained, the switching patterns for each switching device will be stored inside the memory of the MCU as a look-up table.

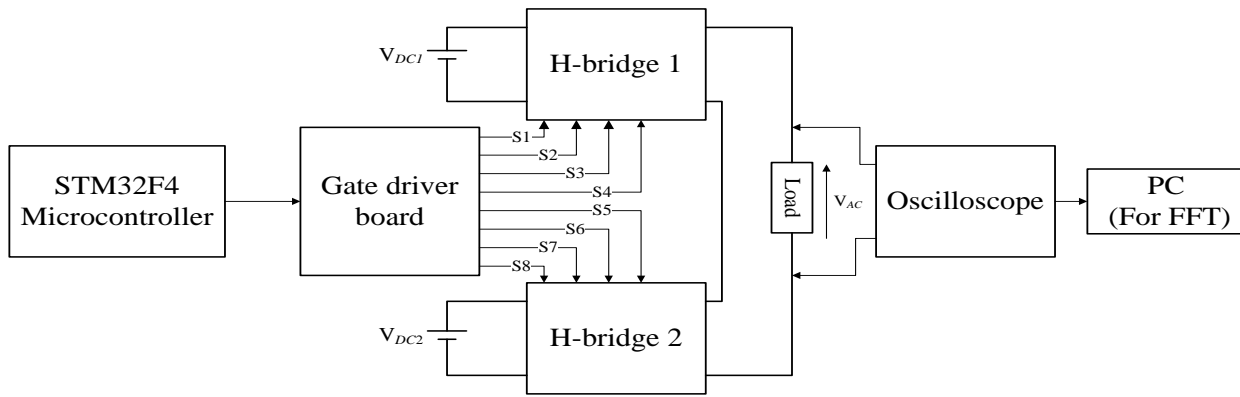


Fig. 11. Block diagram of the hardware setup

The single phase seven level voltage pattern obtained in simulation shown in Fig.8 is experimentally validated and the result is shown in Fig.14 the voltage waveforms generated by the upper and lower cell obtained in simulation presented in Fig.6 and Fig.7 respectively , are also experimentally validated, the results are presented in Fig.12 and Fig.13.

Fig.19 illustrates the FFT analysis of the experimentally obtained voltage waveform; it can be clearly seen that the 3rd and the 5th harmonics were successfully eliminated. This result matches perfectly the simulation result presented in Fig.9. The total harmonic distortion of the experimental voltage waveform is 15.85% which is very close to the simulation result.

Fig.15, Fig.16 Fig.17and Fig.18 illustrates the gating signals generated by the STM32F407 microcontroller unit (MCU) for the switching devices; these figures validate the simulation results presented in Fig.10.

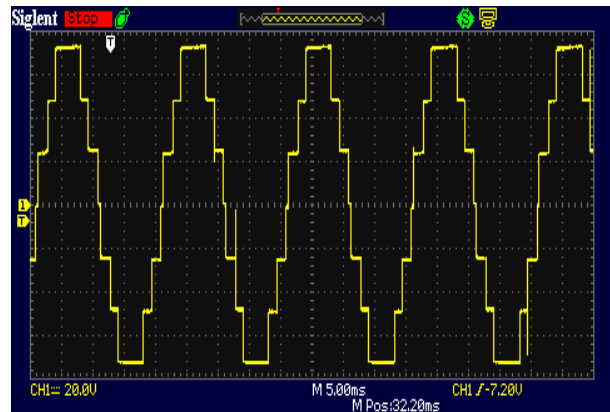


Fig. 14. Output voltage waveform generated the proposed multilevel inverter

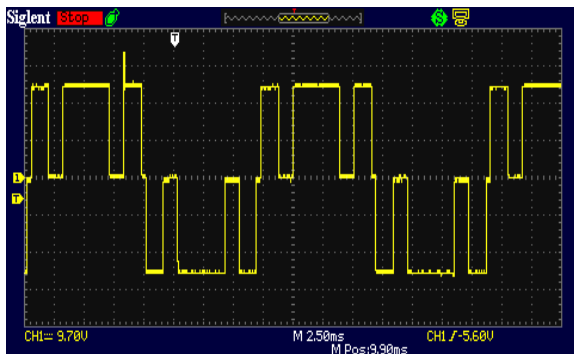


Fig. 12. Output voltage waveform generated by the upper cell

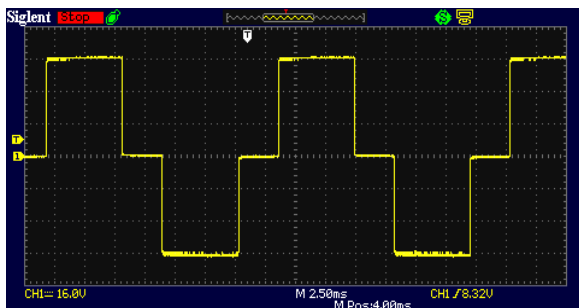


Fig. 13. Output voltage waveform generated by the upper cell

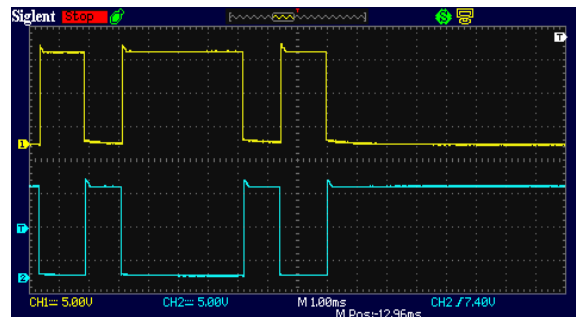


Fig. 15. Control signals generated by the MCU for S1 (yellow trace) and S2 (Blue trace)

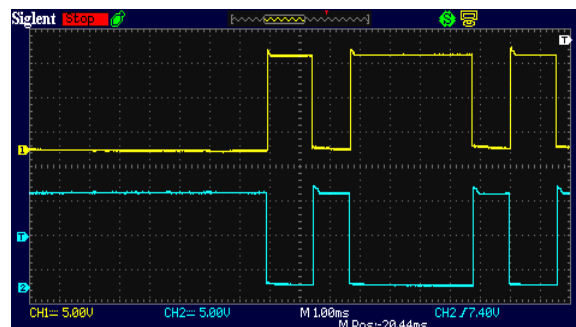


Fig. 16. Control signals generated by the MCU for S3 (yellow trace) and S4 (Blue trace)

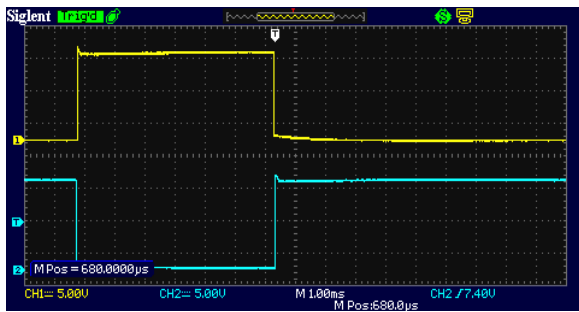


Fig. 17. Control signals generated by the MCU for S5 (yellow trace) and S6 (Blue trace)

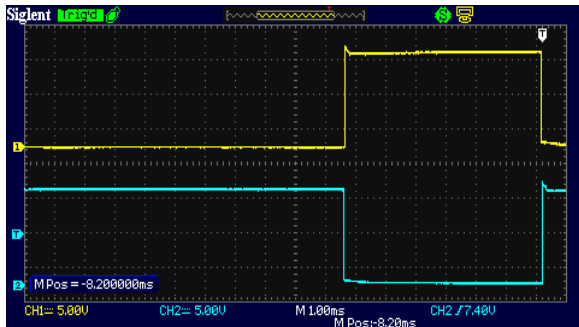


Fig. 18. Control signals generated by the MCU for S7 (yellow trace) and S8 (Blue trace)

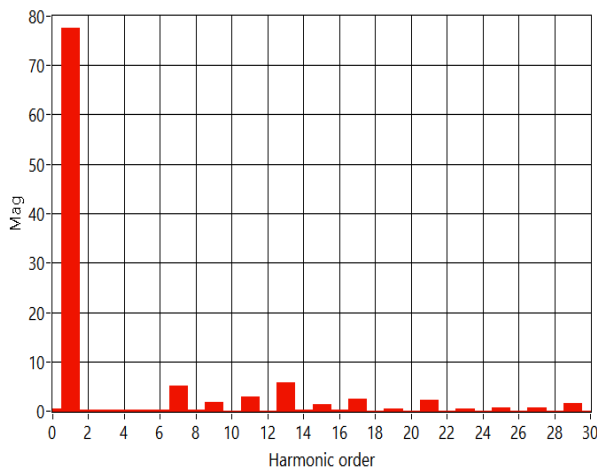


Fig. 19. FFT of 7-level inverter experimental voltage output

VI. CONCLUSION

This paper illustrates the use of differential evolution algorithm in selective harmonic elimination for a single phase seven level voltage source inverter to improve the harmonic quality of the generated output voltage. The proposed multi-level inverter with non-equal DC sources has the advantage of generating multiple voltage levels with less switching components. The differential evolution algorithm is used to solve a set of non-linear equations in order to obtain the optimal switching angles to perform the (SHE) modulation strategy. Optimal switching angles are investigated over the range $r \in [0.2 \ 0.95]$. The total harmonic distortion (THD) was chosen as a performance indicator in order to examine the effectiveness of the proposed algorithm. The validity of the

method has been proven by computer simulation using Matlab/Simulink scientific programming environment and verified by experimental hardware set-up based on STM32F407 microcontroller. The obtained results from the simulation and hardware show a good agreement with the theoretical prediction.

REFERENCES

- [1] H. Wang, Y. F. Liu and P. C. Sen, "A neutral point clamped multilevel topology flow graph and space NPC multilevel topology," 2015 IEEE Energy Conversion Congress and Exposition (ECCE), Montreal, QC, 2015, pp. 3615-3621.
- [2] G. P. Adam, S. J. Finney, O. Ojo and B. W. Williams, "Quasi-two-level and three-level operation of a diode-clamped multilevel inverter using space vector modulation," in IET Power Electronics, vol. 5, no. 5, pp. 542-551, May 2012.
- [3] Z. Chunyan and L. Zhao, "Advanced compensation mode for cascade multilevel static synchronous compensator under unbalanced voltage," in IET Power Electronics, vol. 8, no. 4, pp. 610-617, 4 2015.
- [4] A. Edpuganti and A. K. Rathore, "Optimal Low Switching Frequency Pulsewidth Modulation of Nine-Level Cascade Inverter," in IEEE Transactions on Power Electronics, vol. 30, no. 1, pp. 482-495, Jan. 2015.
- [5] M. Hajizadeh and S. H. Fathi, "Selective harmonic elimination strategy for cascaded H-bridge five-level inverter with arbitrary power sharing among the cells," in IET Power Electronics, vol. 9, no. 1, pp. 95-101, 1 20 2016.
- [6] C. I. Odeh and D. B. N. Nnadi, "Single-phase 9-level hybridised cascaded multilevel inverter," in IET Power Electronics, vol. 6, no. 3, pp. 468-477, March 2013.
- [7] B. Karami, R. Barzegarkhoo, A. Abrishamifar and M. Samizadeh, "A switched-capacitor multilevel inverter for high AC power systems with reduced ripple loss using SPWM technique," Power Electronics, Drives Systems & Technologies Conference (PEDSTC), 2015 6th, Tehran, 2015, pp. 627-632.
- [8] K. C. Jana and S. K. Biswas, "Generalised switching scheme for a space vector pulse-width modulation-based N-level inverter with reduced switching frequency and harmonics," in IET Power Electronics, vol. 8, no. 12, pp. 2377-2385, 12 2015.
- [9] T. Mistry, S. K. Bhatta, A. K. Senapati and A. Agarwal, "Performance improvement of induction motor by Selective Harmonic Elimination (SHE) using Newton Raphson (N-R) method," 2015 International Conference on Energy Systems and Applications, Pune, India, 2015, pp. 364-369.
- [10] Erkan Deniz, Omur Aydogmus, Zafer Aydogmus, Implementation of ANN-based Selective Harmonic Elimination PWM using Hybrid Genetic Algorithm-based optimization, Measurement, Volume 85, May 2016, Pages 32-42
- [11] M. Gnana Sundari, M. Rajaram, Sujatha Balaraman, Application of improved firefly algorithm for programmed PWM in multilevel inverter with adjustable DC sources, Applied Soft Computing, Volume 41, April 2016, Pages 169-179
- [12] Shimi Sudha Letha, Tilak Thakur, Jagdish Kumar, Harmonic elimination of a photo-voltaic based cascaded H-bridge multilevel inverter using PSO (particle swarm optimization) for induction motor drive, Energy, Volume 107, 15 July 2016, Pages 335-346,
- [13] S. Das and P. N. Suganthan, "Differential Evolution: A Survey of the State-of-the-Art," in IEEE Transactions on Evolutionary Computation, vol. 15, no. 1, pp. 4-31, Feb. 2011.
- [14] Z. Salam, A. M. Amjad and A. Majed, "Using Differential Evolution to Solve the Harmonic Elimination Pulse Width Modulation for Five Level Cascaded Multilevel Voltage Source Inverter," Artificial Intelligence, Modelling and Simulation (AIMS), 2013 1st International Conference on, Kota Kinabalu, 2013, pp. 43-48.
- [15] A.Hiendro, "Multiple switching patterns for SHEPWM inverters using Differential evolution algorithms" in International Journal of Power Electronics and Drive Systems, vol.1, no2, pp94-103,Dec 2011

- [16] K. K. Gupta, A. Ranjan, P. Bhatnagar, L. K. Sahu and S. Jain, "Multilevel Inverter Topologies With Reduced Device Count: A Review," in *IEEE Transactions on Power Electronics*, vol. 31, no. 1, pp. 135-151, Jan. 2016.
- [17] J. Han, T. Yang, D. Peng, T. Wang and G. Yao, "Model predictive control for asymmetrical cascaded H-Bridge multilevel grid-connected inverter with flying capacitor," *IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society*, Dallas, TX, 2014, pp. 1611-1616
- [18] R. Vijayakumar, C. Devalalitha, A. Nachiappan and R. Mazhuvendhi, "Selective harmonic elimination PWM method using two level inverter by differential evolution optimization technique," *Science Engineering and Management Research (ICSEMR), 2014 International Conference on*, Chennai, 2014, pp. 1-6.
- [19] M. I. Mohd Rashid, A. Hiendro and M. Anwari, "Optimal HE-PWM inverter switching patterns using differential evolution algorithm," *Power and Energy (PECon), 2012 IEEE International Conference on*, Kota Kinabalu, 2012, pp. 32-37.
- [20] C. Sun, H. Zhou and L. Chen, "Improved differential evolution algorithms," *Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on*, Zhangjiajie, 2012, pp. 142-145.

Human Face Classification using Genetic Algorithm

Tania Akter Setu

Dept. of Computer Science and Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh, Bangladesh

Dr. Md. Mijanur Rahman

Dept. of Computer Science and Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh, Bangladesh

Abstract—The paper presents a precise scheme for the development of a human face classification system based human emotion using the genetic algorithm (GA). The main focus is to detect the human face and its facial features and classify the human face based on emotion, but not the interest of face recognition. This research proposed to combine the genetic algorithm and neural network (GANN) for classification approach. There are two way for combining genetic algorithm and neural networks, such as supportive approach and collaborative approach. This research proposed the supportive approach to developing an emotion-based classification system. The proposed system received frontal face image of human as input pattern and detected face and its facial feature regions, such as, mouth (or lip), nose, and eyes. By the analysis of human face, it is seen that most of the emotional changes of the face occurs on eyes and lip. Therefore, two facial feature regions (such as lip and eyes) have been used for emotion-based classification. The GA has been used to optimize the facial features and finally the neural network has been used to classify facial features. To justify the effectiveness of the system, several images were tested. The achievement of this research is higher accuracy rate (about 96.42%) for human frontal face classification based on emotion.

Keywords—Face Detection; Facial Feature Extraction; Genetic Algorithm; Neural Network

I. INTRODUCTION

The human face plays a central role in social interaction; hence it is not surprising that automatic facial information processing is an important and highly active subfield of pattern recognition research [1]. In the vision technology area, researchers have started to investigate and develop human face processing systems. Due to the complexity of face recognition, detecting a human face and its facial features and classify the human face base on emotion without identifying the person is of interest [2]. In recent years, there has been a growing interest in improving all aspects of the interaction between humans and computers especially in the area of human

emotion recognition by observing facial expressions. The universally accepted categories of emotion, as applied in human-computer interaction are Sad, Anger, Joy, Fear, Disgust (or Dislike) and Surprise [3]. Emotions related to facial expressions. Hence , the features based on the position of the face. Hence, several methods have been proposed to classify emotions. Mase proposed emotion recognition systems that use directions of facial muscles. Muscle movements were extracted use of optical flow with 11 windows method place in the face [4]. For classification, K-nearest neighbor rule was uses with an accuracy of 80% with happy, anger, disgust, surprise emotions [5].Yacoub proposed the same method instead of muscle action, he uses the edge of mouth, eyes and eyebrows, into a frame, mid-level representation, classify the emotions [6]. Black et al. proposed a parametric model. In this model to extract the shape and movement of eyes, mouth, eyebrows, into a mid and high-level representation of facial expression with 80% of accuracy [6]. Ekman proposed a geometric model in which to extract shape and appearance of a lip, nasolabial furrow and wrinkles with 82% accuracy [7]. M. Karthigayan, M. Rizon, R. Nagarajan and Sazali Yaacob proposed a method of Genetic Algorithm and Neural Network for Face Emotion Recognition [3].This research focus on finding, segmenting and classifying human faces, actually includes three parts: human face detection, facial feature segmentations and classification. The goal is to find the face region of a person in an image that is dominated by the upper half of the body, and to segment this face region into four parts: the face region, eyes region, mouth region and nose region. From Segmenting it optimizes the feature value using Genetic algorithm. Then classify face image base on Emotion by Neural Network.

II. METHODOLOGY

As shown in Figure 1, methodological steps for combining genetic algorithm and neural network to classify the facial features based human emotion.

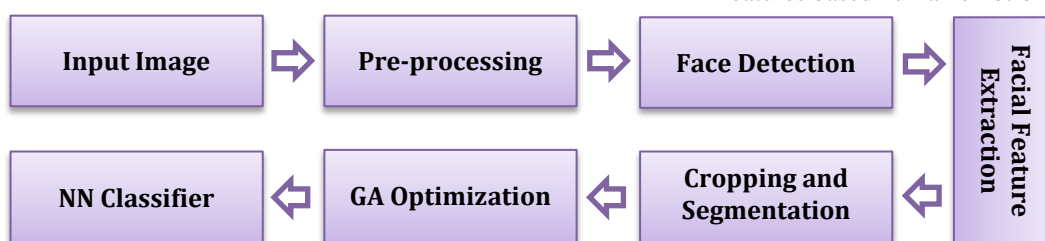


Fig. 1. Block Diagram of the Proposed GANN Classifier

A. Face Image Acquisition

The process of getting the image from any source, especially hardware is called as image acquisition. For image acquisition use a digital camera. In the image processing it is impossible without image receiving/acquisition. The sweetest Acquisition process is a digital camera into various formats such as Bitmap, JPEG, GIF and TIFF etc. and collects image from Google Image.

B. Image Preprocessing

The image preprocessing includes smoothing or filtering and gray-scale conversion. The purpose of smoothing is to reduce noise and improve the visual quality of the image often; smoothing is referred to as **filtering**. For this purpose of filtering we have used Gaussian Filter. The equation – 1 expresss Gaussian function. If the image is not noisy it is not necessary to filtering.

$$G_{\sigma}(x, y) = \frac{1}{2\pi\sigma^2} EXP^{-\frac{x^2+y^2}{2\sigma^2}} \dots\dots\dots(1)$$

Filtering is not suitable for all images. Then convert RGB image into Gray Scale image.

C. Face Detection and Feature Extraction

Feature is very significant to any object detection algorithm. The computer vision object detector of Matlab 2013 has been used in this research. The Viola Jones algorithm used for selecting the facial features [8]. There are a

lot of features, such as eyes, nose, the topology of eye and nose, can be used for face detection. In Viola Jones face detection, a very simple and straightforward feature has been used. Each feature obtained by subtracting white areas from the black areas. The area means the summation of all the pixels gray value within the rectangle. A special representation of image, named integral image, has been used for calculating these features. At first, the facial region will detect then other parts of the face. This research identifies the human facial feature regions such as face, nose, eyes and lip. It also computes the boundary box value which performs multi-scale object detection on the input image and returns M-by-4 matrix.

D. Cropping and Segmentation

From detecting feature, cropping the Eyes and lip region according to the BBOX value. Image segmentation is typically used to identify objects or other relevant information in digital images. Edge detection of an image which converts in a binary image. For image segmentation and edge detection has been used Sobel operator of Gradient Based Method[9].

E. GANN Face Classification

In this research, eyes and lip are used for human face classification. After detecting the human face, it cropped and segmented the eyes and mouth part as individual segments by using edge detection and then combining the genetic algorithm and neural network (GANN) for classification. The overall process is shown in Figure 2.

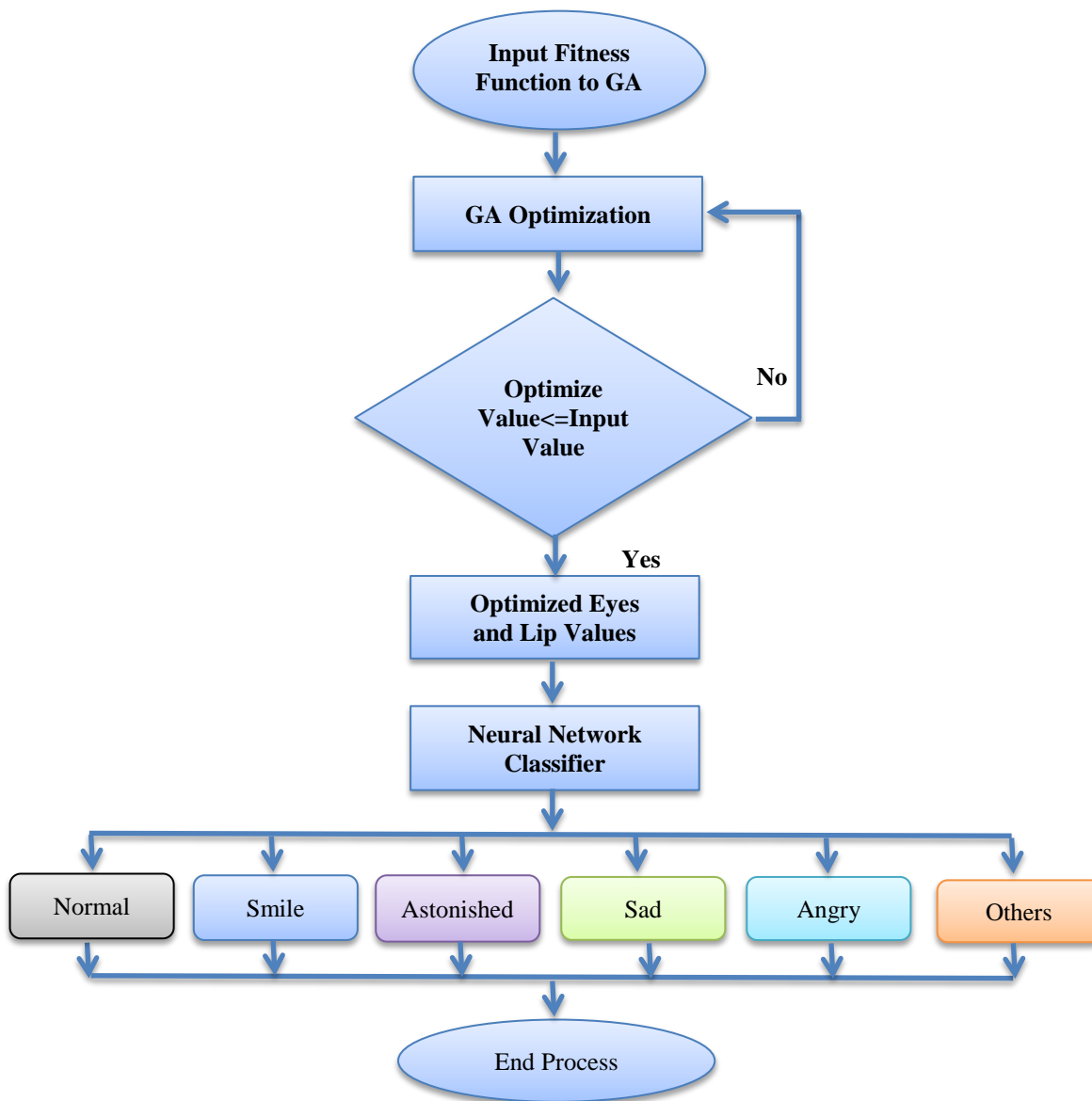


Fig. 2. Flow Chart of GANN Based Face Classification

The height and width of the mouth and eyes white pixel are calculated; so, it is measured from the top and bottom row through the X coordinate and also measure the left and right columns. After segmentation, the shape of eyes and mouth region are looked like an ellipse. The ellipse has two axes such as, the major axis and minor axis. This task is done by using the equations given below:

$$a = (y_{max} - y_{min})/2; \quad [y_{max} - y_{min} = \text{mouth width and } a = \text{major axis}]$$

$$b = (x_{max} - x_{min})/2; \quad [x_{max} - x_{min} = \text{mouth height and } b = \text{minor axis}]$$

The research proposed to combine genetic algorithm and neural network (GANN) for classification. In this research, the supportive approach for GANN has been used.

1) GA Optimization

GA is better than conventional AI. It is more robust. GA is a heuristic Search algorithm. They do not break easily even if the inputs changed slightly, or in the presence of reasonable noise. A genetic algorithm may offer significant benefits over more typical search of optimization techniques (linear programming, heuristic, depth-first, breath-first, and praxis) [10]. The region of eyes and lip consider as irregular ellipse. The region of eyes and lip are calculated by ellipse area equations. GA uses ellipse area calculation equation as a fitness function and this equation is given below:

$$Area = 3.1416 * a * b$$

GA takes this irregular ellipse's major axis and minor axis as input. GA Optimizes the major axis and minor axis of the irregular ellipse and provides a regular ellipse major axis and minor axis value, as shown in Figure 3.

For GA optimization, it uses another function called the condition function where the optimize area is less than or equal to actual area. GA individually optimizes the left eyes, right eyes and mouth major axis and minor axis. The GA optimization uses the mean value of eyes and the ratio of eyes and mouth.

TABLE I. GA PARAMETER

TolFun	1.0000e-08
Display	'iter'
Population Size	10

The mean value of eyes and the ratio of eyes and mouth are calculated using the following measurement equation (2), (3), (4), (5):

$$\begin{aligned} \text{Mean Eyes Major Axis} &= (\text{right eye major axis} + \text{left eyes major axis})/2 \dots\dots\dots(2) \\ \text{Mean Eyes Minor Axis} &= (\text{right eye minor axis} + \text{left eyes minor axis})/2 \dots\dots\dots(3) \\ \text{Eyes ratio} &= (\text{Mean Eyes Major Axis} / \text{Mean Eyes Minor Axis}) \dots\dots\dots(4) \\ \text{Mouth ratio} &= (\text{Major Axis} / \text{Major Axis}) \dots\dots\dots(5) \end{aligned}$$

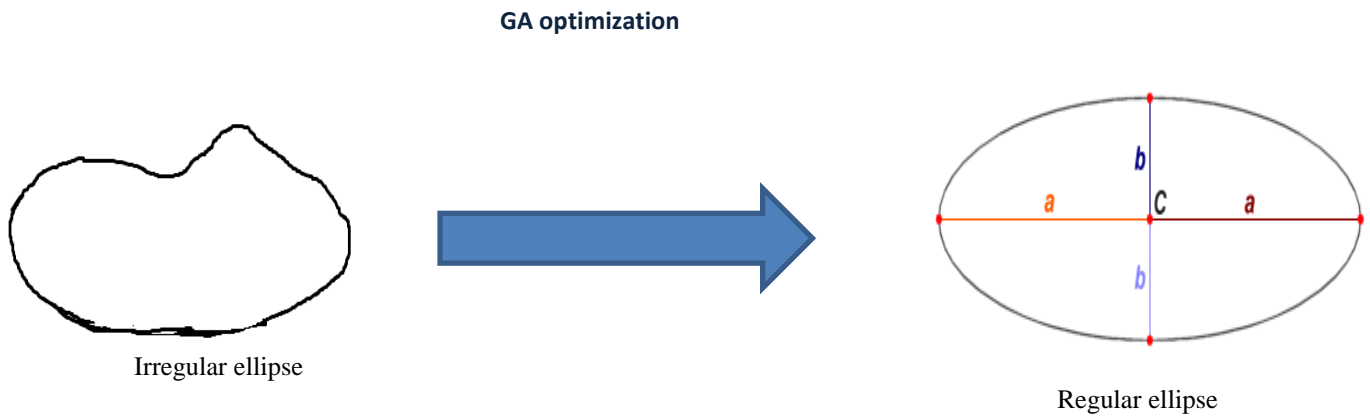


Fig. 3. Converting an irregular ellipse into a regular ellipse

2) Neural Network Classifier

This research uses the ratio value of eyes and mouth for classification of the human face. The proposed system classifies the facial images into six categories based on these values such as normal face, smile face, astonished face, angry face and sad face. The images that do not match in these five classes belong to other class. Therefore, the measurement includes five input pattern and five target pattern. The feed forward neural network with gradient decent adaptive learning algorithm is used for training the input pattern. The network has five input, two hidden and five output layers of twenty five neurons. The tangent sigmoid (tansig) is used as a layer transfer function.

There are several stopping criteria of the network, like as, maximum epochs required, performance goal meet, minimum gradient reach, validation check etc.

III. CLASSIFICATION RESULTS

Table 3 shows the measured ratio and GA optimized ratio value of mouth and eyes. Total 15 images are tested to measure the performance of the classification system. The

developed system achieved the better result in face classification. Table 4 shows the performance of the system based on GA.

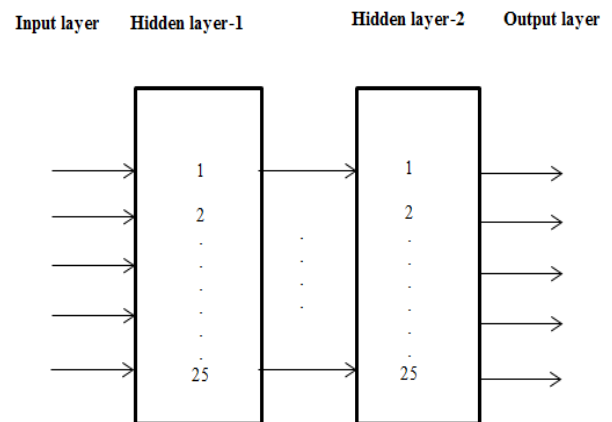


Fig. 4. 5x25x25x5 Neural Network Structure Block Diagram

TABLE II. DETAILS ABOUT NEURAL NETWORK

NN Structure	NN Type	Maximum Epoch	Training Algorithm	Transfer function
5×25×25×5	Feed Forward	1500	Gradient Decent with adaptive learning rate (traingda)	Tangent sigmoid (tansig)

TABLE III. MEAN AND GA OPTIMIZED VALUE OF EYES AND MOUTH AND THEIR CLASS OF EMOTION

Image file	For eyes				For mouth				Class
	Mean Eyes Major Axis	Mean Eyes Minor Axis	Mean Eyes Ratio	GA optimized Eyes Ratio	Mouth Major Axis	Mouth Minor Axis	Mouth Ratio	GA Optimized Mouth Ratio	
n1	20	7	2.8571	3.0800	42.5000	17.5000	2.4286	4.4268	Normal
n2	5.5000	5	1.1000	1.6885	39	13.5000	2.8889	3.4962	Normal
n3	19.5000	7	2.7857	2.8992	38.5000	18	2.1389	3.2521	Normal
s1	5.5000	4	1.3750	2.0435	25.5000	15.5000	1.6452	2.4533	Smile
s2	24	11	2.1818	2.1860	39	19	2.0526	2.3428	Smile
s3	10	4	2.5000	2.0526	27.5000	18	1.5278	2.6598	Smile
s4	102	56	1.8214	2.4591	161.5000	97	1.6649	2.7444	Smile
w1	15.5000	16	0.9688	1.4465	30.5000	23	1.3261	1.4367	Astonished
w2	12	4	3	2.8684	28	17.5000	1.6000	1.7460	Astonished
an1	19	6	3.1667	4.1110	31	23	1.3478	1.3232	Angry
an2	30	9	2.36	3.33	50	22.50	2.22	1.90	Angry
sad1	29	13.500	2.1481	3.1691	49.5000	27	1.8333	2.9662	Sad
sad2	11.5000	3.5000	3.2857	3.2487	27.5000	16.5000	1.6667	2.5187	Sad
sad3	19.5	7	2.78	3.25	55.5000	27.5000	2.0182	3.0317	Sad

TABLE IV. PERFORMANCE TABLE

Image File	Belong Class	Number of Test	Number of Proper Classification	Number of the Wrong Classification	Accuracy (%)
n1	Normal	10	9	1	90%
n2	Normal	10	10	0	100%
n3	Normal	10	10	0	100%
s1	Smile	10	10	0	100%
s2	Smile	10	10	0	100%
s3	Smile	10	10	0	100%
s4	Smile	10	10	0	100%
w1	Astonished	10	10	0	100%
w2	Astonished	10	9	1	90%
an1	Angry	10	10	0	100%
an2	Angry	10	7	3	70%
sad1	Sad	10	10	0	100%
sad2	Sad	10	10	0	100%
sad3	Sad	10	10	0	100%
Total		140	135	5	96.42%

IV. CONCLUSION

The developed system classified the frontal face of human based on human emotion, like normal, smile, anger, sad and astonished. The achievement of this research is higher accuracy rate for human frontal face classification based on emotion. This research proposed a new fitness function of genetic algorithm and made a ratio based classification of the human face. The developed system achieved the classification performance rate is 96.42%. This system works only frontal face of single image and it's not work by side view of Image. Sometimes it is overlapped when one class is very much close to another class.

ACKNOWLEDGEMENT

The author would like to thanks to the Ministry of Information and Communication Technology of Bangladesh for giving financial support by providing MOICT-Fellowship.

REFERANCES

- [1] J. Daugman, "Face and Gesture Recognition: An Overview," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 675-676, July 1997
- [2] Qing Gu "Finding and Segmenting Human Faces" 2008.
- [3] M. Karthigayan, M. Rizon, R. Nagarajan and Sazali Yaacob (2008). Genetic Algorithm and Neural Network for Face Emotion Recognition, Affective Computing, Jimmy Or (Ed.), ISBN: 978-3-902613-23-3
- [4] Mase K. Recognition of facial expression from optical flow. IEICE Trans., E. 74(10):3474-3483, October 1991.
- [5] Yacoob, Y., Davis, L. Computing spatio-temporal representations of human faces. Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on , 21-23 June 1994 PP: 70 -75.
- [6] Black, M. J. and Yacoob, Y. Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion. In Proceedings of the International Conference on Computer Vision, pages 374-381. IEEE Computer Society, Cambridge, MA, 1995.
- [7] Tian, Ying-li, Kanade, T. and Cohn, "Recognizing lower face action units for facial expression Analysis", IEEE Transaction on Automatic Face and Gesture Recognition, march, 2000, pp.484-490.
- [8] Paul Viola and Michel J. Jones: "Rapid Object Detection Using a Boosted Cascade of Simple Features", *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.511-518, 2001.
- [9] Poonam Dhankhar, Neha Sahu "Edge Based Human Face Detection Using Matlab" CSE ITM University Gurgaon-Haryana, Proceedings of IRF International Conference, 16th February 2014.
- [10] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html

An Example-based Super-Resolution Algorithm for Multi-Spectral Remote Sensing Images

W. Jino Hans

Assistant Professor

Department of Electronics and Communication Engineering,
SSN College of Engineering,
Kanchipuram, India

Venkateswaran N

Professor

Department of Electronics and Communication Engineering,
SSN College of Engineering,
Kanchipuram, India

Lysiya Merlin.S

Student, M.E communication systems

Department of Electronics and Communication Engineering,
SSN College of Engineering,
Kanchipuram, India

Divya Priya T

Student, M.E communication systems

Department of Electronics and Communication Engineering,
SSN College of Engineering,
Kanchipuram, India

Abstract—This paper proposes an example-based super-resolution algorithm for multi-spectral remote sensing images. The underlying idea of this algorithm is to learn a matrix-based implicit prior from a set of high-resolution training examples to model the relation between LR and HR images. The matrix-based implicit prior is learned as a regression operator using conjugate decent method. The direct relation between LR and HR image is obtained from the regression operator and it is used to super-resolve low-resolution multi-spectral remote sensing images. A detailed performance evaluation is carried out to validate the strength of the proposed algorithm.

Keywords—Remote sensing Super-resolution; Image-pair analysis; Regression operators

I. INTRODUCTION

Remote sensing is vital for various application such as decision support for disaster management, weather monitoring and surveillance of land [1]. The data provided by Geo Information System (GIS) is significant for remote sensing. In recent years, sophisticated imaging devices and state-of-the-art technologies have been continuously deployed in the Earth observation system to provide precise data for remote sensing. The need to have a high spatial resolution for satellite images is twofold: apart from the improvement in visual inspection of a larger dataset precisely, it also plays a vital role in the post-processing steps such as feature extraction and segmentation of objects from the image. However, despite using sophisticated high-resolution (HR) imaging devices for satellite imagery, the captured images will inherit a poor spatial resolution due to the larger distance between the sensors and the sensed object. cartosat series of satellites are used to regularly monitor earth for disaster management.. These satellites have evolved over past few decades and currently the Cartosat-3 with the advanced imaging device is in existence. The panchromatic (PAN) and multi-spectral (MS) imaging devices are deployed in these satellites to provide progressive imaging. The PAN images have high spatial and low spectral resolution whereas MS images have high spectral but low spatial resolution. The

spatial resolution of an MS image captured by cartosat-2 series satellite will be approximately 2.5 meters/pixel.

The modern image sensor element used in MS imaging device is typically a charge-coupled device (CCD) or a complementary metal-oxide-semiconductor (CMOS) active-pixel sensor. The image signals are captured by the sensor elements that are typically arranged in a two-dimensional array. The size of sensor element or the number of sensor element present in a unit area determines the spatial resolution of an image. The spatial resolution of MS image is significantly less due to the limited dynamic range of CCD sensors. An imaging device with deficient sensor elements will generate low-resolution (LR) images with blocky and displeasing visual artifacts due to aliasing effect. However, deploying more sensor elements to increase spatial resolution will incur additional cost.

Moreover, the limitation to deploy high precision optics in imaging device is diploid. In addition to the cost incurred due to the increase in sensor elements, the ceaseless demand to improve the spatial resolution cannot be catered by the state-of-the-art camera technologies. For instance, reducing the pixel will increase the spatial resolution but will introduce shot noise [2]. Similarly increasing the number of pixels in a unit area by increasing the chip size can increase the spatial resolution. However, increasing the chip size will increase the capacitance which results in undesired artifacts [3]. Due to the inherited limitations, the spatial resolution of MS images will be poor. Anyhow, in many applications including disaster management, rescue operations, resource surveying, etc. precise geo spatial information is required. Henceforth, it is significant to use an effective post-process technique such as image super-resolution (SR) approach to improve the spatial resolution of MS images. The need to improve the spatial resolution of remote sensing imagery have garnered special interest by researchers and have witnessed diverse SR algorithms [4-12].

Spurred by the need to improve the spatial resolution of Landsat images, Tsai et al. presented a conventional multi-image SR algorithm in frequency domain [4]. Classical SR algorithm requires multiple frames of the same scene with exact registration to super-resolve a LR image [4-7]. To overcome this, a wide variety of learning-based single image SR algorithm have been proposed [8-12]. SR from single LR satellite image is a challenging task as the problem is severely ill-posed. However, learning-based SR algorithms can effectively handle the ill-posed problem by learning an efficient prior to model the relation between low and HR training image patches. The prior required to handle the ill-posed problem can be either explicit or implicit. Explicit priors use a mathematical energy functional of an image class such as primal sketches, Field of Experts (FoE) [13], Gradient profile [14] etc. to model the relation between LR images with its HR counterpart. In contrary, implicit priors are learned from the training image pairs and give rise to a family of SR algorithms called as Example-based SR algorithms [15]. It requires a collection high-quality example images and synthetically generated LR images to learn the image-pair prior information. The correspondence between LR images with its corresponding HR image is learned as an implicit prior.

The implicit prior can be learned either by a direct mapping approach or an indirect mapping approach depending on the patch reconstruction strategy used. The indirect mapping approaches employ nearest neighbor embedding algorithm [16,17], which requires an exhaustive search to find the nearest neighbor which makes it computationally expensive for practical applications like satellite remote sensing. Direct mapping approaches will learn the relation between LR and HR image as a regression function, thereby computationally it will be efficient to super-resolve remote sensing images [18]. Despite, most of the conventional regression based SR algorithms vectorizes the image patches which results in loss of image-level information while learning the implicit prior. To address this recently a few matrix-based implicit priors have been reported [19] [20], which avoids the vectorization step and learn the implicit priors as a matrix-based regression operator. The regression operator establishes a direct mapping between the training image patches and can be effectively used to reconstruct the HR image.

In this paper, we propose an example-based SR algorithm to super-resolve spatially under-sampled cartosat-2 series MS images by learning an efficient matrix-based implicit prior from a set of HR-MS images. The proposed matrix-based implicit prior will preserve the structural similarities in the image thereby will not introduce any unpleasant artifacts.

The reminder of this paper is as follows. A brief discussion on implicit prior is presented in section 2. In section 3, the methodology of the proposed example-based SR algorithm for MS remote sensing image is presented. The performance of the proposed algorithm is evaluated and the results are reported in section 4 and finally, section 5 concludes the paper.

II. A BRIEF DESCRIPTION ON IMPLICIT PRIORS

The fine details that are explicitly missed during the degradation process are estimated by an example-based SR algorithm. Fig. 1 illustrates the process of example-based SR algorithm.

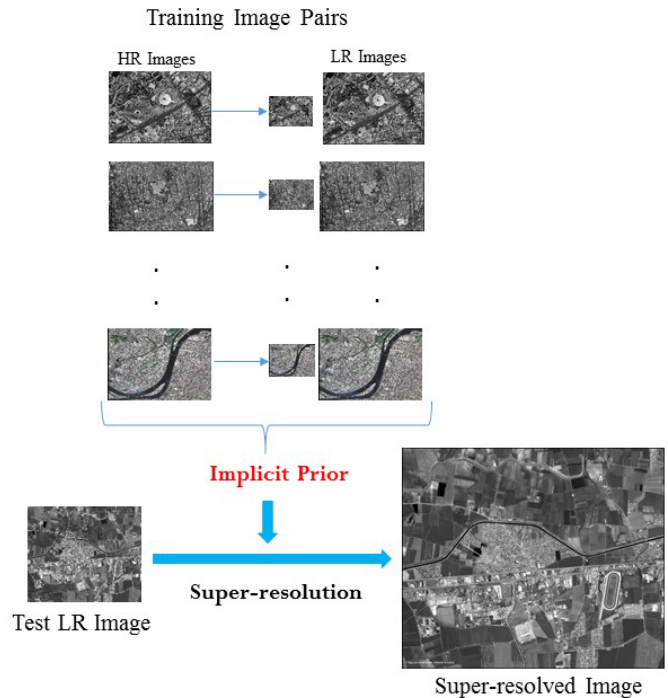


Fig. 1. An overview of Example-based SR Algorithm

Though the problem is severely ill-posed, efficient priors are used for regularizing the solution. For an example-based SR algorithms, the prior will be learned from a set of training examples itself and hence they are named as implicit priors. Training examples will be a collection of LR images and its corresponding synthetically generated LR images. An implicit prior which models the relation between LR image patches to its corresponding HR image patches can be learned as a regression function. In most of the state-of-the-art SR approaches, the regression function will be learned by vectorizing the training image patch-pairs.

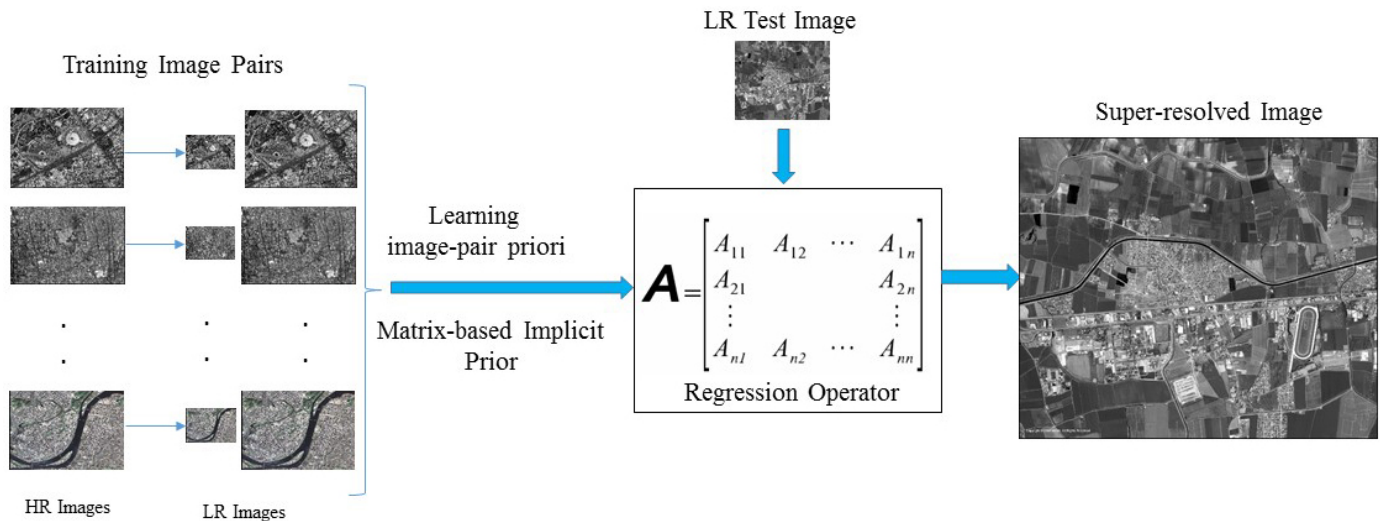


Fig. 2. An overview of the proposed example-based super-resolution algorithm for multi-spectral remote sensing images

In vector-based implicit priors, the correspondence between LR-HR image pair is learned from the feature vectors and hence instead of learning the image-level correspondence; feature-level correspondence is learned. In matrix-based implicit prior, the image patches will be preserved as matrix itself and therefore the image-level structural information will be preserved in the image. It is vital to preserve the structural information in an MS remote sensing image as it will possess a lot of HF details. In literature, quite a few matrix-based implicit prior have been reported. In these methods, the correspondence between LR and HR image patches will be learned as linear matrix-based regression operator.

A. Matrix-based Implicit Prior

The matrix-based implicit prior is learned as regression operator as follows. Let a set of training example patch-pair be denoted as $P = (x_i, y_i)_{i=1}^n$ such that x and y are the LR and HR patches of size $m \times m$ respectively. A linear matrix-based regression operator $R: \in \mathbb{R}^{m \times m}$ serves as an implicit prior, such that for an image patch-pair (x, y)

$$y = R \cdot x \quad (1)$$

If the image patches are assumed to be full rank, then the regression operator can be defined as

$$R = yx^{-1} \quad (2)$$

However, as the regression operator will be learned from a collection of LR-HR patch-pairs, it is required to find a suitable optimal regression operator by solving a least square regression problem. The optimal regression operator can serve as an implicit prior to learn the correspondence between LR and HR image patches.

III. METHODOLOGY OF THE PROPOSED ALGORITHM

The overview of the proposed SR methodology is shown in Fig. 2. The example-based SR algorithm to super-resolve MS images captured by cartosat-2 series satellite will typically have two phases, viz. training and reconstruction phase. In training phase, the required prior information is learned as a matrix-based implicit prior which is performed offline. In the

reconstruction phase, the learned implicit prior is used to reconstruct the HR image.

A. Training Phase

In the training phase, high-quality MS images are collected from a remote sensing database (For instance, to super-resolve existing cartosat 2 series MS images with a spatial resolution of 2 meters, cartosat-3 images with a spatial resolution of 1 meter is collected). These HR images are synthetically degraded to obtain the LR images. The degradation process includes a blur operator which is modeled by the movement of sensor element and a decimation operator, which corresponds to the insufficient sensor elements. Also, atmospheric noises can degrade the quality of satellite images. A set of HR image and its corresponding LR image form the training examples. Let the set of training examples is given by

$$T = \{X_i, Y_i\}_{i=1}^n \quad (3)$$

Where X and Y represents the LR and HR training examples respectively. Let K patches of size $m \times m$ are extracted from the training examples from the same location such that a set of patch-pairs is represented as

$$P = \{x_i, y_i\}_{i=1}^n \quad (4)$$

Algorithm 1: Learning the regression Operator

Input: Training sample set, $T = \{X_i, Y_i\}_{i=1}^n$
Output: Optimal regression operator, R^*
Step (1): Obtain the patch pairs,
 $P = \{x_i, y_i\}_{i=1}^n$
Step(2): Calculate the initial estimate,

$$R_j = \sum_{i=1}^n (y_i x_i^{-1})_{j \neq i}$$

Step(3): Calculate the optimal regression operator

$$R^* = \underset{R_j}{\operatorname{argmin}} \left\| y_i - R_j x_i \right\|^2 + g \left\| R_j - \tilde{R}_j \right\|_F^2$$

Output: Optimal regression operator, R^*

As the patch-pairs are extracted from the same location, the linear regression model can be adopted to relate the LR and HR patch-pairs. Therefore

$$y = Rx \quad (5)$$

The matrix-based implicit prior is learned from the training patch-pairs by solving a least square regression as follows. The objective function to learn the optimal regression operator is given by

$$R^* = \operatorname{argmin}_R \|y_i - Rx_i\|_F^2 \quad (6)$$

In the above equation, let the initial estimate of the regression operator is obtained by taking the inverse of x , such that

$$R_j = (y_j x_j^{-1})_{j \neq i} \quad (7)$$

Let the global constraint to estimate the regression operator is given by

$$R^* = \operatorname{argmin}_{R_j} \|y_i - R_j x_i\|_F^2 + g \|R_j - \tilde{R}_j\|_F^2 \quad (8)$$

The above optimization problem to find the optimal regression operator is solved by conjugate gradient decent method. The term $g \|R_j - \tilde{R}_j\|_F^2$ is the priori for the optimization problem. This is an iterative approach and the update equation for the iteration is given by [21],

$$R^{i+1} = R^i + \varepsilon [S^T \cdot E_t + g(R_j - \tilde{R}_j)] \quad (9)$$

Where $E_t = y_i - Rx_i$ is the error due to the i^{th} iteration and R^i is the learned regression operator after i^{th} iteration. The optimal regression operator is used to reconstruct the HR image. Algorithm 1 summarizes the steps involved to learn the optimal regression operator.

B. Reconstruction Phase

In the reconstruction phase, the LR MS cartosat images are super-resolved using the matrix-based implicit prior which is learned as a regression operator given by Eq. (6). The test LR MS image is up-scaled by an interpolator by a scale-factor s . Non-overlapping patches of size $m \times m$ are extracted from the interpolated image. The collection of the extracted patches is represented as a set $T = \{p_{lr}^i\}_{i=1}^n$. All the test LR patches are super-resolved using the matrix-based regression operator, given by

$$p_{hr} = R^* p_{lr} \quad (10)$$

Input: Optimal regression operator R^* , LR Test set, T
Output: Super-resolved HR image, H
Step (1): Merge the LR test patches,
 $T = \{p_{lr}^i\}_{i=1}^n$
Step (2): Obtain the HR patches using regression operator obtained from Algorithm-1,
 $p_{hr} = R^* p_{lr}$
Step(3): Merge the super-resolved test patches,
 $H = \{p_{hr}^i\}_{i=1}^n$
Output: Super-resolved HR image, H

The super-resolved patches are merged to obtain the super-resolved HR image H . The steps to super-resolve a LR MS image is summarized in Algorithm 2.

Algorithm 2: SR Reconstruction

IV. RESULTS AND DISCUSSION

The effectiveness of the proposed SR algorithm to super-resolve LR multi-spectral image is evaluated on a set of remote sensing images captured by cartosat-1 satellite as shown in Fig. 3. All the experiments are simulated in MATLAB using a personal computer with Intel core-i5-2400 @ 2.7 GHz processor with 4 GB RAM. To generate training examples, HR multi-spectral images captured by COMSAT-1 are collected. Sample training images are shown in Fig.4. These HR images are downgraded with a scale-factor s using bi-cubic interpolation to synthetically generate the LR images. In all the experiments, the patch-size is 11×11 and the LR images are super-resolved by a scaling factor of 2 and 4. The test images are shown in Fig. 3 are super-resolved by various state-of-the-art SR algorithms such as Yang et al.'s sparse representation based approach [22], and Dong et al.'s non-local autoregressive modeling (NARM) [23]. The results of the above algorithms are obtained using the source code available on the author's webpage.

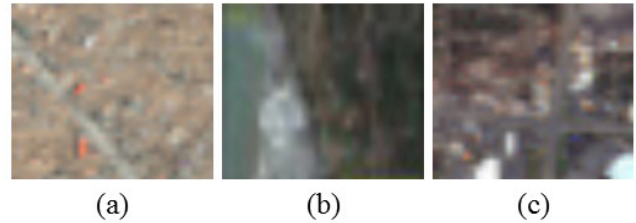


Fig. 3. Low-resolution multi-spectral test images (a) MS-1 (b) MS-2 (c) MS-3



Fig. 4. Sample training example images

The performance of the algorithm is examined by the experimental results obtained by the proposed algorithm. The effectiveness of the proposed algorithm is a measure of visual experience obtained from the reconstructed image. The reconstructed image is evaluated both qualitatively and quantitatively to assess its effectiveness.

Qualitative evaluation of SR depends on a few attributes of the reconstructed SR image. The image is visually inspected for its naturalness and sharpness to assess the quality of the reconstructed image. The sharpness of an image is assessed based on the high-frequency details present in it. It is desired that the SR algorithm should not introduce any counterfeit HF details. Similarly, image naturalness is attributed to the distortions and artifacts present in the image. If the fine-details in the image are not preserved, it will introduce jaggy and ringing and staircase artifacts. These artifacts will severely affect the quality of the image. These attributes in the images can be evaluated by visual comparison of the images.

Fig. 5 depicts a visual comparison for three multi-spectral images with other state-of-the-art SR approach. It can be seen from the super-resolved results for the three test multispectral images MS-1, MS-2 and MS-3 in Fig. 5, the results of the proposed algorithm is better in terms of visual fidelity. The proposed algorithm reconstructs an SR image with minimal jaggy and ringing artifacts compared with other approaches depicted in Fig. 5 (b & c).

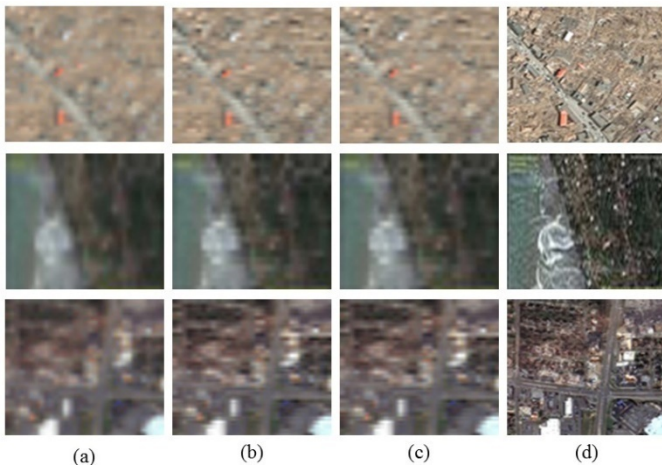


Fig. 5. Visual comparison for three multi-spectral images with state-of-the-art SR approaches (a) LR image (b-d) reconstructed SR image by Yang et al.'s method, Dong et al.'s method and proposed method respectively

The quantitative measure to evaluate the quality of the reconstructed image is figured by the PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index measure). A high PSNR score indicates that the magnified image is free from distortions and is more likely to carry HF details. SSIM value [24] (typically close to 1) indicates the similarity structure between the reconstructed image and its ground truth.

The PSNR of an image is defined by,

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE_{xy}} \right)$$

Where $MSE_{x,y} = \frac{\|x-y\|^2}{W*H}$, W is the width of the image patches x and y , H is height of both the image patches.

The SSIM of the reconstructed image is obtained using

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Where,

$$\mu_x = \frac{1}{W * H \sum_{i=1}^{W*H} x_i} ; \mu_y = \frac{1}{W * H \sum_{i=1}^{W*H} y_i}$$

$$\sigma_x = \frac{1}{W * H - 1 \sum_{i=1}^{W*H} ((x_i - \mu_x)^2)^{\frac{1}{2}}}$$

$$\sigma_y = \frac{1}{W * H - 1 \sum_{i=1}^{W*H} ((y_i - \mu_y)^2)^{\frac{1}{2}}}$$

c_1 and c_2 are constants.

TABLE I. A SUMMARY OF QUANTITATIVE EVALUATION (PSNR/SSIM) FOR MULTI-SPECTRAL IMAGES

Test Image	Yang et al.'s		Dong et al.'s		Proposed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MS-1	21.42	0.5677	21.48	0.5811	21.78	0.6877
MS-2	22.94	0.6789	22.91	0.6821	23.11	0.7347
MS-3	21.87	0.7211	21.92	0.7396	22.16	0.7919

Table I summarizes the quantitative comparison of the proposed method with various SR algorithms on multispectral LR images. The results tabulated in Table-1 shows that the proposed SR algorithm has the highest quantitative measures compared with other state-of-the-art algorithm. From Table-1, it is evident from the PSNR index that the proposed algorithm reconstructs the HR image with minimum distortions and the high SSIM index validates that the image-level information is preserved by the proposed matrix-based implicit prior.

V. CONCLUSION

In this paper, an example-based SR algorithm to super-resolve multi-spectral remote sensing image is presented. The proposed SR algorithm will learn a matrix-based implicit prior to map the correspondence between LR and HR images. The implicit prior is learned as a regression operator using the conjugate decent method. The learned matrix-based implicit prior is effectively used to super-resolve clean LR multi-spectral remote sensing images. In the future, the proposed algorithm will be extended to super-resolve noisy MS remote sensing images. The proposed algorithm is evaluated on clean images. Qualitative and quantitative experiments on various remote sensing images validates the efficacy of the proposed algorithm.

REFERENCES

- [1] V. Jayaraman, "India's Earth Observation Missions: Traversing through experiences of bilateral, regional and international cooperation," 58th IAC (International Astronautical Congress), International Space Expo, Hyderabad, India, Sept. 24-28, 2007, IAC-07-B1.1.02
- [2] H. Stark & P Oskui 1989, 'High Resolution Image Recovery from Image-plane Arrays using Convex Projections', Journal of Optical Society of America, Vol. 6, pp. 1715-1726.
- [3] T. Komatsu, K. Aizawa, T. Igarashi, & T. Saito 1993, 'Signal Processing Based Method for Acquiring Very High Resolution Image with Multiple Cameras and Its Theoretical Analysis', in Proc. IEE-I, pp. 19 - 25.
- [4] R.Y.Tsai and T.S Huang, 'Multiframe Image Restoration and Registration' in Advances in Computer vision and Image Processing, pp. 317-339. JAI Press Inc.,1984
- [5] M. T. Merino and J. Nunez, "Super-resolution of remotely sensed images with variable-pixel linear reconstruction,"IEEE Trans. Geosci Remote Sens., vol. 45, no. 5, pp. 1446-1457, May 2007.

- [6] .Núñez and M.Merino, "Super-resolution of remotely sensed images using drizzle and wavelets," in Proc. 25th Asian Conf. Remote Sens., Chiang Mai, Thailand, 2004, pp. 262–269.
- [7] F. Li, X. Jia, D. Fraser, and A. Lambert, "Super resolution for remotesensing images based on a universal hidden Markov tree model," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 3, pp. 1270–1278, Mar. 2010.
- [8] Y. Zhang, Y. Du, F. Ling, S. Fang, and X. Li, "Example-based super-resolution land cover mapping using support vector regression," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 4, pp. 1271–1283, Apr. 2014.
- [9] X. Xu, Y. Zhong, L. Zhang, and H. Zhang, "Sub-pixel mapping based on a MAP model with multiple shifted hyperspectral imagery," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 6, no. 2, Part: 2, pp. 580–593, Apr. 2013.
- [10] X. Li, Y. Du, and F. Ling, "Super-resolution mapping of forests with bitemporal different spatial resolution images based on the spatial-temporal Markov random field," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 1, pp. 29–39, Jan. 2014.
- [11] Y. Gu, Y. Zhang, and J. Zhang, "Integration of spatial-spectral information for resolution enhancement in hyperspectral images," IEEE Trans. Geosci. Remote Sens., vol. 46, no. 5, pp. 1347–1358, May 2008.
- [12] Y. Ge, S. Li, and V. C. Lakhan, "Development and testing of a subpixel mapping algorithm," IEEE Trans. Geosci. Remote Sens., vol. 47, no. 7, pp. 2155–2164, Jul. 2009.
- [13] Roth S & Black M J 2005, "Fields of Experts: A Framework for Learning Image Priors", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, USA.
- [14] Sun J, Sun J, Xx Z B, and Shum H Y 2008, 'Image super-resolution using gradient profile prior', in IEEE Conference on Computer Vision and Pattern Recognition, USA.
- [15] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based superresolution," IEEE Comput. Graph. Appl., vol. Vol. 22, pp. 56–65, March 2002.
- [16] Chang H, Yeung D Y, and Xiong Y 2004, 'Super-Resolution Through Neighbor Embedding', in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 275–282.
- [17] Hong C, Dit-Yan Y & Yimin X 2004, 'Super-resolution through neighbor embedding', Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, USA.
- [18] Hu J and Luo Y 2014, 'Single-image superresolution based on local regression and nonlocal self-similarity', Journal of Electronic Imaging, vol. 23, no. 3, pp. 1--14.
- [19] Y. Tang and Y. Yuan, "Image pair analysis with Matrix-value operator," IEEE Transactions on Cybernetics, January 2015
- [20] W. Jino Hans, N. Venkateswaran, Srinath Narayanan, Sandeep Ramachandran "An Example based Super-resolution Algorithm for Selfie Images" The Scientific World Journal, Hindawi Publishing Corporation, ISSN No. 1537-744X (Online), Article ID 8306342, 12 pages, 2016, doi:10.1155/2016/8306342
- [21] H. Y. Zhiliang Zhu, Fangda Guo and C. Chen, "Fast single image superresolution via self-example learning and sparse representation," IEEE Transactions on Multimedia, vol. Vol. 16, pp. 2178–2190, Dec 2014.
- [22] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," Trans. Img. Proc., vol. Vol. 19, pp. 2861–2873, nov 2010.
- [23] W. Dong, L. Zhang, R. Lukac, and G. Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," IEEE Transactions on Image Processing, vol. Vol. 22, pp. 1382–394, 2013
- [24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on image processing, vol. Vol. 13, pp. 600–612, 2004

Fitness Proportionate Random Vector Selection based DE Algorithm (*FPRVDE*)

Qamar Abbas

Department of Computing and
Technology
Iqra University
Islamabad,44000, Pakistan

Jamil Ahmad

Computer Science Department,
Abasyn University,
Islamabad,44000, Pakistan

Hajira Jabeen

Department of Computing and
Technology
Iqra University
Islamabad,44000, Pakistan

Abstract—Differential Evolution (DE) is a simple, powerful and easy to use global optimization algorithm. DE has been studied in detail by many researchers in the past years. In DE algorithm trial vector generation strategies have a significant influence on its performance. This research studies that whether performance of DE algorithm can be improved by incorporating selection advancement in effective trial vector generation strategies. A novel advancement in DE trial vector generation strategies is proposed in this research to speeds up the convergence speed of DE algorithm. The proposed fitness proportion based random vector selection DE (FPRVDE) is based on the proportion of individual fitness mechanism. FPRVDE reduces the role of poor performing individuals to enhance its performance capability of DE algorithm. To form a trial vector using FPRVDE, individual based on the proportion of their fitness are selected. FPRVDE mechanism is applied to most commonly used set of DE variants. A comprehensive set of multidimensional function optimization problems is used to access the performance of FPRVDE. Experimental result shows that proposed approach accelerates DE algorithm.

Keywords—Differential Evolution; Fitness Proportion; Trial vector generation; Mutation; Optimization

I. INTRODUCTION

Differential evolution (DE), proposed by Storn and Price [1] is a stochastic population based evolutionary algorithm. The advantage of DE over other evolutionary algorithms is that it is simple, easy to use, speedy and greater probability of finding the global optima for function optimization [2] [3]. DE has been successfully used in various real life fields like Electrical power systems [4], electromagnetism [5], control systems [6], Bioinformatics [7], chemical engineering [8], image processing [9], artificial neural networks [10], signal processing [11] etc.

DE is a population based algorithm where a population of potential solutions is randomly initialized within an n-dimensional search space. All potential solutions are equally likely to be selected as parent in DE algorithm. The candidate solutions evolve themselves by exploring the entire search space over time to locate the optima of the objective function [12]. New vector is generated by adding the weighted difference between two population vectors to a third vector at each generation of DE algorithm [13]. Many algorithms are used for numerical benchmark optimization but DE has shown better performance than Genetic Algorithm and Particle Swarm Optimization for such problems [14] [15]. There are

many parameters in DE algorithm like Population ‘NP’, mutation probability ‘F’ and Crossover ‘CR’. The DE algorithm mutation variants are formed by the linear combination of existing population members. The trial vector and target vector forms the mutant vector in DE. Throughout this paper x_i denotes the target vector (or current vector), u_i represents the trial vector and v_i as a mutant vector. In DE algorithm, different mutation schemes are used to create the trial vector by using any combination of current, best and random vectors. The behavior of DE algorithm is influenced by the selection of mutation strategy and crossover scheme along with their control parameters: mutation probability ‘F’ and Crossover rate ‘CR’ [16] [17]. DE mutation strategies can be formed by the combinations of current vector, random vector (s), better vector and best vector. In any mutation strategy the order, number and name of vector(s) are very important. Throughout the following analysis $x_g^{r_k}$ states the r_k th random vector for g th generation, v_g^i will refer to i th component of donor vector at g th generation, x_g^{best} states the best vector at g th generation, x_g^i refer to current vector at g th generation. The most commonly used DE variants given in table-1 have been used by many researchers in their research work [18] [19].

TABLE I. COMMONLY USED DE VARIANTS

S.No	Name	Equation
V ₁	DE/rand/1/bin	$v_g^i = x_g^{r1} + F(x_g^{r2} - x_g^{r3})$
V ₂	DE/best/1/bin	$v_g^i = x_g^{best} + F(x_g^{r1} - x_g^{r2})$
V ₃	DE/rand/2/bin	$v_g^i = x_g^{r1} + F(x_g^{r2} - x_g^{r3}) + F(x_g^{r4} - x_g^{r5})$
V ₄	DE/best/2/bin	$v_g^i = x_g^{best} + F(x_g^{r1} - x_g^{r2}) + F(x_g^{r3} - x_g^{r4})$
V ₅	DE/rand to best/1/bin	$v_g^i = x_g^i + F(x_g^{best} - x_g^i) + F(x_g^{r1} - x_g^{r2})$
V ₆	DE/rand/1/exp	$v_g^i = x_g^{r1} + F(x_g^{r2} - x_g^{r3})$
V ₇	DE/best/1/exp	$v_g^i = x_g^{best} + F(x_g^{r1} - x_g^{r2})$

V ₈	DE/rand/2/ exp	$v_g^i = x_g^{r1} + F(x_g^{r2} - x_g^{r3}) + F(x_g^{r4} - x_g^{r5})$
V ₉	DE/best/2/ exp	$v_g^i = x_g^{best} + F(x_g^{r1} - x_g^{r2}) + F(x_g^{r3} - x_g^{r4})$
V ₁₀	DE/rand to best/1/ exp	$v_g^i = x_g^i + F(x_g^{best} - x_g^i) + F(x_g^{r1} - x_g^{r2})$

Qin et al. [20] have introduce self adaptive version of DE algorithm in their research work. They have control parameter self adaption as well as strategy self adaption in their paper. The memory concept based on finite learning period is used for self adaption. A strategy pool consisting of four commonly used mutation strategies “DE/rand/1/bin”, “DE/rand/2/bin”, “DE/rand-to-best/2/bin” and “DE/current-to-rand/1” is created for strategy self adaption. Zaharie [21] has introduced a parameter adaption scheme in this research that focuses on the choice of parameters by maintaining the diversity in the population. He has used $\lambda \in [0, 1]$ as a coefficient of convex combination between random selected population and the best population member. Variance at component level for each population individual is calculated to maintain diversity in the population and population evolution is controlled by the evolution of variance. Liu and Lampinen [22] has introduced very famous parameter adaption method in their research work. Their parameter adaption scheme is based on the fuzzy controller designed for adaption of F and CR control parameters of DE algorithm. They have used two fuzzy logic controllers (FLC) to implement to adaptive the values of F and CR control parameters based on fuzzy control actions. Brest et al. [23] have introduced the idea of self adaption to choose the suitable values of DE control parameters F and CR. They have evolved both control parameters at individual level so that better individuals may survive and propagate to next generation. The evolution of F and CR is based on two predefined probabilities $\tau_1 = 0.1$, $\tau_2 = 0.1$ respectively. Zhang and Sanderson [24] have proposed adaptive differential evolution with optional external archive (JADE) parameter adaption based algorithm in their research work. They have used new strategy “DE/current to pbest/1” that is based on the conventional “DE/current to best/1” strategy. They generate new population with the union of current population and external archive containing inferior fit population members. They have adopted control parameter F by using Cauchy distribution and CR by using the normal distribution. They have also used two variants of JADE with name rand-JADE and nonaJADE. In rand-JADE instead of DE/current-to-pbest, DE/rand/1 is implemented & variant nona-JADE does not adopt adaptive parameter control instead it set values of both parameters to 0.5. Mallipeddi et al. [25] have introduced an ensemble of control parameter and mutations, crossover strategies in DE algorithm. A set of predefined values of F and CR control parameter values in used to form parameter pool and binomial, exponential versions of “DE/current-to-rand/1/bin” & JADE mutation strategies are used to form a strategy pool in their research work. A trial vector in EPSDE is generated based on assigned mutation strategy and assigned control parameter value from their corresponding pools. Wang, Cai and Zhand [26] have introduced a composite trial vector generation scheme in DE algorithm. They have use three various combination of F and CR $[F= 1.0, Cr= 0.1]$, $[F=$

$1.0, Cr= 0.9]$ and $[F= 0.8, Cr= 0.2]$ that forms a strategy pool in composite DE (CoDE). Strategy pool in CoDE is formed by using four commonly used mutation strategies “rand/1/bin”, “current-to-rand/1” and “rand/2/bin”. A combination of control parameter values pool and a strategy from the strategy pool is selected to generate a trial vector against each target vector in the current population. Gong et al. [27] have introduced enhanced DE in their research work that is based on the strategy adaption mechanism (SaM). Further they ensembled SaM with JADE to form a strategy pool that is based on the convention mutation strategies of DE algorithm. DE/rand-to-pbest” with archive, “DE/current-to-pbest” without archive, “DE/current-to-pbest” with archive, “DE/rand-to-pbest” without archive are used to form a strategy pool in enhanced DE. Das et al. [28] have proposed a neighbourhood base mutation strategy in their research work. They have used family of variants of “DE/target-to-best/1/bin” conventional mutation strategy based on the neighbourhood mechanism. A local and global neighbourhood scheme brings equilibrium in the exploration and exploitation without any extra load of other parameters. To generate a donor vector α and β are used as a scaling factors for local and global neighbor respectively. Their proposed version has significant performance when compared with other DE variants. MinhazulIslam et al. [29] have introduced a novel crossover and mutation strategies in their research work. They have used a new mutation strategy “DE/current-to-gr best/1” that utilizes q% of best population members and is based on the conventional strategy “DE/current-to-best/1”. The new crossover scheme in their research work is pbest crossover that utilizes p-top-ranked individual components. The third important aspect in their research is statistical distribution based adaption in DE control parameters F and CR.

DE mutation strategies are generated by using combination of random and/or best vector(s). In random selection best and worst members have same probability of selection as a parent. The worst parent may lead to worst child. A novel variant of DE is proposed in this research that will be helpful in avoiding the selection of bad performing individuals. The proposed variant will select those individuals having at least some level of fitness that will be helpful in enhancing the convergence speed and searching capability of DE algorithm. It is important to mention that fitter parent does not mean best parent fitter means better than worst performing members. This proposed variant will prove to be a significant addition in DE algorithm. The residue of the paper consist of the DE algorithm that is discussed in section-II, various crossover schemes are given in section-III of the paper, section-IV contains the proposed DE variant, the detail of benchmark functions is given in section-V of this paper and section-VI contains the results and discussion.

II. DE ALGORITHM

DE is a population based algorithm that consists of NP members where each population member is a point in a D-dimensional continuous hyperspace. D-dimensional real-valued parameter vectors in DE algorithm are initialized randomly in search of optimal solution. Vector $x_{i,G}$ represents a population member at Gth generation where $i=1, 2, 3, \dots, NP$,

NP. In DE algorithm various operators like mutation, selection and crossover are used to create offspring population.

Mutation: In mutation a donor vector $v_{i,G+1}$ is generated against each target vector $x_{i,G}$ for i^{th} population member at each g^{th} iteration of DE algorithm. According to conventional DE mutation strategy *DE/rand/1* generates mutant vector $v_{i,G+1}$ by using the following equation

$$v_{i,G+1} = x_{r_1,G} + F(x_{r_2,G} - x_{r_3,G}) \dots\dots\dots(1)$$

Where $x_{r_1,G}$, $x_{r_2,G}$ and $x_{r_3,G}$ are randomly chosen population members such that $i \neq r_1 \neq r_2 \neq r_3$ and F is a difference vector scaling factor.

Crossover: In DE crossover operation the dimensions of target vector $x_{i,G}$ and a donor vector $v_{i,G+1}$ are swapped that is controlled by crossover control parameter CR. Crossover is used to enhance diversity in the DE population members. Binomial and exponential are two main crossover schemes of DE algorithm [16] [30] [31] that differ from each other in the way of number of mutation vector components distribution. Binomial vector generates the trial vector $u_{i,G} = (u_{i,1,g}, u_{i,2,g}, \dots, u_{i,D,g})$ by using following equation

$$u_{i,G} = \begin{cases} v_{i,j,G} & \text{if } (randj(0,1) \leq CR \text{ or } j = j_{rand}) \\ x_{i,j} & \text{otherwise} \end{cases} \dots\dots\dots(2)$$

Where j_{rand} is a random selected dimension from $[1, D]$, $v_{i,j,G}$ is the mutant vector; and $randj(0,1)$ a random value between 0 and 1, Control parameter CR takes value from $(0, 1]$. Whereas Exponential crossover scheme generates the trial vector $u_{i,G}$ by using following equation

$$u_{i,G} = \begin{cases} v_{i,j,G} & \text{for } j = \langle l \rangle_D + \langle l+1 \rangle_D + \dots + \langle l+L-1 \rangle_D \\ x_{i,j} & \text{otherwise} \end{cases} \dots\dots\dots(3)$$

Where modulo function is represented as $\langle \rangle_D$ with modulus D; $i = 1, 2, 3 \dots NP$ and $j = 1, 2, 3 \dots D$. A random number L is generated from $[1, D]$ and to represent a starting index l is chosen randomly from $[1, D]$.

Selection: In selection operation it is decided that whether target vector $x_{i,G}$ target vector or trial vector $u_{i,G}$ survives into next generation. The selection is done by using greedy selection based on the fitness of trial vector and target vector. The equation of selection operation is DE is given as follows

$$x_{i,G+1} = \begin{cases} u_{i,G+1} & \text{if } (f(u_{i,G+1}) < f(x_{i,G})) \\ x_{i,G} & \text{otherwise} \end{cases} \dots\dots\dots(4)$$

Where $f(u_{i,G+1})$ is the fitness value of trial vector $u_{i,G}$ and $f(x_{i,G})$ is the fitness value of target vector $x_{i,G}$

III. PROPOSED FITNESS PROPORTIONATE RANDOM VECTOR SELECTION BASED DE ALGORITHM (FPRVDE)

The Proposed DE variant is based on the fitness proportionate random vector selection method that is used to select members from the parent population to generate the offspring population. Trial vector in FPRVDE is generated by selecting random individuals using fitness proportionate selection criteria. Fitness proportionate selection mechanism allocates the region to each individual according to its fitness value. To apply this method fitness of all individuals in the current population is calculated. Then population members are sorted based on the calculated fitness by maintaining population index of each target vector. After sorting population members, cumulative proportion C of each population member i is calculated. Then we iterate sorted population cumulative proportion that is greater than or equal to a random value r generated in the interval $(0,1)$ to select as a fitness proportionate random vector. Then these fitness proportionate random vectors (*FPRV'S*) are used to form a trial vector against each target vector in the population. The fitness proportionate criteria will ensue that selected population members have some fitness level and are not the bad individuals of the current population. Fitness proportionate selection method helps in ignoring the poor performing population individuals as poor performing candidates may lead towards poor offspring's. The exceptional fit and inferior fit candidates will have equal chance of selection in *FPRVDE* version that will incorporate plenty of diversity in the population. The poor performing population members can generate poor offspring that might increase the convergence speed of DE algorithms while better performing population members will generate healthy offspring that might be helpful in improving the convergence speed of DE algorithm. The proposed variant maintains randomness in selecting the individuals since randomness is the important property in EC algorithm. *FPRVDE* ignores the worst performing members having about no fitness level and for other population member including weak, average and best performing population members have equal chance of selection in forming trial vector. If generated number is close to zero then less fitter population member be selected and if the random value is close to 1 then one of the best performing individual will be selected that incorporates reasonable randomness in *FPRVDE* version.

1. Generate the initial population $P_G = \{X_{1,G}, \dots, X_{NP,G}\}$ for generation $G=0$ and randomly initialize each population member $X_{i,G} = \{x_{i,G}^1, \dots, x_{i,G}^D\}$ where $i = 1, \dots, NP$
2. FOR $i = 1$ to NP
Calculate fitness $f(X_{i,G})$ for each population member $X_{i,G}$
END FOR
3. WHILE the stopping criterion is not true

Step 3.1 Mutation Step

/ Start of FPRVDE vectors selection */*

Step 3.1.1 FPRVDE vectors selection

Sort calculated fitness in ascending order by maintaining the index of each target vector $X_{i,G}$
FOR $i = 1$ to NP
Calculate the cumulative fitness proportion C of each target vector $X_{i,G}$
END FOR

END FOR

FOR $i = 1$ to number of FPRVDE vectors

- Generate a random number r in the interval $(0,1)$
- Select population member i by iterating through sorted population until we reach at C that is greater than or equal to r
- Return i^{th} member index to be used as one of FPRVDE vectors in mutation strategy

END FOR

/ End of FPRVDE vectors selection */*

FOR $i = 1$ to NP

For the i^{th} target vector $X_{i,G}$ generate a donor vector $V_{i,G} = \{v_{i,G}^1, \dots, v_{i,G}^D\}$ with the specified mutation strategy (Any Strategy from table-II) with FPRVDE vectors

END FOR

Step 3.2 Crossover Step

FOR $i = 1$ to NP

For the i^{th} target vector $X_{i,G}$ generate a trial vector $U_{i,G} = \{u_{i,G}^1, \dots, u_{i,G}^D\}$ with the specified crossover scheme (Equation-2 or Equation-3)

END FOR

Step 3.3 Selection Step

FOR $i=1$ to NP

Evaluate the trial vector $U_{i,G}$ against the target vector $X_{i,G}$ with fitness function f

IF $f(U_{i,G}) \leq f(X_{i,G})$, THEN

$$X_{i,G+1} = U_{i,G}, f(X_{i,G}) = f(U_{i,G})$$

IF $f(U_{i,G}) \leq f(X_{best,G})$, THEN

$$X_{best,G+1} = U_{i,G},$$

$$f(X_{best,G}) = f(U_{i,G})$$

END IF

END IF

END FOR

Step # 3.4 increment generation number $G=G+1$

Step 4. END WHILE

Fig. 1. Pseudocode of FPRVDE Algorithm

TABLE II. FPRVDE VERSIONS OF COMMONLY USED DE VARIANTS

S.No	Name	Equation
FPRVDE ₁	FPRVDE /rand/1/bin	$v_g^i = x_g^{fprv1} + F(x_g^{fprv2} - x_g^{fprv3})$
FPRVDE ₂	FPRVDE /best/1/bin	$v_g^i = x_g^{best} + F(x_g^{fprv1} - x_g^{fprv2})$
FPRVDE ₃	FPRVDE /rand/2/bin	$v_g^i = x_g^{fprv1} + F(x_g^{fprv2} - x_g^{fprv3}) + F(x_g^{fprv4} - x_g^{fprv5})$
FPRVDE ₄	FPRVDE /best/2/bin	$v_g^i = x_g^{best} + F(x_g^{fprv1} - x_g^{fprv2}) + F(x_g^{fprv3} - x_g^{fprv4})$
FPRVDE ₅	FPRVDE /rand to best/1/bin	$v_g^i = x_g^i + F(x_g^{best} - x_g^i) + F(x_g^{fprv1} - x_g^{fprv2})$
FPRVDE ₆	FPRVDE /rand/1/exp	$v_g^i = x_g^{fprv1} + F(x_g^{fprv2} - x_g^{fprv3})$
FPRVDE ₇	FPRVDE /best/1/exp	$v_g^i = x_g^{best} + F(x_g^{fprv1} - x_g^{fprv2})$
FPRVDE ₈	FPRVDE /rand/2/exp	$v_g^i = x_g^{fprv1} + F(x_g^{fprv2} - x_g^{fprv3}) + F(x_g^{fprv4} - x_g^{fprv5})$
FPRVDE ₉	FPRVDE /best/2/exp	$v_g^i = x_g^{best} + F(x_g^{fprv1} - x_g^{fprv2}) + F(x_g^{fprv3} - x_g^{fprv4})$
FPRVDE ₁₀	FPRVDE /rand to best/1/exp	$v_g^i = x_g^i + F(x_g^{best} - x_g^i) + F(x_g^{fprv1} - x_g^{fprv2})$

The mutation strategies given in table-II are FPRVDE mutation strategy versions corresponding to each commonly used DE mutation strategy given in table-I. Mutation strategies given in table-II used FPRVDE method to select each random vector in the current population.

IV. TEST FUNCTIONS AND EXPERIMENTAL RESULTS

A comprehensive set of 37 N-dimensional benchmark functions is used to evaluate the performance of proposed FPRVDE and commonly used set of DE variants. These benchmark functions are commonly used multidimensional global optimization problems. The necessary detail of these functions is given in table-III (a, b). The convergence speed of DE algorithm and the proposed variation are measured by using one of the most commonly used performance metric that is number of function calls (NFC) [32]. The acceleration rate is also calculated. Acceleration rate (AR) that is based on NFC is used to compare convergence speed of algorithm. Here AR for DE algorithm and proposed FPRVDE variation is defined by the following formula

$$AR = \frac{NFC_{DE}}{NFC_{FPRVDE}} \dots\dots\dots(5)$$

$AR > 1$ means that FPRVDE is faster than DE, $AR < 1$ means that FPRVDE is slower than DE and $AR = 1$ will shows that DE and FPRVDE have same convergence speed. Further the average acceleration rate and is calculated using equation (6) for the suit of functions used in this research. Average acceleration rate

$$A.A.R = AR_{ave} = \frac{1}{n} \sum_{i=1}^n AR_i \dots\dots\dots(6)$$

Experimental results reported in this section are generated for number of functions calls over a test suit of N-dimensional functions. NFC is the most common performance metric used in evolutionary algorithms [19] [33]. The results are generated using Exponential and binomial crossover schemes over 30 independent trials. NFC experimental results are reported in tables (II-VII) with best values as boldfaces. Convergence graphs of some functions are contained in Figure-2 and Figure-3 for DE and FPRVDE. Commonly used DE variants given in table-I as DE₁-DE₁₀ are corresponding AFPRVDE variants are reported in table-II as FPRVDE₁- FPRVDE₁₀. Experimental results are generated by using control parameter N_p (Population Size) with value 30 and dimensions are taken as 10D, 20D and 30D.. Control parameter F and CR values used are F=0.5, CR=0.9 [16] [30] [34]. Number of function calls value is calculated by taking 30 trials for both DE and FPRVDE for maximum NFC $10^4 * DIM$ [35]. To find out NFC, VTR value is set to 0.0001 and Max-NFC values are 100,000; 200,000 and 300,000 for 10D, 20D and 30D respectively for both DE and FPRVDE in all functions.

TABLE III. (a)TEST BENCHMARK FUNCTION

Function	Name of Function (type)	Equation	Search Space	Optima
f_1	Sphere model (Separable, Multimodal)	$f(x) = \sum_{i=0}^n x_i^2$	$-5.12 \leq x_i \leq 5.12$	0
f_2	Axis parallel hyperellipsoid (Separable, Unimodal)	$f(x) = \sum_{i=0}^n i \cdot x_i^2$	$-5.12 \leq x_i \leq 5.12$	0
f_3	Schwefel's problem 1.2 (Non-Separable, Unimodal)	$f(x) = \sum_{i=0}^n \left(\sum_{j=0}^n x_j \right)^2$	$-65 \leq x_j \leq 65$	0
f_4	Rosenbrock's valley (Non-Separable, Unimodal)	$f(x) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i)^2 + (1 - x_i)^2]$	$-30 \leq x_i \leq 30$	0
f_5	Rastrigin's function (Separable, Multimodal)	$f(x) = 10n + \sum_{i=1}^n (x_i^2 - 10 \cos(2\pi x_i))$	$-5.12 \leq x_i \leq 5.12$	0
f_6	Griewangk's function (Non-Separable, Multimodal)	$f(x) = \sum_{i=1}^n \left(\frac{x_i^2}{4000} - \prod_{i=1}^n \cos\left(\frac{x_i}{\sqrt{i}}\right) \right) + 1$	$-600 \leq x_i \leq 600$	0
f_7	Sum of different power (Non-Separable, Multimodal)	$f(x) = \sum_{i=1}^n x_i ^{(i+1)}$	$-1 \leq x_i \leq 1$	0
f_8	Ackley's path function (Non-Separable, Multimodal)	$f(x) = -20 \exp \left(-0.2 \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}} \right) - \exp \left(\frac{\sum_{i=1}^n \cos(2\pi x_i)}{n} \right) + 20 + e$	$-32 \leq x_i \leq 32$	0
f_9	Levy function (Separable, Multimodal)	$0.1 \left[\sin^2(3\pi x_1) + \sum_{i=1}^{n-1} (x_i - 1)^2 \times (1 + \sin^2(3\pi x_{i+1})) + (x_n - 1)(1 + \sin^2(2\pi x_n)) \right]$	$-10 \leq x_i \leq 10$	0
f_{10}	Zakharov function (Non-Separable, Multimodal)	$f(x) = \sum_{i=1}^n x_i^2 + \left(\sum_{i=1}^n 0.5ix_i \right)^2 + \left(\sum_{i=1}^n 0.5ix_i \right)^4$	$-5 \leq x_i \leq 10$	0
f_{11}	Schwefel's problem 2.22 (Non-Separable, Unimodal)	$f(x) = \sum_{i=1}^n x_i + \prod_{i=1}^n x_i $	$-10 \leq x_i \leq 10$	0

f_{12}	Step function (Separable, Unimodal)	$f(x) = \sum_{i=1}^n (\lfloor x_i + 0.5 \rfloor)^2$	$-100 \leq x_i \leq 100$	0
f_{13}	Quartic function, i.e., noise (Separable, Unimodal)	$f(x) = \sum_{i=1}^n ix_i^4 + \text{random}[0,1)$	$-1.28 \leq x_i \leq 1.28$	0
f_{14}	De Jong's function 4 (no noise) (Separable, Unimodal)	$f(x) = \sum_{i=1}^n ix_i^4$	$-1.28 \leq x_i \leq 1.28$	0
f_{15}	Alpine function (Separable, Multimodal)	$f(x) = \sum_{i=1}^n x_i \sin(x_i) + 0.1x_i $	$-10 \leq x_i \leq 10$	0
f_{16}	Pathological function (Non-Separable, Multimodal)	$f(x) = \sum_{i=1}^{n-1} \left(0.5 + \frac{\sin^2 \sqrt{(100x_i^2 + x_{i+1}^2)} - 0.5}{1 + 0.001(x_i^2 - 2x_ix_{i+1} + x_{i+1}^2)} \right)$	$-100 \leq x_i \leq 100$	0
f_{17}	Inverted cosine wave function (Non-Separable, Multimodal)	$f(x) = -\sum_{i=1}^{n-1} \left(\exp\left(\frac{-(x_i^2 + x_{i+1}^2 + 0.5x_ix_{i+1})}{8}\right) \times \cos\left(4\sqrt{x_i^2 + x_{i+1}^2 + 0.5x_ix_{i+1}}\right) \right)$	$-5 \leq x_i \leq 5$	$-n+1$

TABLE III. (b)TEST BENCHMARK FUNCTION

f_{18}	Exponential problem (Non-Separable, Multimodal)	$f(x) = -\exp\left(-0.5 \sum_{i=1}^n x_i^2\right)$	$-1 \leq x_i \leq 1$	-1
f_{19}	Levy and Montalvo Problem (Separable, Multimodal)	$f(x) = \left(\frac{\pi}{n}\right) \left(10 \sin^2(\pi y_1) + \sum_{i=1}^{n-1} (y_i - 1)^2 [1 + 10 \sin 2(\pi y_{i+1})] \right) + (y_n - 1)^2$ where $y_i = 1 + \frac{1}{4}(x_i + 1)$	$-10 \leq x_i \leq 10$	0
f_{20}	Neumaier 2 Problem (Separable, Unimodal)	$f(x) = \sum_{i=1}^n (x_i - 1)^2 - \sum_{i=2}^n (x_i x_{i-1})$	$-n^2 \leq x_i \leq n^2$	0
f_{21}	Salomon Problem (Non-Separable, Multimodal)	$f(x) = 1 - \cos(2\pi \ x\) + 0.1 \ x\ $ where $\ x\ = \sqrt{\sum_{i=1}^n x_i^2}$	$-100 \leq x_i \leq 100$	0
f_{22}	Cosine Mixture (Separable, Multimodal)	$f(x) = -0.1 \sum_{i=1}^n \cos(5\pi x_i) + \sum_{i=1}^n x_i^2$	$-1 \leq x_i \leq 1$	$-0.1x(n)$
f_{23}	Cigar (Separable, Multimodal)	$f(x) = x_1^2 + 100000 \sum_{i=1}^n x_i^2$	$-10 \leq x_i \leq 10$	0
f_{24}	Function '15' (Separable, Multimodal)	$f(x) = \sum_{i=1}^{n-1} [0.2x_i^2 + 0.1x_i^2 \sin(2x_i)]$	$-10 \leq x_i \leq 10$	0
f_{25}	Dixon Price (Non-Separable, Unimodal)	$f(x) = (x_1 - 1)^2 + \sum_{i=2}^n i.(2x_i^2 - x_{i-1})^2$	$-10 \leq x_i \leq 10$	0
f_{26}	Ellipse Function (Separable, Unimodal)	$f(x) = \sum_{i=1}^n (10^{6(\frac{i-1}{n-1})} . x_i^2)$	$-100 \leq x_i \leq 100$	0
f_{27}	Tablet Function (Separable, Unimodal)	$f(x) = 10^4 x_1^2 + \sum_{i=2}^n x_i^2$	$-100 \leq x_i \leq 100$	0
f_{28}	Schewel (Separable, Multimodal)	$f(x) = \sum_{i=1}^n ((x_i - x_i^2)^2 + (x_i - 1)^2)$	$-32 \leq x_i \leq 32$	0
f_{29}	Deflected Corrugated Spring (Separable, Multimodal)	$f(x) = 0.1 \sum_{i=1}^n \left((x_i - \alpha)^2 - \cos\left(K \sqrt{\sum_{i=1}^n ((x_i - \alpha)^2)}\right) \right)$	$0 \leq x_i \leq 10$ $K=5$ $\alpha=5$ $x_i \in [0, 2\alpha]$	0
f_{30}	Mishra 1 global optimization problem (Non-Separable, Multimodal)	$f(x) = (1 + x_n)^{x_n}$ where $x_n = n - \sum_{i=1}^{n-1} x_i$	$0 \leq x_i \leq 1$	2
f_{31}	Mishra 2 global optimization problem (Non-Separable, Multimodal)	$f(x) = (1 + x_n)^{x_n}$ where $x_n = n - \sum_{i=1}^{n-1} \frac{(x_i + x_{i+1})}{2}$	$0 \leq x_i \leq 1$	2

f_{32}	MultiModal global optimization problem (Separable, Multimodal)	$f(x) = \left(\sum_{i=1}^n x_i \right) \left(\prod_{i=1}^n x_i \right)$	$-10 \leq x_i \leq 10$	0
f_{33}	Plateau global optimization problem (Separable, Multimodal)	$f(x) = 30 + \sum_{i=1}^n \lfloor x_i \rfloor$	$-5.12 \leq x_i \leq 5.12$	30
f_{34}	Quintic global optimization problem (Separable, Multimodal)	$f(x) = \sum_{i=1}^n x_i^5 - 3x_i^4 + 4x_i^3 + 2x_i^2 - 10x_i - 4 $	$-10 \leq x_i \leq 10$	-1
f_{35}	Stochastic global optimization problem (Separable, Multimodal)	$f(x) = \sum_{i=1}^n \varepsilon_i \left x_i - \frac{1}{i} \right $	$-5 \leq x_i \leq 5$	0
f_{36}	Stretched V global optimization problem (Non-Separable, Multimodal)	$f(x) = \sum_{i=1}^{n-1} t^{1/4} \left[\sin(50t^{0.1}) + 1 \right]^2$ where $t = x_{i+1}^2 + x_i^2$	$-10 \leq x_i \leq 10$	0
f_{37}	XinSheYang (Non-Separable, Multimodal)	$f(x) = \frac{\left(\sum_{i=1}^n x_i \right)}{e^{\sum_{i=1}^n \sin(x_i^2)}}$	$-2\pi \leq x_i \leq 2\pi$	0

TABLE IV. 10D-NUMBER OF FUNCTIONS CALLS (NFC) AND ACCELERATION RATE (AR) OF DE AND FPRVDE VARIANTS

Function	rand/1/bin			best/1/bin			rand-to-best/1/bin			rand/2/bin			best/2/bin		
	DE ₁	FPRVDE ₁	A.R	DE ₂	FPRVDE ₂	A.R	DE ₃	FPRVDE ₃	A.R	DE ₄	FPRVDE ₄	A.R	DE ₅	FPRVDE ₅	A.R
f_1	217.03	177.20	1.22	104.83	91.57	1.14	439.70	344.13	1.28	273.07	224.27	1.22	102.70	88.10	1.17
f_2	206.97	169.67	1.22	103.97	89.23	1.17	400.77	321.87	1.25	254.87	207.20	1.23	103.07	87.47	1.18
f_3	1284.50	896.50	1.43	454.63	344.73	1.32	2973.07	2083.50	1.43	1668.87	1172.60	1.42	422.20	297.93	1.42
f_4	3078.30	2895.93	1.06	1305.44	1188.86	1.10	6460.33	4813.97	1.34	3665.60	2736.50	1.34	2827.29	1272.43	2.22
f_5	1427.69	605.80	2.36	-	-	-	5974.03	2277.79	2.62	2326.00	-	-	19333.00	13036.60	1.48
f_6	1955.07	693.13	2.82	356.50	-	-	12929.19	4091.82	3.16	1988.00	-	-	2964.86	1111.50	2.67
f_7	113.00	89.30	1.27	57.43	47.00	1.22	228.33	174.20	1.31	144.93	118.23	1.23	55.93	48.47	1.15
f_8	1928.69	573.50	3.36	-	-	-	7251.46	2819.07	2.57	3313.00	889.50	3.72	23812.33	18253.50	1.30
f_9	460.03	250.80	1.83	227.80	131.97	1.73	908.70	632.73	1.44	600.93	469.13	1.28	224.23	156.23	1.44
f_{10}	925.10	663.47	1.39	363.03	258.18	1.41	2052.00	1449.60	1.42	1167.30	826.13	1.41	338.43	226.40	1.49
f_{11}	397.23	320.23	1.24	195.30	200.89	0.97	819.43	639.13	1.28	520.43	398.20	1.31	200.33	164.83	1.22
f_{12}	148.43	106.93	1.39	71.17	57.24	1.24	304.10	196.30	1.55	187.43	131.10	1.43	69.17	57.32	1.21
f_{13}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f_{14}	93.43	71.70	1.30	45.60	34.83	1.31	177.90	130.87	1.36	113.30	83.70	1.35	45.50	34.43	1.32
f_{15}	18182.10	5376.40	3.38	16026.50	-	-	30839.63	15202.65	2.03	23863.37	8666.27	2.75	19874.47	17748.30	1.12
f_{16}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f_{17}	3766.47	889.50	4.23	490.00	-	-	21796.43	6421.33	3.39	10837.17	1260.00	8.60	9887.10	14297.04	0.69
f_{18}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f_{19}	221.87	169.17	1.31	101.83	84.57	1.20	478.30	363.40	1.32	290.40	221.63	1.31	111.17	92.73	1.20
f_{20}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f_{21}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f_{22}	219.20	177.23	1.24	105.93	90.39	1.17	481.80	364.67	1.32	292.33	232.43	1.26	105.23	89.73	1.17
f_{23}	407.60	330.70	1.23	193.27	169.20	1.14	844.73	651.33	1.30	511.90	416.37	1.23	189.97	160.73	1.18
f_{24}	216.53	173.00	1.25	103.60	88.90	1.17	439.60	342.00	1.29	268.57	221.20	1.21	100.80	84.77	1.19
f_{25}	-	-	-	203.00	-	-	-	-	-	-	-	-	234.00	15373.25	-
f_{26}	306.97	248.80	1.23	145.20	126.60	1.15	633.60	488.97	1.30	383.97	312.57	1.23	143.57	122.17	1.18
f_{27}	320.83	258.69	1.24	154.23	131.69	1.17	654.10	507.77	1.29	406.87	316.04	1.29	154.27	126.63	1.22
f_{28}	383.97	301.90	1.27	169.70	137.45	1.23	829.17	617.83	1.34	492.37	374.27	1.32	164.53	141.30	1.16
f_{29}	1537.37	849.27	1.81	1031.77	482.54	2.14	2632.30	2113.17	1.25	2133.60	1394.90	1.53	950.33	725.07	1.31
f_{30}	194.40	192.14	1.01	85.70	82.86	1.03	282.70	267.43	1.06	143.53	131.00	1.10	114.27	109.27	1.05
f_{31}	211.20	208.17	1.01	91.33	88.63	1.03	319.07	298.23	1.07	156.93	144.27	1.09	121.73	120.33	1.01
f_{32}	41.47	19.53	2.12	26.47	12.47	2.12	88.40	21.80	4.06	64.20	11.63	5.52	28.43	12.47	2.28
f_{33}	5.40	5.27	1.03	3.20	2.77	1.16	6.40	5.60	1.14	3.77	3.23	1.16	4.03	3.47	1.16
f_{34}	773.57	548.87	1.41	262.67	204.83	1.28	3417.97	1687.70	2.03	1008.60	644.93	1.56	624.57	1336.86	0.47
f_{35}	2017.40	902.47	2.24	400.50	453.00	0.88	-	-	-	24284.62	2791.10	8.70	413.83	445.68	0.93
f_{36}	31.60	23.20	1.36	26.00	6.14	4.23	46.97	31.00	1.52	38.13	22.95	1.66	32.70	25.53	1.28
f_{37}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

TABLE V. 10D-NUMBER OF FUNCTIONS CALLS (NFC) AND ACCELERATION RATE (AR) OF DE AND FPRVDE VARIANTS

Function	rand/1/exp			best/1/exp			rand-to-best/1/exp			rand/2/exp			best/2/exp		
	DE ₆	FPRVDE ₆	A.R	DE ₇	FPRVDE ₇	A.R	DE ₈	FPRVDE ₈	A.R	DE ₉	FPRVDE ₉	A.R	DE ₁₀	FPRVDE ₁₀	A.R
f ₁	195.63	184.57	1.06	140.73	134.20	1.05	266.77	247.83	1.08	216.20	207.03	1.04	145.77	136.97	1.06
f ₂	199.53	186.67	1.07	144.00	131.00	1.10	272.13	257.50	1.06	226.00	209.77	1.08	152.47	137.90	1.11
f ₃	1450.60	1248.77	1.16	808.47	663.65	1.22	2261.67	1968.10	1.15	1657.70	1412.10	1.17	765.83	594.90	1.29
f ₄	4457.47	4462.95	1.00	2813.57	1804.00	1.56	12168.87	6357.00	1.91	6153.20	3775.80	1.63	4416.00	2824.20	1.56
f ₅	389.14	345.07	1.13	261.53	225.00	1.16	587.17	537.03	1.09	469.00	426.00	1.10	1689.00	1597.00	1.06
f ₆	806.63	611.93	1.32	457.50	-	-	1736.17	1367.93	1.27	1137.37	818.40	1.39	904.08	1052.44	0.86
f ₇	122.60	115.30	1.06	88.47	75.30	1.17	163.13	146.03	1.12	134.87	122.10	1.10	95.93	83.83	1.14
f ₈	516.76	448.08	1.15	343.58	-	-	766.59	722.63	1.06	609.00	554.43	1.10	2055.00	2150.50	0.96
f ₉	140.03	118.97	1.18	108.77	98.97	1.10	196.67	164.10	1.20	160.57	155.50	1.03	117.97	108.43	1.09
f ₁₀	2040.73	1741.89	1.17	1163.27	858.67	1.35	3158.57	2738.23	1.15	2276.43	1694.73	1.34	1170.00	1044.23	1.12
f ₁₁	362.70	337.37	1.08	264.17	293.80	0.90	499.93	463.20	1.08	411.03	376.46	1.09	279.17	258.57	1.08
f ₁₂	129.07	118.70	1.09	95.47	87.37	1.09	174.27	161.40	1.08	147.87	130.77	1.13	99.50	88.10	1.13
f ₁₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₄	82.17	70.30	1.17	61.20	52.37	1.17	108.60	93.60	1.16	90.10	75.10	1.20	61.73	51.03	1.21
f ₁₅	8878.67	4080.68	2.18	6114.27	1666.29	3.67	7031.57	7788.83	0.90	7241.00	5948.28	1.22	6573.27	8645.10	0.76
f ₁₆	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₇	1026.23	804.15	1.28	501.27	416.50	1.20	2546.67	1811.28	1.41	1576.90	1072.00	1.47	2044.37	1781.07	1.15
f ₁₈	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₉	181.13	166.03	1.09	129.17	118.93	1.09	248.47	233.57	1.06	208.60	193.67	1.08	148.33	126.54	1.17
f ₂₀	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₂₁	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₂₂	191.27	178.67	1.07	136.53	128.17	1.07	268.03	253.20	1.06	221.50	207.23	1.07	147.93	137.60	1.08
f ₂₃	360.00	338.60	1.06	265.10	241.83	1.10	496.73	465.80	1.07	412.33	383.87	1.07	271.40	249.63	1.09
f ₂₄	191.37	180.23	1.06	138.43	130.93	1.06	259.43	247.63	1.05	217.30	200.87	1.08	145.07	134.30	1.08
f ₂₅	4753.55	3579.67	-	443.83	295.00	-	13984.04	8131.43	-	4080.56	2929.33	-	659.00	1738.25	-
f ₂₆	274.80	255.53	1.08	197.47	181.57	1.09	368.93	351.57	1.05	304.13	292.03	1.04	205.93	187.43	1.10
f ₂₇	293.07	290.00	1.01	211.70	191.80	1.10	405.17	356.03	1.14	329.60	286.18	1.15	221.27	190.43	1.16
f ₂₈	437.73	396.00	1.11	272.67	235.76	1.16	646.23	566.73	1.14	495.80	418.31	1.19	281.37	244.57	1.15
f ₂₉	248.10	213.90	1.16	202.67	206.17	0.98	352.57	305.67	1.15	240.20	272.57	0.88	174.40	198.70	0.88
f ₃₀	237.83	241.70	0.98	160.40	156.43	1.03	302.57	293.17	1.03	221.43	213.53	1.04	214.33	209.23	1.02
f ₃₁	250.33	252.23	0.99	170.90	166.18	1.03	314.57	312.30	1.01	237.07	225.87	1.05	224.83	222.73	1.01
f ₃₂	39.00	26.90	1.45	32.60	18.80	1.73	49.33	28.63	1.72	44.00	21.03	2.09	33.80	22.33	1.51
f ₃₃	6.73	5.93	1.13	4.90	4.70	1.04	7.30	6.67	1.10	5.47	5.27	1.04	5.93	5.50	1.08
f ₃₄	595.07	491.72	1.21	354.93	305.33	1.16	1101.90	873.30	1.26	660.27	567.66	1.16	985.30	1649.53	0.60
f ₃₅	5082.63	2678.23	1.90	819.12	713.54	1.15	-	-	-	-	9189.07	-	889.03	798.93	1.11
f ₃₆	102.37	57.96	1.77	72.37	26.25	2.76	61.07	82.25	0.74	88.93	45.00	1.98	90.60	60.93	1.49
f ₃₇	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

TABLE VI. 20D-NUMBER OF FUNCTIONS CALLS (NFC) AND ACCELERATION RATE (AR) OF DE AND FPRVDE VARIANTS

Function	rand/1/bin			best/1/bin			rand-to-best/1/bin			rand/2/bin			best/2/bin		
	DE ₁	FPRVDE ₁	A.R	DE ₂	FPRVDE ₂	A.R	DE ₃	FPRVDE ₃	A.R	DE ₄	FPRVDE ₄	A.R	DE ₅	FPRVDE ₅	A.R
f ₁	621.73	377.93	1.65	217.83	183.33	1.19	2189.40	1095.57	2.00	1042.77	654.30	1.59	201.07	159.97	1.26
f ₂	632.63	407.80	1.55	235.27	199.77	1.18	2103.73	1133.80	1.86	1031.43	679.50	1.52	212.30	175.70	1.21
f ₃	19967.33	6367.70	3.14	3036.20	1869.57	1.62	105853.73	31443.03	3.37	37375.10	14276.20	2.62	2350.90	1338.07	1.76
f ₄	9581.83	7559.90	1.27	3168.92	2730.91	1.16	37431.93	17153.43	2.18	16060.03	9343.85	1.72	3243.67	3148.86	1.03
f ₅	20874.18	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₆	1323.57	632.20	2.09	330.46	270.11	1.22	11182.82	2834.08	3.95	2365.38	1143.36	2.07	519.19	536.38	0.97
f ₇	261.57	132.67	1.97	94.33	66.82	1.41	883.63	383.20	2.31	438.83	240.27	1.83	80.20	63.90	1.26
f ₈	26503.47	-	-	-	-	-	-	-	-	120792.00	-	-	-	-	-
f ₉	590.90	413.20	1.43	248.53	228.27	1.09	2223.93	1000.23	2.22	1070.60	614.67	1.74	312.53	187.63	1.67
f ₁₀	7023.60	3550.40	1.98	1922.50	1421.86	1.35	26725.13	12014.97	2.22	11749.37	6397.53	1.84	1611.03	1028.97	1.57
f ₁₁	1093.77	662.45	1.65	403.03	-	-	3791.53	1968.19	1.93	1916.67	1123.67	1.71	388.37	293.23	1.32
f ₁₂	416.40	217.83	1.91	151.25	106.33	1.42	1494.87	558.60	2.68	739.93	349.64	2.12	144.47	96.38	1.50
f ₁₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₄	342.57	192.37	1.78	118.47	96.43	1.23	1188.87	536.67	2.22	557.57	325.80	1.71	104.77	83.10	1.26
f ₁₅	31342.53	27903.97	1.12	20916.47	12814.90	1.63	35370.63	29618.83	1.19	48597.90	34810.60	1.40	30609.34	13474.37	2.27
f ₁₆	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₇	110766.40	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₈	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₉	629.50	356.67	1.76	208.14	164.64	1.26	2488.83	1107.17	2.25	1158.63	653.80	1.77	202.43	170.22	1.19
f ₂₀	70111.12	82374.67	-	46788.86	40745.32	-	66372.56	69646.32	-	84393.60	84237.05	-	42102.07	33848.90	-
f ₂₁	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₂₂	644.30	376.87	1.71	219.83	176.73	1.24	2518.40	1180.20	2.13	1171.70	690.83	1.70	211.59	167.62	1.26
f ₂₃	1114.40	692.57	1.61	393.40	329.90	1.19	3969.23	2029.77	1.96	1890.80	1188.63	1.59	357.00	293.37	1.22
f ₂₄	606.67	373.00	1.63	216.03	178.10	1.21	2117.20	1085.00	1.95	1028.23	643.60	1.60	194.13	158.27	1.23
f ₂₅	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₂₆	841.20	527.07	1.60	298.67	249.77	1.20	3000.63	1539.57	1.95	1428.30	900.00	1.59	270.07	220.90	1.22
f ₂₇	862.27	535.55	1.61	309.83	251.53	1.23	3020.43	1544.30	1.96	1458.93	893.54	1.63	280.20	226.37	1.24
f ₂₈	1094.80	726.00	1.51	364.67	314.93	1.16	4289.73	1937.90	2.21	1927.87	1084.20	1.78	340.40	343.23	0.99
f ₂₉	2406.07	2429.43	0.99	1968.70	1011.03	1.95	8503.43	4641.87	1.83	5674.47	2957.97	1.92	1643.43	1320.17	1.24
f ₃₀	403.57	386.44	1.04	159.87	155.08	1.03	769.40	663.36	1.16	332.50	284.81	1.17	211.13	208.57	1.01
f ₃₁	426.53	404.00	1.06	166.97	161.53	1.03	839.37	702.86	1.19	347.87	300.78	1.16	220.23	216.17	1.02
f ₃₂	84.40	24.03	3.51	40.47	8.39	4.82	209.33	24.90	8.41	139.43	9.93	14.04	53.07	11.90	4.46
f ₃₃	2.17	2.00	1.08	1.90	1.70	1.12	2.10	2.20	0.95	2.43	1.53	1.59	1.77	1.80	0.98
f ₃₄	3095.53	1101.33	2.81	538.87	403.82	1.33	42457.20	5154.20	8.24	5264.00	1862.53	2.83	893.77	539.40	1.66
f ₃₅	-	4099.13	-	921.00	-	-	-	-	-	-	-	-	1618.00	-	-
f ₃₆	20.10	27.40	0.73	33.37	11.07	3.02	34.30	27.41	1.25	30.90	18.85	1.64	31.37	32.27	0.97
f ₃₇	108.13	20.43	5.29	74.80	8.50	8.80	99.37	50.00	1.99	146.70	12.60	11.64	93.50	62.50	1.50

TABLE VII. 20D-NUMBER OF FUNCTIONS CALLS (NFC) AND ACCELERATION RATE (AR) OF DE AND FPRVDE VARIANTS

Function	rand/1/exp			best/1/exp			rand-to-best/1/exp			rand/2/exp			best/2/exp		
	DE ₆	FPRVDE ₆	A.R	DE ₇	FPRVDE ₇	A.R	DE ₈	FPRVDE ₈	A.R	DE ₉	FPRVDE ₉	A.R	DE ₁₀	FPRVDE ₁₀	A.R
f ₁	422.57	397.20	1.06	331.80	318.97	1.04	576.47	546.63	1.05	498.23	477.47	1.04	332.07	318.37	1.04
f ₂	466.00	432.90	1.08	364.50	345.47	1.06	633.27	596.87	1.06	547.50	524.37	1.04	369.37	347.60	1.06
f ₃	6577.87	5523.23	1.19	3841.90	3335.84	1.15	10667.17	9143.73	1.17	7775.13	6758.10	1.15	3347.33	2803.83	1.19
f ₄	22255.50	20637.14	1.08	18310.23	16057.75	1.14	36948.63	31304.63	1.18	54445.10	52531.39	1.04	19624.30	9034.67	2.17
f ₅	808.17	717.82	1.13	629.45	567.00	1.11	1238.53	1131.90	1.09	1046.12	986.28	1.06	-	-	-
f ₆	1048.63	863.56	1.21	598.09	540.42	1.11	1987.50	1543.14	1.29	1344.82	954.78	1.41	807.17	712.77	1.13
f ₇	237.93	226.77	1.05	170.77	141.72	1.20	316.47	290.40	1.09	260.57	238.67	1.09	184.03	164.13	1.12
f ₈	1052.27	906.88	1.16	798.50	-	-	1573.57	1479.97	1.06	1333.45	1221.15	1.09	-	-	-
f ₉	232.47	214.07	1.09	191.37	182.07	1.05	319.90	298.07	1.07	278.43	259.77	1.07	192.20	184.43	1.04
f ₁₀	53471.30	72995.09	0.73	48702.13	-	-	66895.20	70378.48	0.95	59634.30	54576.00	1.09	47413.70	64533.50	0.73
f ₁₁	763.83	708.14	1.08	607.00	567.00	1.07	1057.77	976.50	1.08	914.37	844.47	1.08	621.53	585.00	1.06
f ₁₂	279.33	256.10	1.09	223.57	210.77	1.06	377.53	343.90	1.10	334.47	311.57	1.07	225.00	210.50	1.07
f ₁₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₄	204.50	178.23	1.15	162.13	150.63	1.08	273.47	242.47	1.13	242.90	212.53	1.14	160.40	140.33	1.14
f ₁₅	5473.67	4826.07	1.13	7207.83	8564.00	0.84	5199.23	8918.20	0.58	4477.77	5936.53	0.75	7005.37	5273.17	1.33
f ₁₆	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₇	2373.00	-	-	1286.42	-	-	5219.20	-	-	3647.53	-	-	7633.33	-	-
f ₁₈	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₁₉	365.27	335.53	1.09	291.53	270.13	1.08	504.70	469.83	1.07	434.03	408.70	1.06	314.13	273.89	1.15
f ₂₀	5309.73	6038.67	-	4429.33	8324.33	-	7208.03	7188.70	-	7141.77	6977.07	-	6038.87	3666.87	-
f ₂₁	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₂₂	412.00	386.07	1.07	326.30	312.60	1.04	571.30	543.03	1.05	499.27	472.70	1.06	338.67	318.30	1.06
f ₂₃	747.50	711.13	1.05	593.77	567.20	1.05	1036.50	977.50	1.06	891.60	847.07	1.05	596.17	566.10	1.05
f ₂₄	412.27	389.80	1.06	326.70	312.97	1.04	564.40	539.97	1.05	491.03	465.33	1.06	327.40	307.23	1.07
f ₂₅	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
f ₂₆	575.77	541.83	1.06	450.43	434.10	1.04	791.07	746.73	1.06	678.53	651.40	1.04	455.30	433.17	1.05
f ₂₇	593.97	588.81	1.01	471.80	434.90	1.08	820.90	738.20	1.11	709.23	648.07	1.09	477.10	429.67	1.11
f ₂₈	1289.80	1243.29	1.04	727.70	626.69	1.16	2204.03	2060.70	1.07	1595.73	1295.64	1.23	741.50	662.37	1.12
f ₂₉	291.67	240.73	1.21	208.37	215.87	0.97	396.60	312.10	1.27	355.60	285.80	1.24	229.53	187.27	1.23
f ₃₀	524.37	524.61	1.00	405.23	402.53	1.01	666.47	653.39	1.02	541.47	514.61	1.05	524.57	528.87	0.99
f ₃₁	536.57	542.73	0.99	415.37	408.18	1.02	680.97	661.90	1.03	555.87	531.04	1.05	540.00	542.50	1.00
f ₃₂	69.87	45.13	1.55	52.73	29.80	1.77	75.00	43.67	1.72	73.07	28.93	2.53	56.03	33.57	1.67
f ₃₃	3.23	2.13	1.52	2.13	1.93	1.10	1.93	2.03	0.95	2.23	2.47	0.91	2.63	2.07	1.27
f ₃₄	1262.93	1087.59	1.16	818.10	722.00	1.13	2230.93	1848.17	1.21	1493.03	1317.50	1.13	2706.10	2832.20	0.96
f ₃₅	-	837.33	-	3407.56	2860.25	1.19	-	-	-	-	-	-	4525.07	5111.53	0.89
f ₃₆	205.87	131.76	1.56	217.87	33.45	6.51	147.87	152.07	0.97	138.63	116.21	1.19	197.33	111.23	1.77
f ₃₇	11.23	8.20	1.37	10.57	3.27	3.23	14.13	7.57	1.87	10.83	4.27	2.54	10.87	8.30	1.31

TABLE VIII. 30D-NUMBER OF FUNCTIONS CALLS (NFC) AND ACCELERATION RATE (AR) OF DE AND FPRVDE VARIANTS

Function	rand/1/bin			best/1/bin			rand-to-best/1/bin			rand/2/bin			best/2/bin		
	DE ₁	FPRVDE ₁	A.R	DE ₂	FPRVDE ₂	A.R	DE ₃	FPRVDE ₃	A.R	DE ₄	FPRVDE ₄	A.R	DE ₅	FPRVDE ₅	A.R
<i>f</i> ₁	1280.50	573.77	2.23	341.60	279.37	1.22	7559.47	2118.40	3.57	2631.80	1209.93	2.18	291.17	237.93	1.22
<i>f</i> ₂	1361.63	640.70	2.13	378.13	309.50	1.22	7553.47	2294.87	3.29	2761.40	1326.23	2.08	325.77	264.37	1.23
<i>f</i> ₃	181964.47	20298.43	8.96	9644.17	5503.69	1.75	-	140235.53	-	-	61517.93	-	6959.47	3422.93	2.03
<i>f</i> ₄	21064.43	13081.17	1.61	5721.52	4751.62	1.20	154358.63	34695.17	4.45	42251.63	18119.76	2.33	4869.52	4828.53	1.01
<i>f</i> ₅	152395.82	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₆	2089.50	828.86	2.52	481.21	391.60	1.23	19376.46	3412.00	5.68	4366.75	1782.19	2.45	520.71	345.57	1.51
<i>f</i> ₇	519.67	160.53	3.24	124.47	90.50	1.38	2704.80	546.63	4.95	1005.13	364.30	2.76	103.00	72.30	1.42
<i>f</i> ₈	196031.86	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₉	993.10	520.27	1.91	306.21	235.78	1.30	5285.90	1567.47	3.37	2164.73	1072.07	2.02	297.67	197.20	1.51
<i>f</i> ₁₀	25485.97	8829.60	2.89	5023.00	3570.09	1.41	165938.20	37317.07	4.45	51840.63	19882.03	2.61	3920.97	2483.30	1.58
<i>f</i> ₁₁	2217.33	981.00	2.26	615.20	379.07	1.62	12663.63	3693.00	3.43	4731.63	1944.93	2.43	568.30	402.70	1.41
<i>f</i> ₁₂	873.60	317.77	2.75	235.33	-	-	5559.50	1012.47	5.49	1925.33	626.35	3.07	207.00	-	-
<i>f</i> ₁₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₁₄	831.00	315.40	2.63	200.23	167.00	1.20	5180.97	1129.93	4.59	1699.23	684.27	2.48	171.50	147.33	1.16
<i>f</i> ₁₅	38066.27	30367.10	1.25	18757.17	15014.47	1.25	45907.17	44991.87	1.02	56486.73	40908.67	1.38	30295.93	26086.03	1.16
<i>f</i> ₁₆	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₁₇	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₁₈	770.03	146.73	-	33.17	26.90	-	3933.87	804.00	-	575.70	152.27	-	65.83	55.83	-
<i>f</i> ₁₉	1334.93	520.17	2.57	321.34	255.84	1.26	9707.03	2109.23	4.60	3025.37	1221.10	2.48	293.00	263.24	1.11
<i>f</i> ₂₀	113110.00	125993.48	-	100333.25	96550.33	-	121766.11	153223.33	-	104572.71	120941.94	-	105852.37	61268.83	-
<i>f</i> ₂₁	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₂₂	1375.87	575.20	2.39	351.93	274.50	1.28	9378.57	2254.67	4.16	3062.33	1266.47	2.42	311.77	238.42	1.31
<i>f</i> ₂₃	2312.27	1027.77	2.25	602.13	493.23	1.22	13693.33	3813.50	3.59	4765.67	2161.47	2.20	521.67	417.93	1.25
<i>f</i> ₂₄	1252.67	559.70	2.24	332.97	273.33	1.22	7435.80	2092.00	3.55	2592.93	1199.77	2.16	289.33	234.97	1.23
<i>f</i> ₂₅	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₂₆	1735.97	781.40	2.22	460.63	380.93	1.21	10468.83	2874.80	3.64	3585.30	1667.67	2.15	400.03	325.77	1.23
<i>f</i> ₂₇	1767.03	789.43	2.24	475.27	384.70	1.24	10449.20	2889.17	3.62	3638.47	1675.37	2.17	410.67	327.63	1.25
<i>f</i> ₂₈	2370.10	1207.93	1.96	606.70	564.70	1.07	15999.60	3732.33	4.29	4986.23	2046.57	2.44	580.53	577.83	1.00
<i>f</i> ₂₉	3831.17	3826.07	1.00	2841.30	1220.33	2.33	26153.20	7957.40	3.29	11228.73	7069.67	1.59	2530.53	1245.53	2.03
<i>f</i> ₃₀	639.27	583.45	1.10	226.73	-	-	1572.60	1173.90	1.34	574.27	-	-	297.10	-	-
<i>f</i> ₃₁	666.53	607.88	1.10	230.70	-	-	1649.27	1246.65	1.32	601.47	8271.75	0.07	302.87	-	-
<i>f</i> ₃₂	146.47	30.00	4.88	53.17	9.93	5.35	557.23	28.47	19.57	344.60	3.57	96.62	53.70	10.27	5.23
<i>f</i> ₃₃	1.73	2.23	0.78	1.57	1.57	1.00	1.77	1.80	0.98	2.10	1.77	1.19	1.60	1.53	1.04
<i>f</i> ₃₄	7678.30	1600.07	4.80	809.90	600.46	1.35	270693.67	8136.90	33.27	15289.67	3204.57	4.77	1079.87	604.57	1.79
<i>f</i> ₃₅	-	10549.60	-	-	-	-	-	-	-	-	-	-	9549.81	-	-
<i>f</i> ₃₆	34.63	21.26	1.63	18.57	9.00	2.06	28.40	19.38	1.47	26.07	30.50	0.85	40.53	30.50	1.33
<i>f</i> ₃₇	11.43	13.03	0.88	12.40	12.83	0.97	12.67	12.27	1.03	14.40	12.20	1.18	13.60	16.60	0.82

TABLE IX. 30D-NUMBER OF FUNCTIONS CALLS (NFC) AND ACCELERATION RATE (AR) OF DE AND FPRVDE VARIANTS

Function	rand/1/exp			best/1/exp			rand-to-best/1/exp			rand/2/exp			best/2/exp		
	DE ₆	FPRVDE ₆	A.R	DE ₇	FPRVDE ₇	A.R	DE ₈	FPRVDE ₈	A.R	DE ₉	FPRVDE ₉	A.R	DE ₁₀	FPRVDE ₁₀	A.R
<i>f</i> ₁	652.70	615.13	1.06	536.53	524.47	1.02	889.40	852.87	1.04	793.97	764.50	1.04	535.17	510.37	1.05
<i>f</i> ₂	741.70	701.00	1.06	606.80	588.00	1.03	1015.70	970.93	1.05	892.90	865.57	1.03	600.33	575.13	1.04
<i>f</i> ₃	14352.63	12284.50	1.17	8940.07	8221.74	1.09	23114.43	20119.90	1.15	17615.53	15490.23	1.14	7457.13	6480.43	1.15
<i>f</i> ₄	42169.90	29721.00	1.42	52822.07	77288.78	0.68	71640.67	67045.45	1.07	123743.13	121478.96	1.02	81340.60	-	-
<i>f</i> ₅	1248.69	-	-	1021.88	-	-	1881.83	-	-	1669.85	-	-	-	-	-
<i>f</i> ₆	1188.27	1259.54	0.94	828.50	762.11	1.09	2602.27	1994.13	1.30	1691.17	1291.81	1.31	895.03	912.73	0.98
<i>f</i> ₇	373.83	328.97	1.14	250.73	215.61	1.16	480.73	462.17	1.04	395.23	366.57	1.08	273.33	236.90	1.15
<i>f</i> ₈	1590.63	1393.92	1.14	1305.67	-	-	2371.17	2220.00	1.07	2077.05	1955.88	1.06	-	-	-
<i>f</i> ₉	343.00	327.03	1.05	289.07	282.03	1.02	469.43	441.50	1.06	420.43	389.93	1.08	286.70	282.80	1.01
<i>f</i> ₁₀	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₁₁	1179.90	1092.79	1.08	972.90	868.76	1.12	1641.87	1517.75	1.08	1451.77	1347.72	1.08	992.50	925.97	1.07
<i>f</i> ₁₂	434.67	400.60	1.09	363.10	341.10	1.06	588.00	539.17	1.09	538.10	487.53	1.10	360.93	340.17	1.06
<i>f</i> ₁₃	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₁₄	336.03	299.97	1.12	280.37	261.50	1.07	460.10	403.20	1.14	403.10	368.90	1.09	271.40	245.60	1.11
<i>f</i> ₁₅	3847.77	4484.40	0.86	6088.57	4451.20	1.37	5945.83	5311.87	1.12	4496.60	5532.27	0.81	4770.70	4713.03	1.01
<i>f</i> ₁₆	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₁₇	3591.03	2610.68	1.38	2260.57	1512.00	1.50	8373.57	6694.38	1.25	5792.07	4362.94	1.33	26192.61	37701.14	0.69
<i>f</i> ₁₈	104.00	100.10	-	88.03	79.87	-	112.67	109.43	-	104.50	103.07	-	126.70	127.73	-
<i>f</i> ₁₉	548.77	511.40	1.07	452.00	433.77	1.04	762.90	716.20	1.07	672.50	642.83	1.05	472.07	438.79	1.08
<i>f</i> ₂₀	10754.00	12386.73	-	11588.73	17553.27	-	12538.70	7911.57	-	12754.63	16314.73	-	7084.37	7028.27	-
<i>f</i> ₂₁	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₂₂	637.87	603.60	1.06	527.13	511.50	1.03	893.37	852.40	1.05	796.03	758.27	1.05	536.90	513.23	1.05
<i>f</i> ₂₃	1154.83	1089.77	1.06	947.80	915.33	1.04	1586.70	1509.67	1.05	1404.73	1346.47	1.04	941.67	904.53	1.04
<i>f</i> ₂₄	643.83	609.87	1.06	531.10	512.57	1.04	876.60	836.90	1.05	778.30	744.43	1.05	525.00	502.00	1.05
<i>f</i> ₂₅	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>f</i> ₂₆	882.70	837.10	1.05	724.60	705.10	1.03	1214.77	1155.73	1.05	1079.67	1032.33	1.05	718.00	694.30	1.03
<i>f</i> ₂₇	906.47	816.67	1.11	748.13	707.42	1.06	1244.70	1124.53	1.11	1099.50	1022.73	1.08	744.20	687.00	1.08
<i>f</i> ₂₈	2318.87	2299.57	1.01	1400.00	1145.25	1.22	3705.17	5384.86	0.69	3282.33	7086.22	0.46	1353.97	1293.57	1.05
<i>f</i> ₂₉	295.77	317.23	0.93	330.00	247.27	1.33	435.47	368.10	1.18	404.60	350.20	1.16	222.00	187.43	1.18
<i>f</i> ₃₀	819.83	830.43	0.99	662.03	-	-	1038.23	1015.00	1.02	878.27	-	-	857.30	-	-
<i>f</i> ₃₁	832.13	840.67	0.99	674.37	-	-	1056.53	1027.76	1.03	887.87	-	-	874.43	1617.12	0.54
<i>f</i> ₃₂	85.77	870.33	0.10	73.33	18.40	3.99	100.23	256.57	0.39	94.97	17.70	5.37	74.40	43.70	1.70
<i>f</i> ₃₃	2.20	1.93	1.14	1.70	1.67	1.02	1.70	1.83	0.93	1.60	1.33	1.20	1.30	1.50	0.87
<i>f</i> ₃₄	1901.30	1674.28	1.14	1317.20	1209.78	1.09	3374.67	2895.40	1.17	2437.50	2149.50	1.13	4528.57	6533.87	0.69
<i>f</i> ₃₅	16811.78	-	-	-	-	-	-	-	-	-	-	-	16593.92	9860.13	1.68
<i>f</i> ₃₆	320.90	148.04	2.17	192.10	74.50	2.58	267.70	256.50	1.04	226.70	99.67	2.27	246.37	215.17	1.15
<i>f</i> ₃₇	11.43	13.03	-	12.40	12.83	-	12.67	12.27	-	14.40	12.20	-	13.60	16.60	-

Experimental results given in tables (IV-IX) are obtained using the parameter settings given in section-IV of this paper. Results of number of function calls and acceleration rate performance parameters of each DE strategy and FPRVDE strategy are reported. For easy analysis, results of 10D, 20D and 30D are reported in separate tables. Tables IV-V contains 10-Dimensions results, tables VI-VII contains 20D results and tables VIII-IX contains 30D performance results of benchmark functions for each DE strategy in table-I and its corresponding FPRVDE strategy given in table-II. The best result of the NFC's for each function and each strategy pair (DE/FPRVDE) are highlighted in **boldface** and summary is reported in table-X. Few entries in the tables (IV, IX) are filled with dash symbol where variants fails to find any solution like none of the variant of DE or FPRVDE find any solution of function f_{16} and vice versa. For simple analysis the detailed summary of research results is presented in table-X. Table-IX contains the best values of NFC and average acceleration rate (A.A.R) for each DE and its corresponding FPRVDE mutation strategy. It is obvious from the research result that FPRVDE₁ has better convergence speed in most of the cases for 10D, 20D and 30D results. The average of average acceleration rate of FPRVDE mutation strategies than DE mutation strategies is **1.75**.

The performances of all FPRVDE (FPRVDE₁.... FPRVDE₁₀) mutation strategies dominate over all DE (DE₁.... DE₁₀) strategies for all variables (dimensions) for Separable functions $f_1, f_2, f_9, f_{19}, f_{22}, f_{23}, f_{24}, f_{26}, f_{27}$; non-Separable functions f_3, f_4, f_7, f_{14} ; unimodal functions; unimodal functions $f_2, f_3, f_4, f_{14}, f_{26}, f_{27}$ and multimodal functions $f_1, f_7, f_9, f_{19}, f_{22}, f_{23}, f_{24}$. For other separable/non-separable and unimodal/multimodal functions FPRVDE mutation strategies are better than DE strategies in most of the cases and DE strategies are better in few cases only.

None of the DE/FPRVDE variant reaches to VTR using the current parameter setting for separable & unimodal function f_{13} and for non-separable & multimodal functions f_{16}, f_{21} . For functions f_{20}, f_{37} , none of the algorithm reaches to VTR with few variables (10D) but converges to VTR with more variables (20D, 30D). A non-separable and multimodal function f_{18} converges to VTR only for more variables (30D).

Now we discuss the cases where the performance of DE mutation strategies is better than FPRVDE strategies. DE₁ is better than FPRVDE₁ for functions $f_5, f_8, f_{17}, f_{29}, f_{36}$ (for 10 variables) and functions f_5, f_8, f_{33}, f_{37} (for 20 variables); DE₂ is better than FPRVDE₂ for functions $f_{11}, f_{15}, f_{17}, f_{25}, f_{35}$ (for 10

variables), functions f_{11}, f_{33}, f_{35} (for 20 variables) and functions f_{30}, f_{31}, f_{37} (for 30 variables); DE₃ is better than FPRVDE₃ for a separable & multimodal function f_{33} (for 30 variables); DE₄ is better than FPRVDE₄ for function f_5 (for 10 variables), function f_8 (for 20 variables) and functions f_{30}, f_{31}, f_{36} (for 30 variables); DE₅ is better than FPRVDE₅ for functions $f_{17}, f_{34}, f_{35}, f_{36}$ (for 10 variables), functions $f_6, f_{28}, f_{33}, f_{35}, f_{36}$ (for 20 variables) and functions $f_{12}, f_{30}, f_{31}, f_{35}, f_{37}$ (for 30 variables); DE₆ is better than FPRVDE₆ for functions f_{30}, f_{31} (for 10 variables), functions f_{10}, f_{17}, f_{31} (for 20 variables) and functions $f_5, f_{15}, f_{29}, f_{30}, f_{31}, f_{32}, f_{35}$ (for 30 variables); DE₇ is better than FPRVDE₇ for functions f_6, f_8, f_{11}, f_{29} (for 10 variables), functions $f_8, f_{10}, f_{15}, f_{17}, f_{29}$ (for 20 variables) and functions $f_4, f_5, f_8, f_{30}, f_{31}$ (for 30 variables); DE₈ is better than FPRVDE₈ for functions f_{15}, f_{36} (for 10 variables), functions $f_{10}, f_{15}, f_{17}, f_{33}, f_{36}$ (for 20 variables) and functions $f_5, f_{28}, f_{32}, f_{33}$ (for 30 variables); DE₉ is better than FPRVDE₉ for functions f_{29} (for 10 variables), functions f_{15}, f_{17}, f_{33} (for 20 variables) and functions $f_5, f_{15}, f_{29}, f_{30}, f_{31}$ (for 30 variables); DE₁₀ is better than FPRVDE₁₀ for functions $f_6, f_8, f_{15}, f_{25}, f_{29}, f_{34}$ (for 10 variables), functions $f_{10}, f_{17}, f_{30}, f_{34}, f_{35}$ (for 20 variables) and functions $f_6, f_{17}, f_{30}, f_{31}, f_{33}, f_{34}, f_{37}$ (for 30 variables);

It can be summarized from number of functions calls parameter that in few cases FPRVDE is significantly worse and in most of the cases the performance of FPRVDE is better than DE algorithm variants for 10D, 20D and 30D respectively. Considering the acceleration rate parameter FPRVDE variants are faster as compared to DE variants in most of the cases while in few cases FPRVDE variants has significantly worst performance than DE variants.

Convergence graph are shown in Figure-2 and Figure -3 that contains average of best values of population members obtained at specific iterations during the evolutionary process. Convergence graph contains the fitness value against iteration in each sub-graph in Figure-2 and Figure-3 for 5000 iterations and 10 dimensions. From convergence graphs it can be observed that FPRVDE variants performs better than DE variants in most of the cases; in few cases FPRVDE has comparable performance while in some cases DE performance is better than FPRVDE. Convergence graph of few functions f_{15} (a₁-j₁), f_{17} (a₂-j₂), f_{35} (a₃-j₃), f_{37} (a₄-j₄) are given for 10 DE mutation strategies (DE₁....DE₁₀) and their corresponding FPRVDE (FPRVDE₁.... FPRVDE₁₀) mutation strategies. Convergence graphs of DE mutation strategies and its corresponding FPRVDE strategies are given in each sub-graph (a_i....j_i) in Figure-2 and Figure-3.

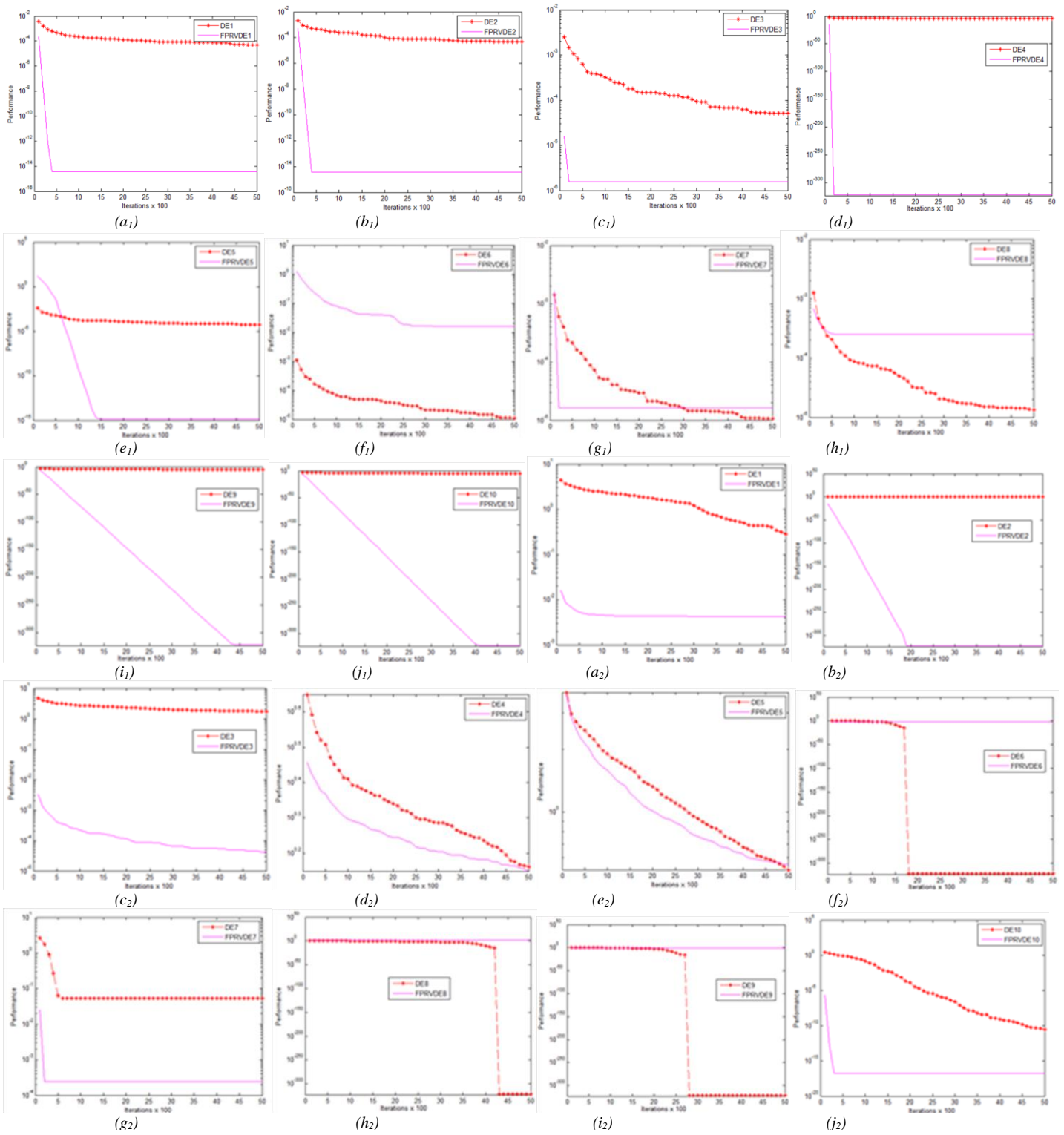


Fig. 2. DE and FPRVDE variants 10D Convergence graphs for function f_{15} (a₁-j₁), f_{17} (a₂-j₂)

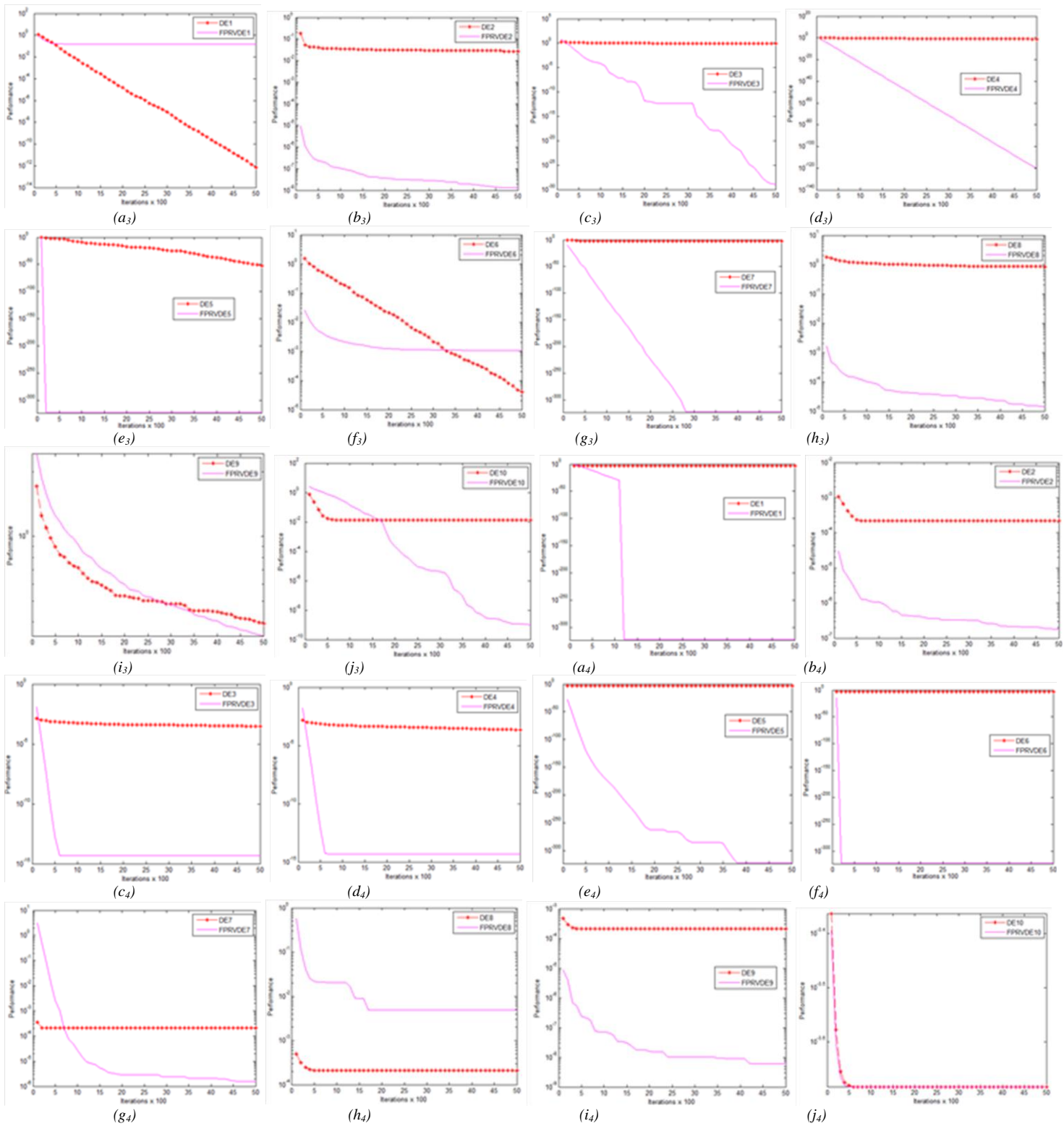


Fig. 3. DE and FPRVDE 10D Convergence graphs for function $f_{35}(a_3-j_3)$, $f_{37}(a_4-j_4)$

TABLE XI. SUMMARY OF RESULTS NUMBER OF FUNCTIONS CALLS (NFC)
BEST AND AVERAGE ACCELERATION RATE (A.A.R)

10D			20D			30D		
Variant	Best	A.A.R	Variant	Best	A.A.R	Variant	Best	A.A.R
DE ₁	0	1.68	DE ₁	5	1.83	DE ₁	2	2.48
FPRVDE ₁	31		FPRVDE ₁	24		FPRVDE ₁	28	
DE ₂	6	1.39	DE ₂	2	1.77	DE ₂	3	1.61
FPRVDE ₂	23		FPRVDE ₂	25		FPRVDE ₂	23	
DE ₃	0	1.68	DE ₃	0	2.50	DE ₃	0	5.19
FPRVDE ₃	29		FPRVDE ₃	28		FPRVDE ₃	29	
DE ₄	2	2.12	DE ₄	1	2.57	DE ₄	3	5.98
FPRVDE ₄	28		FPRVDE ₄	27		FPRVDE ₄	25	
DE ₅	3	1.30	DE ₅	5	1.42	DE ₅	4	1.52
FPRVDE ₅	28		FPRVDE ₅	23		FPRVDE ₅	22	
DE ₆	2	1.21	DE ₆	3	1.13	DE ₆	8	1.08
FPRVDE ₆	29		FPRVDE ₆	28		FPRVDE ₆	22	
DE ₇	4	1.30	DE ₇	5	1.31	DE ₇	5	1.17
FPRVDE ₇	25		FPRVDE ₇	24		FPRVDE ₇	22	
DE ₈	2	1.15	DE ₈	5	1.10	DE ₈	4	1.05
FPRVDE ₈	28		FPRVDE ₈	25		FPRVDE ₈	26	
DE ₉	1	1.21	DE ₉	2	1.14	DE ₉	5	1.28
FPRVDE ₉	30		FPRVDE ₉	28		FPRVDE ₉	23	
DE ₁₀	5	1.10	DE ₁₀	5	1.16	DE ₁₀	7	1.05
FPRVDE ₁₀	26		FPRVDE ₁₀	24		FPRVDE ₁₀	21	

V. CONCLUSION AND FUTURE WORK

Trial vector has a key role in generating offspring/child population in DE algorithm. Different vectors like *random*, *best* and *current* are commonly used vectors to generate child population. Although random vector selection method is less biased and generates the more diverse population but has slow convergence to reach to a specific value VTR or optimal value. In this research a novel fitness proportionate based selection in random selecting random vectors used in DE (FPRVDE) mutation strategy is introduced. FPRVDE advancement is applied on most commonly used DE mutation strategies given in table-I. FPRVDE approach selects parent vectors by generating a random vector following the approach of fitness proportionate selection criteria. Performance of FPRVDE is accessed by taking a comprehensive set of multidimensional function optimization problems given in appendix section of this paper. Research result shows that FPRVDE variation approach enhances convergence speed of DE algorithm by maintaining appropriate altitude of diversity. The proposed approach ignores the poor performing individuals in generating the trial vector. NFC and AR performance parameters are used compare the performance of commonly used DE mutation strategies and FPRVDE approach. Research results shows that the convergence speed of FPRVDE approach is better than DE approach. NFC of FPRVDE mutation strategies is better than DE mutation strategies for most functions and various dimensions. Acceleration rate of FPRVDE mutation strategies for various functions and various mutation strategies is better than DE mutation strategies. In this research an effort is made to work in other directions of DE algorithm that will prove to a significant addition in DE research work. The future challenges of this research work can be the deep insight of FPRVDE along with its parameter settings and to explore the proposed approach in the other dimensions of research.

REFERENCES

- [1] R Storn and K Price, "Differential evolution—A simple and efficient adaptive scheme for global optimization over continuous spaces," CA, Berkeley, Tech. Rep TR-95-012, 1995.
- [2] J. Brest and et. al., "Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems," IEEE Transaction on Evolutionary Computing, vol. 10, no. 6, pp. 646-657, December 2006.
- [3] K. Price, R. M. Storn, and J. A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization (Natural Computing Series), 1st ed. New York:, USA: Springer-Verlag, 2005.
- [4] A. A. Abou El Ela, M. A. Abido, and S. R. Spea, "Optimal power flow using differential evolution algorithm," Electric Power Systems Research, vol. 80, no. 7, pp. 878-885, 2010.
- [5] K. A. Michalski, "Electromagnetic imaging of circular-cylindrical conductors and tunnels using a differential evolution algorithm," Microwave and Optical Technology Letters, vol. 27, no. 5, pp. 330-334, 2000.
- [6] R. Senkerik and et al., "Chaos driven Differential Evolution in the task of chaos control optimization," in IEEE Congress on Evolutionary Computation, Barcelona, Spain, 2010, pp. 1-8.
- [7] N. Noman and H. Iba, "Inference of gene regulatory networks using s-system and differential evolution," in In Proceedings of the 2005 conference on Genetic and evolutionary computation, 2005, pp. 439-446.
- [8] B. V. Babu, P. G. Chakole, and J. H. S. Mubeen, "Multiobjective differential evolution (MODE) for optimization of adiabatic styrene reactor," Chemical Engineering Science, vol. 60, no. 17, pp. 4822-4837, 2005.
- [9] M. Ali, C. W. Ahn, and M. Pant, "A robust image watermarking technique using SVD and differential evolution in DCT domain," Optik - International Journal for Light and Electron Optics, vol. 125, no. 1, pp. 428-434, January 2014.
- [10] N. Chauhan, V. Ravi, and D. K. Chandra, "Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks," Expert Systems with Applications, vol. 36, no. 4, pp. 7659-7665, May 2009.
- [11] B. Singh, B. Dhillon, and Y. S. Brar, "A Hybrid Differential Evolution Method for the Design of IIR Digital Filter.," International Journal on Signal & Image Processing, vol. 4, no. 1, pp. 1-10, January 2013.
- [12] X. Yao, Y. Liu, and G. Lin, "Evolutionary Programming Made Faster," IEEE Transaction on Evolutionary Computing, vol. 3, no. 2, pp. 82-102, July 1999.
- [13] S. F. P. Saramago G. T. T. Oliveira, "A Contribution to the Study About Differential Evolution," Ciencia & Engenharia, vol. 16, no. 1/2, pp. 1-8, 2007.
- [14] S. Das, A. Abraham, and A. Konar, "Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives," Studies in Computational Intelligence, vol. 11, no. 6, pp. 1-38, 2008.
- [15] X. Xu and Y. Li, "Comparison between Particle Swarm Optimization, Differential Evolution and Multi-parents Crossover," in IEEE International Conference on Computational Intelligence and Security, 2007, pp. 124-127.
- [16] R Storn and Price K, "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," Journal of Global Optimization, no. 11, pp. 341-359, 1997.
- [17] S. Das, Ajith Abraham, Uday K Chakraborty, and Amit Konar, "Differential Evolution Using a Neighborhood-Based Mutation Operator," IEEE Transactions on Evolutionary Computing, vol. 13, no. 3, pp. 526-553, June 2009.
- [18] A. Ghosh and et. al, "An improved differential evolution algorithm with fitness-based adaptation of the control parameters," Information Sciences, vol. 181, pp. 3749-3765, 2011.
- [19] S. M. Islam and et. al, "An Adaptive Differential Evolution Algorithm With Novel Mutation and Crossover Strategies for Global Numerical Optimization," IEEE transactions on systems, man, and cybernetics, vol. 42, no. 2, pp. 482-500, April 2012.

- [20] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 13, pp. 398-417, 2009.
- [21] D. Zaharie, "Control of population diversity and adaptation in differential evolution algorithms," in *In proceeding of MENDEL*, vol. 9, 2003, pp. 41-46.
- [22] J. Liu and J. Lampinen, "A Fuzzy Adaptive Differential Evolution Algorithm," *Soft Computing*, vol. 9, pp. 448-462, 2005.
- [23] J. Brest and et al., "Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 6, pp. 646-657, 2006.
- [24] J. Zhang and A. C. Sanderson, "JADE: Adaptive Differential Evolution With Optional External Archive," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 945-958, 2009.
- [25] R. Mallipeddi and et al., "Differential evolution algorithm with ensemble of parameters and mutation strategies," *Applied Soft Computing*, vol. 11, no. 2, pp. 1679-1696, 2011.
- [26] Y. Wang, Z. Cai, and Q. Zhang, "Differential Evolution With Composite Trial Vector Generation Strategies and Control Parameters," *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 55-66, 2011.
- [27] W. Gong and et al., "Enhanced differential evolution with adaptive strategies for numerical optimization," *IEEE Transactions on Systems, Man, and Cybernetics- Part B: Cybernetics*, vol. 41, no. 2, pp. 397-413, 2011.
- [28] S. Das and et al., "Differential Evolution Using a Neighborhood-Based Mutation Operator," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 3, pp. 526-553, 2009.
- [29] S. Minhazul-Islam and et al., "An Adaptive Differential Evolution Algorithm With Novel Mutation and Crossover Strategies for Global Numerical Optimization," *IEEE Transactions on Systems, Man, and Cybernetics-PART B: Cybernetics*, vol. 42, no. 2, pp. 482-500, 2012.
- [30] E. Mezura-Montes, J V Reyes, and C A Coello Coello, "A Comparative Study of Differential Evolution Variants for Global Optimization," in *Genetic and Evolutionary Computation Conference(GECCO)*, Washington, USA, 2006, pp. 485-492.
- [31] M. Ali, M. Pant, and A. Abraham, "Simplex Differential Evolution," *Acta Polytechnica Hungarica*, vol. 6, no. 5, pp. 95-115, 2009.
- [32] S. Rahnamayan, H. R. Tizhoosh, and M. M. A. Salama, "Opposition-Based Differential Evolution," *IEEE transactions on evolutionary computation*, vol. 12, no. 1, pp. 64-79, February 2008.
- [33] Y. Zhou, X. Li, and L. Gao, "A differential evolution algorithm with intersect mutation operator," *Applied Soft Computing*, vol. 13, pp. 390-401, 2013.
- [34] J. Brest, "Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 10, pp. 646-657, December 2006.
- [35] Y. Wang, Z. Cai, and Q. Zhang, "Enhancing the search ability of differential evolution through orthogonal crossover," *Information Sciences*, vol. 185, pp. 153-177, 2012.

Internet of Things based Expert System for Smart Agriculture

Raheela Shahzadi

Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Javed Ferzund

Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Muhammad Tausif

Department of Computer Science
COMSATS Institute of information Technology
Vehari, Pakistan

Muhammad Asif Suryani

Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Abstract—Agriculture sector is evolving with the advent of the information and communication technology. Efforts are being made to enhance the productivity and reduce losses by using the state of the art technology and equipment. As most of the farmers are unaware of the technology and latest practices, many expert systems have been developed in the world to facilitate the farmers. However, these expert systems rely on the stored knowledge base. We propose an expert system based on the Internet of Things (IoT) that will use the input data collected in real time. It will help to take proactive and preventive actions to minimize the losses due to diseases and insects/pests.

Keywords—Internet of Things; Smart Agriculture; Cotton; Plant Diseases; Wireless Sensor Network

I. INTRODUCTION

Internet of Things (IoT) is a broad term that describes the interconnection of different daily life objects through the internet. In the concept of IoT every object is connected with each other through a unique identifier so that it can transfer data over the network without a human to the human interaction [1, 2]. IoT has referred as a network of everyday objects having ubiquitous computing. The ubiquity of the objects has increased by integrating every object with embedded system for interaction [21]. It connects human and devices through a highly distributed network. IoT is basically the world wide interconnection of devices. The aim of IoT is to connect every person and every object through the internet. In IoT ,every object is assigned a unique identifier, so that every object is accessible through the internet [22][23].

Every object in the IoT has the following three capabilities: awareness, representation, and interaction. Awareness is the ability of the smart objects to understand and sense other objects. Representation is the ability of the objects to present, according to the programming concept. Interaction is the ability to communicate with each other

The IoT is evolving, growing and becoming popular day by day; in the today's world, around 5 billion objects have connected through the internet. In 2020, it has estimated that near about 50 billion objects will be connected to the internet [24]. IoT is providing tremendous opportunities for novel

applications, which is now widely used in many aspects of life such as intelligent home monitoring system, products supply chain management, precision agriculture and much more.

Every object in IoT is addressable, recognizable, readable and locatable through the internet by using RFID (Radio Frequency Identification), Wireless Sensor Network (WSN) or other means. The concept of IoT is using many in different domains such as; precision agriculture [1, 2], products supply chain management [3], Smart Grid [4] , environmental monitoring [5], cloud computing [6] and many more. IoT is gaining much importance these days as every object in the network will become a computer [7]. The idea of IoT has become successful due to the invention of recent technologies like sensors, RFID and WSN.

Pakistan is an agricultural country. Although the industry and services sector has transformed Pakistan into a diversified country, still a major part of GDP is contributed by the agriculture sector. The foreign exchange of Pakistan has depended on agricultural products. More than half of the population of Pakistan lives in rural areas and major source of earnings of this population has based on agriculture. Most of the industry in Pakistan is also dependent on agriculture like textile industry, sugar industry, flour industry, juice industry, furniture industry, dairy industry, etc. [2].

Farmers in Pakistan are not aware of the technology and lack agricultural knowledge. They rely on traditional methods and practices. However, the agriculture field in the advanced world has evolved a lot due to the advances in technology and equipment. Pakistan faces huge losses in agriculture due to the following factors.

- Delayed sowing and poor seed quality.
- Environmental hazards.
- Insect and disease attacks.
- Unplanned irrigation and water losses.
- Untimely harvesting.
- Misuse of fertilizer and insecticides.

- Lack of machinery and equipment.
- Mishandling of ripened crops.

Cash crops have a major share in the economy of Pakistan and cotton crop is very important among them. It is also called as 'white gold'. Heavy losses occur every year due to poor farming practices, attack of pests at different stages and attack of diseases. According to a survey [8], 38 percent loss of cotton crop occurred due to the attack of insects in 2013. According to another survey, due to viral attack, 15 percent loss of cotton crop occurs every year. Cotton crop is affected by some of bacterial and fungal diseases as well as pests and insects. Temperature and humidity requirements for the cotton crop are different from other crops before and after sowing. The timing of spray of insecticides, pesticides, and application of fertilizer also affects the crop growth. So, the continuous monitoring is required to keep the crop healthy.

Most of the farmers in Pakistan are illiterate, and they are unaware of the latest research in the field of agriculture. The farmers normally take guidance from agriculture experts and other experienced farmers. However, the experts are not always available every time and everywhere [9]. So, expert systems have been developed for different crops, fruits and vegetables in the world. Basically an expert system (ES) is a computer program which solves the problems as human being solves the problem. It is a tool which generates output using its knowledge base, so it replicates the behavior of the human. The ES can pinpoint the problems as well as figure out the solutions. It combines the same domain knowledge of different experts. In the ES accumulation of knowledge from different sources is a very important factor.

The ES can provide output whenever it has given input. It means it should be easily available to the farmers. In this paper, we present an Expert System based on the concept of Internet of Things. Sensors will collect data and automatically send it to the ES. The ES will process the information and send the results or decisions to the farmer's mobile phone. In this way, crops can continuously monitor, and timely decisions can be taken. It will help to minimize the losses due to sudden disease and pest attacks through timely proactive and preventive actions. The proposed ES is initially developed for the Cotton crop and evaluated by the farmer community.

The proposed system can be used for.

- Efficient Crop Management. Irrigation Control.
- Environment Warnings and Guidance.
- Optimal usage of fertilizers, insecticides and pesticides

The rest of the paper has organized as follow, in section 2, we discuss the related work. In section 3 we describe proposed ES based on IoT, Section 4 we describe implementation and evaluation of proposed ES based on IoT, In section 5 we describe the results, In section 6, we elaborate the validity of proposed solution and, In section 7, we conclude the whole work and describe future work.

II. RELATED WORK

Fan TongKe [10] proposed the smart agriculture based on

cloud computing. The author presented the architecture for the smart agriculture based upon the concept of the IoT and cloud computing. Agriculture information cloud was combined with Internet of Things to achieve the dynamic distribution of resources and balance of the load.

Ji-Chun Zhao et al. [7] studied the applications of IoT in agriculture. The authors proposed a monitoring system based on internet and wireless sensor networks. An information management system was designed to provide the data for research in agriculture. The authors developed software for monitoring of the fields like data acquisition about the fields, data processing models, and system configuration module. The developed application provides accurate control for the monitoring of the green house.

Agrawal and Lal Das [2] discussed the possible future applications and challenges faced by the IoT technology. They presented some key challenges in IoT applications such as: standards, privacy, security, authentication and identification, trust and ownership, integration, coordination, and regulation. They stated that the use of RFID (Radio Frequency Identification), Wireless Sensor Network (WSN) and mobile communication technologies would reduce the gap between theoretical and practical implementations of IoT applications.

Chen and Jin [11] proposed the 'Digital Agriculture' based upon IoT. The working of the digital agriculture is divided into two steps: in the first phase, the information about the temperature, the wind, the soil contents, etc. is collected by different sensors. In the second phase, ZigBee transfers information. The agricultural products has labeled with EPC code. The EPC code reader reads the code of the products.

Li Li et al. [4] discussed the application of smart and Wi-Fi based Wireless Sensor Network in IoT. The authors discussed the applications of IoT-based upon Wi-Fi, WSN and smart grid. Smart grid provides the intelligent data collection application, improving reliability of data collection and providing accurate information. IoT provides the intelligent environment monitoring application; water data and air data collected through sensors and sent to server for further processing. They proposed the concept of the precision agriculture. The authors stated that new WSN technology is better as compared to ZigBee.

Hussain et al. [12] proposed the application of Internet of Things (IoT) technology in animal stock chain management. By using the RFID technology, anyone can be tracked or monitored. They discussed some operational principle of IoT. RFID technology is used for the unique identification of objects; each object in the RFID is labeled with EPC code. The authors proposed the use of this technology for maintaining all records for livestock management.

Kosmatos et al. [13] proposed the architecture based on the RFID and smart objects. RFID objects will perform the primitive functionality in the proposed architecture while the smart objects will perform the complex functionality. The architecture was proposed based on the integration of RFID and smart objects. RFID tags have widely used for the identification of objects. So the RFID is used in the proposed architecture for tracking of the objects. The authors used

service oriented architecture and semantic model-driven approach in the proposed architecture.

Carvin et al. [14] proposed the ubiquitous cognitive management system based on IoT and ubiquitous computing. They presented the problems as well solutions of problems. The basic idea was to use ambient intelligence provided by smart objects to serve the human, improving communication by context information.

Zhou and Zhou [15] proposed a management model based on IoT for visualization and traceability of agricultural products. The aim was to ensure food safety and promote sustainable development of modern agriculture. The authors used the products logistic information along with Internet of Things for effective products supply chain management.

Prasad et al. [16] proposed an expert system for the diagnosis of pests, diseases, and disorders in mango. The system had developed in ESTA (Expert System Shell for Text Animation). In the proposed system, the first step is the knowledge acquisition; the second step is the diagnosis of disease based on the input. They briefly described the type of mango diseases and recommendations for the disease control on the basis by visual symptoms.

Sarma et al. [17] proposed an expert system for diagnosis of disease in rice plant in India. The purpose of the expert system is to assist farmers in solving the problems. The first step is the development of the knowledge base in the form of condition rules. The proposed system is easy to use and will be useful to those people who are unable to get the assistance of some agriculture expert.

Kaliuday et al. [18] proposed rule based expert system for the prevention of pest diseases in rice and wheat crops. They built an expert system (ES) called AgPest for the diagnosis of pest disease in wheat and rice. They developed AgPest in CLIPS. It consists of the IF then else rule for finding the disease. They formulated the rules about the diseases of wheat and rice from different online sources.

Negied [19] proposed the expert system for the protection of the wheat yield in Egypt. The proposed system is developed using following steps; the first step is the problem identification of the domain, the second step is information . They developed the system in MATLAB. The proposed system is helpful for improving the yield of wheat crop and for providing assistance to the farmers in the remote areas.

Kaur et al. [20] proposed the expert system for the detection and diagnosis of the leaf diseases in cereals. It is quite difficult for the farmers to identify the leaf diseases without the assistance of the experts. For the identification of the diseases, they proposed image comparison techniques in JAVA. They used techniques like affine transformation and edge detection for this purpose. It is web based expert system so that it can be accessible from any web-enabled system

A. Expert System in Agriculture

Expert system (ES) is the branch of artificial intelligence that deals with the development of computer programs which can solve the problems as the human beings solve the problems [25]. The application of expert system in agriculture is

increasing widely since many years. A number of expert systems have been developed in the field of agriculture such as

- AMRAPALIKA is the ES for the diagnosis of disease, pest and disorders in Indian mango [16]
- An expert system for diagnosis of disease in Indian rice plant[4].
- CITEEX: An expert system for citrus crop management[26].
- CUPTTEX: An expert system proposed for the management of cucumber[27].
- An expert system for the olive crop diseases and weed identification in Spain[9].
- LIMEX: An expert system for the management of lime crop[28].
- CALEX: An expert for the diagnosis of peach and nectarine disorders[29].
- CITPATH: An expert system for the diagnosis of fungal disease in citrus fruit[30].

The first step in the development of any ES is the problem identification. For example, if we are developing the ES for cotton crop then symptoms are identified. The description of the diseases can be textual or in the form of images. After that rules are formulated based on the concept of if then else structure.

Most of the farmers in the remote areas are illiterate not having proper knowledge of dealing with diseases. Some diseases of the crops are difficult to distinguish because two or more diseases have the same symptoms. So, it creates problems for the farmers. This problem can overcome with the help of ES by combining the knowledge of different experts in one application. Most of the researchers are trying to develop the ES for fulfilling the needs of the farmers. If farmers. The farmer gets the assistance in time, the productivity rate of the crops will increase

III. IOT BASED EXPERT SYSTEM

For overcoming the problems of agriculture, we develop an initial framework based on IoT. The proposed solution consists of three main components the first component is the deployment of sensors in the field; we deploy soil sensors, humidity sensors, and temperature and leaf wetness sensors in the fields. Sensors collect the data and send it to sever, on the serve side we deploy the expert system, which processes the data and send the recommendations to the farmers about crops.

A. Deployment of Sensors

The sensors have deployed in the fields for the collection of data about the environment, humidity, soil moisture and leaf wetness. For the collection of data waspmote agriculture sensor board is used because it is specially designed for handling agriculture activities. The sensor board consists of AT mega 1281 microprocessor and 2GB micro SD –card. Every sensor board consists of four different types of sensors, the soil sensor, humidity sensor, temperature sensor and leaf wetness sensors.

We use three soil sensors, temperature sensors, humidity sensors and leaf wetness receptively for a precise and accurate measure of soil contents, environmental temperature, the humidity level in the environment and leaf wetness at the same time. The communication module XBee-802.15.4 is present in wasp-mote agriculture board. It can communicate with microcontroller at the rate of 38400 bps. The range of transmission is near about 500 meters. The gateway is the bridge between sensor nodes and server. It can communicate wirelessly with the sensor and through USB port with a computer. We conducted this experiment under controlled environmental conditions. The overall architecture of sensors communication has described in Fig 1

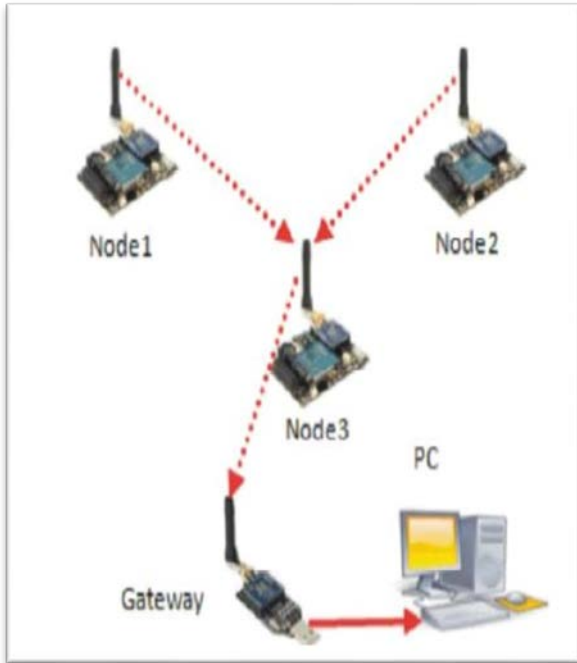


Fig. 1. Communication of Sensors Data in IoT Based ES

B. IoT- Based Expert System for Cotton Crop

IoT-based Expert System is different from traditional Expert Systems regarding inputs. It uses real-time input data gathered with the help of sensors. The sensor nodes send data to the gateway after the defined interval of time. The server receives data through the USB port. For storing and copying the cool term software, is used. The expert system deploys in the server process the data and sends the recommendation to the farmer cell phone. For solving this problem, the expert system based on the concept of IoT has proposed in Fig 2.

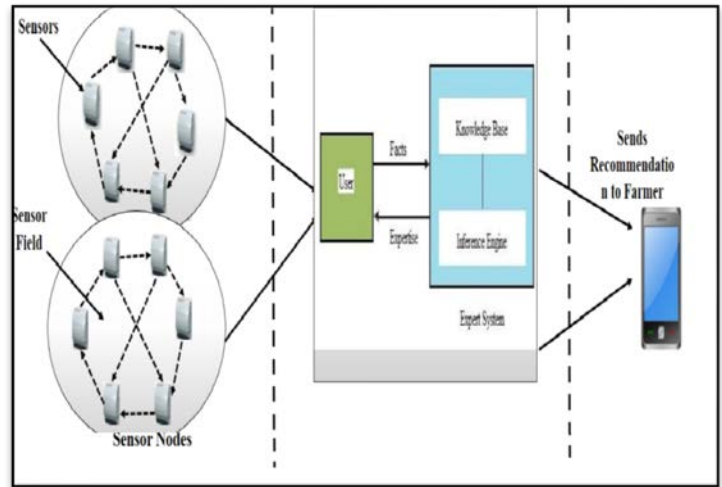


Fig. 2. IoT Based Expert System

The rest of the working of the proposed ES is similar to the traditional Expert Systems. It is implemented using CLIPS (C Language Integrated Production System) developed by NASA [31]. CLIPS is a C based instead of LISP and supports three programming approaches: rule-based, object oriented and procedural. It is portable, extendable, can be easily integrated and supports interactive development. It also has features for verification and validation of expert systems. The proposed expert system consists of the following main components.

- Knowledge Base
- Inference Engine Agenda
- Working Memory
- Explanation Facility
- User Interface

The structure of the expert system has shown in Fig 3.

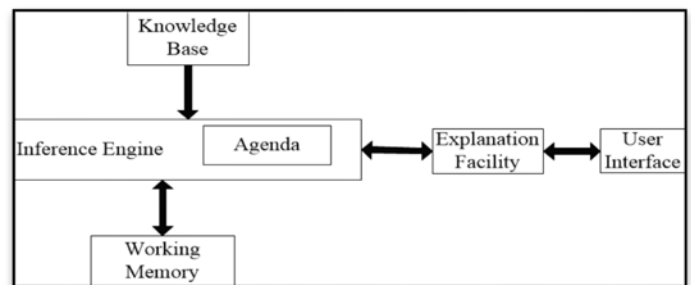


Fig. 3. Expert System in CLIP

The first step in the development of expert system is the knowledge acquisition of the domain. The most important thing in the knowledge acquisition is what type of knowledge we require for the expert system. In the proposed system, we need data about pests, insects, diseases, weeds and growth environment required for cotton crops. The knowledge acquisition can have accomplished in three ways.

- Experts of the domain are interviewed.

- Research articles about the domain are reviewed.
- Information is obtained by field observations.

Field observation may not be much stronger, as being computer scientist we are unaware about the diseases. Another thing which makes the field observation weak is that, all types of diseases, pests, weeds and insects are not spotted on the cotton crop at the same time. We gather the majority of the information by conducting interviews with experts. We get information about the type of diseases, causes of diseases, symptoms of diseases, the insects which attack on the cotton crop, the weeds which destroy the cotton crop, and the insects which spread the disease from one plant to another.

Different sensors collect the data, soil sensors collect the data about soil condition, soil moisture and soil content, while weather sensors collect the data about humidity and temperature. Sensors send the data to the server, the server decides about the diseases on the basis of the fact list which is used for the training. In CLIPS, fact list, rule list, and agenda with the activation list kept in memory. All the facts have based on simple if then else logic. The sensors collect data and send to the server, on the server side, we deploy the expert system which processes the data and analyzes the data and send the recommendation to the farmer about crops.

C. Server Send Recommendation to Farmer

The server processes the sensor data, after processing it send the recommendation to the farmer cell phone. So, for farmer convenience, we develop an android app for farmers. Farmers install the android app on their phone. The server sends the recommendation to the farmer cell phone. The server sends the recommendation in English; farmer can convert the recommendation to the Urdu or Punjab in his convenient language.

IV. IMPLEMENTATION AND EVALUATION

This section describes the implementation and evaluation of the expert system for SA. The sensors collect the real-time data and send to the server. On the server side, ES have deployed for extracting the information from sensor data.

A. Expert System for CLIPS

In this section, we implement different fact list of insect diagnosis, pest diagnosis, weeds diagnosis and irrigation scheduling.

Table 1 describes the different fact lists of insects, pest symptoms and recommendation for the attack of insect pest.

Table 2 describes the different fact lists of weed symptoms and recommendations for the attack of weeds.

Table 3 describes the different fact lists of sucking insects and recommendation for the attack of sucking insects.

The irrigation of crops depends on several factors like soil moisture, soil type, depth of root zone and the environment. Every soil has different physical properties and textures like coarse soil, medium texture, soil and heavy fine textured soil. The water capacity of every soul is different, so the amount of irrigation also varies according to the texture of the soil. Environmental fluctuations are also important factors for scheduling of irrigation. The cotton crop has a specific limit of water depletion, if the water gets depleted more than the specified threshold limit, then irrigation should be applied. To calculate the minimum flow of irrigation q (in cubic meters per hour), we used Equation 1. In this equation, Dg is gross application dose, A is an area, I is interval of day, T is operating hour per day and 10 is a constant for hectare. In equation 1 we are presenting a formula for calculating the irrigation dose.

$$q=10A*Dg/I+T.....(1)$$

Table 4 describes the irrigation scheduling in the cotton crops.

V. RESULTS

We conduct this experiment in the cotton fields Sahiwal from June 2015 to December 2015. The server processes the data and sends the recommendation to the farmer. The proposed ES provide diagnose diseases, attack of weeds and attack of pests, it provides the pesticide recommendation for weeds, diseases, and pests. It provides predication of diseases based on sensor data. It provides irrigation scheduling based on temperature and soil contents; it can also provide the dose of irrigation.

After that sensors send data to the server, on the server side, we deploy the expert system which processes the sensor data and sends the recommendation to the farmers.

Table 5 describes the comparison of the expert system, which has presented during different eras. In the previous expert system, the user manually inputs the symbol of diseases, and they use web-based, ontology based and expert system tool for the development of expert systems. In previous literature, they did not use the concept of IoT for the collection of data. In our proposed solution, we develop an expert system based on the concept of IoT. Different kind of sensors is deployed in the fields which monitor the crops, and send the data to the server, server process the data and send a recommendation to the farmer.

Fig 4 represents the relationship between temperature and humidity sensor data, temperature, and humidity, inversely proportional to each other. If temperature increase then humidity level decrease. We are just representing the 117 recording of sensor data. By taking these sensor data, we are scheduling the automatic irrigation.

TABLE I. INSECT SYMPTOMS AND INSECTICIDES RECOMMENDATION

Insects Symptoms	Insect Diagnosed	Insecticides Recommendation
If location=underside of leaves and body color=yellowish and wing color=white then insectpest diagnosed.	IF ?insectpest= Whitefly	Then (?insecticides= Polo500SC ^ ?dose= 250ml) ^ (?insecticides= Confidor200SL ^ ?dose= 250ml) ^ (?insecticides= Mospilan200SP ^ ?dose= 5gm) ^ (?insecticides= Danitol30EC ^ ?dose= 200ml) And Use Neem Leaf Extract
If temp=warm and shape=spindle shaped and wings=elongated then Thrips diagnosed.	IF ?insectpest=Thrips	Then (?insecticides= Confidor200SL ^ ?dose= 80ml) ^ (?insecticides= Confidor70WS ^ ?dose= 5gm/kg seed) ^ (?insecticides= Mospilan 20SP ^ ?dose= 5gm) ^ (?insecticides= Thiodan 35EC ^ ?dose= 600ml)
If leaves curl downward=yes and color=yellowish then Jassid diagnosed	IF ?insectpest=Jassid	Then (?insecticides= Baythroid TM 525EC ^ ?dose= 100ml) ^ (?insecticides= Nurelle D 505EC ^ ?dose= 500ml)

TABLE II. WEEDS SYMPTOMS AND PESTICIDES RECOMMENDATION

Weeds Symptoms	Weeds Diagnosed	Herbicide Recommendation
If stem type=slender and structure=smooth and height=24 inch and leafcolor=yellowgreenandleaf structure=flat then sedge weed diagnosed	If ?weeds= sedges	Then ?herbicide=Stomp 330EC ?dose= 1000ml -50ml ?time = In drilling method

TABLE III. WORMS SYMPTOMS AND INSECTICIDES RECOMMENDATION

Worms Symptoms	Worms Diagnosed	Insecticides Recommendation
If symptoms =chewed holes and caterpillars=white and larvae= yellow then Cotton bollworm diagnosed.	If ?insect=American BollWorm	Then (?insecticides= Procalim 019EC ^?dose= 200ml) ^ (?insecticides= Larvin 80DF ^?dose= 450 gm) ^ (?insecticides= Tracer 240SC ^ ?dose= 80ml) ^ (?insecticides= Shogan1.8EC ^ ?dose= 250ml) ^ (?insecticides= Deltaphos 360EC ^?dose= 700 ml) (?insecticides= Match 050EC ^?dose= 800 ml)
If color= light brown and rain in August= Yes and Rain in September= Yes then Pink BollWorm diagnosed.	If ?insect=Pink BollWorm	Then (?insecticides= Deltaphos 360EC ^ ?dose= 600ml) ^2.5EC ^ ?dose= 400ml) ^ (?insecticides= Talstar10 EC ^ ?dose= 250ml) (?insecticides= Karate
if rainfall= high and time >=July and time <=September and wings= four and streak =one white then Spotted Boll Worms Diagnosed.	If ?insect=Spotted bollworms	Then (?insecticides= Deltaphos 36EC ^ ?dose= 600ml) ^ (?insecticides= Karate 2.5EC ^ ?dose= 400ml) ^ (?insecticides= Match 50EC ^ ?dose= 800ml) ^ (?insecticides= Talstar 10EC ^ ?dose= 250ml) ^ (?insecticides= Sumi Alpha 110EC ^ ?dose= 200ml)

TABLE IV. IRRIGATION SCHEDULING IN COTTON CROP BASED ON EXPERT SYSTEM

	Fact Regarding Scheduling	Irrigation Scheduling
1	If crop=cotton and area=1.5ha and growing season > =August and growing season <= December and Soil Texture=99mm/m Then	Then Irrigation method= Pressured piped surface method
2	If month=August and Time=beginning of August and pre sowing irrigation=0.6m	Then Irrigation After=2 days and irrigation=crop establishment
3	If month=August and Time=8 August	Then Irrigation After=2 days and irrigation=425m3
4	If month=August and Time=16 August	Then Irrigation After=2 days and irrigation=425m3
5	If month=August and Time=24 August	Then Irrigation After=2 days and irrigation=425m3
6	If month= September and Time=1 September	Then Irrigation After=5days and irrigation=891m3
7	If month= September and Time=1 September	Then Irrigation After=5days and irrigation=891m3
8	If month= September and Time=11 September	Then

		Irrigation After=5days and irrigation=891m3
9	If month= September and Time=22 September	Then Irrigation After=5days and irrigation=891m3
10	If month= October and Time=2 October	Then Irrigation After=5days and irrigation=75m3
11	If month= October and Time=11 October	Then Irrigation After=5days and irrigation=75m3
12	If month= October and Time=21 October	Then Irrigation After=5days and irrigation=75m3
13	If month= October and Time=31 October	Then Irrigation After=5days and irrigation=75m3
14	If month= November and Time=13 November	Then Irrigation After=5days and irrigation=75m3
15	If month= November and Time=26 November	Then Irrigation After=5days and irrigation=75m3

TABLE V. COMPARISON OF PROPOSED IOT BASED EXPERT SYSTEM WITH OTHER ES

Output	Testing	Recommendation	Diagnosis	Technology Used	Crops	Authors
User input symptoms	Tested by Experts	Providing Control Measures	Identifying Pests based on morphology of insects	Web based expert system and	Soybean	1.Sharma et al.[32]
User input different symptoms	Nil	Provide diagnosis about diseases, symptoms, chemical control and	Provide information about diseases, symptoms, chemical control and	Web based expert system ID3 , algorithm , optimization algorithm	Tomato	2. Pradesh et al.[33]
Nil	Nil	Nil	Diagnosis in agriculture	Ontology based expert system , CSS, PHP	Framework for Agriculture	3. Qirui et al.[34]
User input different symptoms	Nil	Recommendation for wheat diseases	Diagnosis of diseases in wheat	Web Based Expert System	Wheat	4.Fahad et al.[35]
User input different symptoms.	Tested by agriculture engineers, researchers and extension officers.	Provide recommendation about variety selection, land preparation,	Used for variety selection, land preparation, irrigation, planting,	Small talk	Wheat	5. Edrees et al.[36]
User Input different symptoms.	System is validated and verified.	It provides recommendation for the diseases.	Used for diagnosis and identification of diseases.	Microsoft Visual Basic	Cereals	6. Andujar et al.[37]
User input different symptom of diseases.	System is validated and verified.	It provides picture based diseases recommendation.	Used for diagnosis of diseases.	Web based and fuzzy logic based approach.	Oilseeds	7. Kolhe et al.[9]

8. Marchal et al.[38]	Oliveoil	Computer vision based technique for the diagnosis of diseases.	Used for the diagnosis of diseases based on images	It classifies the diseases into different categories.	System is tested and verified in chemical lab.	Cameras capture image under the static environment in lab.
9. Kaloudis et al.[39]	Identification of insects in forests.	Expert system developed in EXSYS Professional, Ver. 5.1.0,	Predicting of attack of insects,	It diagnosis different insects categories.	System is tested and verified by multiple users.	User input characteristic and symptoms of insects.
10. Mansinghet all.[40]	Coffee	Expert system developed	Diagnosis of diseases and pests in coffee	It diagnosis diseases provides appropriate solutions.	System is tested and verified by multiple users.	User Input characteristic and symptoms of diseases and pests.
11. Andujar et al.[9]	Olive Crops	A standalone expert system	Identification of harmful organisms like insects , diseases and weeds	It provides images after the identification of insects.	System has verified and validated by technician and students.	User manually input different characteristics and symptoms of diseases.

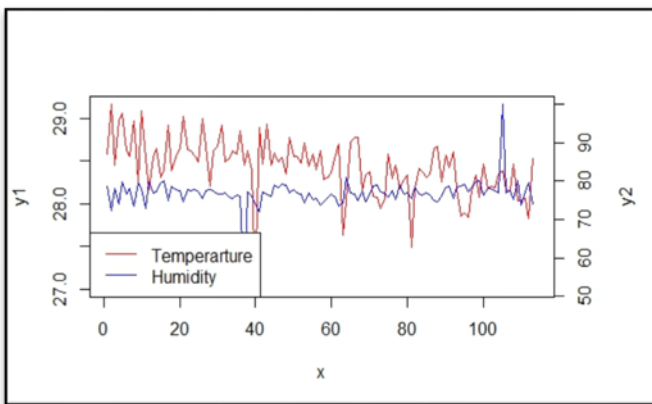


Fig. 4. Comparison of Temperature Sensor and Humidity Sensors Data

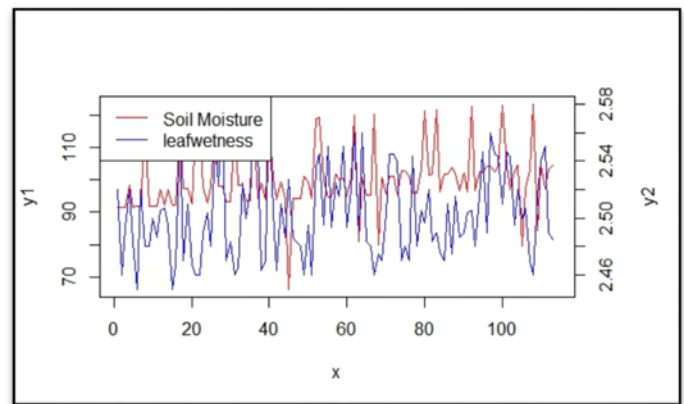


Fig. 5. Comparison of Soil Moisture Sensor and Leaf Wetness Sensor Data

Fig 5 represents the relationship between soil moisture and leaf wetness data; these are directly proportional to each other. If the soil moisture increases then leaf wetness will increase automatically.

VI. VALIDITY OF PROPOSED SOLUTION

Our proposed solution is not so much costly as one may think due to the costly deployments of the sensors and actuators in the field. It is a fact that farmers invest heavily on electricity, fertilizers, insecticides, and pesticides. They are using these resources efficiently because they are unaware of the actual needs of the cotton crop. This investment can be minimized by one-time investment on sensors. The sensors and cameras can monitor the cotton crop 24/7 and provide input for

proactive actions and optimal use of resources like water, fertilizer, insecticides and pesticides. Before deploying the IoT-based ES, we conduct the survey from users either they are willing to accept the IoT-based ES for agriculture. In our survey 100 different respondents like farmers, experts of agriculture participate in it. Farmers in Pakistan are illiterate, so we conduct surveys to ensure framework feasibility of its implementation, to evaluate the need of the system, after effects and their tradeoff while using IoT-based ES.

VII. CONCLUSION

In this paper, we have presented an ES for Cotton crop based on the concept of IoT. We tried to develop an initial frame for IoT-based agriculture. We developed an IoT-based ES. It's based ES consists of three modules; the first part consists of the deployment of WSN in the cotton fields. WSN has used for the monitoring of the cotton crop condition. The Waspnote agriculture sensor board has used for the monitoring of the cotton crop condition. It consists of temperature sensors, humidity sensors, leaf wetness sensors and soil sensors.

In the concept of the IoT, the server should send the commands to the actuators of the fields, so the actuators of fields can take appropriate decisions. The sever should be intelligent enough to take decisions appropriately. For this purpose, we deploy an ES so that it can make decisions automatically. In Table 5 describes that different ES developed during a different era, but in this paper, we combine the IoT and ES. The sensors send the data to the server on the server side; we deployed the ES, which process and analyzes the sensor data. The data is fed to the ES that analyses it using the knowledge base and produces findings and recommendations. The ES consists of user interface, knowledge base and inference engine. On the server side, we deploy the concept of smart irrigation. Sensors monitor the soil moisture, leaf wetness, temperature, and humidity level in the environment and send the recommendation to the farmer about the irrigation in the cotton crop. In this paper, we ES for identification of different weeds, pests and different insects which attack on the cotton crop. These findings are sent to the farmer's mobile phone for taking necessary actions in the field. We proposed an initial framework for the working of Smart Agriculture (SA).

Before developing the concept of SA we conduct the survey and ask form user either the proposed system will be accepted by users or not. In this survey, we also ask from farmers and experts the flaws of the current system and whether they are satisfied the working of current agriculture or not. After that proposed system was evaluated by 100 different users like farmers and experts of Agri domain and 65 percent of respondents are satisfied with the working of SA and they are willing to accept the concept of SA. As we know farmers are illiterate in Pakistan so that we get 65 percent results. By deploying the IoT-based ES the productivity rate of the cash crops can be increased and problem of farmers also be reduced The proposed system was evaluated by 100 experts from the field and was found helpful for the farmers.

VIII. FUTURE WORK

In this paper, we present an initial framework for the diagnosis weeds, insects and different pests in cotton crop. We

also deployed the concept of Smart Irrigation in cotton crop. For this purpose, we deploy the ES.

In the future we will try to deploy the actuators in the fields and we enhance the functionality of server by deploying genetic algorithm, artificial neural network and digital image processing techniques on the server. We can diagnose the diseases in a better way if we deploy the cameras in the fields.

REFERENCE

- [1] Morais, Raul, A. Valente, and C. Seródio. "A wireless sensor network for smart irrigation and environmental monitoring: A position article." In 5th European federation for information technology in agriculture, food and environment and 3rd world congress on computers in agriculture and natural resources (EFITA/WCCA), pp.45-850. 2005.
- [2] Agrawal, Sarita, and Manik Lal Das. "Internet of Things—A paradigm shift of future Internet applications." In Engineering (NUICON), 2011 Nirma University International Conference on, pp.1-7. IEEE, 2011.
- [3] Hu, Xiangyu, and Songrong Qian. "IoT application system with crop growth models in facility agriculture." In 2011 6th International Conference on Computer Sciences and Convergence Information Technology ICCIT. 2011.
- [4] Li, Li, Hu Xiaoguang, Chen Ke, and He Ketai. "The applications of WiFi-based wireless sensor network in internet of things and smart grid." In Industrial Electronics and Applications ICIEA, 2011 6th IEEE Conference on, pp. 789-793. IEEE, 2011.
- [5] Tuli, Anupriya, Nitasha Hasteer, Mukesh Sharma, and Ankur Bansal. "Framework to leverage cloud for the modernization of the Indian agriculture system." In Electro/Information Technology (EIT), 2014 IEEE International Conference on, pp. 109-115. IEEE, 2014.
- [6] Liu, Yuxi, and Guohui Zhou. "Key technologies and applications of internet of things." In Intelligent Computation Technology and Automation (ICICTA), 2012 Fifth International Conference on, pp. 197-200. IEEE, 2012.
- [7] Zhao, Ji-chun, Jun-feng Zhang, Yu Feng, and Jian-xin Guo. "The study and application of the IOT technology in agriculture." In Computer Science and Information Technology ICCSIT, 2010 3rd IEEE International Conference on, vol. 2, pp. 462-465. IEEE, 2010.
- [8] Jhuria, Manoj, Ajit Kumar, and Rushikesh Borse. "Image processing for smart farming: Detection of disease and fruit grading." In Image Information Processing (ICIIP), 2013 IEEE Second International Conference on, pp.21-526. IEEE, 2013.
- [9] González-Andújar, José Luis. "Expert system for pests, diseases and weeds identification in olive crops." *Expert Systems with Applications* 36, no. 2,pp 3278-3283 ,2009.
- [10] TongKe, Fan. "Smart Agriculture Based on Cloud Computing and IOT." *Journal of Convergence Information Technology* 8, no. 2 ,2013.
- [11] Chen, Xian-Yi, and Zhi-Gang Jin. "Research on key technology and applications for internet of things." *Physics Procedia* 33, pp. 561-566, 2011.
- [12] Talpur, Mir Sajjad Hussain, Murtaza Hussain Shaikh, and Hira Sajjad Talpur. "Relevance of Internet of Things in Animal Stocks Chain Management in Pakistan's Perspectives." *International Journal of Information and Education Technology* 2, no. 1 ,2012.
- [13] Evangelos A, Kosmatos, Tselikas Nikolaos D, and Boucouvalas Anthony C. "Integrating RFIDs and smart objects into a UnifiedInternet of Things architecture." *Advances in Internet of Things* 2011 ,2011.
- [14] Carvin, Denis, Philippe Owezarski, and Pascal Berthou. "Managing the upcoming ubiquitous computing." In Proceedings of the 8th International Conference on Network and Service Management, pp. 1276-280. International Federation for Information Processing, 2012.
- [15] Prasad, Rajkishore, Kumar Rajeev Ranjan, and A. K. Sinha. "AMRAPALIKA: An expert system for the diagnosis of pests, diseases, and disorders in Indian mango." *Knowledge-Based Systems*, Vol. 19, no. 1,pp. 9-21,2006.
- [16] Sarma, Shikhar Kr, Kh Robindro Singh, and Abhijeet Singh. "An Expert System for diagnosis of diseases in Rice Plant." *International Journal of Artificial Intelligence*, Vol. 1, no. 1 ,pp. 26-31,2010, ,

- [17] Ballea, Kaliuday, D. Satyanvesh, N. V. S. S. P. Sampath, K. T. N. Varma, and P. K. Baruah. "Agpest: An efficient rule-based expert system to prevent pest diseases of rice and wheat crops." In Intelligent Systems and Control ISCO, 2014 IEEE 8th International Conference on, pp. 262-268. IEEE, 2014.
- [18] Negid, N. K. "Expert System for Wheat Yields Protection in Egypt ESWYP." International Journal of Innovative Technology and Exploring Engineering IJITEE, ISSN ,pp. 2278-3075,2014.
- [19] Kaura, Ramanjeet, Salam Dina, and P. P. S. Pannub. "Expert System to Detect and Diagnose the Leaf Diseases of Cereals." Int J of Current Engineering and Technology , Vol. 3, no. 4 , 2013.
- [20] Kortuem, Gerd, Fahim Kawsar, Daniel Fitton, and Vasughi Sundramoorthy. "Smart objects as building blocks for the internet of things." Internet Computing, IEEE 14, no. 1 ,pp.~ 44-51,2010.
- [21] Coetzee, Louis, and Johan Eksteen. "The Internet of Things-promise for the future? An introduction." In IST-Africa Conference Proceedings, 2011, pp. 1-9. IEEE, 2011.
- [22] Ashton, Kevin. "That 'internet of things' thing." RFID Journal, Vol. 22, no. 7 ,pp.97-114,2009.
- [23] Chase, Jim. "The evolution of the internet of things." Texas Instruments ,2013.
- [24] Travis, J. W., E. Rajotte, R. Bankert, K. D. Hickey, L. A. Hull, V. Eby, P. H. Heinemann, R. Crassweller, and J. McClure. "Penn State apple orchard consultant: design and function of the pest management module." In III International Symposium on Computer Modelling in Fruit Research and Orchard Management 313, pp. 209-214. 1992.
- [25] Salah, A., H. Hassan, K. Tawfik, I. Ibrahim, and H. Farahat. "CITEX: an expert system for citrus crop management." In Proceedings of the Second National Expert Systems and Development Workshop ESADW-93. 1993.
- [26] El-Dessouki, A., S. Edrees, and S. El-Azhari. "CUPTEX: An integrated expert system for crop management of cucumber." ESADW-93, May ,1993.
- [27] Mohammed, M., K. El-Arby, and A. Rafea. "LIMEX: an integrated expert system for lime crop management." In Second IFAC/IFIP Enr AGENG Workshop on AI in Agriculture, Netherlands. 1995.
- [28] Plant, R. E., F. G. Zalom, J. A. Young, and R. E. Rice. "CALEX/Peaches, an expert system for the diagnosis of peach and nectarine disorders." HortScience ,Vol. 24, no. 4 ,1989.
- [29] Ferguson, James J., Fedro S. Zazueta, and Juan I. Valiente. "Citpath: diagnostic and hypertext software for fungal diseases of citrus foliage and fruit." HortScience,Vol. 30, no. 4 ,pp.~ 899-899,1995.
- [30] Rani, Mercy Nesa, and Thangaswamy Rajesh. "Comparative Analysis on Software's used in Expert System with Special Reference to Agriculture." MANAGEMENT, Vol.2, no. 8.
- [31] Saini, Harvinder S., Raj Kamal, and A. N. Sharma. "Web based fuzzy expert system for integrated pest management in soybean." International Journal of Information Technology ,Vol. 8, no. 1 ,pp.55-74, 2002.
- [32] Babu, MS Prasad. "A web based tomato crop expert information system based on artificial intelligence and machine learning algorithms." ,2010.
- [33] Qirui, Yin. "Kaas-based intelligent service model in agricultural expert system." In Consumer Electronics, Communications and Networks CECNet, 2012 2nd International Conference on, pp. 2678-2680. IEEE, 2012.
- [34] Khan, Fahad Shahbaz, Saad Razzaq, Kashif Irfan, Fahad Maqbool, Ahmad Farid, Inam Illahi, and T. Ul Amin. "Dr. Wheat: a Web-based expert system for diagnosis of diseases and pests in Pakistani wheat." In Proceedings of the World Congress on Engineering, Vol. 1, pp. 2-4. 2008.
- [35] Edrees, Soliman A., Ahmed Rafea, Ibrahim Fathy, and Mohamed Yahia. "NEPER: a multiple strategy wheat expert system." Computers and electronics in agriculture , Vol. 40, no. 1,pp. 27-43,2007.
- [36] Gonzalez-Andujar, J. L., C. Fernandez-Quintanilla, J. Izquierdo, and J. M. Urbano. "SIMCE: An expert system for seedling weed identification in cereals." Computers and electronics in agriculture,Vol. 54, no. 2 ,pp.115-123,2006.
- [37] Marchal, P. Cano, D. Martínez Gila, J. Gámez García, and J. Gómez Ortega. "Expert system based on computer vision to estimate the content of impurities in olive oil samples." Journal of Food Engineering , Vol.119, no. 2 ,pp. 220-228,2013.
- [38] Kaloudis, S., D. Anastopoulos, Constantine P. Yialouris, Nikos A. Lorentzos, and Alexander B. Sideridis. "Insect identification expert system for forest protection." Expert Systems with Applications, Vol. 28, no. 3, pp. 445-452,2005.
- [39] Mansingh, Gunjan, Han Reichgelt, and Kweku-Muata Osei Bryson. "CPEST: An expert system for the management of pests and diseases in the Jamaican coffee industry." Expert systems with Applications ,Vol.32, no. 1,pp.~ 184-192,2007.

Dependency Test: Portraying Pearson's Correlation Coefficient Targeting Activities in Project Scheduling

Jana Shafi

Department of Computer Science,
Prince Sattam bin Abdul Aziz
University,
Saudi Arabia

Amtul Waheed

Department of Computer Science,
Prince Sattam bin Abdul Aziz
University,
Saudi Arabia

Sumaya Sanober

Department of Computer Science,
Prince Sattam bin Abdul Aziz
University,
Saudi Arabia

Abstract—In this paper, we discuss project scheduling with conflicting activity-resources. Several project activities require same resources but, may be scheduled with the certain lapse of time resulting in repeatedly using the same kind of resources for executing dissimilar activities. Due to the frequent usage of same resources multiple times, expenditure become more expensive and project duration extends. The problem is to find out such kind of activities which are developing implicit relations amid them. , we proposed a solution by introducing TVs (Transparent view of Scheduling) model. First, we analyze and enlists activities according to required resources, categorize them and then we segregate dependent and independent activities by indicating a value. Performing Dependency test on activities by using Pearson's Correlation Coefficient (PCC) to calculate the rate of relations among the ordered activities for similar resources. By using this model we can reschedule activities to avoid confusion and disordering of resources without consumption of time and capital.

Keywords—TVS; Transparent; Dependency; PCC; Activity; Resource; Schedule; Project Introduction

I. INTRODUCTION (MOTIVATION)

With the development of project management system in order to direct the project schedules along with the activities and various resources, project scheduling becomes the helpful tool for standardizing and ordering activities as well resources according to specifications [19, 20]. However, project schedules returned by popular, methods are not satisfactory .Sometimes much of the resources are linked with the activities which have nothing to do with them .It is a remarkable part for the developers to pick resources in order to get an appropriate activity. As we have experience in using project schedules every day, the result set returned by estimation of activity resource set is really too big and mostly and merely useless. The relationship between the activities is obvious to users or managers, while it is not for the project schedules [21].

Late delivery of software projects results in huge loss of manpower, industrial efforts and money which discouraged our software industry less or more for accepting challenges for successful projects in recent years[1].Developers, Engineers, and researchers gave a thought of planning all parameters of software management in an organized manner so that the modules of the project become transparent with due time. This outcome called as Project scheduling which minimizes the failure of projects and encourages workers morale[1][2]. For Complex projects mostly engineering tasks take place in

parallel so one work may be interdependent in another work or task. These interdependencies can be understood by schedules only.

An activity – Must have a clear start and a clear stop – Must have a duration that can be forecasted – May require the completion of other activities before it begins – should have some ‘deliverables’ for ease of monitoring.

Resources- they are used to accomplish the project activities. Resources are classified in vast areas such as manpower, finance, and budgeting, inventories, maintenance and services are required for the completion of the task in limited time.

In order to show where the problem is we input the following activity and resources: Activity1”Questionnaire: Public” Activity2 “Demonstrating new software: Public Resources: Vehicles, Chauffer, Maps, Locations.

TABLE I. EXAMPLE OF ACTIVITY RESOURCE SCHEDULING

S.No	Activity	Resources	Dependents
1	Questionnaire Public	Location,Vehical,Chauffer, Map	Yes
2	Demonstrating New Software Public	Location,Vehical,Chauffer, Map	Yes
3	System Configuration	Software and Hardware	No

We can exemplify from above(table1:) that as both the activities requires some common resources for example Vehicles, Chauffer, Maps, and Locations which set both the activities having invisible relation. Let’s analyze the aforesaid activities. We want resources in order to accomplish activity1” Questionnaire: Public” which requires “Vehicles, Chauffer, Maps, Locations” and activity2 “Demonstrating new software: Public” requires the same kind of resources as activity1 as mention above. However, both the activities are totally different.

Colliding resources with more than one activity leads an invisible relation among activities. Colliding resources is the key to the whole problem in prioritizing the activities.

Nothing in this boundless universe exists independently. All objects are related to other activities in various means. We comprehend this activity from the way it relates to other

activities. Regard activity1 "Questionnaire" and activity2 "Demonstration" one of the relations between them are resources "Vehicles, Chauffer, Maps, Locations" in this context. For better understanding the relationships between activities have to be defining before the developers, a tester understands the semantic of each other. In the project management system, schedules are presented by the activity estimation of resources for example "System Configuration" an activity presented by "IT block" resources they dealt with activity resource combinations. However, it is not always necessary that all activities must be related to each other or with any other kind of activities. Activities which carry out in an individual manner are called isolated activities.

II. RELATED WORK

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

With the development of Project Scheduling, hundreds of methods, ways established algorithms. Project scheduling in the recent years observed along heuristics, constraint-resources, metaheuristics, and resource-based constraints, consistency tests are furnished[4,14,18]. E.L. Demeulemeester and W.S. Herroelen study and display depth knowledge of project scheduling by several algorithms and classify them accordingly for better practical examples are computed [24].

J.Alcaraz focused Genetic Algorithm counting resource allocations[10].Dale.F.Cooper suggested project scheduling a problem with multiple constrained resources with an experimental investigation with a set of the project their characteristics is scheduled by each of these heuristics with a variety of priority rule[8,17]. J-H-Cho and Y-D Kim emerges with another simulated annealing algorithm for resource constrained project scheduling proble[9].Christian Artigues presented the flow network model for static and dynamic resource-constrained project scheduling[13,15].Peter Brucker encompasses notations, classifications, models and methods of project scheduling problems[11,12,16]. And R.Kolisich compiled a survey on deterministic project scheduling remarking net present value maximization and make span minimization[7].The search algorithm for the resource constrained project scheduling problem with an interval is defined with a solution by representing resource flows extending the disjunctive graph model for shop scheduling problems by Poppenborg and Kust in 2016.

Execution of task by generalizing precedence relationships and assigning resources for completion of the task in given deadline is deliberated by Bianco et al. for defining a problem of leveling resources[25].To decrease the resource consumption a method is introduced by author Naffaf et

al.(2016) which consist of three mixed-integer mathematical programs and an adapted satisfiability test for the cumulative constraint.

Pearson correlation coefficient is a formula designated in a statistical test between two variables to determine how strong that relationship is. The value "0" indicates no linear relationship and value "1" indicates a positive linear relationship and "-1" indicates a negative relationship. R denotes Pearson correlation coefficient value [22 , 23].

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n \sum x^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}} \quad (1)$$

III. PROBLEM STATEMENT

The main issue address in this paper is how to identify the relations between the various activities which carried out in project scheduling later in project development. The resources are viewed as independent or interdependent. Currently, a challenge is to know when prioritizing resources using an activity, how many activities are implicitly interconnected to each other which should be recognized to identify the dependent activities.

IV. PROPOSED SYSTEM

In this paper, we concentrate on the framework of activities along with computation and manipulation in order to identify, sort as well categorize dependent and independent activities.

A. Paper Organization

In the paper Section IV will introduce the system of "Transparent view of scheduling"TVS , depicts its various classification and its formal model.Section V & VI implements TVs and manipulates results via Pearson correlation coefficient.

Schedule's point of view can be understood by TVS Model which is able to provide a transparent view of interdependent and independent tasks. TVS provides a computational framework for network activities. TVS abandons complication at the workplace. Thus in TVS everything (activity) belongs to some or the other category which separates various activities from to be get confused or left behind. As we are using mathematical computation to be more focused on points

With each task efforts and duration of time are allocated and thus a task is a part of a network that aware the software team to meet the product delivery deadline.

Fig.1 proper scheduling is essential for the project which an experienced team can do [1] and must include

- a) Tasks must be created inside the network as shown in.
- b) Efforts and timings are allocated to each task
- c) Interdependencies between tasks must be transparent
- d) Resources must be allocated for the targeted work



Fig. 1. Activity Network

Work done for both optimistic and pessimistic scheduling in order to get more realistic parameters for the project to proceed.

Fig.2 Transparent view of scheduling enables specification viewers as well as programmers and clients to instantly associate their activity in concerning categories to carry out process interdependently or independently according to the requirements and resources.

TVS includes

a) Outsource Activities-Those activities which include all external resources,external components,external behavior and resources other than usual which are or will affect our activities in near or far future.For Example:Power-shortage,Politics,Rate ofInflation etc

b) Insider-Activities which can be completed within organization including manpower, resources, and Coordination.(internal resources)

c) Dependent-Some insider and outsource activities cannot be completed without each other. These are the proportional activities which change with the variance of each other.

For example:-Detail Marketing Plans sales tax, Bond insurance etc

For example:Clients,Distributors,advertisement etc.

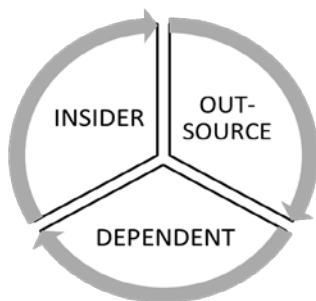


Fig. 2. Transparent view of scheduling

V. MODALITIES WITH EXAMPLE

A. As we already know that we have TVS classes as Inside,Outsider,Dependent which we abbreviated for our convenience in the following way.

- Inside-I_n,
- OutSource-O_s,
- Dependent-D_e

B. Formal Model

Figure.3 portrays how the activities are implicitly dependent on each other for resources.

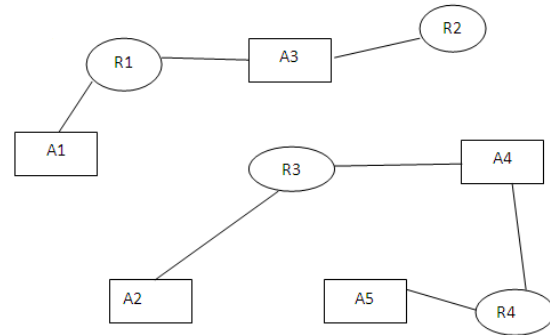


Fig. 3. Resource-Activity Relationship Model

Definition 1: A keyword ‘I_s’ is a set of activities which Include internal resources.

Definition 2: A concept of activities C in a given schedule is Presented by a square vertex of the graph.

Definition 3: A relation collection in a given domain R is a Set of related activities. It represented by arch of the graph.

Definition 4: A graph of an activity relation G is a set of Vectors of the form (A, R) where A is activity Set and is a relation.

Definition 5: An activity relation subgraph G_p is a subgraph of G.

Definition 6: Keyword ‘O_s’ is a set of activities belongs to outside resources.

Definition 7: An activity-resource pair set RksetI_sO_s is a set of activities of the form(Kis,Kos,Risos) where Kis and Kos where Kis refers to Insider,Kos-Outsource Rqisos-is a relation between Ki and Ko.

Definition 8:An activity-resource pair candidate set CRksetp is a set an activity-resource pair set and every Rqisos presents an arc of G_p .

Definition 9:A result set U(Kis,Kos,Rqisos is the returned Dependency value when we correlate an the activity of CRksetp into TVS.

C. Demonstration

We are going to demonstrate how to classify activities and to get a correlation coefficient of ‘I_n’ and ‘O_s’.In figure 3,

-0.933	- 0.200	0.871	0.040	0.187
-0.933	- 0.200	0.871	0.040	0.187
M_X : -0.933	M_Y : -0.200	SUM: 1.867	SUM: 4.800	SUM: 0.400

B. Key

- a) I_s : I_s Values
- b) O_s : O_s Values
- c) M_I : Mean of I_s Values
- d) M_O : Mean of O_s Values
- e) $I_s - M_I$ & $O_s - M_O$: Deviation scores
- f) $(I_s - M_I)^2$ & $(O_s - M_O)^2$: Deviation Squared
- g) $(I_s - M_I)(O_s - M_O)$: Product of Deviation Scores

C. Result Details & Calculation

I_s Values

$$\sum = 28$$

$$\text{Mean} = 0.933$$

$$\sum(I_s - M_I)^2 = SSI = 1.867$$

O_s Values

$$\sum = 6$$

$$\text{Mean} = 0.2$$

$$\sum(O_s - M_O)^2 = SSO = 4.8$$

I_s and O_s Combined

$$N = 30$$

$$\sum(I_s - M_I)(O_s - M_O) = 0.4$$

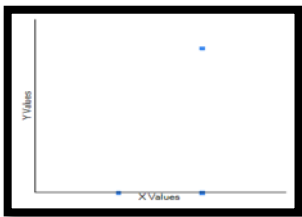
R Calculation

$$r = \frac{\sum((I_s - M_I)(O_s - M_O))}{\sqrt{((SSI)(SSO))}}$$

$$r = 0.4 / \sqrt{((1.867)(4.8))} = 0.1336$$

Meta Numeric (cross-check)

$$r = 0.1336$$



Depicting Correlation Graph

Result clarifies that insider is not dependent on outsource And vice-versa.

The value of R is 0.1336. Although technically a positive correlation, the relationship between variables is weak (nb. the nearer the value is to zero, the weaker the relationship).The value of R², the coefficient of determination, is 0.0178.

D. Performance Analysis

a) First, the five text boxes represent the calculations that would be required if you were to calculate the R value in stages.

b) Second, there is more than one way to calculate the R value, but these are all mathematically equivalent.

c) Third, in the "Result Details & Calculations" box, you'll find what we've called a cross-check value, which is the R value calculated using an algorithm supplied by the [Meta Numeric](#) statistical library. This should be identical to the value that we've calculated.

d) Forth, In the graph Depicting Correlation also we showed you that as the result is zero means the Dependability between Insider and Outsider activities is not proved hence no correlation.

e) Hence this experiment proved that how we can weight our activities as well classify them as dependent or independent

VII. APPLICATION

The concept can be useful to accurately correlate activities according to their nature which gives perfect scheduling/sequence to carry out in an organization.as well activities can be assigned independently or dependently to the team or individuals according to the requirements.

PCC Result directly classifies the existence of activities correlation by computing values positive , negative and neutral.PCC result is represented in graphical view for enhanced perceptive of determining dependent or independent activities.

TVs model is efficient in communication among team leaders who are responsible for project modules .This also helps in controlling cost as usage of resources is clearly defined.The concept is useful in managing changes in schedules which become an easy job and it can adjust according to the environment, requirements and time.Project Schedule ,module's responsibilities are simply assigned and allotted with a time limit for carrying out the project on time. TVs serves as a virtual model to track progress according to plan , measure quality , cost control in real-time schedules. Project schedules can be structured in a hierarchical form which enables in easily identifying the relationship among activities. The model portrays flow of activities and resources simultaneously which assist in debugging errors.

VIII. SUMMARY & CONTRIBUTION

In the project scheduling, estimation of activity-resource is recorded and prioritize. We call the activity defining the resource and its relations among various activities.Here the question arises is how to define the activity resource? Whether the relation should be activity -activity- resource? The answer is no.

The paper proposes "TVS" activity-resource-activity correlation base scheduling since it takes an advantage of activity-resource estimation and achieves whole correlation value of activities. Defining schedule of activity resource can be standardized by correlating them and weighing them.

Initially, we explore and procure activities according to essential resources, sort out and segregate Insider (dependent) and outsource(independent)activities by denoting with a value 0 for existence and 1 for non-existence.

Performing Dependency test on activities by using Pearson's Correlation Coefficient (PCC) to calculate the rate of relations among the ordered activities for similar resources.

The Key idea in Project scheduling is the sequence of events according to their category Insider(I_s), Outsider(O_s) and Dependent(De). This paper presents a close view of complex, time-taking Project Scheduling activities, an approach through which many activities can be weighed accordingly and we can easily correlate the dependency.

IX. FURTHER FUTURE WORK

Project development, scheduling and various accurate operations can be improve and manipulate in future work. The transparent view of scheduling can lead to an era of automatic generalizing events and sequences of Project scheduling engrave with computations for accuracy.

REFERENCES

- [1] Software Engineering: A Practitioner's Approach, 6/e Roger S Pressman, R.S. Pressman and Associates
- [2] Software Engineering:-Ivan Marsic,Rutgers,the state University of New Jersey,September 10,2012
- [3] Pearson product-moment correlation coefficient from Wikipedia, the free encyclopedia
- [4] V. Ahuja V. Thiruvengadam, (2004),"Project scheduling and monitoring: current research status", Construction Innovation, Vol. 4 Iss 1 pp. 19 – 31
- [5] Interpretation of the Correlation Coefficient: A Basic Review RICHARD TAYLOR, EDD, RDCS JDMS 1:35-39,January/February 1990
- [6] A survey of variants and extensions of the resource-constrained project scheduling problem, Sönke Hartmann,1,DirkBriskomb,doi:10.1016/j.ejor.2009.11.005
- [7] An integrated survey of deterministic project scheduling, Volume 29, Issue 3, June 2001, Pages 249–272 R Kolischa, , R Padmanb, , .
- [8] Heuristics for Scheduling Resource-Constrained Projects: An Experimental Investigation Dale F. Cooper Royal Holloway College (University of London),Published Online: July 1, 1976 Page Range: 1186 – 1194Journal of the Operational Research Society(1997)48,736–744.
- [9] A simulated annealing algorithm for resource constrained project scheduling problemsJ-H Choi and Y-D Kim2February 2001, Volume 102, Issue 1-4, pp 83-109
- [10] A Robust Genetic Algorithm for Resource Allocation in Project Scheduling J. Alcaraz, C. Maroto
- [11] Resource-constrained project scheduling: Notation, classification, models, and methods Peter Bruckera, 1, 1, Andreas Drexlb, , Rolf Möhringc, 2,2,Klaus Neumannd, 3, 3, Erwin Pesche, 4, 4
- [12] Bianco L, Caramia M (2011) A new lower bound for the resource-constrained project scheduling problem with generalized precedence relations. *Comput Oper Res* 38:14–20
- [13] Cavalcante CCB, de Souza CC, Savelsbergh MWP, Wang Y, Wolsey LA (2001) Scheduling projects with labor constraints. *Discrete Appl Math* 112(1–3):27–52
- [14] de Reyck B, Demeulemeester EL, Herroelen WS (1999) Algorithms for scheduling projects with generalized precedence relations. In Węglarz, pp 77–106
- [15] Drezet L-E, Billaut J-C (2008) A project scheduling problem with labour constraints and time-dependent activities requirements. *Eur J Oper Res* 112((1):217–225
- [16] Hartmann S (1999) Project scheduling under limited resources: models, methods, and applications. Number 478 in *lecture notes in economics and mathematical systems*. Springer, Berlin
- [17] Hartmann S, Briskorn D (2010) A survey of variants and extensions of the resource-constrained project scheduling problem. *Eur J Oper Res* 207:1–14
- [18] Klein R (2000) Project scheduling with time-varying resource constraints. *Int J Prod Res* 38:3937–3952
- [19] Węglarz, J (eds) (1999) Project scheduling: recent models, algorithms, and applications. Kluwer, Dordrecht
- [20] Editorial “Project Management and Scheduling” Rainer Kolisch1 · Erik Demeulemeester2 · Rubén Ruiz Garcia3 · Vincent T Kindt4 · Jan Węglarz5 Published online: 3 March 2016 © Springer-Verlag Berlin Heidelberg 2016
- [21] J. Józefowska, M. Mika, R. Różycki, G. Waligóra, and J. Węglarz, “A heuristic approach to allocating the continuous resource in discrete-continuous scheduling problems to minimize the makespan”, *Journal of Scheduling* 5 (6), 487–499 (2002).
- [22] Park, E., and Lee, Y. J., 2001, Estimates of the standard deviation of Spearman's rank correlation coefficients with dependent observations: *Comm. Statist. Simul.*, v.30, no.1, p. 129-142.
- [23] Robinson, P. M., 1977, Estimation of a time series model from unequally spaced data: *Stochast. Proc. Appl.*, v.6, no.1, p. 9-24.
- [24] E.L. Demeulemeester and W.S. Herroelen, *Project Scheduling – A Research Handbook*, Kluwer, Boston, 2002
- [25] Bianco L, Caramia M, Giordani S (2016) Resource leveling in project scheduling with generalized precedence relationships and variable execution intensities. doi:10.1007/s00291-016-0435-1 (this issue)

Comparison of Digital Signature Algorithm and Authentication Schemes for H.264 Compressed Video

Ramzi Haddaji

Electrical department

National Engineering School of Monastir, University of
Monastir, Tunisia

Laboratory of Electronic and Microelectronic, University of
Monastir, Tunisia
Monastir, Tunisia

Samia Bouaziz

Electrical department

National Engineering School of Monastir, University of
Monastir, Tunisia

Laboratory of Electronic and Microelectronic, University of
Monastir, Tunisia

Raouf Ouni

Mathematical department

National Engineering School of Monastir, University of
Monastir, Tunisia

Faculty of sciences, Tunis El-Manar University
Tunis, Tunisia

Abdellatif Mtibaa

Electrical department

National Engineering School of Monastir, University of
Monastir, Tunisia

Laboratory of Electronic and Microelectronic, University
of Monastir, Tunisia
Monastir, Tunisia

Abstract—In this paper we present the advantages of the elliptic curve cryptography for the implementations of the electronic signature algorithms “elliptic curve digital signature algorithm, ECDSA”, compared with “the digital signature algorithm, DSA”, for the signing and authentication of H.264 compressed videos. Also, we compared the strength and add-time of these algorithms on a database containing several videos sequences.

Keywords—*Elliptic curve cryptography; H.264; DSA (Digital signature algorithm); ECDSA (Elliptic Curve Digital Signature Algorithm); Implementation*

I. INTRODUCTION

The media industry has witnessed a phenomenal and unprecedented explosion in the recent decade. Communication, technology and media have transcended all boundaries, and the entire global community seems to have been brought together into one unified whole. Therefore in this era of evolving communication, different types of business related to media such as IPTV, Voice IP and videoconferencing, have also found solid grounds, these must be secured to protect privacy and to prevent from hackers [1].

Certain implementation security aspects of video are authentication, data integrity and confidentiality.

Authentication is the act of verifying a claim of identity.

Data integrity in information security means maintaining and assuring the accuracy and completeness of data over its entire life-cycle. This means that data cannot be modified in an unauthorized or undetected manner.

Confidentiality is the property, that information is not made available or disclosed to unauthorized individuals, entities, or processes [2].

The multimedia information including video data has some special characteristics like high capacity, redundancy and high correlation among pixels which leads us to choose the type of video encoding on which we will work. This brings us to use H.264 given the advantage that provides this type as size standpoint and video quality [3].

In this paper we focus our work in the authentication aspect which is verified using the signature algorithms. We compare the implementation of the most known two signature algorithms DSA, digital signature algorithm, and ECDSA, elliptic curve digital signature algorithm [4]-[5].

As this type of data requires memory space, the process of electronic signature is not used directly on the video but rather on what we call the hash of this one. A hash function known also a one-way function is a cryptographic tool which produce a fixed size fingerprint regardless of the size of the input [4].

The remaining of this paper is organized as follows. In Sect. 2, we recall properties and give some example of hash functions. In section 3 and 4, we describe the signature algorithms DSA and ECDSA. H.264 encoding is briefly described in section 5. Performance evaluation and comparative results of our implementation are given in detail in Sect. 6. Finally, some conclusions are made.

II. HASH FUNCTION

Cryptographic hash function plays an important role in the world of cryptography. They are employed in many applications for digital signatures, data integrity, message authentication, and key derivation. Secure Hash Algorithm (SHA-1) specifies which generates condensed of message called message digest. Hash functions takes a message of variable length as input and produce a fixed length string as output referred to as hash code or simply hash of the input message. The basic idea of cryptographic hash function is use

of hash code as compact and non ambiguous image of message from which latter cannot be deduced. The term non ambiguous refers to the fact that the hash code can be as it was uniquely identifiable with the source message. For this reason it is also called as digital finger print of the message. The hash functions [4]- [6] are classified into keyed and unkeyed hash function; the keyed hash functions are used in the Message Authentication Code (MAC) whose specification are dictates two distinct inputs a message and a secret key. The unkeyed hash function have there categories hash function based on block ciphers, modular arithmetic and customized hash function. The hash functions have one-way property; given n and an input M , computing $H(M) = n$, must be easy and given n , it is hard to compute M such that $H(M) = n$. The type of attacks are the collision attack (find two message $M = M'$ with $H(M) = H(M')$), the preimage attacks (given a random value γ , find a message M with $H(M) = \gamma$) and the second preimage attack (given a message M , find a message $M = M'$ with $H(M) = H(M')$) [7].

The most common used family of hash functions are SHA and MD families. The SHA-1 is required for use with the digital signature algorithm as specified in Digital Signature Standard (DSS) and whenever a secure hash algorithm is required. Both the transmitter and intended receiver of a message in computing and verifying a digital signature uses the SHA-1 [7]-[8]. It is necessary to ensure the security of digital signature algorithm, when a message of any length is input, the SHA produces m bits output called Message Digest (MD). The MD is then used in the digital signature algorithm. Signing the MD using the private key rather than the message often improved efficiency of the process because the MD is usually much smaller than the message. The same MD should be obtained by the verifier using the user public key when the received version of the message is used as input to SHA.

In the recent years much progress has been made in the design of practical one-way hashing algorithms which is efficient for implementation by both hardware and software. Noteworthy work includes the MD family which consist of three algorithms MD2, MD4, MD5 [9]-[10]-[11]-[12]. In our work we are interested of MD5 [11]-[12], which is the most adapted hash function in the authentication and signature of video data. Let begin by a brief description of MD5 which is developed by Ron Rivest, a much more detailed description can be found in RFC 1321 [11]. MD5 works by first padding the message until it is a multiple of 512 bits long. Padding is done as follows:

- 1) Append a '1' bit to the message.
- 2) Append '0' bits until the message is 64 bits shorter than a multiple of 512 bits.
- 3) Append a 64-bit representation of the message's original length.

The state of MD5 is kept in four 32-bit words, A, B, C, and D, all of which are initialized to magic constant values. MD5 processes the message in 512-bit blocks. As we process the i th block of message, we update A_{i-1} , B_{i-1} , C_{i-1} , and D_{i-1} to A_i , B_i , C_i , and D_i . The output of MD5, a 128 bit value, is the final state of A, B, C, and D concatenated. For each block of message, we have four rounds of updates. Each round

updates one of the four 32-bit words A, B, C, or D four times. (For a total of sixteen updates per block of message.) Initially on each round, $A_i \leftarrow A_{i-1}$, $B_i \leftarrow B_{i-1}$, etc. Each of the updates is something similar to $A_i \leftarrow B_i + (A_i + F(B_i; C_i; D_i) + M_i + T_i \lll s)$, where F is a function, M_i is the i th block of the message, and T_i and s are magic constants. (The symbol \lll means "rotate left".) At the end of each round, we finish by updating all of the values one last time, namely: $A_i \leftarrow A_i + A_{i-1}$, $B_i \leftarrow B_i + B_{i-1}$, etc.

The maximum security depends on the length of message digest generated by the hash functions which is limited by the size of input to the algorithm. It also shows how the modification is done with satisfying the properties like compression, preimage resistance, and collision resistance. The simulation results show that proposed scheme provides better security than the existing one, in figure 1 we illustrate the diagram of a general hash function.

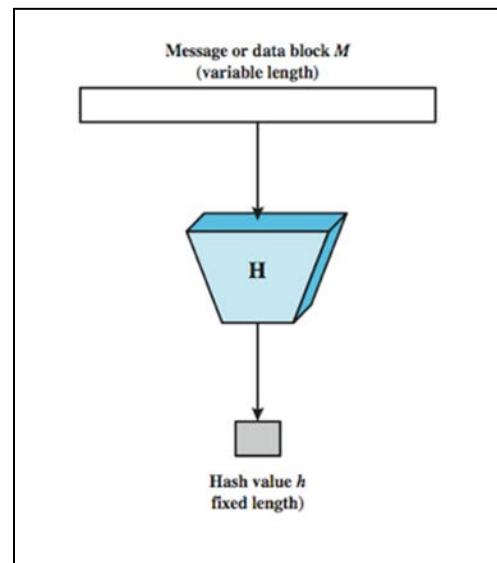


Fig. 1. Diagram of the hash function

III. DIGITAL SIGNATURE ALGORITHM

Digital signature is a mechanism by which a message is authenticated which means proving that a message is effectively coming from a given sender, much like a physical signature on a paper document. For instance, let suppose that Alice wants to digitally sign a message to Bob. To do so, she uses her private-key to encrypt the message; she then sends the message along with her public-key (typically, the public key is attached to the signed message). Since Alice's public-key is the only key that can decrypt that message, a successful decryption constitutes a Digital Signature Verification, and meaning that there is no doubt that it is Alice's private key that encrypted the message [13].

The DSA was proposed in August 1991 by the U.S. National Institute of Standards and Technology (NIST) and became a U.S. Federal Information Processing Standard (FIPS 186) in 1993. The FIPS 186 standard is also referred to as the Digital Signature Standard (DSS). The DSA was the first digital signature scheme accepted as legally binding by a government. The algorithm is a variant of the ElGamal

signature scheme. It exploits small subgroups in \mathbb{Z}_p^* in order to decrease the size of signatures. The key generation, signature generation, and signature verification procedures for DSA are given next.

DSA key generation. Each entity A does the following:

1. Select a prime q such that $2159 < q < 2160$.
2. a 1024-bit prime number p with the property that $q \mid p-1$. (The DSS mandates that p be a prime such that $2^{511+64t} < p < 2^{512+64t}$ where $0 \leq t \leq 8$ then I is a I prime.)
3. Select an element $h \in \mathbb{Z}_p^*$ and compute $g = h^{p-1} |q \bmod p$ repeat until $g \geq I$. (g is a generator of the unique cyclic group of order $q \in \mathbb{Z}_p^*$)
4. Select a random integer x in the interval $[1; q-1]$.
5. Compute $y = g^x \bmod p$
6. The public key is $(p; q; g; y)$; And the private key is x .

DSA signature generation. To sign a message

m , A does the following:

1. Select a random integer k in the interval $[1; q-1]$.
2. Compute $r = (g^k \bmod p) \bmod q$
3. Compute $k^{-1} \bmod q$
4. Compute $s = k^{-1} \{h(m) + xr\} \bmod q$ where h is the Hashed message.
5. If $s = 0$ then go to step 1. (If $s = 0$, then $s^{-1} \bmod q$ does not exist; s^{-1} is required in step 3 of signature verification.)
6. The signature for the message m is the pair of integers $(r; s)$.

DSA signature verification. To verify A's signature

$(r; s)$ on m , B should:

1. Obtain an authentic copy of A's public key $(p; q; g; y)$.
2. Verify that r and s are integers in the interval $[1; q-1]$.
3. Compute $s^{-1} \bmod q$ and $h(m)$.
4. Compute $u_1 = h(m)w \bmod q$ and $u_2 = rw \bmod q$
5. Compute $v = (g^{u_1} g^{u_2} \bmod p) \bmod q$.
6. Accept the signature if and only if $v = r$.

Since r and s are each integers less than q , DSA signatures are 320 bits in size. The security of the DSA relies on two distinct but related discrete logarithm problems. One is the discrete logarithm problem in \mathbb{Z}_p^* where the number field sieve algorithm [4] applies; this algorithm has a sub exponential running time. More precisely, the running time of the algorithm is $O(\exp(c + o(1))(\ln p)^{1/3}(\ln(\ln p))^{2/3})$, where $c \cong 1,923$, and $\ln(n)$ denotes the natural logarithm function. If p is a 1024-bit prime, then the precedent expression represents an infeasible amount of computation; thus the DSA is currently not vulnerable to this attack. The second discrete logarithm problem works to the base g given p, q, g , and y , find x such that $y \equiv gx \pmod{p}$. For large p (e.g., 1024-bits), the best algorithm known for this problem is the Pollard rho-method [4]-[6], and takes about $\sqrt{\pi q/2}$ (2) steps. If $q \approx 2^{160}$, then the expression (2) represents an infeasible amount of computation; thus the DSA is not

vulnerable to this attack. However, note that there are two primary security parameters for DSA, the size of p and the size of q . Increasing one without a corresponding increase in the other will not result in an effective increase in security. In figure 2, we illustrate the digital signature process.

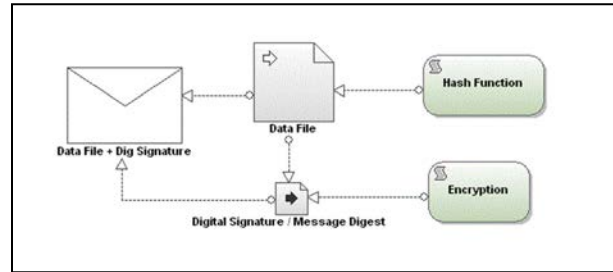


Fig. 2. Digital signature process

IV. ELLIPTIC CURVE DIGITAL SIGNATURE ALGORITHM

A. Elliptic Curve Cryptography

The theory of elliptic curves is deep and an enormous amount of research has been done on elliptic curve cryptography during the past twenty years or so. Therefore, it is impossible to present an extensive review of the field here and only subjects which are the most relevant are discussed in the following. Interested readers are referred to [14], for example, for further information.

All elliptic curve cryptosystems are based on an operation called elliptic curve point multiplication which is defined as $Q = kP$ where k is an integer and Q and P are points on an elliptic curve. A point is represented with two coordinates as (x, y) . The reason why elliptic curve point multiplication is used in cryptosystem is that it is relatively easy to compute but its inverse operation called elliptic curve discrete logarithm problem, that is finding k if P and Q are known, is considered impossible to solve with present computational resources if parameters are chosen correctly. Thus, elliptic curve discrete logarithm problem can be compared, for example, to integer factorization problem which is used in the popular RSA cryptosystems [4]. There is, however, a notable difference because sub-exponential algorithms for solving elliptic curve discrete logarithm problem are not known and, therefore, key lengths can be shorter than in RSA. Elliptic curve point multiplication is computed by using two principal operations; namely, point addition and point doubling. Point addition is the operation $P_3 = P_1 + P_2$, where P_i are points on an elliptic curve. Point doubling is the operation $P_3 = 2P_1$. In this design, point multiplication is computed with the so-called Montgomery's ladder. Elliptic curves used in cryptosystems are defined over finite fields denoted by

$GF(q)$ where q is the number of elements in the field. It is commonly preferred especially in hardware implementations to use binary field $GF(2^m)$ s where an element of the field is presented with m bits. In this design, the field $GF(2^{163})$ is used and it is constructed by using normal basis. Arithmetic operations are computed as follows:

- Addition $a + b$ is computed with a bitwise exclusive-or (XOR).

- Multiplication $a \times b$ is computed as presented by Wang et al. in [15]. This multiplier structure is referred to as Massey-Omura multiplier in the paper [16].
- Squaring a^2 is simply a cyclical rotation of the bit vector representing a .
- Finding an inverse element a^{-1} such that $a^{-1}a = 1$ is performed as suggested by Itoh and Tsujii in [17] and it is called henceforth Itoh-Tsujii inversion. One Itoh-Tsujii inversion requires 9 multiplications and 162 squarings if $m = 163$.

Point representation with two coordinates as (x, y) is referred to as the affine coordinate representation. When points are represented in affine coordinates, both point addition and point doubling require inversion in $GF(2^m)$. Inversion is by far the most expensive operation and, thus, it is advantageous to trade inversions for multiplications. This can be done by representing points with projective coordinates as (X, Y, Z) ; that is, with three coordinates. Mappings between these two representations are performed as $(x, y, 1)$ and $(X/Z, Y/Z)$. As can be seen, the mapping from affine to projective coordinates does not require any operations but the mapping from projective to affine coordinates requires two multiplications and one inversion. Using projective coordinates is very advantageous because point additions and point doublings can be performed without inversions and the total number of inversions in elliptic curve point multiplication is therefore one. A very efficient algorithm for computing (1) on elliptic curves over $GF(2^m)$ was presented in [18] by Julio Lopez and Ricardo Dahab. The authors of [18] shows that it suffices to consider only the x-coordinate and the y-coordinate can be recovered in the end [18]. This leads to a very efficient algorithm with projective coordinates. Point addition $(X_3, Z_3) = (X_1, Z_1) + (X_2, Z_2)$ can be computed as follows:

$$Z_3 = (X_1Z_2 + X_2Z_1)^2, X_3 = xZ_3 + X_1Z_2X_2Z_1 \quad (2)$$

where x is the x-coordinate of the base point P . The cost of point addition is four multiplications, two additions and one squaring. Point doubling $(X_3, Z_3) = 2(X_1, Z_1)$ is even simpler

$$X_3 = X_1^4 + a_6Z_1^4, Z_3 = X_1^2Z_1^2 \quad (3)$$

where a_6 is a fixed curve parameter. Thus, point doubling costs two multiplications, four squarings and one addition. The y-coordinate is recovered in the end by

computing $x_1 = X_1/Z_1$ and $x_2 = X_2/Z_2$ and then by using the formula:

$$y_1 = \frac{(x_1 + x)((x_1 + x)(x_2 + x) + x^2 + y)}{x} + y \quad (4)$$

where (x, y) is the base point P . This can be computed with one inversion, ten multiplications, six additions and one squaring.

B. ECDSA

ECDSA is a standard of ANSI, IEEE, and NIST, among others. The following description is based on Johnson and others' presentation in [19]. The algorithm operates so that first the user, who is commonly called Alice or A for short,

generates two keys, private and public, by performing a key pair generation procedure. Then, she publishes her public key. Alice signs a message by performing a signature generation procedure after which she sends both the message and the attached signature to the receiver who is called Bob, or B for short. Bob can verify the signature on the message by first getting Alice's public key and then by performing the signature verification procedure. Key pair generation, signature generation and signature verification are consider in the following sections.

Key Pair Generation.

Private and public key for an identity A is generated as follows:

$$d \in_R [1, n - 1] Q = dG \quad (5)$$

Where $d \in_R [1, n - 1]$ means that d is an integer selected at random from the interval $[1, n - 1]$. The integer d is A's private key and Q is A's public key. The computation of (5) requires generation of one random integer and computation of one elliptic curve point multiplication.

Signature Generation.

In order to generate a signature for a message M the identity A computes

$$\begin{aligned} k \in_R [1, n - 1] r &= [kG]_x \pmod{n} \\ &= H(M)s \\ &= k^{-1}(e + dr) \pmod{n} \end{aligned} \quad (6)$$

A's signature on M is (r, s) . The notation $[kG]_x$ denotes the x-coordinate of the result point of kG . Notice that A uses his/her private key d in the signature generation. Thus, other identities cannot produce the same signature without knowing d . Signing a message requires generation of one random integer, computation of one elliptic curve point multiplication and one hashing. In addition, modular inversion, addition and multiplication are required.

Signature Verification.

Identity B verifies A's signature (r, s) on the message M by computing

$$\begin{aligned} e &= H(M)w \\ &= s^{-1} \pmod{n} u_1 \\ &= ew \pmod{n} u_2 \\ &= rw \pmod{n} v \\ &= [u_1G + u_2Q]_x \pmod{n} \end{aligned} \quad (7)$$

where Q is A's public key and thus known by B. If $v = r$, B accepts the signature, otherwise (s)he rejects it. Verification requires one hashing and two elliptic curve point multiplications which are combined with a single elliptic curve point addition. Modular inversion and two multiplications are needed, as well.

V. H.264/AVC COMPRESSED VIDEO

An H.264 video encoder is mainly comprised of motion estimation, motion compensation, intra frame prediction, discrete cosine transformation, quantization and entropy encoding [20]. Figure 3 shown block diagram of H.264 Encoder. The brief overview of H.264 block is as follows. Encoder has intra prediction mode, which removes spatial

redundancy from the frame. The feedback path of the decoder module is an access point, which is used to decode intra predicted frame correctly. It works on different intra mode to remove spatial redundant data from the reference frame.

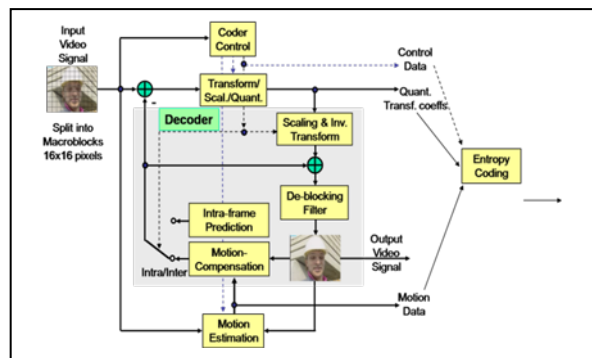


Fig. 3. Diagram block of H.264 encoder

Depending on the H.264 profile, different types of frames such as I-frames, P-frames and B-frames, may be used by an encoder. An I-frame, or intra frame, is a self-contained frame that can be independently decoded without any reference to other images. The first image in a video sequence is always an I-frame. I-frames are needed as starting points for new viewers or resynchronization points if the transmitted bit stream is damaged. I-frames can be used to implement fast-forward, rewind and other random access functions. An encoder will automatically insert I-frames at regular intervals or on demand if new clients are expected to join in viewing a stream. The drawback of I-frames is that they consume much more bits, but on the other hand, they do not generate many artifacts. A P-frame, which stands for predictive inter frame, makes references to parts of earlier I and/or P frame(s) to code the frame. P-frames usually require fewer bits than I-frames, but a drawback is that they are very sensitive to transmission errors because of the complex dependency on earlier P and I reference frames. A B-frame, or bi-predictive inter frame, is a frame that makes references to both an earlier reference frame and a future frame.

In the figure.4, we give a sequence example of I, B and P frames.

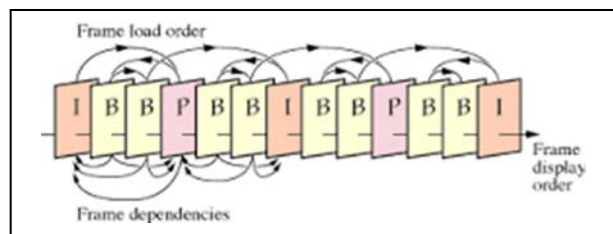


Fig. 4. Sequence of I, B and P frames

An H.264 encoder generated up to 50% fewer bits per second for a sample video sequence than an MPEG-4 encoder with motion compensation. In figure 5 the H.264 encoder was at least three times more efficient than an MPEG-4 encoder with no motion compensation and at least six times more efficient than Motion JPEG.

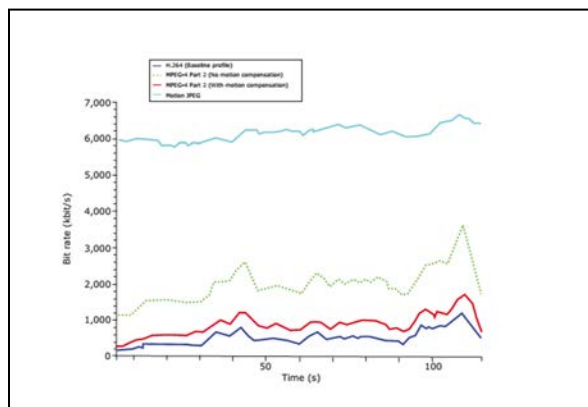


Fig. 5. Comparison of Bit rate of different encoders

VI. EXPERIMENTAL RESULTS

In this section we give the results of the comparison we do between DSA and ECDSA used for the signing of large number of H.264 video. We use MATLAB on a 64-bit Intel Core I7-4500U CPU 2.4 GHz, 6 G RAM machine to implement DSA and ECDSA signatures generation scheme and to test their performances. Our results are given below. Experimental results are given in this section to demonstrate the benefit of using the ECDSA based on the elliptic curve cryptography. These benefits can be seen in the gain of the time and the smallest size of the key in the implementation. We used DSA and ECDSA to sign the hashing output of some H.264 videos. Here below we give some results of our experimental results.

A. Comparison of the speed of hash function

We start by selecting the appropriate hash function to use for the videos signing. For this purpose we have compared the speed of the implementation of the most commonly used hash functions. There are several techniques in which are based the construction of hash functions. For example include the SHA-1 function. The choice of the hash function for the signature depends on the nature of the document to be signed. In the figure.6 we compare the speed of the main existent hash functions. For the rest and for signing the videos with real time constraint we used the MD5 function view the advantage that provides this function with respect to speed.

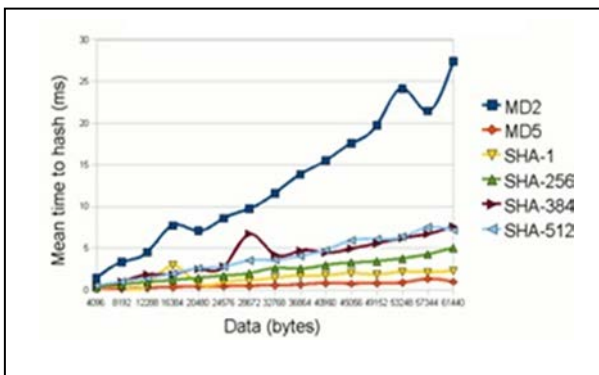


Fig. 6. Speed of secure hash functions

B. DSA vs ECDSA

In the table I below given by NIST, we give a comparison of the key size between DSA and ECDSA for a given level of security. We can see that the key size is very small in the case of ECDSA over DSA which can be an advantage in applications where we have real-time and memory constraints.

TABLE I. COMPARISON OF THE KEY SIZE

Security (bit)	DSA –Size of the key	ECDSA-Size of the key
80	1024	160
112	2048	224
128	3072	256
192	7680	384
256	15360	512

In the figure 7 we illustrate the time to break DSA and ECDSA depending on the size of the key.

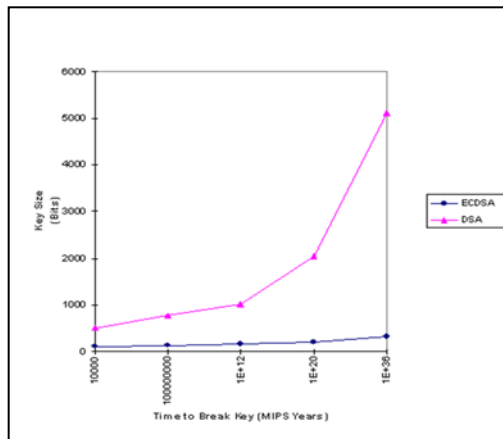


Fig. 7. Comparison of time to break DSA and ECDSA

We also compare the speed and space that requires the hardware implementation of both electronic signing protocols in figure 8 below we can see the advantage of using ECDSA. If we implement our algorithms using VLSI cores, whether in relation to the space used in number of gates or speed, ECDSA differs greatly from its rival DSA.

Hardware comparison: 128-bit security level		
mode	DSA	ECDSA
Space-optimized (same clock speed)	(VLSI Cores) 184 ms 50,000 gates	(Gmbschädl) 29 ms (16 ms for Koblitz curve) 6,660 gates
Speed-optimized (same clock speed)	(VLSI Cores) 110 ms 189,200 gates	(Orlando and Paar) 1.3 ms 80,100 gates

Fig. 8. Hardware comparison of space and time of DSA and ECDSA

Also in the figures 9, 10, 11 and 12, below we show the difference in the shape of histograms in the case of two H.264 videos using in the first two figure fig.9 and fig.10 the DSA protocol and the other two figures fig.11 and fig.12 the protocol ECDSA. We can notice the difference in scope between the two cases of the presented histograms which is due to reduced key size.

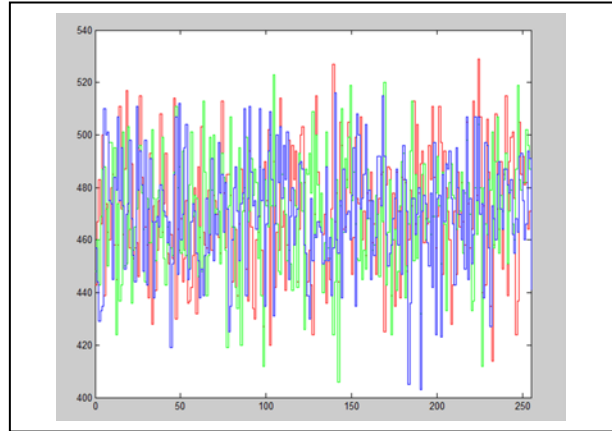


Fig. 9. Histogram of hashing and signed video 1 with DSA

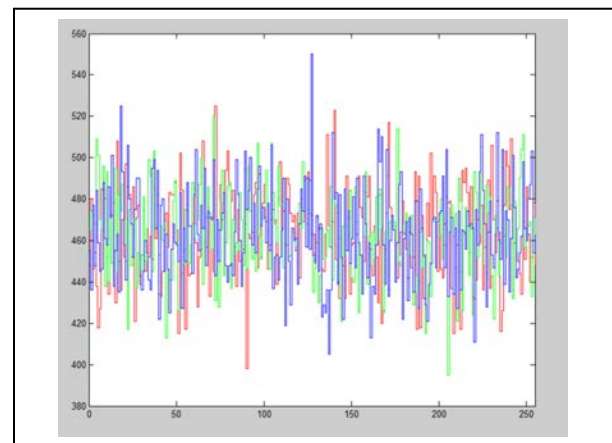


Fig. 10. Histogram of hashing and signed video 2 with DSA

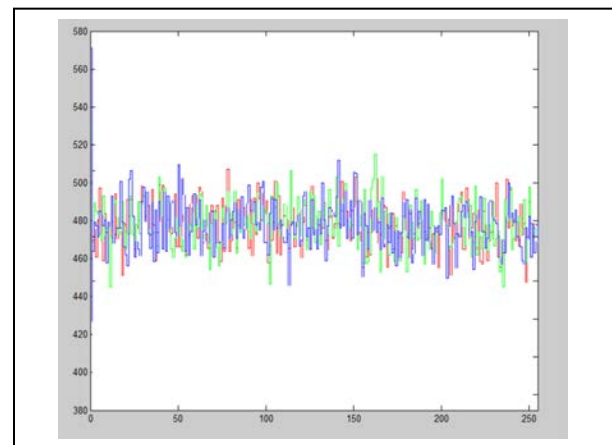


Fig. 11. Histogram of hashing and signed video 1 with ECDSA

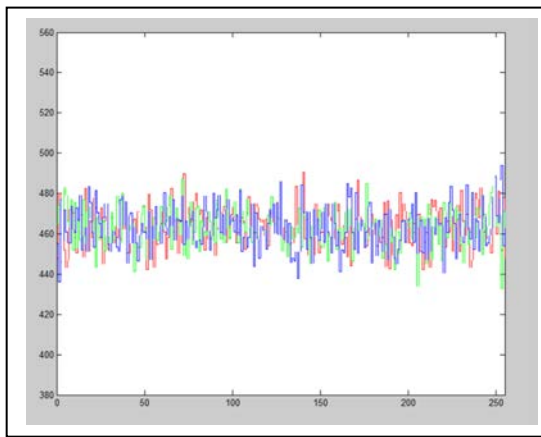


Fig. 12. Histogram of hashing and signed video 2 with ECDSA

We also compared the time of the signature process in second of these two algorithms depending on the size for a library containing a large number of H.264 videos. The speed of ECDSA over DSA is clearly denoted in figure 13 despite the growth in the size of the videos.

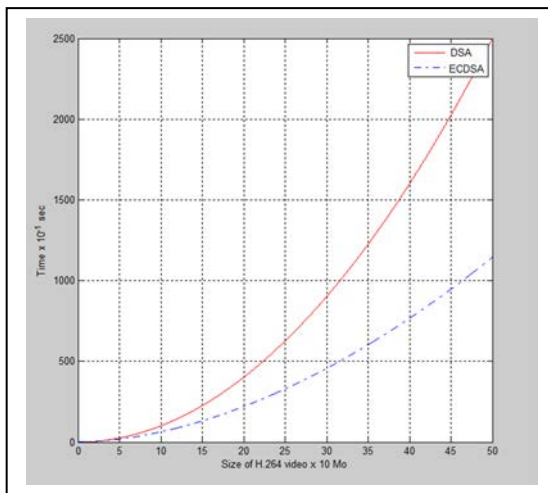


Fig. 13. Comparison of timing signing scheme- DSA vs ECDSA

VII. CONCLUSION

In this paper we compare the performance of two famous methods for electronic signing DSA and ECDSA in order to sign H.264 videos. We studied their speed, the number of gates used in the hardware implementation and the histograms' distribution of the some signed and hashed videos by MD5 function in the cases of these two algorithms.

REFERENCES

- [1] T. Plevyak, V. Sahin, Next Generation Telecommunications Networks, Services, And Management (Wiley-Ieee, 2010).
- [2] M. Krause, H. F Tipton, Information Security Management Handbook. (6th ed. Auerbach Publications, CRC Press LLC, 2010).
- [3] I. E. Richardson, The H.264 Advanced Video Compression Standard 2nd Edition (Wiley, 2010).
- [4] A. Menezes, P. Van Oorschot, S. Vanstone, Handbook of applied cryptography, (CRC Press, 1996).
- [5] Q. Zhang , Z. Li And C. Song, The Improvement Of Digital Signature Algorithm Based On Elliptic Curve Cryptography, 2011 Ieee Artificial Intelligence, Management Science And Electronic Commerce (Aimsec), Pp-1689 – 1691.
- [6] P. Williams, Applied Cryptography (John Wiley & Sons, 1996).
- [7] M. Stevens, Attacks on Hash Functions and Applications, Ph.D. Thesis, Dept. Computer Engineering, University of Leiden, Netherland, 2012.
- [8] FIPS 180-4, Secure Hash Standard (SHS)– Current version of the Secure Hash Standard (SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512), 2012.
- [9] R. L. Rivest, The MD4 Message Digest Algorithm, 1990 CRYPTO Lecture Notes in Computer Science, vol. 537, Springer, pp. 303–311.
- [10] R. L. Rivest, The MD4 Message-Digest Algorithm, Internet Request for Comments, 1990, RFC 1186; obsolete by RFC 1320.
- [11] R. L. Rivest, The MD5 Message-Digest Algorithm, Internet Request for Comments, 1992, RFC 1321.)
- [12] S. Yu and K. Aoki , Finding Preimages in Full MD5 Faster than Exhaustive Search, 2009 Advances in Cryptology - EUROCRYPT 2009, Volume 5479 of the series Lecture Notes in Computer Science pp 134-15.
- [13] M. O. Rabin, Digitalized Signatures, Foundations of Secure Computation (Richard A. Demillo, David P. Dobkin, Anita K. Jones, and Richard J. Lipton, eds.), Academic Press, 1978, pp. 155–168.
- [14] H. Cohen. G. Frey, R. Avanzi, C. Doche, T. Lange, K. Nguyen and F. Vercauteren, Handbook of Elliptic and Hyperelliptic Curve Cryptography, Dis. Math.Its App. 1st Edition, (2005).
- [15] Y. Wang, Task Parallel Implementation of Matrix Multiplication on Multi-socket Multi-core Architectures, Algorithms and Architectures for Parallel Processing, 2015 15th International Conference, ICA3PP 2015, Zhangjiajie, China, Proceedings, Part III.
- [16] A. Reyhani-Masoleh , M. A. Hasan, A new construction of Massey-Omura parallel multiplier over GF(2m), 2002 IEEE Transactions on Computers (Volume:51 , Issue: 5) PP-511-520.
- [17] R. K. Kodali , C. N. Amanchi ; S. Kumar ; L. Boppana, FPGA implementation of Itoh-Tsujii inversion algorithm, 2014 Recent Advances and Innovations in , Engineering (ICRAIE), 2014, pp 1 – 5.
- [18] J. López, R. Dahab, Fast Multiplication on Elliptic Curves Over GF(2m) without precomputation, 2002 Cryptographic Hardware and Embedded Systems, Volume 1717 of the series Lecture Notes in Computer Science pp 316-327.
- [19] D. Johnson, A. Menezes, S. Vanstone, The Elliptic Curve Digital Signature Algorithm (ECDSA), 2001 International Journal of Information Security, Volume 1, Issue 1, pp 36-63.
- [20] Itu, H.264: Advanced video coding for generic audiovisual services, International Telecommunication Union, 2003.

A Novel Information Retrieval Approach using Query Expansion and Spectral-based

Sara Alnofaie, Mohammed Dahab, Mahmoud Kamal

Computer Science
King Abdul-Aziz University
Jeddah, Saudi Arabia

Abstract—Most of the information retrieval (IR) models rank the documents by computing a score using only the lexicographical query terms or frequency information of the query terms in the document. These models have a limitation as they do not consider the terms proximity in the document or the term-mismatch or both of the two. The terms proximity information is an important factor that determines the relatedness of the document to the query. The ranking functions of the Spectral-Based Information Retrieval Model (SBIRM) consider the query terms frequency and proximity in the document by comparing the signals of the query terms in the spectral domain instead of the spatial domain using Discrete Wavelet Transform (DWT). The query expansion (QE) approaches are used to overcome the word-mismatch problem by adding terms to query, which have related meaning with the query. The QE approaches are divided to statistical approach Kullback-Leibler divergence (KLD) and semantic approach P-WNET that uses WordNet. These approaches enhance the performance. Based on the foregoing considerations, the objective of this research is to build an efficient QESBIRM that combines QE and proximity SBIRM by implementing the SBIRM using the DWT and KLD or P-WNET. The experiments conducted to test and evaluate the QESBIRM using Text Retrieval Conference (TREC) dataset. The result shows that the SBIRM with the KLD or P-WNET model outperform the SBIRM model in precision (P@), R-precision, Geometric Mean Average Precision (GMAP) and Mean Average Precision (MAP).

Keywords—Information Retrieval; Discrete Wavelet Transform; Query Expansion; Term Signal; Spectral Based Retrieval Method

I. INTRODUCTION

Many ranking functions or similar functions such as Cosine and Okapi do not take into consideration the query terms proximity. Proximity-based ranking functions based on the supposition, when the query terms closeness to each other, the document becomes more relevant to the query [1]. The document that contains the query terms in one sentence or paragraph is more related than the document, which includes the query terms that far from each other. In a document, the closeness of the query terms is a significant factor as much as their frequency that must not ignore in the information retrieval (IR) model.

The Spectral-Based Information Retrieval Model (SBIRM) ranks the documents according to document scores that combine the frequency and proximity of the query terms [2]. It compares the terms of the query in the spectral domain instead

of the spatial domain to take proximity in consideration without computing many comparisons. It creates a signal for a term, which maps the term frequency and position into the frequency domain and time domain respectively. To score the documents in SBIRM, compare the query terms spectrum that obtained by performing a mathematical transform such as Fourier Transform (FT) [3], Discrete Cosine Transform (DCT) [4] or Discrete Wavelet Transform (DWT) [5].

The conventional IR model lexicographic matches the query terms with the documents collection. In natural language, two terms can be lexicographically different although they are semantically similar. Therefore, directly matching the user query, which can include terms that are not present in documents leads to failure to retrieve the related documents that have other words with the same meaning. The query expansion (QE) approaches overcome vocabulary mismatch issues and enhance the performance of the retrieval by expanding the query with additional relevant terms without users' intervention. The query is expanded by subjoining either statistically related terms to the terms of the original query or semantically related terms chosen from some lexical database. Some statistical QE approaches in [6, 7, 8, 9] and semantic QE approaches in [10, 11, 12, 13, 14, 15, 16, 17] expand a query outperform IR model that ignores the proximity.

This research aims to design a QESBIRM that can retrieve the document relevant to the query terms using a proximity base IR model and QE techniques. This model combines two models: first, the SBIRM model using the DWT [5] that takes the proximity factor in its ranking function, and second, the statistical QE and semantic QE which overcomes vocabulary mismatch. With this merging, one can benefit from proximity ranking function and extend the query with more informative terms to enhance the performance of the IR model.

A thorough literature review will be presented along with a discussion of the proposed model in section two of this paper. The experiment is described in section three followed by results analysis. The conclusions and suggestions for future work will be outlined at the end of the paper.

II. LITERATURE REVIEW

A. Proximity-Base Information Retrieval Model

The proximity-base Model assumption is based on the fact that the document is extremely relevant to the query when the query terms occur near to each other. It uses spatial location information as a new factor to compute the document score in

information retrieval rather than touching the surface of the document by counting the query terms. The shortest substring retrieval model is one of the proximity-base Model proposed by Clarke in [18]. In this model, the document scores based on the shortest substring of text in the document that matches the query. This is done by creating a data structure called a Generalized Concordance List (GCL). These GCLs contain the query terms position in the document. This model does not consider term frequency in the documents when computing the document score although it is an important factor. It also takes long query time to create GCL and do not compute the score to the document that contains one term.

In the fuzzy proximity model [19], the document score is computed using the fuzzy proximity degree of the query terms appearance. The drawbacks of this model are that all the query terms have to occur in the document. If one query term does not occur or query terms are away from each other more than closeness parameter, the document score becomes zero. In addition, the model does not consider the frequency of the query term in the document.

Some research combines the proximity information to frequency scoring function [20, 21, 22]. The proximity IR model [20, 21] does not improve the performance significantly while the BM25P model [22] improves the performance but it is sensitive to the window size.

The Markov Random Field model considers Full Independence, Sequential Dependence, and Full Dependence between query terms but it is also sensitive to the window size [23].

In the proximity model, each query term positions is compared with the other query terms to calculate the document score. Subsequently, the comparisons number grows combinatorially if the query terms number grows [24, 25, 26]. This problem was overcome in SBIRM [2] by comparing the terms of the query in the domain of the spectral. In addition, the previous proximity models measure the proximity of the query terms only in specific region or window while SBIRM measures the proximity of the query terms in the whole document.

Briefly, the SBIRM steps are: first, the term signal is created. Then, the term signals transform into term spectra by using a spectral transform. After that, all terms spectral signal in is stored in each document. Next, the query terms signal is retrieved for every document. Finally, the document score is obtained by combing the spectra of the query terms. In the spectral domain, the query term frequency and position are represented by magnitude and phase values.

Park et al. [3] used the FT in SBIRM model. This model called Fourier Domain Scoring (FDS). Unfortunately, the FDS has a large index storage space [27]. To overcome this problem, the SBIRM use the DCT to perform document ranking [4]. The SBRM high precision still achieved by this model. The frequency information is extracted from the signal as a whole using the FT and DCT transforms.

Many data mining problems use the Wavelets transform as efficient and effective solutions [28] because it has properties [29] such as multiresolution decomposition structure.

Therefore, Park and others used the DWT in SBIRM [5]. The DWT in document ranking is able to concentrate at different resolutions on the signal portions [5]. The signal is break into wavelets of different scales and positions, so that it can analyze the patterns of the terms in the document at various resolutions (whole, halves, quarters, or eighths).

Using the signal concept as representation model with DWT led to improvement in the performance of text mining tasks like document clustering [30], document classification [31, 32, 33] and recommender system on Twitter [34].

B. Automatic Query Expansion Approaches

In respect of information retrieval application, there is a long history for the QE. The experimental and scientific reached by this application reached to maturity especially in laboratory settings like Text Retrieval Conference (TREC). The QE is a process of broadening the query terms using words that share statistical relationships or meaning with query terms. Usually, the queries consist of two or three terms, which are sometimes not enough to understand the expectations of the end user and fail to express topic of search. Various approaches used to expand the query over IR model that ignore the proximity information. Some of this approaches use an external resource or use target corpus or both.

The target corpus approaches is classified to local and global. The global approaches analyze the whole corpus to explore terms that co-occurred. When the terms co-occurs frequently with query term, they are consider as related terms. One of the global approaches constructs automatically in the indexing stage and named as co-occurrence thesaurus. On the other hand, the local approaches use the top relevant documents of the initial search results. The global approaches are less effective than local because they relies on the collection frequency features but are irrelevant with the terms of the query [10].

The latent semantic indexing (LSI) is classified as global approach [35]. It computes the singular value decomposition of the term-document matrix to replace the document features with smaller new features set. This new generated features are then used to expand the query. The Rocchio's is one of the sample Local approaches [36]. It expands the query with the top relevant documents terms that re-weights by sum weights of that term in all top relevant documents. Rivas and other are well known to enhance the performance of the IR using Rocchio's with the biomedical dataset [37]. The limitation of this approach is the term weight that reflects the significance of that term to the entire collection instead of its usefulness to the user query. Local approaches based on distribution analysis, which distinguishes between useful expansion terms and bad expansion terms by comparing the appearance in relevant documents with the query with that in all documents. In other words, the score of the appropriate expansion term becomes high when its frequency is high in relevant documents compared with the collection. One of this statistical comparative analysis approach uses a chi-square variant to select the pertinent terms [8]. On the other hand, The Robertson Selection Value approach Uses Swets theory [7]. Carpineto et al. [6] proposed an effective approach that depends on the terms probability distributions in the related

documents and in the corpus. In average, the Kullback-Leibler divergence (KLD) performance outperforms the previous expansion approaches based on distribution analysis when applied to selecting and weighting expansion terms [6]. Amati [9], calculates the divergence between the distributions of the terms using Bose-Einstein statistics (Bo1) and the KLD. In a different study [39], the KLD gave a good performance compared with the Bo1.

The Local context analysis (LCA) [38] is a local approach base on co-occurrence analysis. It computes term co-occurrence degree with whole query terms using co-occurrence information of the top-ranked documents. Pal and Mandar [39] proposed newLCA that tries to improve LCA [39]. The Relevance Models (RM1) is another co-occurrence approach [40]. The LCA, newLCA and RM1 sometimes do not perform at the expected level.

The external resource approaches use esources such as Dictionaries, Thesaurus, WordNet, Ontology and other semantic resources [10]. Many of the works have concentrated on the use of WordNet to improve the IR performance. Many studies extended the query using all synonyms contained in a synset which contains query terms [11, 41, 42]. The rest of other approaches set all synonyms of the synset, which contain query terms as CET. They then use the word sense disambiguate (WSD) approach to determine the right sense synsets. Finally, they consider the synonyms of the right sense synsets as expansion terms. Giannis and others use the most common sense WSD approach [12]. Recently, Meili et al. [14] used the synonym of the synsets that has the same parts-of-speech with query term to extend each query terms. Fang [15] used the Jaccard coefficient to expand the query using synonyms of the synset that contain query terms and have high overlap between its glosses and the query terms glosses. Tyar and Then [16] proposed considering the glosses of the Synonyms, Hypernyms and Hyponyms synsets in the Gloss overlap WSD that using Jaccard coefficient. The drawback of these approaches are usually sensitive to WSD and the expansion terms independently of the content of the corpus and query [10].

The target corpus and external resource approaches first, use corpus as a source of candidate expansion terms (CET). Then, compute semantic similarity score of this CET with query terms using WordNet. Finally, it add the terms, which have a high score to the query. The semantic similarity measure in [17] using edge base counting approaches while in [13], the gloss overlap approach is used. The drawback of edge base counting approach is that it measures semantic similarity between two terms only if they have the same part of speech.

C. Query Expansion and Proximity-Based Information Retrieval Model

Park [3] expands the query using the Rocchio approach over the FDS model. The precision of the expanded query over the FDS model is less than the FDS without expansion [2]. Audeh [43] studied the effect of the QE on the proximity base IR. He uses LSI and WordNet synonyms to extend the query over fuzzy proximity model. The experiment showed an inadequate performance of the QE approaches over the fuzzy proximity model. The WordNet synonyms low performance

can be improved by taking only the right sense instead of all query terms synonyms while the LSI can be improved by considering enough number of pseudo-documents. The fuzzy proximity model is high selectivity model. For some queries; it got less than five documents. Unlike these papers, the current study use better proximity base IR model and good performance query expansion approach. Over the BM25P He et al. [22] expanded the query using KLD QE approach that sometimes leads to a degraded performance. The performance of the MRF model improves by expanding the query using RM1 approach [44]. Unlike these papers, the current work uses better proximity base IR model and good performance query expansion approach.

III. ARCHITECTURE OF PROPOSED MODEL

The QESBIRM used in this work retrieves more relevant documents to the query by using good performance proximity model (SBIRM) and expands it with semantic relevant terms that overcome the mismatch problem. This is achieved by finding semantic similar term using on average the best distribution approach (KLD), the best target corpus and external resource approaches (P-WNET), and finally combining these approaches.

The proposed model is composed of two stages: text preprocessing and indexing stage; and processing query stage. The text preprocessing and indexing stage consist the following steps:

Text preprocessing, create term signals, apply weighting scheme on the term signal, apply wavelet transform on the signal and create an inverted index. It is all done in offline mode. The processing query stage steps are as follow: first, preprocessing the query. Second, apply weighting scheme on the query term. Third, retrieve query terms transformed signal. After that, compute the documents Scores. Then, the retrieved top ranking documents are sent to automatic query expansion model to extract the related terms as expansion features. Finally, the new query is sent to the spectral-based retrieval model to retrieve the final rank documents. The model architecture is shown in Fig. 1. In the following paragraphs more details on each model steps is provided.

D. Text Preprocessing

The text preprocessing is an essential part of any text mining application. At this stage, a combination of four common text-preprocessing methods were used: tokenization, case folding, stop word removal and stemming [2, 31]. First, the tokenization step, which is the task of converting a raw text file into a stream of individual tokens (words) by using spaces and line breaks and removing all punctuation marks, brackets, number and symbols [31].

Next, the case folding step which involves converting the case of every letter in the tokens to a common case. Usually, the lower case is the common case [2]. The following Stop Word Removal step ignores many terms that are not useful such as and, a, and the, in the English language because they are very common. If they are used in a query, nearly all of the documents in the set would return because every document would contain these words. If they are included in the index, the term weight would be very low.

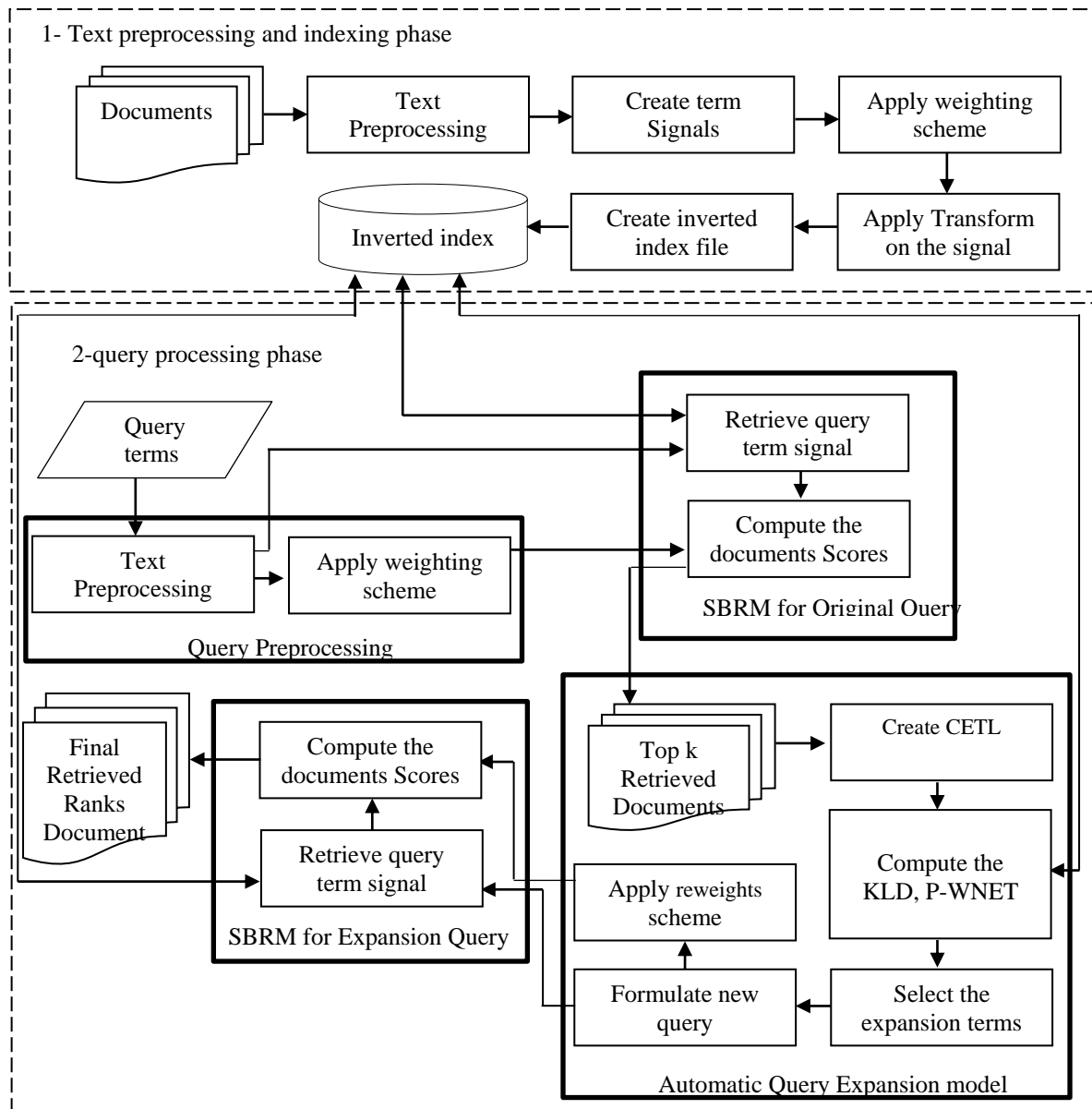


Fig. 1. General Architecture of a proposed model

Therefore, these terms are ignored. By doing this the terms number contained in the document set lexicon is reduced. Therefore, the amount of processing done by the indexer also reduces [2]. Finally, the Stemming step converts each term to its stem by removing all of a term's prefixes and suffixes [2]. The information retrieval model applies a stemming process in text preprocessing because it makes the tasks less dependent on particular forms of words. It also reduces the size of the vocabulary, which might otherwise have to contain all possible word forms [31]. In general, porter2 is the best overall stemming algorithm [45].

E. Term signals

Rather than mapping a document to a vector that contains the count of each word, the SBIRM maps each document into a collection of term signals.

The term signal, introduced by [3], it is a vector representation that displays the spread of the term throughout the document. It shows the term occurrences number in specific partitions or bins within the document.

To create the signal of the term first, divide the document into an equal number of bins. Then, represent the term t signal in document d using (1):

$$\tilde{f}_{d,t} = [\tilde{f}_{d,t,0} \ \tilde{f}_{d,t,1} \dots \ \tilde{f}_{d,t,B-1}] \quad (1)$$

where $\tilde{f}_{d,t,b}$ is the number of times term t occurred in bin b in document d .

For example, document d is divided into eight bins ($B=8$) and they contain two terms "computer" and "data".

Fig. 2. shows how the term signal creates for the terms "computer" and "data". As shown in Fig. 2, "computer" two times occurs in bin_3 , one time in bin_5 , and two times in bin_7 ; "data" occurs one time in bin_0 , one time in bin_2 , three times in bin_5 . The term signals for "computer" and "data" are shown in (2).

$$\tilde{f}_{a,computer} = [0,0,0,2,0,1,0,2] \quad \tilde{f}_{a,data} = [1,0,1,0,0,3,0,0] \quad (2)$$

F. Weighting Scheme

In the index stage, once the term signal created for each term in the corpus, the weighting scheme should apply to minimize the impact of highly common terms or high frequency terms in documents [4]. The BD-ACI-BCA weighting scheme was chosen as document weighting scheme in this experiments, which is shown to be one of the best methods [46]. In term signals, to apply this weighting scheme, the need to modify it to weigh the term signal instead of weighing the term in the document like Vector Space Model. In this work, it is applied to each signal component considering each bin as separate document [4].

$$\omega_{d,t,b} = \frac{1 + \log f_{d,t,b}}{\omega_d} \quad (3)$$

Where $\omega_{d,t,b}$ and $f_{d,t,b}$ is the weight of term t and occurrence number of term t in bin b in document d respectively.

$$\omega_d = (1-s) + s \cdot \frac{\omega'_d}{av_{d \in D} \omega'_d} \quad (4)$$

Where s is the slope parameter (0.7), ω'_d is the document vector norm and $av_{d \in D} \omega'_d$ is the average of the documents vector norm in the collection.

$$\omega_{d,t} = 1 + \log f_{d,t} \quad (5)$$

Where $f_{d,t}$ is the term t occurrence number in document d . In query stage, the following BD-ACI-BCA scheme using to weighting the query term [5]:

$$\omega_{q,t} = (1 + \log(f_{q,t})) \log(1 + f_m / f_t) \quad (6)$$

Where $\omega_{q,t}$ and $f_{q,t}$ are the weight and the frequency of the term t in query q , respectively, f_t is the documents number, which term t occurrence in, f_m is the large value of f_t for all t .

G. Signal Transform

Different signal levels resolution provide by DWT. The DWT is a sequence of high-pass and low-pass filters. The HWT can be described by high-pass filter (wavelet coefficients) is $[1/\sqrt{2} \quad -1/\sqrt{2}]$. While the low-pass filter (scaling function) is $[1/\sqrt{2} \quad 1/\sqrt{2}]$ as appears in Fig. 3[5, 47].

For example, let $\tilde{f}_{a,t} = [3, 0, 0, 1, 1, 0, 0, 0]$ is the term t signal in document d when perform HWT. The Signal Transform will be $W \cdot \tilde{f}_{a,t} = [\frac{5}{\sqrt{8}}, \frac{3}{\sqrt{8}}, \frac{2}{\sqrt{4}}, \frac{1}{\sqrt{4}}, \frac{3}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]$.

The terms positions at many resolutions appear in the transformed signal. Each transformed signal component provides term occurrences information in the specific location. In the first component, $(5/\sqrt{8})$ show that there are five term appears. The term occurrence three times in the signal first half more than the second half as in the second component. There is two more term appearance in the first quarter than the second quarter. As the fourth component appears, there is one more term appearance in the third quarter than in the fourth quarter. The signal eighths compare in the next four components.

H. Inverted index

An inverted index can created to store the word vectors. In this model, the words in each document were represented as:

$$\langle b_1, f_1 \rangle \langle b_2, f_2 \rangle \dots \langle b_y, f_y \rangle \quad (7)$$

Where y is the non-zero bins component, b_a is the bin number and f_a is the spectral value of bin b_a [3].

I. The Document Score

```

For each weighted term signal x in each document d
Repeat
    n ← number of elements in x
    Compute half of x as n/2
    Initialize i to 0
    Initialize j to 0
    Set temp to empty list
    Set result to empty list
    While j < n-3
        temp[i] = (x[j] + x[j+1]) / √2
        temp[i+half] = (x[j] - x[j+1]) / √2
        i ← i + 1
        j ← j + 2
        temp[i] = (x[n-2] + x[n-1]) / √2
        temp[i+half] = (x[n-2] - x[n-1]) / √2
        result ← second have of temp
        x ← first have of temp
    Until number of elements in x = 1
    
```

Fig. 2. Haar wavelet transform

The SBRM [2], [5] compute the document score by using the phase and magnitude information of the query terms transform signal. The phase describes the proximity information while the magnitude value of the component describes the term frequency. To compute the score of the document, let the transformed signal of the query term t in the document d where a number of components B is $\tilde{\zeta}_{d,t} = [\zeta_{d,t,0} \zeta_{d,t,1} \dots \zeta_{d,t,B-1}]$. First, for every spectral component, the magnitude and phase, defined by equation (8) and (9) are respectively calculated.

$$H_{d,t,b} = |\zeta_{d,t,b}| \quad (8)$$

and the phase which defined as

$$\phi_{d,t,b} = \frac{\zeta_{d,t,b}}{H_{d,t,b}} \quad (9)$$

Then, for each component the zero phase precision is calculated using equation (10)

$$\bar{\Phi}_{d,b} = \left| \frac{\sum_{t \in q, H_{d,t,b} \neq 0} \phi_{d,t,b}}{\#q} \right| \quad (10)$$

where q is the query terms and #(q) is the number of the query tokens. The components phases that have zero magnitudes ignores in the zero phase precision ($\bar{\Phi}_{d,b}$) because these phase values mean nothing. After that, the score is computed using equation (11):

$$s_{d,b} = \bar{\Phi}_{d,b} \sum_{t \in Q} H_{d,t,b} \quad (11)$$

Finally, the components scores are combined to obtain the document score:

$$S_d = \|\tilde{s}_d\|_p \quad (12)$$

where $\tilde{s}_d = [s_{d,0} \ s_{d,1} \ \dots \ s_{d,B-1}]$ and $\|\tilde{s}_d\|_p$ is the l^p norm compute by:

$$\|\tilde{s}_d\|_p = \sum_{b=0}^{B-1} |s_{d,b}|^p \quad (13)$$

J. Kullback-Leibler divergence Query Expansion Approach

Carpineto et al. [6] proposed interesting query expansion approaches based on term distribution analysis. They used the KLD concept [48]. The distributions variance between the terms in the top relevant documents collection that is obtained from the first pass retrieval using the query and entire document collection is the base of the scoring function. The query expands with high probability terms in the top related document compared with low probability in the whole set. The KLD score of term in the CET are computed using the equation:

$$KLD(t) = P_R(t) \log \frac{P_R(t)}{P_C(t)} \quad (14)$$

Where $P_R(t)$ is the term t probability in the top relevant documents R, and $P_C(t)$ is the term t probability in the corpus C, given by the following equations:

$$P_R(t) = \frac{\sum_{d \in R} f_{t,d}}{\sum_{d \in R} \sum_{v \in d} f_{v,d}} \quad (15)$$

$$P_C(t) = \frac{\sum_{d \in C} f_{t,d}}{\sum_{d \in C} \sum_{v \in d} f_{v,d}} \quad (16)$$

Where $f_{t,d}$ is the term t frequency in document d.

K. P-WNET Query Expansion approach

The scoring function of the P-WNET approach considers three parameter [13]. First, the semantic similarity between t and q_i using WordNet gloss overlap. Second, the t's rareness in the corpus. Finally, the similarity score of the top relevant document that contains t.

$$Rel_{t,q_i} = \frac{C_{t,q_i}}{C_t + C_{q_i} - C_{t,q_i}} \quad (17)$$

Where C_{t,q_i} is the number of common term between t and q_i definitions and C_t is the number of terms in t definitions.

$$idf_t = \max(0.0001, \log_{10} \frac{N - N_t + 0.5}{N_t + 0.5}) \quad (18)$$

$$s(t,q_i) = Rel_{t,q_i} * idf_t * \sum_{d \in R} \left(\frac{\text{sim}(d,q)}{\max_{d' \in R} \text{sim}(d',q)} \right) \quad (19)$$

$$S(t) = \sum_{q_i \in q} \frac{S(t,q_i)}{1 + S(t,q_i)} \quad (20)$$

L. Reweight scheme

After adding the expansion terms to the authentic query term, the new query must be reweighed. One of the best reweighing schemes is the scheme that is derived from KLD or P-WNET. The weight of the new query is computed using the following equation [13]:

$$\omega_{\text{new}}(t) = \alpha \frac{\omega_{\text{orig}}(t)}{\max_{v \in Q} \omega_{\text{orig}}(v)} + \beta \frac{\text{score}(t)}{\max_{v \in R} \text{score}(v)} \quad (21)$$

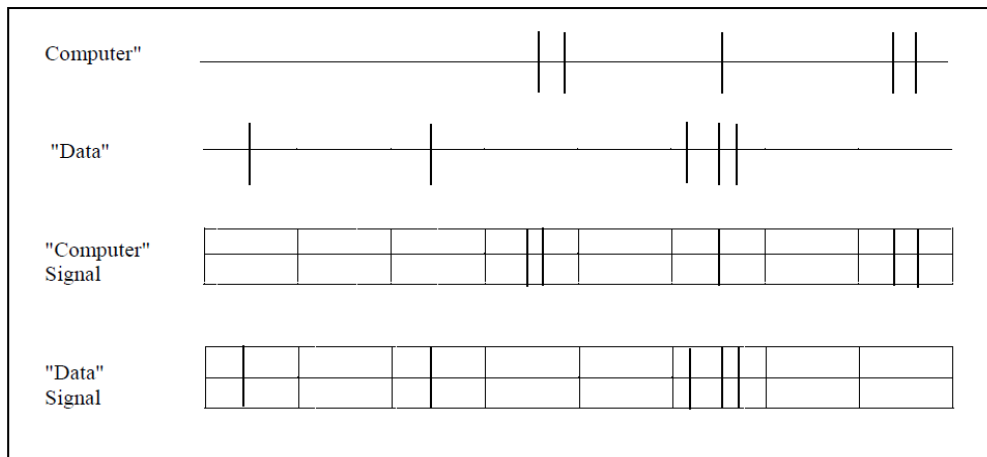


Fig. 3. The example of create the term signals

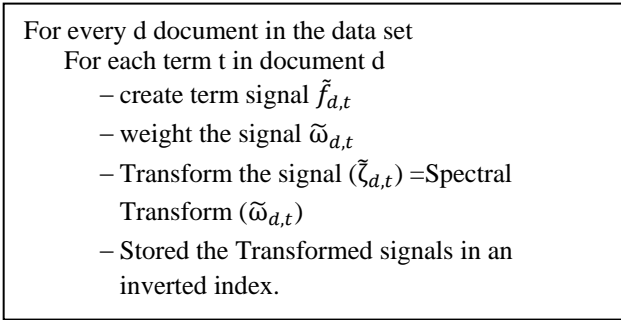


Fig. 4. The text preprocessing and indexing stage steps

Where $w_{orig}(t)$ is the weight of the term t in the original query that normalized using the maximum query terms weight. The $Score(t)$ is the KLD score or P-WNET score of the term t that also normalized using the maximum the terms score in the top document. The steps that are used in the Text preprocessing and indexing stage and the query processing stage appear in Fig. 4 and 5 respectively.

IV. EXPERIMENTS

In this work, python language is used to program the proposed information retrieval model. The TREC dataset was used for the Linguistic Data Consortium (LDC) in Philadelphia, USA. The documents set is the Associated Press disk 2 and Wall Street Journal disk 2 (AP2WSJ2) which consist of 154,443 documents. The query set is the queries number from 51 to 200 from TREC 1, 2, and 3. Queries, which also called ‘topics’ in TREC, have special SGML mark-up tags such as *narr*, *desc* and *title*.

Only the queries title field used contain in average 2.3 word length. Relevance judgments are also part of the TREC collection. In fact, the relevance judgment is marked each document in the documents set as either irrelevant or relevant with every query.

To examine the performance of the QESBIRM, two experiments were conducted using the data set. The first experiment was for SBIRM model, and the second experiment was for QESBIRM that used KLD, P-WNET expansion approach.

The retrieval performance of the QE approach is affected by two parameters. One of the parameter is the top ranked documents number that known as pseudo-relevance set. The second is the informative expansion terms number that, add to the query. The parameters set to $D=10$ and $T=20, 40, 60$ respectively, which perform a good improvement base on the studies in [6, 10, 13].

V. RESULTS

As see in Fig.6 and 7 and Table 1, the SBIRM using KLD or P-WNET improve the performance of the SBIRM. Main reasons behind this are the mismatch issues. It is concluded that the hybrid approach used in this work, i.e. SBIRM using KLD and P-WNET produces high performance in retrieve more relevant document by considering the proximity and expand the query.

VI. CONCLUSIONS

This research studies the impact of extending the query by adding statistical and semantic related terms to the original query terms on proximity base IR model. This is done by combining the SBIRM model with KLD or P-WNET. The QESBIRM using KLD, P-WNET were tested and evaluated. The experiment results show that the QESBIRM using KLD and P-WNET approach outperformed the SBIRM in precision, GMAP and MAP metric.

Recommended future work is to investigate the impact of the other QE approaches and combine KLD and P-WordNet in the SBIRM proximity base IR. It is also of interest to evaluate the model developed with samples written in other languages like Arabic. Another possible research direction is to discover the performance of other proximity base retrieval models with extending the query using QE approaches. Finally, the semantic features can use in text mining models such as text classification and clustering that consider the proximity.

TABLE I. THE PERFORMANCE RESULTS OF THE SBIRM AND QESBIRM

Approach	P@10	P@15	P@20	Map	G-MAP	R-precision
SBIRM	0.439	0.421	0.406	0.232	0.111	0.270
SBIRM with KLD (D=10,T=60)	0.465	0.447	0.421	0.244	0.113	0.271
SBIRM with KLD (D=10,T=40)	0.469	0.448	0.430	0.249	0.115	0.277
SBIRM with KLD (D=10,T=20)	0.467	0.438	0.422	0.252	0.114	0.281
SBIRM with P-WNET (D=10,T=60)	0.459	0.436	0.422	0.251	0.119	0.284
SBIRM with P-WNET (D=10,T=40)	0.472	0.444	0.422	0.245	0.117	0.277
SBIRM with P-WNET (D=10,T=20)	0.467	0.44	0.423	0.249	0.119	0.281

- 1) For each query term $t \in Q$
 - Retrieve inverted list I_t containing Transforming term signals $\{\tilde{\zeta}_{0,t}, \tilde{\zeta}_{1,t}, \dots, \tilde{\zeta}_{d,t}\}$
 - 2) Compute the score for each d document in set using Transform signal $(\zeta_{d,t})$.
 - a) For each magnitudes of the spectral component $\zeta_{d,t,b} \in \tilde{\zeta}_{d,t}$
 - i. Calculate the magnitudes of the signal component using (8)
 - ii. Calculate the unit phase of the signal component using (9)
 - iii. In the spectra of the word signal, For each b component
 - A. Calculate the Zero phase precision using (10)
 - B. Compute the score of the component as (11)
$$s_{d,b} = \phi_{d,b} \sum_{t \in Q} w_{q,t} H_{d,t,b}$$
 - b) Combine component score to obtain document score using (12)
 - 3) Sort the document base on the document score.
 - 4) Select top D retrieved documents.
 - 5) CET list that contains all unique terms of top k retrieved documents.
 - 6) Compute the KLD or P-WNET score for each term in CET equation using (14, 20) respectively.
 - 7) Select T top score terms expansion terms from CET.
 - 8) Add expansion terms to the original query to formulate the new query.
 - 9) Re-weight the new query.
- Repeat step 1, 2 and 3 with the new query.

Fig. 5. The query processing stage steps

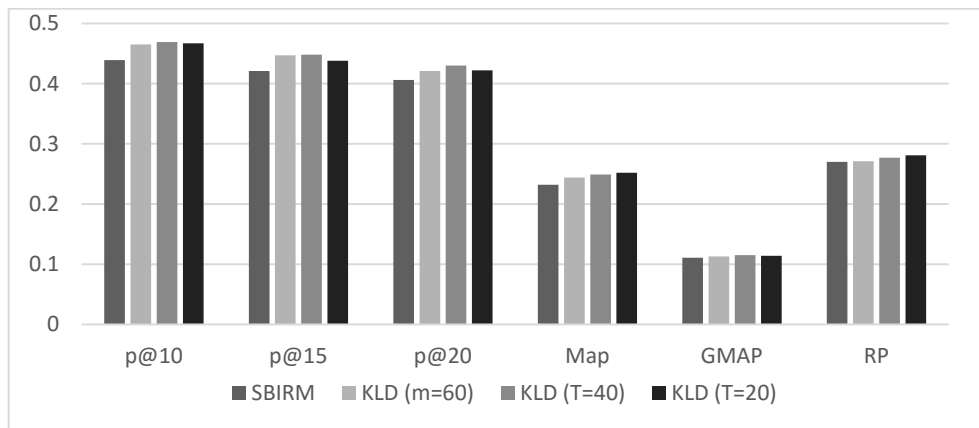


Fig. 6. Comparison of the SBIRM and KLD over the SBIRM

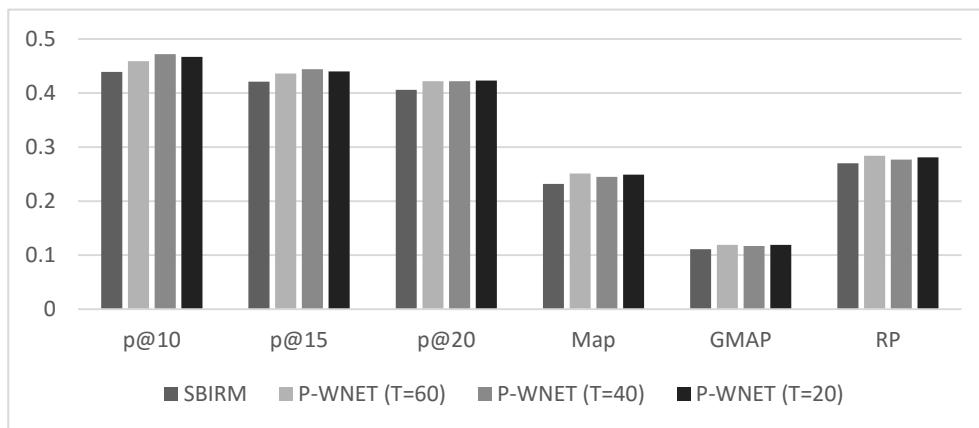


Fig. 7. Comparison of the SBIRM and P-WNET over the SBIRM

REFERENCES

- [1] D. Hawking and P. Thistlewaite, Relevance weighting using distance between term occurrences, Australian National University, 1996.
- [2] P. Laurence, "Spectral Based Information Retrieval," PhD thesis, Electrical and Electronic Engineering Department, Melbourne University, Melbourne, Australia, 2003.
- [3] P. Laurence, K. Ramamohanarao, and M. Palaniswami, "Fourier domain scoring: A novel document ranking method," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no.5, pp. 529-539, 2004.
- [4] P. Laurence, K. Ramamohanarao, and M. Palaniswami, "A novel document ranking method using the discrete cosine transform," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no.1, pp. 130-135, 2005.
- [5] P. Laurence, K. Ramamohanarao, and M. Palaniswami, "A novel document retrieval method using the discrete wavelet transform," ACM Transactions on Information Systems, vol. 23, no.3, pp. 267-298, 2005.
- [6] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, "An information-theoretic approach to automatic query expansion," ACM Transactions on Information Systems, vol. 19, no.1, pp. 1-27, 2001.
- [7] E. Robertson, "On term selection for query expansion," Journal of documentation, vol. 46, no. 4, pp. 359-364, 1990.
- [8] T. Doszkocs, "AID, an associative interactive dictionary for online searching," Online Review, vol. 2, no. 2, pp. 163-173, 1978.
- [9] G. Amati, C. Joost, and V. Rijsbergen, "Probabilistic models for information retrieval based on divergence from randomness," 2003.
- [10] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," ACM Computing Surveys, vol. 44, no. 1, pp. 1-50, 2012.
- [11] A. Barman, J. Sarmah, and S. Sarma, "WordNet Based Information Retrieval System for Assamese," in Proceedings of IEEE 15th International Conference on Computer Modelling and Simulation, pp. 480-484, 2013.
- [12] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E. Milios, "Semantic similarity methods in wordNet and their application to information retrieval on the web," in Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 10-16, 2005.
- [13] D. Pal, M. Mitra, and K. Datta, "Improving query expansion using WordNet," Journal of the Association for Information Science and Technology, vol. 65, no. 12, pp. 2469-2478, 2014.
- [14] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, "Query expansion via wordnet for effective code search," in Proceedings of IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering, pp. 545-549, 2015.
- [15] H. Fang, "A Re-examination of Query Expansion Using Lexical Resources," in Association for Computational Linguistics, pp. 139-147, 2008.
- [16] S. Tyar and M. Than, "Sense-based Information Retrieval System by using Jaccard Coefficient Based WSD Algorithm," in Proceedings of 2015 International Conference on Future Computational Technologies, pp. 197-203, 2015.
- [17] J. Singh and A. Sharan, "Co-occurrence and Semantic Similarity Based Hybrid Approach for Improving Automatic Query Expansion in Information Retrieval," in Proceedings of International Conference on Distributed Computing and Internet Technology, pp. 415-418, 2015.
- [18] C. Clarke and G. Cormack, "Shortest-substring retrieval and ranking," ACM Transactions on Information Systems (TOIS), vol. 18, no. 1, pp. 44-78, 2000.
- [19] M. Beigbeder and A. Mercier, "An information retrieval model using the fuzzy proximity degree of term occurrences," in Proceedings of the 2005 ACM symposium on Applied computing, pp. 1018-1022, 2005.
- [20] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis, "Incorporating term dependency in the DFR framework," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 843-844, 2007.
- [21] S. Butcher C. L. A. Clarke, and B. Lushman, "Term proximity scoring for ad-hoc retrieval on very large text collections," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 621-622, 2006.
- [22] B. He, J. X. Huang, and X. Zhou, "Modeling term proximity for probabilistic information retrieval models," Information Sciences, vol. 181, no. 14, pp. 3017-3031, 2011.
- [23] D. Metzler and W. Croft, "A Markov random field model for term dependencies," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.
- [24] A. El Mahdaouy, E. r. Gaussier, and S. d. O. El Alaoui, "Exploring term proximity statistic for Arabic information retrieval," in 2014 Third IEEE International Colloquium in Information Science and Technology, pp. 272-277, 2014.
- [25] Y. Lv , and C. Zhai. "Positional language models for information retrieval," in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009.
- [26] J. Zhao, J. Huang and B. He. "CRTER: using cross terms to enhance probabilistic information retrieval", In Proceedings of the international ACM SIGIR conference on research and development in information retrieval , pp. 155-164, 2011.
- [27] K. Ramamohanarao and L. A. F. Park, "Spectral-based document retrieval," in Advances in Computer Science-ASIAN 2004. Higher-Level Decision Making: Springer, pp. 407-417, 2004.
- [28] T. Li, Q. Li, S. Zhu, and M. Ogihara, "A survey on wavelet applications in data mining," ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 49-68, 2002.
- [29] J. Walker, A primer on wavelets and their scientific applications, CRC press, 2008.
- [30] h. Almfareji, "Web Document Clustering Using Discrete Wavelet Transforms," M.S. thesis, Computer Science Department, King Abdulaziz University, Jeddah, Saudia Arabia, 2015.
- [31] A. Diwali, M. Kamel, and M. Dahab, "Arabic Text-Based Chat Topic Classification Using Discrete Wavelet Transform," International Journal of Computer Science Issues, vol. 12, no. 2, p. 86, 2015.
- [32] S. Thaicharoen, T. Altman, and K. J. Cios, "Structure-based document model with discrete wavelet transforms and its application to document classification," in Proceedings of the 7th Australasian Data Mining Conference-Volume 87, pp. 209-217, 2008.
- [33] G. Xexéo, J. Souza, P. Castro, and W. Pinheiro, "Using wavelets to classify documents," in Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 272-278, 2008.
- [34] G. Arru, D. Feltoni Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Signal-based user recommendation on twitter," in Proceedings of the 22nd International Conference on World Wide Web Steering Committee/ACM, pp. 941-944, 2013.
- [35] J. Rocchio, "Relevance feedback in information retrieval," In The SMART Retrieval System- Experiments in Automatic Document Processing, pp. 313-323, 1971.
- [36] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, vol. 41, no. 6, pp. 391, 1990.
- [37] A. g. Rivas, E. Iglesias, and L. Borrajo, "Study of query expansion techniques and their application in the biomedical information retrieval," The Scientific World Journal, 2014.
- [38] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," ACM Transactions on Information Systems , vol. 18, no. 1, pp. 79-112, 2000.
- [39] D. Pal, M. Mitra, and K. Datta, "Query expansion using term distribution and term association," arXiv preprint arXiv:1303.0667, 2013.
- [40] V. Lavrenko and W. Croft, "Relevance based language models," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 120-127, 2001.
- [41] R. Nawab, M. Stevenson, and P. Clough, "Retrieving candidate plagiarised documents using query expansion," in Proceedings of

- European Conference on Information Retrieval, Springer Berlin Heidelberg, pp. 207-218, 2012.
- [42] R. Selvi and E. Raj, "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval Using SBIR Algorithm," in Proceedings of IEEE 2014 World Congress in Computing and Communication Technologies, pp. 137-141, 2014.
- [43] B. Audeh, "Experiments on two Query Expansion Approaches for a Proximity-based Information Retrieval Model," in Rencontre des Jeunes Chercheurs en Recherche d'Information 2012, pp. 407-412, 2012.
- [44] M. Lease, "An improved markov random field model for supporting verbose queries." in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 476-483, 2009.
- [45] M. Porter, "An algorithm for suffix stripping," Program, vol. 14, pp. 130-137, 1980.
- [46] J. Zobel and A. Moffat, "Exploring the similarity space," in ACM SIGIR Forum, vol. 32. no. 1, pp. 18-34, 1998.
- [47] R. Hamming and L. Trefethen, "Haar wavelets," 1999.
- [48] T. Cover and J. Thomas, "Elements of information," TheoryWiley, New York, 1991.

A Hybrid Steganography System based on LSB Matching and Replacement

Hazem Hiary and Khair Eddin Sabri
Computer Science Department,
The University of Jordan, Amman, Jordan

Mohammed S. Mohammed and Ahlam Al-Dhamari
Computer Engineering Department,
Hodeidah University, Hodeidah, Yemen

Abstract—This paper proposes a hybrid steganographic approach using the least significant bit (LSB) technique for grayscale images. The proposed approach uses both LSB matching (LSB-M) and LSB replacement to hide the secret data in images. Using hybrid LSB techniques increase the level of security. Thus, attackers cannot easily, if not impossible, extract the secret data. The proposed approach stores two bits in a pixel. The embedding rate can reach up to 1.6 bit per pixel. The proposed approach is evaluated and subjected to various kinds of image processing attacks. The performance of the proposed algorithm is compared with two other relevant techniques; pixel-value differencing (PVD) and Complexity Based LSB-M (CBL). Experimental results indicate that the proposed algorithm outperforms PVD in terms of imperceptibility. Also, it significantly outperforms CBL in two main features; higher embedding rate (ER), and more robust to most common image processing attacks such as median filtering, histogram equalization, and rotation.

Keywords—Steganography; LSB matching; LSB replacement; Embedding capacity; Imperceptibility

I. INTRODUCTION

These days data transmission on the digital communications via the internet confronts a wide range of security issues [1], [2], [3]. Consequently, powerful digital techniques are needed to protect data during its transmission on the internet. One of the great interest solutions used to protect data is steganography [4], [5], [6]. Steganography is the art of invisible communication by hiding a secret message in a digital cover media such as images [7], [8], text [9], audio [10], video [11] and network traffic [12], without being dubious [13], [14]. Because digital images have a great deal of redundant data, there has been an increased interest in utilizing them as cover media for steganographic purposes [15], [16].

Image steganographic algorithms can be classified in terms of the embedding domain into two main classes: spatial domain and frequency domain-based algorithms [17], [18]. Spatial domain-based algorithms conceal secret message straightforwardly in the intensity of pixels of an image, while in frequency domain-based algorithms, the image is firstly transformed into its frequency domain and secret message is then concealed in the transform coefficients [19], [20].

LSB replacement is one of the most well-known methods in the spatial domain [21]. In this method, a secret data is embedded into a cover image by replacing the LSBs of the cover image pixels with secret data bits to get the stego image [22], [23]. For a cover image, LSB replacement increases the even pixels by one or abandons them unaltered, while it decreases

the odd pixels by one or leaves them unchanged [24]. Due to the feeble sensibility of the human visual system (HVS), the presence of the embedded secret data cannot be perceptible. The quality of the stego image presented by LSB replacement may not be tolerable if a large amount of LSB is to be used in the embedding process. As a case, a stego image can accomplish as low as 31.78 dB of the PSNR by utilizing LSB-4 replacement [25]. Numerous credible steganographic methods have been devised for LSB replacement method [26].

LSB matching (LSB-M) method was proposed in [27], which is also called ± 1 embedding [28]. In this scheme, the pixel value of the cover image is increased or decreased randomly by one when the secret bit is not equal to the LSB of the cover image pixel [5]. The LSB-M changes both the histogram of an image and the correlation between adjacent pixels and this helps steganalysis methods to attack this method [29].

In [18], the authors proposed a method called Complexity Based LSB-M (CBL). The method employs the strategy of adaptivity and the use of LSB-M in order to increase the security against attacks. CBL uses a local neighborhood analysis for determination of secure locations of an image and then it uses LSB-M for the embedding purpose.

In this paper, a hybrid approach using both LSB-M and LSB replacement methods is proposed. The approach is an improvement over the CBL method in order to increase both the embedding capacity and the robustness. The proposed approach stores two bits of the secret bits in a pixel. One bit is stored in the seventh bit using LSB-M technique; the other bit is stored in the eighth bit using LSB replacement technique. Thus, the maximum embedding capacity is increased to double (1.6) bits per pixel (bpp) compared to CBL maximum embedding capacity (0.8). This approach does not produce any distortions to be suspected by unauthorized observers and yields lower computational costs in its embedding and extraction processes. Moreover, it provides more robustness against most image processing attacks.

The remainder of this paper is organized as follows: In Section II a number of LSB-M based algorithms are represented. In Section III the embedding and extracting procedures for the proposed algorithm are presented. Experimental results and comparisons between the algorithm and other related algorithms are presented in Section IV. Finally, conclusions and future directions are given in Section V.

II. RELATED WORKS

The literature is worth of the contributions in the field of LSB steganography. The work in [30] proposed pixel-value differencing (PVD) method. The main idea behind PVD is to use the difference of two consecutive pixels of a grayscale image to hide data. In their method, a pixel-value differencing is used to distinguish between edge areas and smooth areas. Consequently, the capacity of embedded data in edge areas is higher than that of smooth areas. Recently, to enlarge the embedding efficiency on PVD method, a lot of methods were proposed by combining PVD and LSB replacement methods, such as [31], [32], [33]. With a slight alteration to the original PVD technique, the side match technique which is based on the correlation of a pixel with its neighboring pixels has additionally been evolved. In [34], the authors presented 2, 3 and 4-sided side match methods by using the correlation of a target pixel with its 2, 3, and 4 neighboring pixels.

Dissimilar to LSB replacement and LSB-M, LSB matching revisited (LSB-MR) uses a pair of pixels as hidden unit instead of one pixel [21]. This method uses grayscale cover images. The embedding process is performed on a cover pixel pair (g_i, g_{i+1}) at a time to embed a secret bit pair (b_i, b_{i+1}) . The corresponding stego pixel pair (g'_i, g'_{i+1}) can be obtained by keeping g_i and g_{i+1} unaltered, or by increasing or decreasing them by one. The method used the function $y = f(g_i, g_{i+1}) = LSB(\lfloor g_i/2 \rfloor + g_{i+1})$ to evaluate whether or not the pixel values g_i and g_{i+1} need alteration. However, on average, the embedding rate (bpp) for both LSB-M and LSB-MR is about 1 bpp, which is poor. A generalized LSB-M scheme (G-LSB-M) was proposed in [35] to generalize the method in [21]. To enhance the level of security of both LSB-M and G-LSB-M, a content adaptive method was proposed in [36]. In this method, if the secret bit does not match the LSB of corresponding cover image pixel, the decision of alteration direction is not arbitrary and is attempted to have the best correlation with the neighboring pixels. In [37], an approach called (ALSBMR) used LSB-MR with adaptive embedding.

In [18], the authors proposed LSB-M adaptive steganography algorithm called Complexity Based LSB-M (CBL). They used an 8-neighborhood of a pixel to determine the complexity region for embedding data in that region. They used LSB-M to embed data. The drawback of CBL algorithm is the low embedding capacity where it can not embed more than one bit in a pixel. This drawback is overcome in the proposed work by using LSB-M and LSB replacement techniques to increase the embedding capacity.

In [38], the authors proposed a data hiding algorithm based on interpolation, LSB substitution, and histogram shifting. In this work interpolation is used to adjust embedding capacity with low image distortion, the embedding process is then applied using LSB substitution and histogram shifting methods. In [39], the LSB substitution is improved by using a bit inversion technique. In this work secret data is hidden after compressing smooth areas of the image losslessly, resulting in fewer number of modified cover image pixels. A bit inversion technique is then applied where certain LSBs of pixels are modified if they occur in a particular pattern. In [40], a semi-reversible data hiding method which employs interpolation and LSB substitution is proposed. Interpolation is first used to scale up and down the cover image before hiding secret data to

achieve high embedding capacity with low image distortion. Then, embedding is done using the LSB substitution method.

III. PROPOSED ALGORITHM

Embedding capacity, visual quality of stego image (imperceptibility), efficiency regarding execution time and the security level (robustness) are four primary criteria that are utilized to evaluate the performance of the steganographic scheme. The proposed algorithm uses both LSB-M and LSB replacement to address these criteria. An adaptive algorithm that is a modification to CBL approach [18] is proposed. Details of embedding and extraction phases are presented in the next lines.

A. Embedding phase

In the embedding phase, the algorithm embeds two bits in the pixels that have complexity value equal or more than a threshold value. One bit is embedded using LSB-M in the seventh bit from the left; another bit is embedded using LSB replacement in the least significant bit. The embedding phase is illustrated in the diagram shown in Fig. 1.

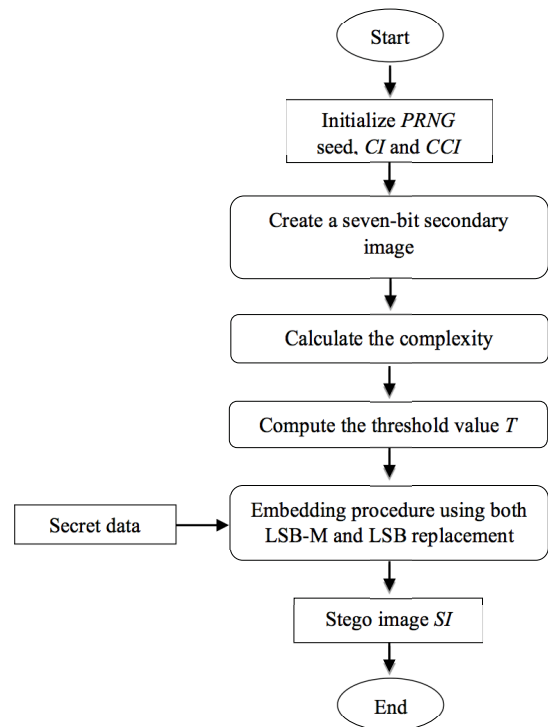


Fig. 1. Flow diagram of the embedding phase

The proposed embedding algorithm is based on CBL algorithm [18]. Step 2 and step 5 in CBL algorithm were modified to achieve more capacity. In step 2, the least significant bit is removed and only the first seven bits are used to create the secondary image. In step 5, both LSB-M and LSB replacement are used to embed two bits of the secret data in the selected pixels. The proposed embedding algorithm is described in the following steps.

Step 1. Initialization: In this step, number of variables are initialized as follows:

```

Set PRNG seed
CI          ← cover image
CCI        ← copy of cover image
[M, N]     ← dimensions of cover image
    
```

where PRNG is a pseudo random number generator, PRNG is initialized by a seed, which is a number that have to be selected and shared between the sender and the receiver.

Step 2. Secondary image formation: In this step, the least significant bit of each pixel in CI is removed and only the first seven bits are used to create the secondary image. This image will be used to compute the complexity values of pixels in the next step. Using the secondary image helps the receiver to get the same complexity value [18]. The following routine creates this image.

```

for each pixel CI(x, y) do
  r ← random number ∈ [0, 1]
  CI(x, y) = bitshift(CI(x, y), -1)
  if CI(x, y) is odd then
    if r ≤ 0.5 then
      CI(x, y) ← CI(x, y) + 1
    else
      CI(x, y) ← CI(x, y) - 1
    end if
  end if
end if
end for
    
```

where bitshift function will shift the pixel bits to the right, so it will remove the least significant bit. For example, if the pixel value is $(215)_{10} = (11010111)_2$, then the pixel value after the shifting process will be $(1101011)_2 = (107)_{10}$.

Step 3. Pixel complexity computation: The complexity of each pixel is computed by adding absolute values of differences of the pixel with its neighbors as follows [18]

$$Complexity(x, y) = \sum_{i=-1}^1 \sum_{j=-1}^1 |CI(x, y) - CI(x+i, y+j)| \quad (1)$$

Figure 2 shows the neighborhood of pixel $CI(x, y)$. The complexity value is an indication of the type of region that the pixel belongs to; edge or smooth region. Where more data can be embedded in edge regions without creating any suspicion. A high complexity value indicates the pixel is located in an edge region, while a low value indicates the pixel is located in a smooth region. Examples of cover images and their complexity values are illustrated in Fig. 3.

$CI(x-1, y-1)$	$CI(x-1, y)$	$CI(x-1, y+1)$
$CI(x, y-1)$	$CI(x, y)$	$CI(x, y+1)$
$CI(x+1, y-1)$	$CI(x+1, y)$	$CI(x+1, y+1)$

Fig. 2. 8-neighbors of pixel $CI(x, y)$

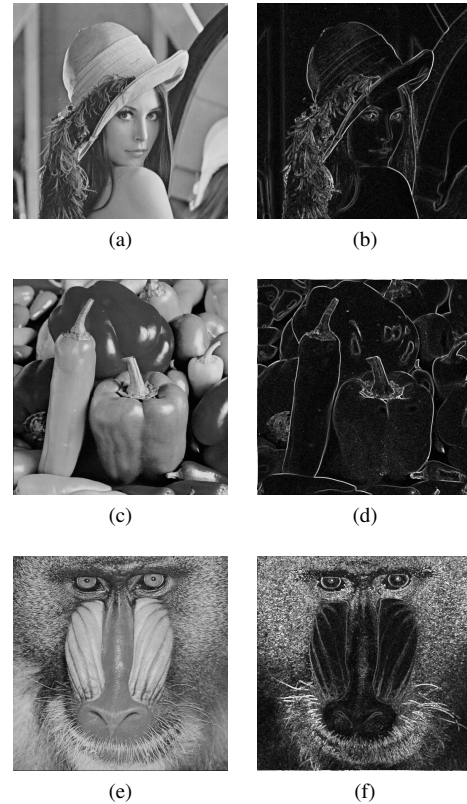


Fig. 3. Cover images (left) and their octonary-complexity values (right)

Step 4. Threshold computation: Complexity values of pixels are compared with a threshold; values greater than or equal to threshold indicate an edge region, and smooth region otherwise. To calculate threshold value, the number of pixels to be embedded (NP) must be computed using $NP = SB/2$, where SB refers to the total number of embedded secret bits. The threshold value T is chosen to make sure that at least NP of the pixels are complex. The following routine shows how T is computed.

```

t_o ← 127
n ← 0
L:
for each pixel Complexity(x, y) do
  if t_o ≥ Complexity(x, y) then
    n ← n + 1
  end if
end for
if n < NP AND t_o ≠ 0 then
  t_o = t_o - 1
  go to L
else
  T = t_o
end if
    
```

where t_o is a temporary value of threshold, which starts with the maximum value of the seven-bit pixel.

Step 5. Embedding: In this step, the secret data are embedded using LSB-M and LSB replacement techniques. Two bits are embedded in each selected pixel; the first one is embedded using LSB-M as in CBL but in the seventh bit, the second bit

is embedded in the least significant bit using LSB replacement.

The following pseudocode presents the embedding procedure, where s is the string of message bits and SI is the stego image.

The last four lines of the algorithm present the proposed modification on this step. The selected pixel value of the secondary image is converted to seven bits binary by using $dec2bin$ function and stored in $conca$. One of the secret bits is concatenated to the seven bits in $conca$ by using $strcat$ function and stored in emb . Finally, the eight bits in emb are converted to decimal value by using $bin2dec$ function and the resulting value is stored in SI .

```

set PRNG
n ← 0
for each pixel CI(x, y) do
    r ← random number ∈ [0, 1]
    if CI(x, y) = 0 OR Complexity(x, y) < T then
        SI(x, y) ← CCI(x, y)
    else
        if CI(x, y) mod 2 ≠ s(n) then
            if r ≤ 0.5 then
                SI(x, y) ← CI(x, y) - 1
            else
                SI(x, y) ← CI(x, y) + 1
            end if
        end if
        n ← n + 1
        conca ← dec2bin(SI(x, y), 7)
        emb ← strcat(conca, s(n))
        SI(x, y) ← bin2dec(emb)
        n ← n + 1
    end if
end for

```

B. Extraction phase

The extraction phase is the same as embedding in the first four steps, except that the stego image SI is used instead of the cover image to create the secondary image and compute the complexity, and a copy of stego image CSI is created to be used in the extraction step. Figure 4 shows the flow diagram of the extracting phase.

After the first four steps are done, the extraction step works by extracting the two least significant bits of each pixel that is labeled as complex. The following pseudocode presents the extraction routine, where $mod\ 4$ is used to extract the two least significant bits as a decimal number, the $dec2bin$ function is used to convert the decimal number back to two binary bits, which are stored in s .

```

n ← 0
for each pixel SI(x, y) do
    if SI(x, y) ≠ 0 AND Complexity(x, y) ≥ T then
        sec ← CSI(x, y) mod 4
        s(n : n + 1) ← dec2bin(sec)
        n ← n + 2
    end if
end for

```

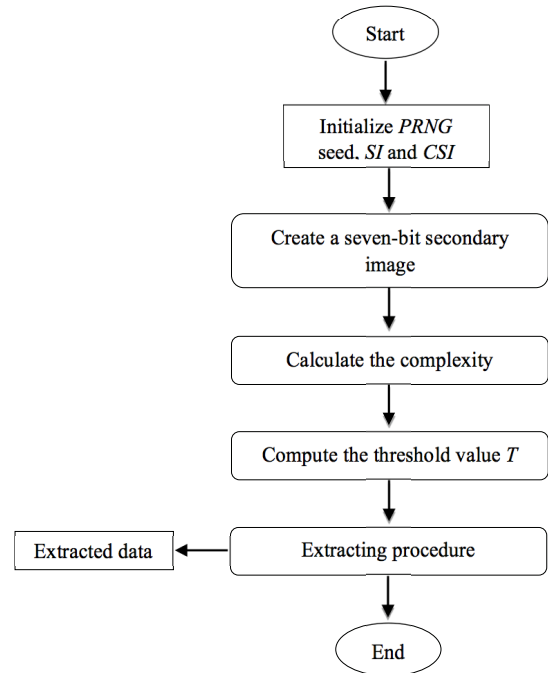


Fig. 4. Flow diagram of the extraction phase

IV. EXPERIMENTAL RESULTS AND COMPARISONS

Here, experimental results are presented to demonstrate the performance of the proposed algorithm. Several grayscale test images of size 512×512 were used from [41], [42]. Sample of these images is shown in Fig. 5.

Generally, the steganographic algorithms can be evaluated by two benchmarks; the embedding rate and the imperceptibility (or the quality of the stego image). The embedding rate (ER) is defined as the number of secret data bits that can be embedded per pixel, it can be calculated as [43], [44]

$$ER = \frac{SB}{M \times N} \quad (bpp) \quad (2)$$

where SB refers to the total number of embedded secret bits, M and N are the width and height of the cover image, respectively. To gauge the imperceptibility or the quality of the stego image, the peak signal-to-noise ratio (PSNR) is used. A high PSNR value indicates a high similarity between the stego and cover images, while a low value demonstrates the opposite. PSNR can be computed as [30], [45]

$$PSNR = 10 \log_{10} \frac{MAX^2}{MSE} \quad (dB) \quad (3)$$

where MAX is the maximum value of the pixel intensity, e.g., $MAX = 255$ for 8-bit grayscale images. MSE is the mean square error, defined as

$$MSE = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (CI(i, j) - SI(i, j))^2 \quad (4)$$

Figure 6 shows stego images for five cover images using the proposed algorithm. The PSNR values range between 47.77dB and 48.13dB when the embedding rate is 0.8 bpp, and between

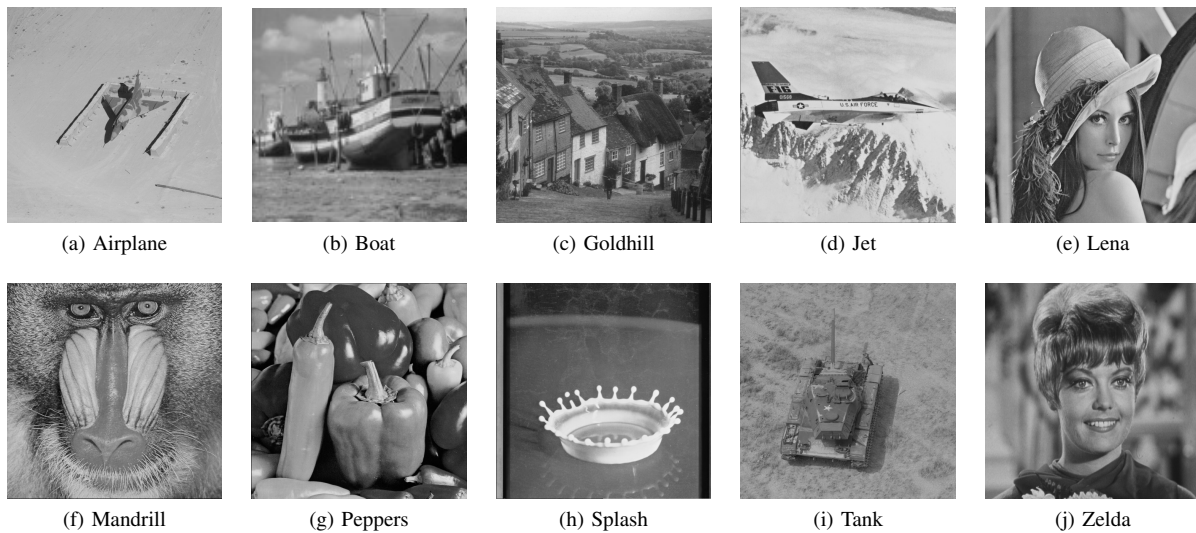


Fig. 5. Sample test images used in the experiments

44.93dB and 45.09dB when the embedding rate is 1.6 bpp. However, in all cases, there is no distortion to be aware of by the human eye.

To evaluate the proposed approach, a comparison was conducted with other approaches: PVD [30] and CBL [18]. Table I presents this comparison, where four embedding rates (0.3, 0.5, 0.8 and 1.6 bpp) were used for each method, and the PSNR average is calculated for the 10 test images.

TABLE I. IMPERCEPTIBILITY AND ER COMPARISON BETWEEN THE PROPOSED APPROACH AGAINST PVD [30] AND CBL [18]

Approach	ER (bpp)	PSNR average (dB)
PVD	0.3	49.72
	0.5	47.30
	0.8	44.81
	1.6	41.83
CBL	0.3	56.16
	0.5	53.95
	0.8	51.94
	1.6	-
Proposed	0.3	52.24
	0.5	50.00
	0.8	47.97
	1.6	45.06

Since CBL cannot embed 1.6 bpp, the corresponding PSNR value is left as (-). It is noticeable that the proposed algorithm in all test images and embedding rates provides better imperceptibility values than PVD. Moreover, compared to CBL, the approach provides double embedding rate. However, average PSNR values are less; this is because CBL only embeds one bit per pixel, while the approach embeds two bits.

To further evaluate the proposed approach, the stego images were subjected to different kinds of image processing attacks, namely JPEG 2000 lossy, sharpening, Gaussian noise, median filter, contrast enhancement, Gaussian filter, histogram equalization, and rotation. Table II shows the average bit correct rate (BCR) [46] of the watermarks after applying these attacks at embedding rate 0.8 bpp. It is worth mentioning that the BCR

can be computed using

$$BCR = \frac{L - \sum_{i=1}^L OB \oplus EB}{L} \quad (5)$$

where OB is the original bit and EB is the extracted bit of the watermark. L is the length of the watermark.

The obtained results show that the proposed algorithm is more robust than CBL against median filtering, histogram equalization and rotation attacks. BCR values after applying JPEG 2000 lossy, sharpening, Gaussian noise and contrast enhancement were similar. All BCR values are low (between 0.48 and 0.63) because the attacks change the secondary image of stego image.

TABLE II. AVERAGE BCR COMPARISON AFTER IMAGE ATTACKS BETWEEN THE PROPOSED APPROACH AND CBL [18]

Attack	CBL	Proposed
JPEG 2000 lossy	0.50	0.50
Sharpening 3×3	0.50	0.50
Gaussian noise (0.001)	0.50	0.50
Median filter 3×3	0.52	0.63
Contrast Enhancement	0.50	0.50
Gaussian filter 3×3	0.50	0.48
Histogram equalization	0.50	0.54
Rotation 35°	0.53	0.58
Rotation 75° w/auto-crop	0.53	0.58

V. CONCLUSIONS AND FUTURE WORK

In this paper, by considering the significance of the embedding efficiency of steganographic algorithms, a hybrid approach is proposed for data hiding with high capacity and robustness. The proposed approach is a modification to CBL to achieve more embedding capacity. The proposed approach uses both LSB-M and LSB replacement techniques to conceal secret data in the least significant two bits of the pixel values. Thus, the proposed approach can achieve more embedding capacity than CBL, which only embeds secret data in the least significant bit of pixel values.

The maximum embedding rate achieved by the proposed approach is 1.6 bpp, which is double of the capacity achieved

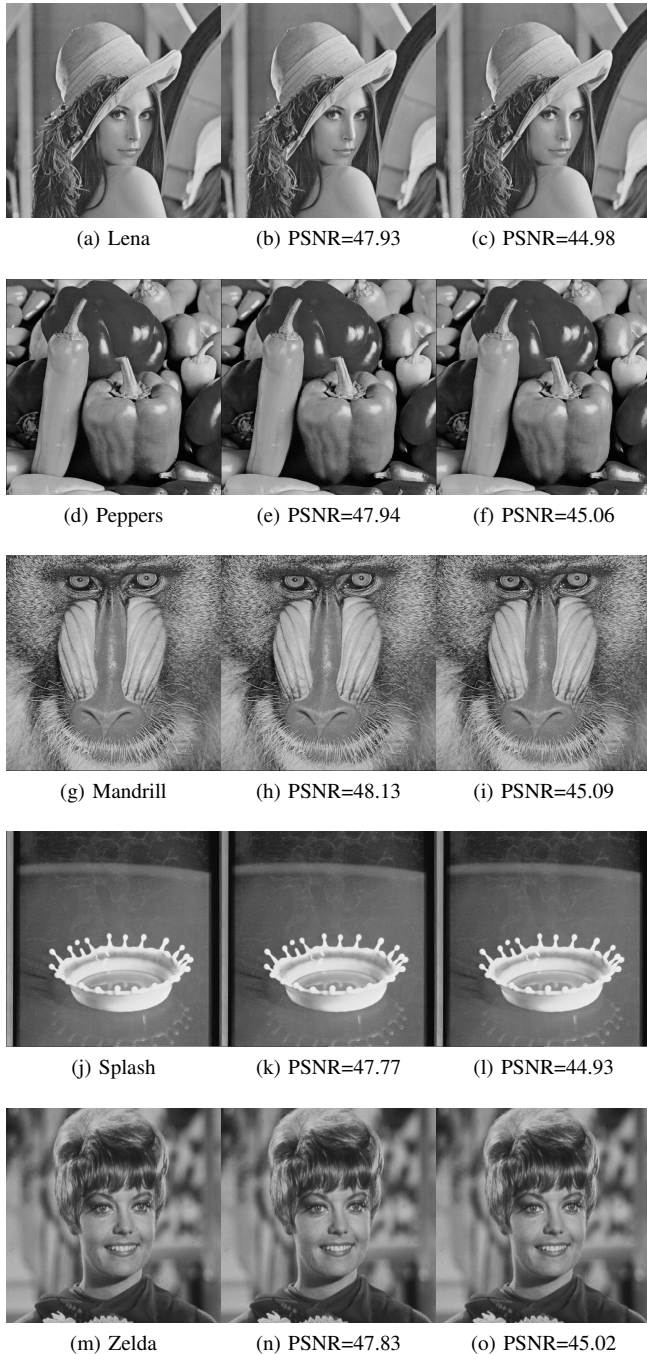


Fig. 6. Cover images (left), stego images with ER=0.8 bpp (middle), and ER=1.6 bpp (right)

by CBL. By comparing the proposed approach to PVD and CBL, the imperceptibility over PVD is improved, embedding capacity is also improved over CBL, without distorting the stego images. The watermarks in the approach are more robust than CBL when subjected to some common attacks.

Since the approach is based on CBL algorithm, future plans include proposing a new algorithm that achieves a higher embedding rate than what is accomplished (1.6 bpp), and improve the imperceptibility.

REFERENCES

- [1] N. Hamid, A. Yahya, R. Ahmad, and O. Al-Qershi, "Image steganography techniques: an overview", *International Journal of Computer Science and Security (IJCSS)*, vol. 6, no. 3, pp. 168–187, 2012.
- [2] C. Gayathri and V. Kalpana, "Study on image steganography techniques", *International Journal of Engineering and Technology (IJET)*, vol. 5, no. 2, pp. 572–577, 2013.
- [3] L. Saini and V. Shrivastava, "A survey of digital watermarking techniques and its applications", *International Journal of Computer Science Trends and Technology (IJCT)*, vol. 2, no. 3, pp. 70–73, 2014.
- [4] A. Tiwari, S. Yadav, and N. Mittal, "A review on different image steganography techniques", *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 3, no. 7, pp. 121–124, 2014.
- [5] F. Huang, Y. Zhong, and J. Huang, "Improved algorithm of edge adaptive image steganography based on LSB matching revisited algorithm", *Lecture Notes in Computer Science, Springer Berlin Heidelberg*, vol. 8389, pp. 19–31, 2014.
- [6] M. Khosravi and A. Naghsh-Nilchi, "A novel joint secret image sharing and robust steganography method using wavelet", *Multimedia systems*, vol. 20, no. 2, pp. 215–226, 2014.
- [7] B. Mohd, S. Abed, B. Na'ami, and T. Hayajneh, "Hierarchical steganography using novel optimum quantization technique", *Signal, Image and Video Processing (SIViP)*, vol. 7, no. 6, pp. 1029–1040, 2013.
- [8] X. Li, W. Zhang, B. Ou, and B. Yang, "A brief review on reversible data hiding: current techniques and future prospects", *Proc. IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, Xi'an, China, 2014, pp. 426–430.
- [9] A. Odeh, K. Elleithy, and M. Faezipour, "Steganography in text by using MS word symbols", *Proc. Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1)*, Bridgeport, CT, USA, 2014, pp. 1–5.
- [10] P. Pathak, A. Chattopadhyay, and A. Nag, "A new audio steganography scheme based on location selection with enhanced security", *Proc. International Conference on Automation, Control, Energy and Systems (ACES)*, Hooghy, India, 2014, pp. 1–4.
- [11] M. Beno, A. George, I. Valarmathi, and S. Swamy, "Hybrid optimization model of video steganography technique with the aid of biorthogonal wavelet transform", *Journal of Theoretical and Applied Information Technology*, vol. 63, no. 1, pp. 190–199, 2014.
- [12] W. Mazurczyk, P. Szaga, and K. Szczypiorski, "Using transcoding for hidden communication in IP telephony", *Multimedia Tools and Applications*, vol. 70, no. 3, pp. 2139–2165, 2014.
- [13] N. Johnson, Z. Duric, and S. Jajodia, "Information hiding: steganography and watermarking—attacks and countermeasures", Kluwer, USA, 2001.
- [14] G. Liu, W. Liu, Y. Dai, and S. Lian "Adaptive steganography based on block complexity and matrix embedding", *Multimedia systems*, vol. 20, no. 2, pp. 227–238, 2014.
- [15] R. Chandramouli, M. Kharrazi, and N. Memon, "Image steganography and steganalysis: concepts and practice", *Lecture Notes in Computer Science, Springer Berlin Heidelberg*, vol. 2939, pp. 35–49, 2004.
- [16] A. Martin, G. Sapiro, and G. Seroussi, "Is image steganography natural?", *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2040–2050, 2005.
- [17] P. Singh and R. Chadha "A survey of digital watermarking techniques, applications and attacks", *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 9, pp.165–175, 2013.
- [18] V. Sabeti, S. Samavi, and S. Shirani, "An adaptive LSB matching steganography based on octonary complexity measure", *Multimedia tools and applications*, vol. 64, no. 3, pp. 777–793, 2013.
- [19] V. Verma and R. Jha, "Improved watermarking technique based on significant difference of lifting wavelet coefficients", *Signal, Image and Video Processing (SIViP)*, vol. 9, no. 6, pp. 1443–1450, 2015.
- [20] V. Sabeti, S. Samavi, M. Mahdavi, and S. Shirani "Steganalysis and payload estimation of embedding in pixel differences using neural networks", *Pattern Recognition*, vol. 43, no. 1, pp. 405–415, 2010.
- [21] C. Sumathi, T. Santanam, and G. Umamaheswari, "A study of various steganographic techniques used for information hiding", *International*

- Journal of Computer Science and Engineering Survey (IJCSSES), vol. 4, no. 6, pp. 9–25, 2014.
- [22] KH. Jung and KY. Yoo, “High-capacity index based data hiding method”, *Multimedia Tools and Applications*, vol. 74, no. 6, pp. 2179–2193, 2015.
- [23] A. Khan, A. Siddiqua, S. Munib, and S. Malik, “A recent survey of reversible watermarking techniques”, *Information Sciences*, vol. 279, pp. 251–272, 2014.
- [24] J. Mielikainen, “LSB matching revisited”, *IEEE Signal Processing Letters*, 2006, vol. 13, no. 5, pp. 285–287, 2006.
- [25] NI. Wu, KC. Wu, and CM. Wang, “Exploring pixel-value differencing and base decomposition for low distortion data embedding”, *Applied Soft Computing*, vol. 12, no. 2, pp. 942–960, 2012.
- [26] X. Liao, Q. Wen, and J. Zhang, “A steganographic method for digital images with four-pixel differencing and modified LSB substitution”, *Journal of visual communication and image representation*, vol. 22, no. 1, pp. 1–8, 2011.
- [27] T. Sharp, “An implementation of key-based digital signal steganography”, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, vol. 2137, pp. 13–26, 2001.
- [28] B. Li, J. He, J. Huang, and Y. Shi, “A survey on image steganography and steganalysis”, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 2, pp. 142–172, 2011.
- [29] Z. Xia, X. Wang, X. Sun, Q. Liu, and N. Xiong, “Steganalysis of LSB matching using differences between nonadjacent pixels”, *Multimedia Tools and Applications*, vol. 75, no. 4, pp. 1947–1962, 2016.
- [30] D. Wu and W. Tsai, “A steganographic method for images by pixel-value differencing”, *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1613–1626, 2003.
- [31] T. Mahjabin, S. Hossain, M. Haque, “A block based data hiding method in images using pixel value differencing and LSB substitution method”. *Proc. International Conference on Computer and Information Technology (ICCIT)*, Chittagong, Bangladesh, 2012, pp. 168–172.
- [32] M. Khodaei and K. Faez, “New adaptive steganographic method using least-significant-bit substitution and pixel-value differencing”, *IET image processing*, vol. 6, no. 6, pp. 677–686, 2012.
- [33] JK. Mandal and D. Das, “A novel invisible watermarking based on cascaded PVD integrated LSB technique”, *Communications in Computer and Information Science*, Springer Berlin Heidelberg, vol. 305, pp. 262–268, 2012.
- [34] C. Chang and H. Tseng, “A steganographic method for digital images using side match”, *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1431–1437, 2004.
- [35] X. Li, B. Yang, D. Cheng, and T. Zeng, “A generalization of LSB matching”, *IEEE Signal Processing Letters*, vol. 16, no. 2, pp. 69–72, 2009.
- [36] C. Wang, X. Li, B. Yang, X. and Lu, C. Liu, “A content-adaptive approach for reducing embedding impact in steganography”. *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, 2010, pp. 1762–1765.
- [37] W. Luo, F. Huang, and J. Huang, “Edge adaptive image steganography based on LSB matching revisited”, *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 201–214, 2010.
- [38] Y. Tsai, Y. Huang, R. Lin, and C. Chan, “An Adjustable Interpolation-based Data Hiding Algorithm Based on LSB Substitution and Histogram Shifting”, *International Journal of Digital Crime and Forensics*, vol. 8, no. 2, pp. 48–61, 2016.
- [39] N. Akhtar, “An LSB Substitution with Bit Inversion Steganography Method”, *Smart Innovation, Systems and Technologies*, Springer India, vol. 43, pp 515–521, 2015.
- [40] K. Jung and K. Yoo, “Steganographic method based on interpolation and LSB substitution of digital images”, *Multimedia Tools and Applications*, vol. 74, no. 6, pp. 2143–2155, 2015.
- [41] USC-SIPI Image Database. <http://sipi.usc.edu/database/database.php>. Accessed 17th September 2016.
- [42] The University of Waterloo-Image repository. <http://links.uwaterloo.ca/Repository.html>. Accessed 17th September 2016.
- [43] M. Subhedar and V. Mankar, “Current status and key issues in image steganography: A survey”, *Computer Science Review*, vol. 13, pp. 95–113, 2014.
- [44] C. Lee and H. Chen, “A novel data hiding scheme based on modulus function”, *Journal of Systems and Software*, vol. 83, no. 5, pp. 832–843, 2010.
- [45] P. Gupta, R. Roy, and S. Changder, “A secure image steganography technique with moderately higher significant bit embedding”. *Proc. International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2014, pp. 1–6.
- [46] O. Okman, and G. Akar, “Quantization index modulation-based image watermarking using digital holography”, *Journal of the Optical Society of America A, Optics, Image Science, and vision*, vol. 24, no. 1, pp. 243–252, 2007.

A Novel High Dimensional and High Speed Data Streams Algorithm: HSDStream

Irshad Ahmed

Department of Computer Science
National University of Computer and
Emerging Sciences, Islamabad, Pakistan

Irfan Ahmed

Department of Computer Engineering
Taif University, Taif, KSA

Waseem Shahzad

Department of Computer Science
National University of Computer and
Emerging Sciences, Islamabad, Pakistan

Abstract—This paper presents a novel high speed clustering scheme for high-dimensional data stream. Data stream clustering has gained importance in different applications, for example, network monitoring, intrusion detection, and real-time sensing. High dimensional stream data is inherently more complex when used for clustering because the evolving nature of the stream data and high dimensionality make it non-trivial. In order to tackle this problem, projected subspace within the high dimensions and limited window sized data per unit of time are used for clustering purpose. We propose a High Speed and Dimensions data stream clustering scheme (HSDStream) which employs exponential moving averages to reduce the size of the memory and speed up the processing of projected subspace data stream. It works in three steps: i) initialization, ii) real-time maintenance of core and outlier micro-clusters, and iii) on-demand offline generation of the final clusters. The proposed algorithm is tested against high dimensional density-based projected clustering (HDDStream) for cluster purity, memory usage, and the cluster sensitivity. Experimental results are obtained for corrected KDD intrusion detection dataset. These results show that HSDStream outperforms the HDDStream in all performance metrics, especially, the memory usage and the processing speed.

Keywords—Evolving data stream; high dimensionality; projected clustering; density-based clustering; micro-clustering

I. INTRODUCTION

The exponential growth in data mining and clustering is an apparent result of the Internet penetration and the use of the network applications. Network applications have become an integral part of our daily life, whether it is related to the academic, research, health care, finance, business, or public service domains.

Data sources are monotonically increasing from past few decades. Additionally, the technological developments in data sensing systems (sensor networks) have resulted in a real-time data with large number of attributes. The large volume of the data together with its high dimensionality has motivated the research in the area of high dimensional data mining and exploration. Data stream is a form of data that continuously evolves reflecting the real-time variation in volume, dimensionality, and correlation. In recent years, a large amount of streaming data, such as network flows, wireless sensor networks data and the multimedia streams have been generated. Analyzing and mining of real-time streaming data have become a hot research topic [1], [2], [3]. Discovery of the patterns hidden in the streaming data imposes great challenges for cluster formation, especially in high dimensional data. By definition, a cluster

is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Data stream clustering algorithms are used to get important information from these streams in real-time. These algorithms search for the clusters that contain streaming objects with a certain degree of similarity across all dimensions. Stream clustering algorithms have special challenges that do not face most other clustering techniques. Storage and time limits are critical for clustering algorithms to perform a fast single-pass over that stream data. In addition to this, the evolving nature of the stream, requires the clustering algorithm to be highly adaptive to the new patterns. Generally, there are two types of stream clustering algorithms: full dimensional and projected or preferred dimension streaming algorithms. Clustering applications in various domains often have very high-dimensional data; the dimension of the data being in the tens, hundreds or thousands, for example, in network streaming, web mining and bioinformatics, respectively. It is often require to focus on a certain subset of dimensions rather than the full dimension space because it requires less memory and render fast processing. In addition to the high dimensionality, real-time high-speed evolution makes it more intractable. Clustering such high-dimensional high-speed datasets is a contemporary challenge. Clustering algorithms must avoid the curse of dimensionality but at the same time should be computationally efficient. Some applications that generate data streams include: telecommunication (call records), network operation centers (log information from network entities), financial market (stock exchange), and day to day business (credit card, automated teller machine (ATM) transactions, etc). In a high dimensional dataset, among many features some attributes can be expected to be irrelevant for any given object of interest. Irrelevant attributes can obscure clusters that are clearly visible when we consider only the relevant subspace of the dataset. Therefore, clusters may be meaningfully defined by some of the available attributes only. The irrelevant attributes interfere with the efforts to find targeted clusters. This problem is become more intensive in streaming data, because it requires a single scan of the data to find the useful attributes for describing a potential cluster for the current object. Moreover, streams are impulsive and the discovered clusters might also evolve over time. High dimensional streaming data clustering is more challenging than the high density or high dimensional data. Among various challenges in clustering high dimensional streaming data [4], following two are the focuses of this paper:

- Processing speed: Data streams arrive continuously, which requires fast and real-time response. The clustering algorithm needs to have processing speed (which comes from low complexity) such that it can handle the speed of data streams in the limited time.
- Memory usage: Large data streams are generated rapidly which need an unlimited memory. Therefore, the clustering algorithm must be optimized for realistic memory constraints.

In this paper, we introduce a novel tuple structure to summarize the high speed high dimensional data stream. This structure not only speed up the process but also requires less memory. Our clustering technique also modifies weights in some definitions of HDDStream, namely, the micro-cluster variance, projected dimensionality, projected distance, and projected radius. In terms of experimental results, we compare our scheme with HDDStream for cluster purity, memory usage, and cluster's sensitivity.

Notations:

Vectors and matrices are represented by bold letters, other notations are explained below:

\mathbb{R}	Set of real numbers
\mathbb{N}	Set of natural numbers
\mathcal{C}	Dataset
N	Window size
ϵ	Radius threshold
\mathcal{D}	Dataset used in initialization phases
α	Exponential weighted average constant
β	Outlier threshold
μ	Number of points threshold
ξ	Variance threshold
ψ_j	j^{th} preferred dimension
π	Projected dimensionality threshold

II. RELATED WORK

In the last few years many research works have been done on high-dimensional data clustering and evolving data streams clustering. There are extensive research works on clustering algorithms for static datasets [5], [6], [4] where some of them have been further extended for evolving data streams. The clusters are formed based on a Euclidean distance function like k -means algorithm [7]. k -mean clustering splits the n d -dimensional points into k cluster ($k < n$). One of the well-known extensions of k -means on data streams is presented by Aggarwal et al. [8]. They propose an algorithm called CluStream based on k -means for clustering evolving data streams. CluStream introduces an online-offline method for clustering data streams. CluStream clustering idea is adopted in the majority of data stream clustering algorithms. Aggarwal et al. extended their work in HPStream [9], which introduces the projected clustering to data streams. In projected clustering high dimensional stream data is partitioned based on the preferred dimensions instead of full the dimensional space. Cao et al. [10] use the density-based clustering without projected dimensions in DenStream algorithm. For streaming data, although a considerable research has tackled the full-space clustering, relatively limited work deals with the subspace clustering. These few researches include [9] HPStream, [11] HDDStream, and [12] SubCMM. A more comprehensive review and classifications are given in survey [13]. In

[11], authors propose a density-based projected clustering scheme for high dimensional data streams called HDDStream. HDDStream works in three phases; an initial phase, in which initial set of core micro-clusters is formed, then online core and outlier clusters' maintenance with projected clustering, and finally, an on-demand offline clustering phase. Compared with the HPStream which requires the fixed number of clusters, the number of clusters in HDDStream is variably adjusted over time, and the clusters can be of arbitrary shape. SubCMM suggests a different way for evaluating stream subspace clustering algorithms by making use of available offline subspace clustering algorithms with the streaming environment to handle the errors caused by emerging, moving, or splitting subspace clusters. A recent, similarity-based Data Stream Classifier (SimC)[14] introduces an insertion/removal policy that adapts evolving data tendency and maintains a representative, small set of clusters. It uses instance based learning techniques to form adaptive clustering algorithm. In [1] clustering method based on a multi-agent system that uses a decentralized bottom-up self-organizing strategy to group similar data points is presented. It uses bio-inspired flocking model to eliminate the need of offline clustering. In [15], authors present a clustering algorithm for stream data with uncertain attributes has been presented in . This scheme works only for low dimensional streaming data. Liu [16] develop HSWStream algorithm. It is a data stream clustering algorithm based on exponential histogram over sliding windows with projected dimensions. Another density-based algorithm D-Stream [17] maps each input data into a grid, computes the density of each grid, and forms the clusters using these grids. In [18], authors propose a scalable algorithm to trace clusters in a high-dimensional data stream. The proposed scheme transforms the problem of multi-dimensional clustering into that of one-dimensional clustering along with a frequent item-set mining technique. This scheme achieves the scalability on the number of dimensions while sacrificing the accuracy of identified clusters. Bellas et al. [19] present an online variant of mixture of probabilistic principal component analyzers (MPPCA) to model and cluster the high dimensional high speed data. But to do so, it is necessary to add a classification step at the end of the online MPPCA algorithm to provide the expected clustering. MuDi-Stream [20] is a hybrid grid-based multi-density clustering algorithm with online-offline phases. In the online phase, it keeps summary information of evolving multi-density data stream in the form of core micro-clusters. The offline phase generates the final clusters using an adapted density-based clustering algorithm. The grid-based method is used as an outlier buffer to handle both noises and multi-density data in order to reduce the merging time of clustering. MuDi-Stream is not suitable for high-dimensional data since the number of empty grids increases which requires longer processing time. SE-Stream [21] is a standard-deviation based projected clustering method to support high dimensional data streams. It forms clusters within subgroups of dimensions and can detect change in the clustering structure during the progression of data streams. SED-Stream [22] is an extension of SE-Stream, in which some selected dimensions are used to represent the clusters to increase the quality of the output clustering. SED-Stream projects any cluster to its discriminative dimensions that are highly relevant to the cluster itself but distinguished from the other clusters. SED-Stream is better than its previous version, SE-Stream, in terms of purity and f-measure. Both SE-

Stream and SED-Stream use fading cluster structure (5-tuple) of the form similar to in section III definition 0 with two extra elements.

This paper presents High Speed and Dimensions data stream clustering scheme (HSDStream) which introduces a novel tuple structure to summarize the high speed high dimensional data stream. This structure not only speed up the process but also requires less memory. Our clustering technique also modifies weights in some definitions of HDDStream, namely, the micro-cluster variance, projected dimensionality, projected distance, and projected radius. In terms of experimental results, we compare our scheme with HDDStream for cluster purity, memory usage, and cluster's sensitivity.

III. PROBLEM FORMULATION

In general, data stream is modeled as an infinite series of points $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, \dots\}$ arriving at discrete time $\{t_1, t_2, \dots, t_i, \dots\}$. Each point \mathbf{p}_i is a vector of dimension d such that $\mathbf{p}_i = \{p_{i,1}, p_{i,2}, \dots, p_{i,d}\}$.

An important characteristic of data streams is that we cannot store all data points. A usual way to overcome this problem is to summarize the data through an appropriate summary structure, often called micro-cluster. A micro-cluster summarizes the time and dimensionality limited stream data in the form of a tuple. When aging is also under consideration, the temporal extension of micro clusters [9] is employed. Recent research works [9], [11] use the following definition of micro-cluster:

Definition 0. (Micro-cluster mc)

A micro-cluster at time t for a set of d -dimensional data points $\mathcal{C} = \{\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{N-1}\}$ arriving at discrete time t_0, t_1, \dots, t_{N-1} , is summarized as $(2d + 1)$ size tuple $mc(t) = \{\mathbf{CF1}(t), \mathbf{CF2}(t), W(t)\}$, where $\mathbf{CF1}(t)$ and $\mathbf{CF2}(t)$ are d dimensional vectors, defined as:

- $\mathbf{CF1}(t)$ is the d -dimensional vector of weighted sum of points $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, \dots\}$ along each dimension, such that for dimension j we have $CF1_j = \sum_{i=0}^{N-1} p_{i,j} f(t - t_i)$, where N is the size of time window, $p_{i,j}$ is the i^{th} point in time window and $f(t - t_i)$ is the weight of the i^{th} point.
- $\mathbf{CF2}(t)$ is the d -dimensional vector of weighted sum of the squares of the points $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, \dots\}$ along each dimension, such that for dimension j we have $CF2_j = \sum_{i=0}^{N-1} p_{i,j}^2 f(t - t_i)$, where N is the size of time window, $p_{i,j}$ is the i^{th} point in time window and $f(t - t_i)$ is the weight of the i^{th} point.
- $W(t)$ is the sum of the weights of data points, mathematically, $W(t) = \sum_{i=0}^{N-1} f(t - t_i)$.

In data streams, since we are more interested in the data within a certain recent time window instead of all historical data, an aging effect has been used for weighted function $W(t)$. Recent works [9], [11] have used conventional exponential fading function $f(t) = 2^{-\lambda t}$, where λ is the decay rate. By using fading function $f(t)$ we need to maintain a memory buffer of time window size for each cluster, because, whenever a new point arrives we need to shift the previous data in the buffer of fixed size. We want to highlight an important

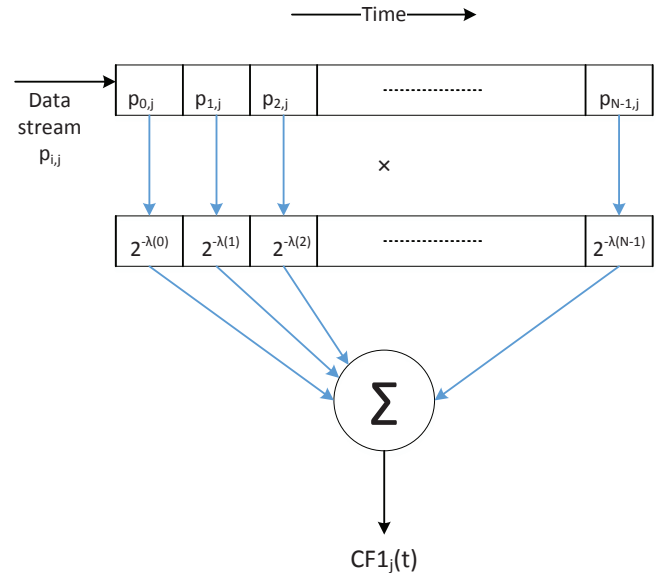


Fig. 1. Practical approach to update micro-cluster tuple

point here that the online update of the tuples [9], [11] of the form $mc(t) = \{\mathbf{CF1}(t) + p, \mathbf{CF2}(t) + p^2, W(t) + 1\}$ is not practically feasible because it leads to monotonically increasing weighted sum data. A practical approach for updating the tuple is shown in Fig. 1. It is obvious that for a fixed size memory shift register, when a new point arrives the old point is discarded. The correct mathematical expression for online update, then, becomes, $mc(t) = \{\mathbf{CF1}(t) - \mathbf{CF1}_{N-1} + p, \mathbf{CF2}(t) - \mathbf{CF2}_{N-1} + p^2, W(t) - f(t - t_{N-1}) + 1\}$, such that for dimension j we have $CF1_{N-1,j} = p_{N-1,j} f(t - t_{N-1})$ and $CF2_{N-1,j} = p_{N-1,j}^2 f(t - t_{N-1})$.

We define micro-cluster as follows:

Definition 1. (Micro-cluster mc) We redefine the micro-cluster as a set of points $\mathcal{C} = \{\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{N-1}\}$ arriving at discrete time points t_0, t_1, \dots, t_{N-1} . The mc is summarized as $(2d + 1)$ size tuple $mc(t) = \{\mathbf{EA1}(t), \mathbf{EA2}(t), W(t)\}$, where $\mathbf{EA1}(t)$ and $\mathbf{EA2}(t)$ are d dimensional vectors, defined as:

- $\mathbf{EA1}(t)$ is the d -dimensional vector of exponential weighted moving average of points $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, \dots\}$ along each dimension, such that for dimension j we have $EA1_j(t) = \alpha p_j(t) + (1 - \alpha)EA1_j(t - 1)$, where $\alpha = 2/(1 + N)$ is a smoothing factor controlled by the size of time window; and $p_j(t)$ is the latest point in time window.
- $\mathbf{EA2}(t)$ is the d -dimensional vector of exponential weighted average of points $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i, \dots\}$ along each dimension, such that for dimension j we have $EA2_j(t) = \alpha p_j^2(t) + (1 - \alpha)EA2_j(t - 1)$.
- $W(t)$ is the sum of the of data points at time t .

In order to formalize aging effect of data we introduce exponential moving average of data stream within a specified

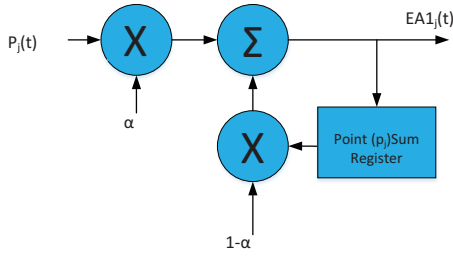


Fig. 2. Exponential moving average based update of micro-cluster tuple

time window. We use exponential weighted moving average in the tuple as decreasing exponential function. Note that, now the calculation of $EA1(t)$ or $EA2(t)$ does not require storage of past values, and only one addition and two multiplications with one memory register (of the size of dimension j) are required to update the tuple at any time instance. Design implementation of our micro-cluster update is shown in Fig. 2.

Data stream contains high dimensional data where each dimension has its own importance. In order to collate the similar points in data stream we use variance along each dimension. The lower the variance the higher the correlation among the points in particular dimension. We use variance as a metric to limit the number of dimensions to preferred dimensions only.

Definition 2. (Preferred Dimension) A dimension j is said to be a preferred dimension if $Var_j(mc) < \xi$, where ξ is the variance threshold and $Var_j(mc)$ is the variance of mc along dimension j , defined as:

$$Var_j(mc) = EA2_j(t) - (EA1_j(t))^2 \quad (1)$$

The preferred dimension helps gather the data points which have preferred dimensions less than a pre-defined threshold. Intuitively, it indicates the similarity across dimensions controlled by the variance threshold (ξ). In conjunction with preferred dimension, we define the preferred dimension vector.

Definition 3. (Preferred Dimension Vector) Every micro-cluster has a preferred dimension vector defined as:

$$\Psi(mc) = \{\psi_1, \psi_2, \dots, \psi_d\} \quad (2)$$

with

$$\psi_j = \begin{cases} \varrho, & Var_j(mc) < \xi; \\ 1, & \text{otherwise.} \end{cases} \quad (3)$$

where $\xi \in \mathbb{R}$, and $\varrho \in \mathbb{R}$ is a constant $\varrho \gg 1$. The number of elements in preferred dimension vector gives the projected dimensionality of the micro-cluster. The term 'projected' differentiates the micro-cluster defined over s projected subspace of the feature space instead of the whole feature space.

Definition 4. (Projected Dimensionality) Let $p \in \mathcal{C}$ and $\xi \in \mathbb{R}$. The number of dimensions j with $Var_j(mc) < \xi$ is called projected dimensionality of mc and denoted by $PDIM(mc)$.

Weighting the dimensions inversely proportional to their variance is not useful because we are only interested in distinguishing between dimensions with low variance and all

other dimensions. Therefore, we use only two-valued weight vector. It can be easily determined from the preferred dimension vector by counting the number of dimensions with the normalization factor ϱ . The intuition of calculating projected dimensionality is to find projected core micro-cluster, i.e., the clusters with some subspace of dimensions instead of all dimensions.

Definition 5. (Projected Radius) Let mc be a micro-cluster, $\xi \in \mathbb{R}$, and $\varrho \in \mathbb{R}$ is a constant $\varrho \gg 1$. The projected radius of mc is given by:

$$r_\Psi(mc) = \sqrt{\sum_{j=1}^d \frac{\psi_j}{\varrho} (EA2_j(t) - (EA1_j(t))^2)} \quad (4)$$

where ϱ normalizes the variance along each dimension. This is the projected radius that takes into account the preferred dimensions of the micro-cluster.

Definition 6. (Projected Distance) Let $p \in \mathcal{D}$ and mc be a projected micro-cluster with dimension preference vector $\Psi(mc)$. The projected distance between p and mc is given by:

$$dist^{proj}(p, mc) = \sqrt{\sum_{j=1}^d \frac{\psi_j}{\xi} (p_j - center_j^{mc})^2} \quad (5)$$

where $center^{mc}$ is the center of micro-cluster mc and is given by $center^{mc} = EA1(t)$.

Now we introduce the notion of core-projected mc which is an essential component of density based clustering. A core-projected mc is a mc that contains at least μ number of points within a projected radius of ϵ with projected dimensionality less than a threshold π .

Definition 7. (Core Projected Micro-cluster) Let $\epsilon, \xi \in \mathbb{R}$ and $\pi, \mu \in \mathbb{N}$. A micro-cluster mc is called a core projected mc if the preference dimensionality of mc is at most π and it contains at least μ points within its projected radius ϵ , formally:

$$\text{CORE}^{proj}(mc) \iff (r_\Psi(mc) < \epsilon) \wedge (W(t) > \mu) \wedge (PDIM < \pi). \quad (6)$$

In other words, a micro-cluster mc is a core projected mc iff:

- (1) $r_\Psi(mc) < \epsilon$
- (2) $W(t) > \mu$
- (3) $PDIM < \pi$

There might be micro-clusters that do not fulfill the above constraints either because their associated number of points is smaller than μ or because their projected dimensionality exceeds π . These micro-clusters are treated as outliers.

Definition 8. (Outlier Micro-cluster) Let $\epsilon, \xi \in \mathbb{R}$ and $\pi, \mu \in \mathbb{N}$. A micro-cluster mc is called a outlier mc , if its projected dimensionality is at least π and its projected radius and ϵ -Neighbors are at most ϵ and μ , respectively, formally:

$$\text{outlier}(mc) \iff (PDIM > \pi) \wedge (r_\Psi(mc) < \epsilon) \wedge (W(t) < \mu). \quad (7)$$

In order to keep update the micro-clusters, i.e., to check for possible conversion of core micro-cluster to outlier micro-cluster and vice versa, we introduce an outlier threshold ($0 < \beta < 1$) such that an outlier micro-cluster becomes a potential core micro-cluster if $W > \beta\mu$ in addition to the conditions in (6). Similarly, a core micro-cluster becomes a potential outlier micro-cluster if $W < \beta\mu$ in addition to the conditions in (7). The micro-cluster can be easily maintained online when a new point arrives in a cluster and other mc need time degradation.

Remark. (Online maintenance) The micro-cluster mc defined in definition 1 holds simple additive property that facilitates the online maintenance.

- If a point p arrives at time t , then the updated tuple is given by $mc(t) = \{\alpha p + (1 - \alpha)EA1(t - 1), \alpha p^2 + (1 - \alpha)EA2(t - 1), W(t - 1) + 1\}$.
- If no point adds in a micro-cluster at time t , then the updated tuple is given by $mc(t) = \{(1 - \alpha)EA1(t - 1), (1 - \alpha)EA2(t - 1), W(t - 1)\}$.

IV. THE HSDSTREAM ALGORITHM

HSDStream algorithm can be divided into three parts: 1) initialization to produce a set of representative core micro-cluster (core-mc) from an initial chunk of data points, 2) online maintenance of core-mc and outlier micro-cluster (outlier-mc), and, 3) the on-demand offline generation of the final clusters.

A. Initialization

In order to get initial set of micro-clusters from a fixed size of data points, we apply density-based projected clustering algorithm, a variant of PreDeCon algorithm [23], which is designed to work for fixed size of high dimensional data. Let \mathcal{D} be a set of initial chunk of d -dimensional data points ($\mathcal{D} \subseteq \mathbb{R}^d$). For each point $p \in \mathcal{D}$, we find a set of ϵ -neighbors $\mathcal{N}_\epsilon(p)$, where ϵ is the radius threshold. In addition to this, we find the neighbors of p with projected distance equal to or less than the ϵ , namely, $\mathcal{N}_\epsilon^{\Psi(p)}(p)$.

Definition 9. (Projected Distance of a Point) Let $p, q \in \mathcal{D}$. The projected distance of a point p with any point q is given by:

$$dist_p(p, q) = \sqrt{\sum_{i=1}^d \frac{\psi_i(p)}{\varrho} (d_i(p) - d_i(q))^2} \quad (8)$$

where $d_i(p)$ is the i^{th} dimension of point p . Note that, in general $dist_p(p, q) \neq dist_p(q, p)$ because of the projected dimension vectors of point p and q . In order to get symmetrical distance between p and q we use maximum of $dist_p(p, q)$ and $dist_p(q, p)$.

A projected core point $o \in \mathcal{D}$ can be defined with the same intuition of projected micro-cluster in definition 7.

$$CORE^{proj}(o) \iff PDIM(\mathcal{N}_\epsilon(o)) \leq \pi \wedge |\mathcal{N}_\epsilon^{\Psi(p)}(o)| \geq \mu \quad (9)$$

The initialization function in algorithm 1 line 5 runs the algorithm for the creation of initial set of mc . It starts by inserting all points in the set $\mathcal{N}_\epsilon(o)$ into a queue. For each point in the queue, it computes all directly projected weighted reachable points and inserts those points into the queue which

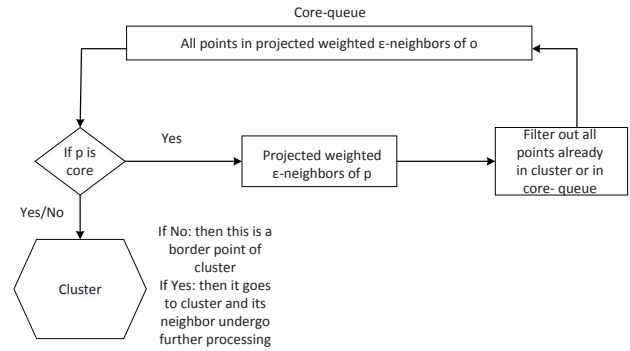


Fig. 3. Generation of initial set of micro-clusters

are still unclassified. This process repeats until the queue is empty and the cluster is computed. The flow chart of algorithm is shown in Fig. 3. Remove all those points belong to calculated cluster from dataset \mathcal{D} and repeat the process for another core point. This process remains continue till all the core points are exhausted.

B. Real-time Maintenance of Micro-clusters

In order to find out the clusters in an evolving real-time data stream, we maintain two groups of micro-clusters, namely, core-mc and outlier-mc in real-time. All the micro-clusters are maintained in a separate memory space. A new point might be assigned to core-mc, outlier-mc, or it may start new outlier-mc depends upon various factor. Sequential process of merging a new point p is described below:

- 1) When a new point arrives, it first becomes the candidate of core-mc (algorithm 1, line 13). The projected dimensionality of each core-mc has been evaluated before and after adding this point p (algorithm 2, line 4). After that, projected distance of p is calculated with those core-mc which still satisfy the projected dimensionality constraint, i.e., after the addition of point p (algorithm 2, line 6). Then, we choose one core-mc which has smallest projected distance from p (algorithm 2, line 9). Finally, the projected radius of chosen core-mc (p included) has been evaluated and checked for upper bound (ϵ) (algorithm 2, line 11). If it satisfies, then point p is assigned to that core-mc (algorithm 2, line 12 using update tuple function in algorithm 4), else it becomes candidate of outlier-mc list.
- 2) When a new point becomes a candidate for an outlier-mc, the projected distance of p with each outlier-mc is evaluated (algorithm 3, line 4). The closest distant outlier-mc is chosen in line 6. The point p becomes the member of that outlier-mc if the projected radius is less than or equal to the radius threshold (ϵ) (algorithm 3, line 9). In order to get long term effect we check the possibility of outlier-mc to core-mc conversion after certain number of points (window size N).
- 3) If point p cannot be added in core-mc or outlier-mc (algorithm 3, line 14) then a new outlier-mc is created with this point being the first element. It may become the seed of future core-mc.

Algorithm 1 HSDStream main

```
1: Initialization
2: initial parameters  $\pi, \xi, \epsilon, N$ 
3:  $datastream = \{p_1, p_2, \dots, p_i, \dots\}$ 
4:  $initialBuffer = readData(numOfInitialPoints)$ 
5:  $core\_mc = initialization\_fn(initialBuffer)$ 
6: for  $i = 1$  to  $numOfMc$  do
7:    $mcTuple = createMcTuples(core\_mc)$ 
   {It creates  $mcTuple = \{EA1(t), EA2(t), W(t)\}$ , an
    $numOfMc \times (2d + 1)$  matrix}
8: end for
9: while Stream has data points do
10:  $windowBuffer = readData(N)$ 
11:   for  $i = 1$  to  $N$  do
12:      $p_i = windowBuffer(i)$  // i-th point from windowBuffer
13:      $[trial\_core, mcTuple] = addpToCoreMc(p_i, mcTuple)$ 
14:     if  $trial\_core == 1$  then
15:       Degrade all outlierTuples
16:     else
17:        $[trial\_outlier, outlierTuple] =$ 
          $addpToOutlierMc(p_i, outlierTuple)$ 
18:     end if
19:     if  $trial\_core == 0$  &&  $trial\_outlier == 0$  then
20:        $newOutlierMc = createOutlierMc(p_i)$ 
21:       update outlierTuple list
22:     end if
23:   end for
   {core-mc to outlier-mc conversion}
24:    $[movedMcTuples, remainingMcTuples] =$ 
      $moveMcTuples(mcTuples)$ 
   {outlier-mc to core-mc conversion}
25:    $[movedOutlierTuples, remainingOutlierTuples] =$ 
      $moveOutlierTuples(outlierTuples)$ 
26:    $updatedMcTuples = remainingMcTuples +$ 
      $movedOutlierTuples$ 
27:    $updatedOutlierTuples = remainingOutlierTuples +$ 
      $movedMcTuples$ 
28: end while
```

Algorithm 2 Add data point to core-mc

```
1:  $addpToCoreMc(p, mcTuples)$ 
2: for  $i = 1$  to  $numOfTuples$  do
3:    $updatedTuples = updateTuple\_fn(p, mcTuple(i))$ 
4:   Calculate updated PDIM // using definition 4
5:   if  $PDIM \leq \pi$  then
6:     Calculate projected distance // using definition 6
7:   end if
8: end for
9:  $core\_mc\_closest = \min(projectedDistances)$ 
10: Calculate projected radius  $r_\Psi(core\_mc\_closest)$  // using definition 5
11: if  $r_\Psi(core\_mc\_closest) < \epsilon$  then
12:    $mcTuple = updateTuple\_fn(p, mcTuple)$ 
13:   Update all other mcTuples with one degradation
14:   return  $trial\_core = 1$ 
15: else
16:   Degrade all mcTuples
17:   return  $trial\_core = 0$ 
18: end if
```

C. Clusters Generation: Offline

The real-time maintained micro-clusters capture the density area and the projected dimensionality of data streams. However, in order to get meaningful clusters, we need to apply some clustering algorithm to get the final result. When a clustering request arrives, a variant of PreDeCon algorithm [23] is applied on the set of real-time maintained core-mc(s) to get the final result of clustering. In density-based PreDeCon, a core point starts a micro-cluster, all the directly connected points and the chain of core points which satisfy ϵ -neighborhood

Algorithm 3 Add data point to outlier-mc

```
1:  $addpToCoreMc(p, outlierTuples)$ 
2: for  $i = 1$  to  $numOfOutlierTuples$  do
3:    $updatedTuples = updateTuple\_fn(p, mcTuple(i))$ 
4:   Calculate projected distance // using definition 6
5: end for
6:  $core\_mc\_closest = \min(projectedDistances)$ 
7: Calculate projected radius  $r_\Psi(outlier\_mc\_closest)$  // using definition 5
8: if  $r_\Psi(outlier\_mc\_closest) < \epsilon$  then
9:    $outlierTuple = updateTuple\_fn(p, outlierTuple)$ 
10:   Update all other outlierTuples with one degradation
11:   return  $trial\_outlier = 1$ 
12: else
13:   Degrade all outlierTuples
14:   return  $trial\_outlier = 0$ 
15: end if
```

Algorithm 4 Update Tuple function

```
1:  $updateTuple\_fn(p, Tuple)$ 
2:  $EA1(t - 1) = Tuple(1 : d)$ 
3:  $EA2(t - 1) = Tuple(d + 1 : 2d)$ 
4:  $W(t - 1) = Tuple(end)$ 
5:  $EA1(t) = \alpha p + (1 - \alpha)EA1(t - 1)$ 
6:  $EA2(t) = \alpha p^2 + (1 - \alpha)EA2(t - 1)$ 
7:  $W(t) = W(t - 1) + 1$ 
8:  $newTuple = \{EA1(t), EA2(t), W(t)\}$ 
```

criteria and maximum dimensionality π become the member of that cluster. During offline on-demand clustering phase, each core-mc acts as core point. Each core-mc is regarded as a virtual point located at the center of core-mc. We use the concept of density connectivity to determine the final clusters, i.e., all the density-connected core-mc(s) form a cluster.

V. DISCUSSION

In this section we highlight issues and challenges in the development of high dimensional data stream clustering in Internet traffic monitoring. We maintain the density with ϵ -neighborhood and minimum number of points μ in a core-mc. When an identical burst of data (in case of attack on network) arrives, outlier-mc(s) are diminished and only one core-mc remains there. In this case, an important entity of core-mc formation i.e., projected dimensionality cannot work because, now $PDIM = d$ and it no longer satisfies the condition $PDIM \leq \pi$. In order to overcome this problem we introduce another condition ORed with the condition $PDIM \leq \pi$ to maintain one core-mc containing exactly similar data. The new condition is $W(t)/N > 90\%$, i.e., if the data points window contains more than 90% points, then no need to check PDIM because the majority of identical data points indicates some abnormal activity on the network being monitored. During real-time maintenance, when a new point arrives and it becomes a part of only one micro-cluster, then, all the other micro-clusters undergo one time degradation. For each existing core-mc, if no new point is merged into it, then the weight of core-mc will decay gradually. If the weight is below $\beta\mu$, then it means that core-mc has become an outlier-mc, it should be deleted and its memory space should be released for new core-mc. Similarly, if the weight is above $\beta\mu$ then it means that the outlier-mc has become a core-mc, it should be deleted and its memory space should be released. Therefore, we need to check the weight of each micro-cluster periodically. We use a fixed time period to perform this check at every time window

interval (N). In this way any outlier-mc automatically vanishes if no point merges in it during N time units.

VI. EXPERIMENTAL EVALUATION

We compare our proposed HSDStream algorithm with HDDStream [11] which is the recent projected clustering algorithm for high dimensional data streams. We use corrected KDD (Knowledge Discovery and Data mining) 1999 [24] Computer Network Intrusion detection dataset which is typically used for the evaluation of stream clustering algorithms. Both algorithms are implemented in MATLAB and run on Intel i5 Dual Core 2.0GHz with 2 GB RAM.

A. Dataset

To evaluate the performance of clustering algorithm we use KDD 1999 Network Intrusion detection dataset. This is the dataset used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. It has been reported that original dataset contains bugs, therefore, we use the corrected dataset available online at <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>. KDD-CUP'99 Network Intrusion Detection stream dataset which has been used earlier [8], [9], [10], [11] to evaluate CluSTREAM, HPStream, DenStream, HDDStream, respectively. This dataset corresponds to the important problem of automatic and real-time detection of network attacks and consists of a series of transmission control protocol (TCP) connection records from two weeks of local area network (LAN) traffic managed by MIT Lincoln Labs. Each record can either corresponds to a normal connection, or an intrusion. Most of the connections in this data set are normal, but occasionally there could be a burst of attacks at certain times. In this dataset, attacks fall into four main categories:

- DOS: denial-of-service e.g., syn flood
- R2L: unauthorized access from a remote machine, e.g., guessing password
- U2R: unauthorized access to local superuser (root) privileges, e.g., various "buffer overflow" attacks
- Probing: surveillance and other probing, e.g., port scanning

The attack-types are further classified into one of 24 types, such as back, buffer_overflow, ftp_write, guess_passwd, imap, ipsweep, spy, and so on. It is obvious that each specific attack type can be treated as a sub-cluster. Also, this data set contains totally 494020 connection records, and each connection record has 42 attributes or dimensions that belongs to one of the continuous (35) or symbolic type (7). In the performance analysis of proposed algorithm we use all 35 continuous attributes.

B. Cluster Quality Evaluation

Traditional full dimensional clustering algorithms, for example, [8] used the sum of square distances (SSQ) to evaluate the clustering quality. However, SSQ is not a good measure in evaluating projected clustering [9] because it is a full

TABLE I. PARAMETER VALUES

Parameter	Value
N	200
π	30
μ	10
β	0.2
ξ	0.002
<i>initialPoints</i>	1000
ϵ	10
H	1

dimensional measure, and full dimensional measures are not very useful for measuring the quality of a projected clustering algorithm. So, as in [9] and [11], we evaluate the clustering quality by the average purity of clusters, which examines the purity of the clusters with respect to the true cluster (class) labels. The purity is defined as the average percentage of the dominant class label in each cluster [10]. Let there are K number of cluster in a cluster set \mathcal{K} at query time such that $k \in \mathcal{K} = \{1, 2, \dots, K\}$.

$$purity(\mathcal{K}) = \frac{\sum_{k=1}^K \frac{|P_k^d|}{|P_k|}}{K} \quad (10)$$

where $|P_k^d|$ is the number of points with dominant class label in cluster k and $|P_k|$ is the number of points in cluster k . Intuition behind the cluster purity is to measure the actual capture of distinct groups of data points which are known to the given dataset. The time span in which we measure the purity is called *Horizon* window H . It is measured in the number of time windows N . In the performance analysis $H = 1$ otherwise stated.

Fig. 4-8 show the cluster purity of HDDStream and HSDStream. In network streaming data, normal traffic packets (or points) are random in nature at any particular time interval, however, a network attack is characterized by bursts of correlated data packets. Therefore, we cannot fit normal traffic packets in a single cluster. We can fine tune the design parameters (α, β, ξ, N) to capture the known types of attacks or even the unknown abnormal traffic patterns. We can see that cluster purity can take values from 0 to 1. Cluster purity for normal network traffic usually varies from 0.5 to 1. It can go below 0.5 if we have more than 50% data points with more than 20% dimensions outside the standard deviation of cluster in a certain time window. Intuitively, cluster purity is low if the cluster contains uncorrelated data or in other words, the normal data traffic. High purity (or purity 1) corresponds to highly correlated data as a result of some network attack. In Fig. 4, *smurf* attack can be seen between 34 – 57 time units (for $N = 200$) which corresponds to data points 7795 to 11489 in the KDD network intrusion database. The network is again under *smurf* attack from 211 to 249 time units. During the time interval from 250 to 365 we encounter with several attacks (*back, ipsweep, nmap, and neptune*) along with correlated normal data so that we can see cluster purity is equal to 1 for this time interval. *Satan* attacks the network from 453 to 455 time units, followed by *smurf* attack which continues till the end of simulations at 495 time units. It can be observed that HDDStream has the same purity graph pattern as HSDStream but with considerably low magnitude. This is due to the large number of core-mc(s) in HDDStream and the fact that percentage purity is inversely proportional to the number of clusters (10). The average cluster purity for HSDStream

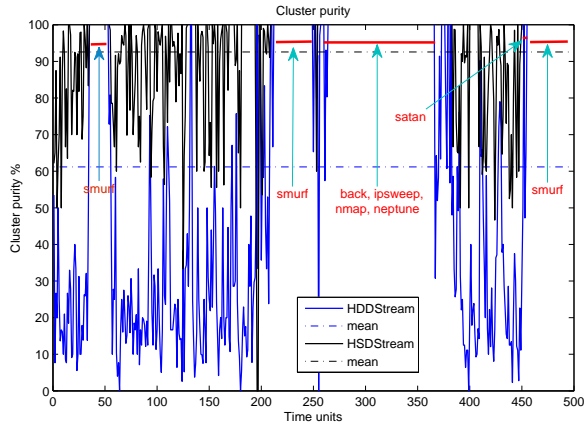


Fig. 4. Cluster purity with default values of parameters

is 92.57% as compared to the 61.18% of HDDStream. Next we illustrate: Why HSDStream has fewer number of clusters compared to HDDStream. Since the velocity of points is same for both schemes, it implies that HSDStream has more points per cluster than the HDDStream. For HSDStream, mean value of points in a window N is given by

$$\begin{aligned} \mathbf{EA1}(n) = & \alpha(1 - \alpha)\mathbf{p}_n^{n-n} + \alpha(1 - \alpha)^{n-n-1}\mathbf{p}_{n-1} \\ & + \alpha(1 - \alpha)^{n-n-2}\mathbf{p}_{n-2} + \dots + \alpha(1 - \alpha)^n\mathbf{p}_0 \end{aligned} \quad (11)$$

where $\alpha = 2/(1 + N)$. Let $N = 200$ and $n = \{0, 1, 2, \dots, 199\}$ with 0 being the first point and 199 is the latest point in a buffer window. Similarly, the mean value of points in window N is given by

$$\begin{aligned} \frac{\mathbf{CF1}(n)}{W} = & \frac{2^{-\lambda(\frac{n-n}{N})}}{W}\mathbf{p}_n + \frac{2^{-\lambda(\frac{n-n-1}{N})}}{W}\mathbf{p}_{n-1} \\ & + \frac{2^{-\lambda(\frac{n-n-2}{N})}}{W}\mathbf{p}_{n-2} + \dots + \frac{2^{-\lambda(\frac{n}{N})}}{W}\mathbf{p}_0 \end{aligned}$$

Substituting the values of parameters, we get $\mathbf{EA1}(199) = 0.01p_{199} + 0.009p_{198} + 0.0098p_{197} + \dots + 0.0014p_0$ and $\mathbf{CF1}/W = 0.0054p_{199} + 0.0054p_{198} + 0.0054p_{197} + \dots + 0.0046p_0$. Thus, for the same point HSDStream gives larger mean value than HDDStream. From equation (5), it is obvious that higher values of mean (center) result in smaller projected distance, hence larger number of points per cluster and fewer number of clusters. ■

Fig. 5 depicts the cluster purity for default values of parameters in bar graph. It can be noticed that HSDStream and HDDStream are equally good in detecting the attacked points but the cluster purity for normal traffic is low in HDDStream because of large number of clusters (low density clusters). Fig. 6 shows the cluster purity with $N = 100$. By decreasing the window size we actually increase the granularity and can capture smaller attacks. The price for this granularity is the more processing for the same amount of data. Again the average value of cluster purity for HSDStream is significantly larger than the HDDStream: 95.23 versus 67.31. Fig. 7 and Fig. 8 show the cluster purities for $N = 300$ and $N = 400$, respectively. We notice that the changing window size has minimal effect on the average cluster purity.

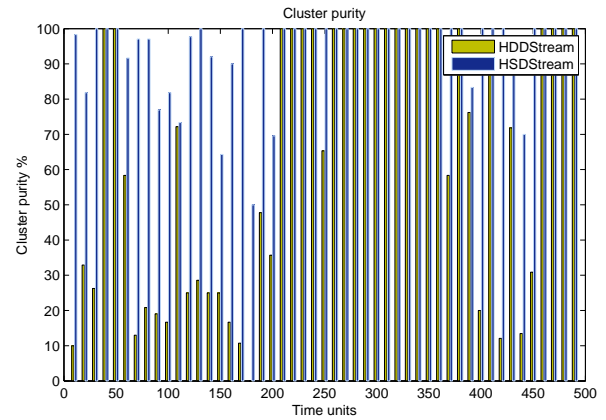


Fig. 5. Cluster purity with default values of parameters

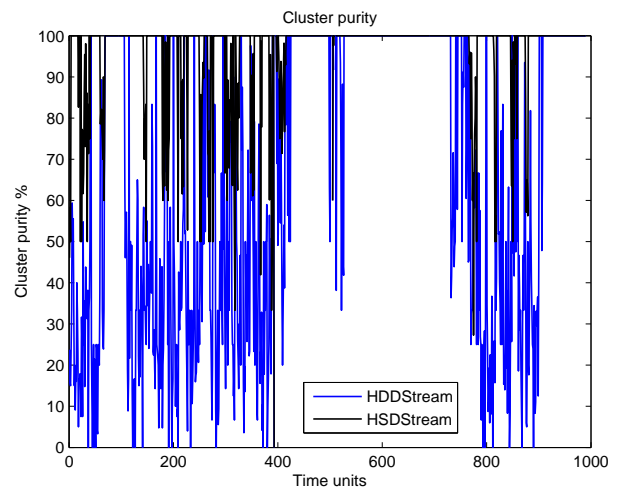


Fig. 6. Cluster purity with $N = 100$

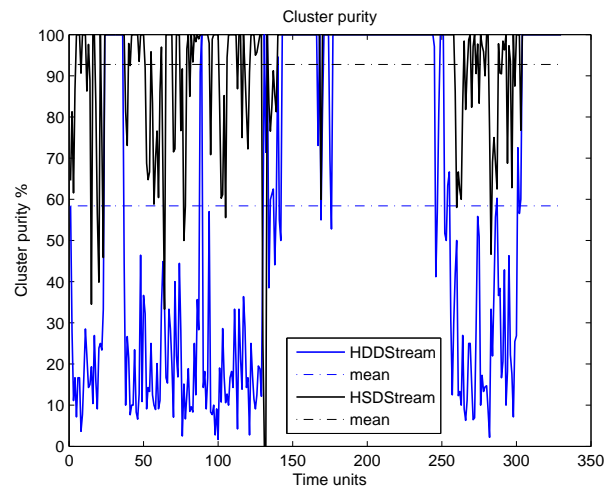


Fig. 7. Cluster purity with $N = 300$

C. Memory Usage

We measure the memory usage as a number of micro-clusters in HDDStream and HSDStream. During the period

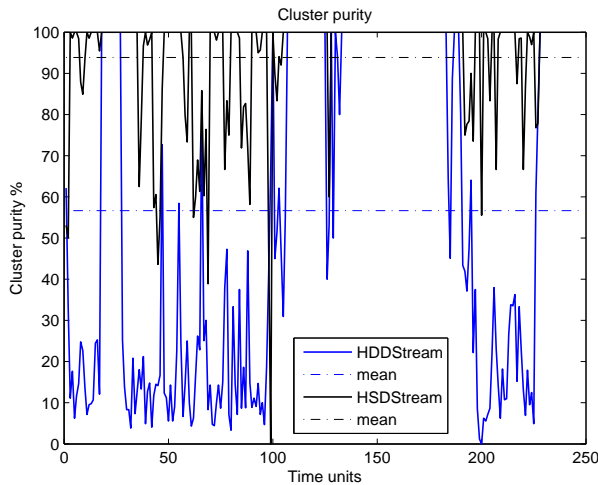


Fig. 8. Cluster purity with $N = 400$

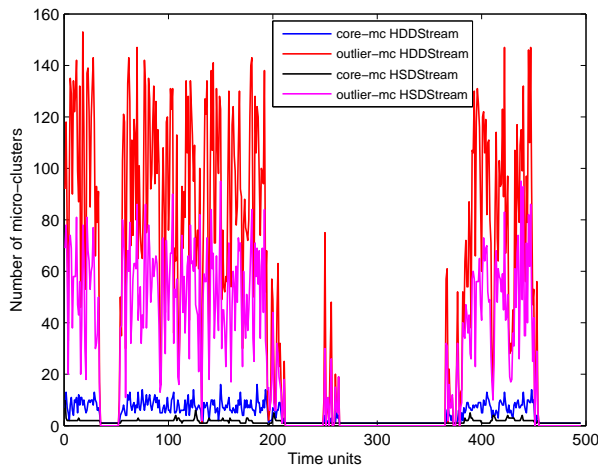


Fig. 9. Number of clusters with default values of parameters

of highly correlated normal data or the network attack, there is only one core-mc containing all the correlated points and no outlier cluster exists. It can be seen from the Figs. 10, 11, 12, and 13 that the total number of clusters is reduced to one during network attacks. When we compare these figures with different window sizes, we can see that there is a gradual increase of number of clusters with increasing number of window size. HSDStream outperforms the HDDStream in terms of memory usage for all window sizes, which is due to our reduced memory sized tuple and high density micro-clusters. Theoretically, the online update of $CF1_j$ requires N number of memory registers (one for each point's j^{th} dimension), whereas, EAI_j needs only one memory register, as shown in Fig. 1 and Fig. 2, respectively.

D. Sensitivity and Delay Analysis

In sensitivity analysis, we show how sensitive the clustering quality is in relevance to the outlier threshold β , and the processing time with different window sizes. In Fig. 15 we see that cluster purity improves with increasing values of

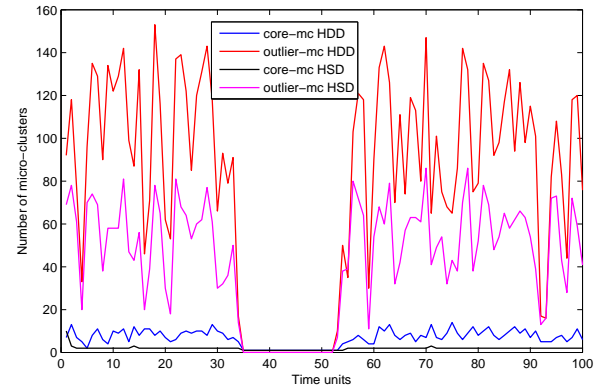


Fig. 10. Number of clusters with default values of parameters with zoom in

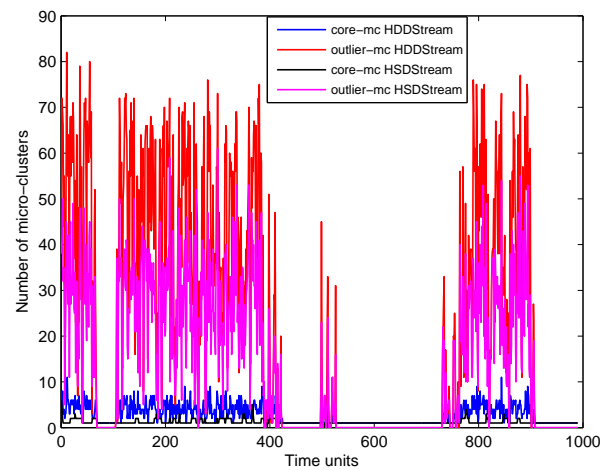


Fig. 11. Number of clusters with $N = 100$

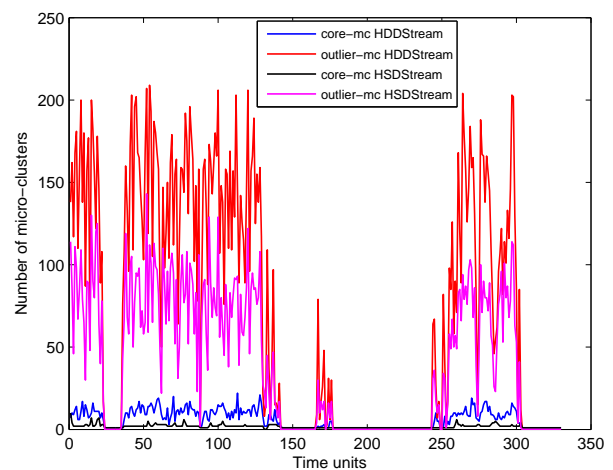


Fig. 12. Number of clusters with $N = 300$

outlier threshold. Outlier threshold controls the limit of the number of points that make it eligible to become core-mc or outlier-mc. After the end of each window size, all micro-

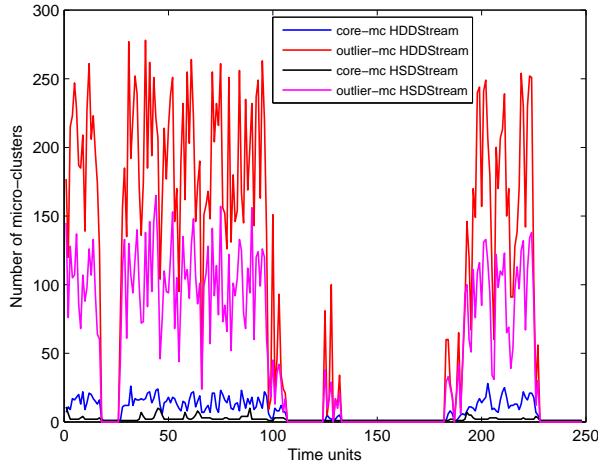


Fig. 13. Number of clusters with $N = 400$

clusters are examined for their eligibility as core or outlier. For small values of β , a cluster remains its current state for the larger time duration making cluster pollute for larger duration. Whereas with high values of β the cluster changes its state more quickly (as soon as it violate the condition $NumOfPoints > or < \beta\mu$) leaving the cluster more pure. Fig. 16 shows an important result that we can decrease the memory usage by increasing the outlier threshold. Higher values of β help remove the outlier points thus reducing the unnecessary core-mc(s). Since the core-mc(s) are small proportion of total number of clusters as shown in Fig. 10, therefore, the total number of clusters do not exhibit significant improvement in Fig. 17. However, the memory usage argument remains still valid because core-mc(s) are highly dense and utilize large proportion of memory.

Finally, we examine the processing time of HDDStream and HSDStream for different window sizes in Fig. 18. This processing time includes the time for the initialization phase and the data collection for the plotting purpose. It can be seen that HSDStream outperforms the HDDStream for all window sizes. This verifies the efficiency of our micro-cluster design in definition 1 where we need only two multipliers and one adder as compared to the conventional micro-cluster defined in definition 0 which requires N number of multipliers and $N - 1$ number of adders with $\lceil \log_2(N) \rceil$ stages delay. For example if $N = 6$, then in order to add 6 numbers, we need 5 adders which incur 3 stages delay as shown in Fig. 14.

VII. CONCLUSION

This paper presents a clustering algorithm for high-dimensional high-density streaming data. We propose a new structure of micro-cluster's tuples. This structure uses exponential weighted averages to reduce the memory usage and decrease the computational complexity. We have compared our scheme with HDDStream with KDD network intrusion detection dataset. The results show that HSDStream give significant improvement over HDDStream in terms of cluster purity, memory usage, and the processing time.

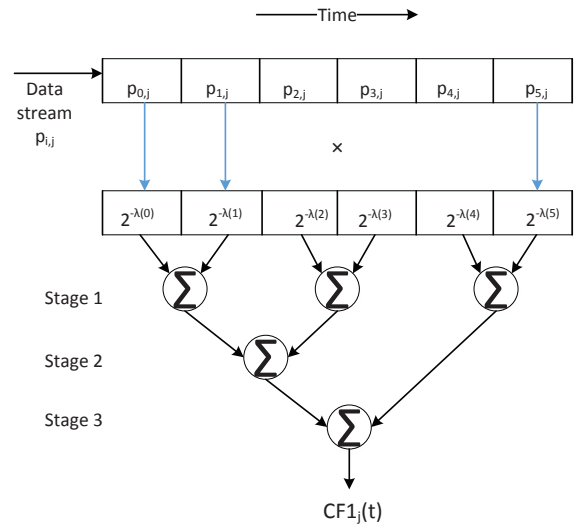


Fig. 14. Processing time delay in conventional micro-cluster update with $N = 6$

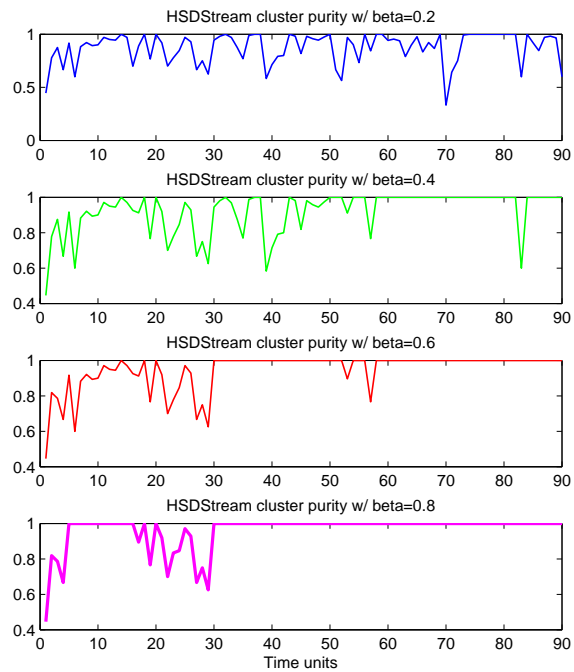


Fig. 15. Cluster purity for different values of β

REFERENCES

- [1] A. Forestiero, C. Pizzuti, and G. Spezzano, "A single pass algorithm for clustering evolving data streams based on swarm intelligence," *Data Mining and Knowledge Discovery*, vol. 26, no. 1, pp. 1–26, Jan. 2013.
- [2] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong, "A survey on enhanced subspace clustering," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 332–397, Mar. 2013.
- [3] C. C. Aggarwal, "A segment-based framework for modeling and mining data streams," *Knowledge and Information Systems*, vol. 30, no. 1, pp.

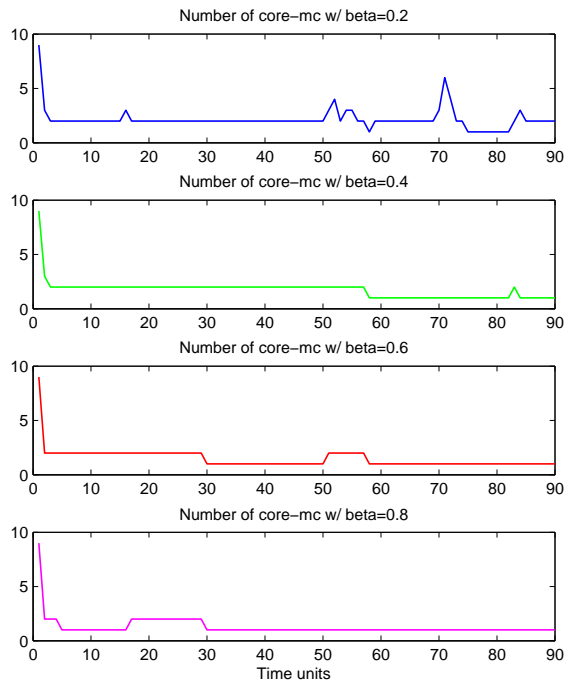


Fig. 16. Number of core-mc in HSDStream versus β

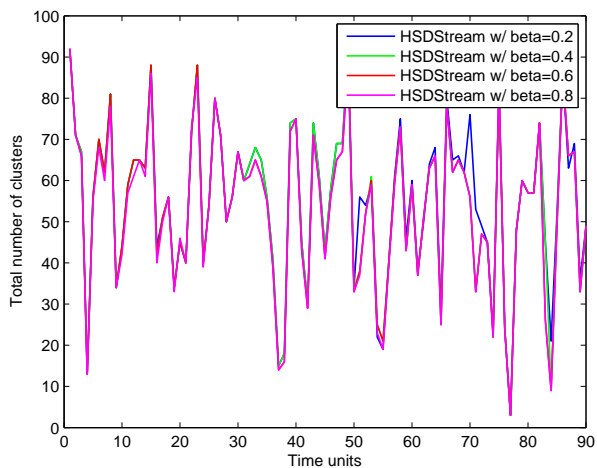


Fig. 17. Total number of clusters in HSDStream versus β

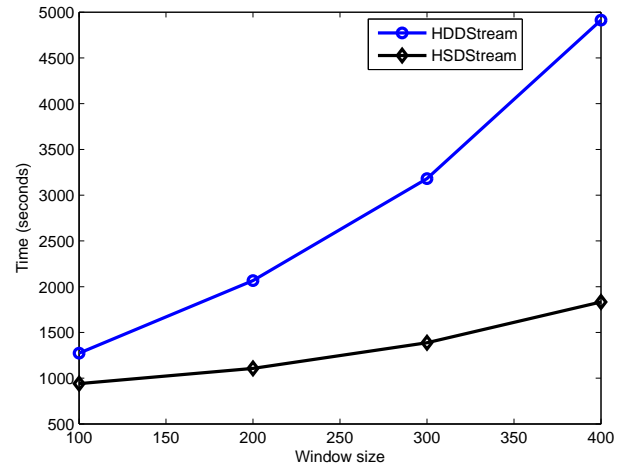


Fig. 18. Processing time for different window sizes

1–29, Jan. 2012.

[4] A. Amini, T. Y. Wah, and H. Saboohi, "On density-based data streams clustering algorithms: A survey," *Journal of Computer Science and Technology*, vol. 29, no. 1, pp. 116–141, 2014.

[5] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[6] H.-P. Kriegel, P. Krger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 1, pp. 1:1–1:58, Mar. 2009.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations." The Regents of the University of California, 1967.

[8] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proceedings of the 29th international conference on Very large data bases-Volume 29*. VLDB Endowment, 2003, pp. 81–92.

[9] —, "A framework for projected clustering of high dimensional data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 852–863.

[10] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *SDM*, vol. 6. SIAM, 2006, pp. 326–337.

[11] I. Ntoutsis, A. Zimek, T. Palpanas, P. Krger, and H.-P. Kriegel, "Density-based projected clustering over high dimensional data streams," in *SDM*. SIAM, 2012, pp. 987–998.

[12] M. Hassani, Y. Kim, S. Choi, and T. Seidl, "Subspace clustering of data streams: new algorithms and effective evaluation measures," *Journal of Intelligent Information Systems*, Jun. 2014.

[13] H.-L. Nguyen, Y.-K. Woon, and W.-K. Ng, "A survey on data stream clustering and classification," *Knowledge and Information Systems*, Dec. 2014.

[14] D. Mena-Torres and J. S. Aguilar-Ruiz, "A similarity-based approach for data stream classification," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4224–4234, Jul. 2014.

[15] C. Jin, J. X. Yu, A. Zhou, and F. Cao, "Efficient clustering of uncertain data streams," *Knowledge and Information Systems*, vol. 40, no. 3, pp. 509–539, Sep. 2014.

[16] W. Liu and J. OuYang, "Clustering algorithm for high dimensional data stream over sliding windows." IEEE, Nov. 2011, pp. 1537–1542.

[17] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Density-based clustering of data streams at multiple resolutions," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 3, pp. 14:1–14:28, Jul. 2009.

[18] J. W. Lee, N. H. Park, and W. S. Lee, "Efficiently tracing clusters over high-dimensional on-line data streams," *Data & Knowledge Engineering*, vol. 68, no. 3, pp. 362–379, Mar. 2009.

[19] A. Bellas, C. Bouveyron, M. Cottrell, and J. Lacaille, "Model-based clustering of high-dimensional data streams with online mixture of probabilistic PCA," *Advances in Data Analysis and Classification*, vol. 7, no. 3, pp. 281–300, May 2013.

[20] A. Amini, H. Saboohi, T. Herawan, and T. Y. Wah, "MuDi-stream: A multi density clustering algorithm for evolving data stream," *Journal of Network and Computer Applications*, Dec. 2014.

[21] R. Chairukwattana, T. Kangkachit, T. Rakthanmanon, and K. Waiyamai, "SE-stream: Dimension projection for evolution-based clustering of high dimensional data streams," in *Knowledge and Systems Engineering*, V. N. Huynh, T. Denooux, D. H. Tran, A. C. Le, and S. B. Pham,

- Eds. Cham: Springer International Publishing, 2014, vol. 245, pp. 365–376.
- [22] K. Waiyamai, T. Kangkachit, T. Rakthanmanon, and R. Chairukwattana, “SED-stream: Discriminative dimension selection for evolution-based clustering of high dimensional data streams,” *Int. J. Intell. Syst. Technol. Appl.*, vol. 13, no. 3, pp. 187–201, Oct. 2014.
- [23] C. Bohm, K. Railing, H.-P. Kriegel, and P. Kroger, “Density connected clustering with local subspace preferences,” in *Fourth IEEE International Conference on Data Mining, 2004. ICDM '04*, Nov. 2004, pp. 27–34.
- [24] “kdd-cup-1999-computer-network-intrusion-detection.” [Online]. Available: <http://www.sigkdd.org/kdd-cup-1999-computer-network-intrusion-detection>

Anti-noise Capability Improvement of Minimum Energy Combination Method for SSVEP Detection

Omar Trigui

Advanced Technologies for Medicine and Signals 'ATMS', King Abdulaziz University (KAU) Jeddah, Saudi Arabia
ENIS, Sfax University, Tunisia

Wassim Zouch

Mohamed Ben Messaoud

Advanced Technologies for Medicine and Signals 'ATMS',
ENIS, Sfax University, Tunisia

Abstract— Minimum energy combination (MEC) is a widely used method for frequency recognition in steady state visual evoked potential based BCI systems. Although it can reach acceptable performances, this method remains sensitive to noise. This paper introduces a new technique for the improvement of the MEC method allowing ameliorating its Anti-noise capability. The Empirical mode decomposition (EMD) and the moving average filter were used to separate noise from relevant signals. The results show that the proposed BCI system has a higher accuracy than systems based on Canonical Correlation Analysis (CCA) or Multivariate Synchronization Index (MSI). In fact, the system achieves an average accuracy of about 99% using real data measured from five subjects by means of the EPOC EMOTIVE headset with three visual stimuli. Also by using four commands, the system accuracy reaches 91.78% with an information-transfer rate of about 27.18 bits/min.

Keywords— Brain-Computer Interface; Steady State Visual Evoked Potential; Minimum Energy Combination; Empirical Mode Decomposition.

I. INTRODUCTION

Severely paralyzed people with neuromuscular disorders lose the majority or the totality of their movement and expression abilities. This is the case of people with locked-in syndrome, Amyotrophic lateral sclerosis and Spinal cord injury. A Brain-Computer Interface (BCI) is a tool for mobility, communication and control assistance which can provide them with the possibility to interact with their surroundings [1]. The principle of a BCI is to detect the brain activity from the scalp and convert it into commands to control devices such as prosthesis and computers. The control is done only by thought without any apparent movement. The electroencephalography (EEG) is usually used in BCI field for brain activity measurements. This is mainly due to its time resolution efficient for real-time applications, its low cost compared to other technics and the possibility to wear an EEG headset everywhere.

The P300 evoked potentials, the Event-Related Synchronization and Desynchronization (ERS/ERD) and the Steady State Visual Evoked Potential (SSVEP) are the most promising EEG brain activity patterns. SSVEPs are near-sinusoidal waveforms from the occipital area reflecting a visual stimulation [2]. SSVEP-based BCI systems offer many

advantages: the small number of required electrodes which makes the equipment cheaper, the no need of a tiring training, the suitability for almost any person and any environment and the better resistance face to noise and artifacts compared with other brain responses [3] [4]. Also, the performances reached by these systems are very encouraging. For example, F. Gembler et al. [5] have made an experiment where SSVEP was used to distinguish one among four possible commands. Ten subjects from different age ranges participated in the study and the data were acquired from 8 channels with a sampling rate of 128 Hz. The average accuracy was about 93% which makes the system functional and useful.

Several processing methods were presented to distinguish the target at which the subject gazes [6]. For instance, the Minimum Energy Combination (MEC) method proposed by O. Friman et al. [7] estimates the signal to noise ratio (SNR) corresponding to the stimuli frequencies then selects the frequency that maximizes this quantity. This method was exploited by N. Chumerin et al. [8] to create an SSVEP-based BCI game. The task was to navigate an avatar through a maze using four commands. The average accuracy of six subjects was about 82.4% which is considered acceptable. Moreover, Z. Lin et al. [8] use a frequency recognition method based on the Canonical Correlation Analysis (CCA). The essence of this method is to extract the correlation coefficients between the EEG signal and the reference signals then to select the frequency which maximizes this coefficient. Both MEC and CCA methods offer good performances thanks to the multi-channel and multi-harmonics properties. G. Hakvoort et al. [9] emphasize the usefulness of the multi-channel criterion by making a comparison between the CCA as a multi-channel based method and the power spectral density analyses (PSDA) as a mono-channel based one. The average accuracy of seven subjects was 47.81% using PSDA and 78.12 using CCA in the discrimination among seven different frequencies. This result clearly demonstrates the importance of multi-channel techniques. On the other hand, despite the solid mathematical foundations of the CCA and the MEC methods, a comparison between them shows that the CCA appears to have widely superior performances. For example, N. Mora et al. [10] make an experiment where five subjects are asked to gaze at one of four possible LEDs flickering with different stimulation frequencies. Data is measured from six different channels then processed using different discrimination methods. Results

show that CCA method ensures a higher accuracy rate by about 13% than MEC method. In another study results indicate that the high sensitivity to the noise level of MEC method leads to a lower accuracy rate [11].

Recently, Y. Zhang et al. [12] designed a new recognition method based on the multivariate synchronization index (MSI). The idea is to calculate the synchronization indexes reflecting the similarity between the EEG data and reference signals similar to those used in CCA method. Following the same idea, the frequency which maximizes the synchronization index is chosen.

The paper presents a novel amelioration of MEC method for achieving better resistance to noise. The Empirical mode decomposition and the moving average are used to reject irrelevant signals. The remaining signals located at the stimuli and harmonics frequency band are used to recognize the target.

The following section describes the origin of the noise in the SSVEP signal. Then the recognition methods are presented. The new amelioration of MEC method is presented in section four. To validate the improvement of the modified method, a comparison with the three methods is carried out. Finally, section 6 concludes the paper.

II. SSVEP AND NOISE SOURCES

The SSVEP occurs when a subject gazes at a light source target flickering with a fixed frequency. It can be detected as a signal power increase at the same frequency of the stimulus. In addition, a number of harmonic frequencies multiple of the principal frequency can be detected [13]. To exploit this phenomenon, several targets are presented in front of the subject. Each target flickers with a unique frequency. Besides one command is assigned to each target. To execute a command, the subject has to gaze directly at the appropriate target. Due to the effect of cortical magnification, the quality of SSVEP increases if the subject gazes directly at a stimulus object located in the center of his vision field [14] [15]. Likewise, the retinal cones distribution shows that the foveola located in the center of the visual field is more sensitive to the light. Therefore, as it was proven by A. González-Mendoza et al. [16], the amplitude of the SSVEP increases proportionally with the area size of the visual stimulus. Consequently, the noise level increases if the stimulus does not exist in the center of the vision field or if the light intensity is not sufficient. Also, the choice of a wrong stimulus color can weaken the power of the SSVEP response [17]. Before performing ameliorations on the MEC method, the stimulus parameters effect on the performances of the processing methods is studied. A comparison of MEC method versus other ones for different noise levels is investigated.

III. MATERIALS AND EXPERIMENTS

Five healthy male volunteers participated in the study. Their ages are 28, 30, 30, 34, and 58. None of the subjects had visual or neurological disorders or a previous experience with the BCI systems. Subjects were asked to sit in a comfortable chair in a room with low noise and luminance level and to avoid any movement. The Epoc Emotiv headset was used to acquire the EEG signal from the scalp. Data is sampled at 128 Hz within a bandwidth from 0.2 to 45 Hz with a digital notch

filter at 50 Hz and 60 Hz. The choice of this equipment is due to its short preparation time and low price which are important factors to bring BCI systems into daily life. Furthermore, Epoc Emotiv headset had shown good performances in SSVEP based BCI as in Y. Liu et al. [18] study where the accuracy rate reached $95.83 \pm 3.59\%$ with online application. To cover the maximum zone of the occipital area, data were obtained from electrodes T7, T8, P7, P8, O1, and O2 according to the 10/20 international system. In the present study, the chosen application is the wheelchair navigation command. In the first experiment, 3 commands are used: one to move forward, one to turn left, and one to turn right. Thus, the stimulation system is composed of three LEDs positioned on the left, top, and right sides of a computer screen. The LEDs are flickering with fixed frequencies: 8, 9, and 10 Hz respectively and the subject sits 0.6 m far from them. In the second experiment, in order to improve the information-transfer rate of the system, another command is added to the system allowing to move backward. Its appropriate target which flickers at 11 Hz is positioned at the bottom side of the screen. To ensure the control of the stimulation frequencies with precision, an electronically device based on the STM8 microcontroller was used. Subjects were asked to follow the scenario of the figure 1.

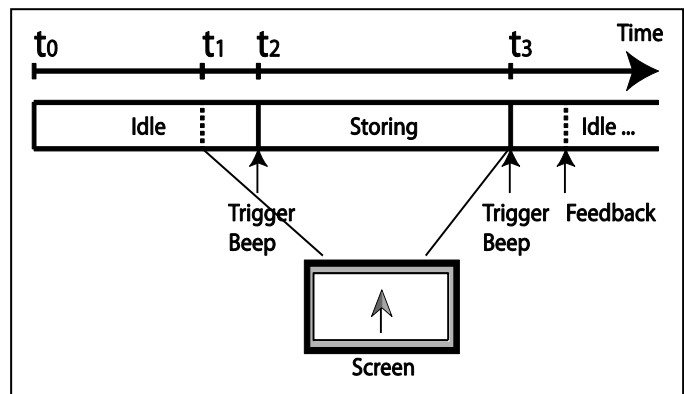


Fig. 1. Experimental Scenarios.

The experiment consists of 18 trials. Each one lasts between 3 and 8 seconds. The first period is an idle one where EEG signals are not used. At $t = t_1$ a beep sound indicates the beginning of the trial accompanied by an arrow pointing the LED to gaze at. At $t = t_2$ data start to be sent to the processing bloc. At $t = t_3$ another beep sound triggers indicating the end of the trial. The feedback is shown on the screen as an increment in the value of true or false trials counters.

IV. METHODS

The recognition techniques in SSVEP-based BCIs are based on the estimation of a coefficient that reflects the power of a stimulus frequency in the EEG signal. The frequency that maximizes this coefficient is considered as the frequency of the selected target.

A. MEC method in the SSVEP-based BCI

MEC based method uses the SNR as a clue of the stimulation frequency. The diagram of figure 2 shows the different steps of the method.

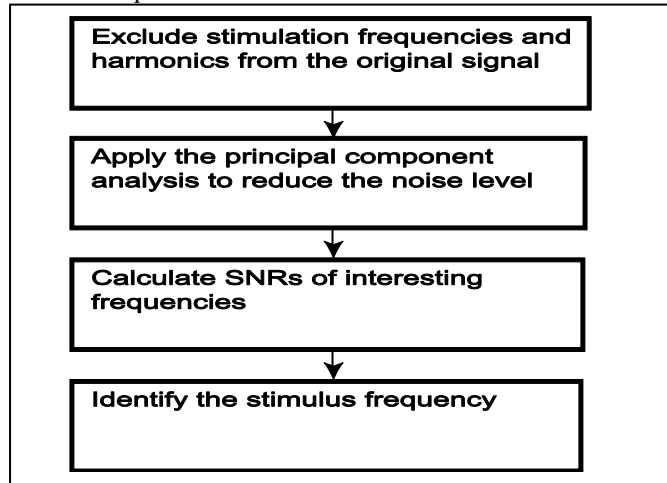


Fig. 2. Block diagram of the MEC method.

The EEG signal $y_i(t)$ is represented by a linear model following the equation (1).

$$y_i(t) = \sum_{k=1}^{N_h+1} a_{i,k} \sin(2\pi kft + \phi_{i,k}) + \sum_j (b_{i,j} + z_j(t)) + e_i(t) \quad (1)$$

Where i is the unique electrode identifier, N_h is the number of harmonics, and f is the stimulation frequency.

The model is decomposed into the sum of three quantities. The first one defines the frequencies of interest where $a_{i,k}$ and $\phi_{i,k}$ are respectively the specific amplitudes and phases, the second is a nuisance signals $z_j(t)$ such as the artifacts where $b_{i,j}$ is the weight factor, and the third quantity represents the noise.

The aim of the two first steps of the block diagram is to reduce the noise level and to increase the interesting frequencies level. First of all, the frequencies of interest are eliminated by projecting the matrix Y of the EEG signal onto the orthogonal complement of the matrix X containing a pair of $\sin(2\pi kft)$ and $\cos(2\pi kft)$ in its columns.

$$\tilde{Y} = Y - X(X^T X)^{-1} X^T Y \quad (2)$$

Where

\tilde{Y} is the matrix of uninteresting signals,

$$X = \begin{bmatrix} \sin\left(2\pi\frac{1}{F_s}f_1\right) & \cos\left(2\pi\frac{1}{F_s}f_1\right) & \dots & \sin\left(2\pi\frac{1}{F_s}N_h f_n\right) & \cos\left(2\pi\frac{1}{F_s}N_h f_n\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sin\left(2\pi\frac{N_t}{F_s}f_1\right) & \cos\left(2\pi\frac{N_t}{F_s}f_1\right) & \dots & \sin\left(2\pi\frac{N_t}{F_s}N_h f_n\right) & \cos\left(2\pi\frac{N_t}{F_s}N_h f_n\right) \end{bmatrix}$$

F_s is the sampling frequency, N_t is the number of samples, and $f_1 \dots f_n$ are the stimulation frequencies.

In the second step, the principal component analysis (PCA) is utilized to find a linear combination minimizing the variance of the matrix \tilde{Y} . The application of this linear combination on the original matrix Y allows the creation of the matrix S with a reduced noise level.

In the following steps, the SNR values, as described in equation 3, are measured then the stimulus frequency that maximizes the SNR is considered as the frequency of interest.

$$SNR(f) = P(f) / \sigma(f) \quad (3)$$

Where $P(f)$ is the signal power function and $\sigma(f)$ is an estimation of the noise power.

B. CCA method in the SSVEP-based BCI

The CCA allows to compare two groups of variables in order to know if they describe the same phenomenon. In the SSVEP-based BCI, the CCA is used for a comparison between the multi-channel EEG signals and a reference signal R_{f_i} including the stimulus frequencies and the harmonics.

$$R_{f_i} = \begin{pmatrix} \sin(2\pi f_i(t)) \\ \cos(2\pi f_i(t)) \\ \vdots \\ \sin(2\pi N_h f_i(t)) \\ \cos(2\pi N_h f_i(t)) \end{pmatrix} \quad (4)$$

Figure 3 depicts the different steps of the CCA based method. The value of the frequency of interest is the same as the reference frequency that maximizes the correlation coefficient.

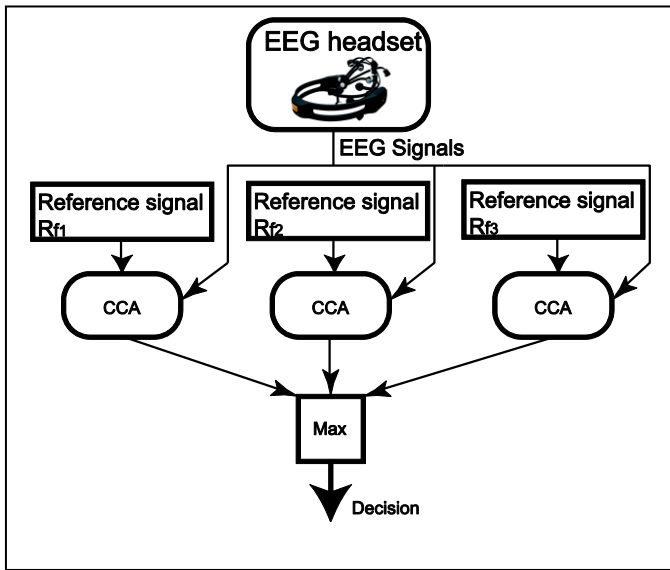


Fig. 3. Block diagram of the CCA method.

C. MSI method in the SSVEP-based BCI

The MSI method estimates the synchronization index between the EEG signal Y and the reference signals R_{f_i} .

The correlation matrix between two sets of data Y and R is given by the equation (5):

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} \quad (5)$$

Where:

$$C_{xx} = \frac{1}{M} YY^T, C_{xy} = C_{yx} = \frac{1}{M} YR^T, C_{yy} = \frac{1}{M} RR$$

and M is the number of samples.

To reduce the effect of the autocorrelation, the following linear transformation is applied to the matrix C producing the matrix \check{C} .

$$\check{C} = UCU^T \quad (6)$$

$$\text{Where: } U = \begin{bmatrix} C_{xx}^{-\frac{1}{2}} & 0 \\ 0 & C_{yy}^{-\frac{1}{2}} \end{bmatrix}$$

A normalization of the eigenvalues λ_i of the matrix \check{C} is given by:

$$\lambda'_i = \frac{\lambda_i}{\sum_{j=1}^P \lambda_j} \quad (7)$$

Where: P is the number of eigenvalues.

Finally, the synchronization index S is defined by the equation (8).

$$S = 1 + \frac{\sum_{i=1}^P \lambda'_i \log(\lambda'_i)}{\log(P)} \quad (8)$$

The S quantity tends towards 0 when Y and R are increasingly uncorrelated and towards 1 when Y and R are increasingly correlated. Consequently, the frequency of the reference which has the maximum synchronization index is considered as the frequency of interest.

D. MEC method amelioration

The EEG signals have a poor signal to noise ratio [19] which makes the brain activity patterns difficult to be detected. Nevertheless, it is not always possible to discriminate between the different mental tasks.

The goal of this improvement is to reduce the noise sensitivity of the MEC method. Considering that interesting signals are composed of stimulation frequencies and harmonics, and the rest is noise. The noise sensitivity effect can be caused by the fact that a part of the noise is considered as relevant or a part of the interesting signal is considered as irrelevant. In both cases, the problem is in the separation between the noise and the interesting signal. Thus, this problem can be localized in the first step of the MEC method.

The idea of the improvement is to use the empirical mode decomposition (EMD) to divide the EEG signal into useful and noise signals instead of using the matrix projection.

The EMD was firstly proposed by Huang et al. [20] as an efficient method to evaluate the frequency and amplitude of time-series with excellent time resolution. It divides the original signal into some Intrinsic Mode Functions (IMF), which are different scales of oscillation components, and a residue. The sifting process described as follows leads to extract each IMF and the residue.

Initially, the first residue and the first difference take the value of the initial EEG data.

- Step1: Locate the local maxima and minima of the difference.
- Step2: Calculate the lower and the upper envelop using these extrema.
- Step3: Calculate the mean by averaging the upper and the lower envelop.

- Step4: Calculate the new difference by subtracting the mean from the previous difference.
- Step5: If the stopping criteria are satisfactory, then the last calculated difference is an IMF, otherwise go to step1 and continue the process step by step.
- Step6: Calculate the new residue by subtracting the last IMF from the previous residue.
- Step7: If the new residue is not isometric then repeat the process from step1. Otherwise, the sifting process is ended, and the last found residue is considered as the final residue of the process.

The stopping criteria are a compound of two conditions. First, the number of extrema and the number of zero crossing must be either equal or differ by one at most. Second, the standard deviation (SD) is smaller than a predetermined value.

$$SD = \frac{\sum_{t=0}^T |h_{k-1}(t) - h_k(t)|^2}{\sum_{t=0}^T h_{k-1}^2(t)} \quad (9)$$

Where: h_i are the vectors of differences; k is the differences counter and T is the period of considered samples.

Figure 4 illustrates the result of the sifting process for a SSVEP signal during the period of 2s.

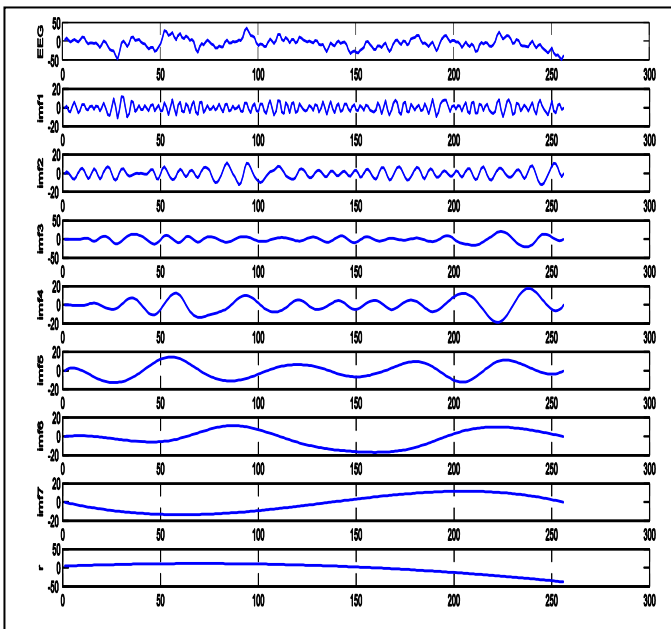


Fig. 4. Decomposition of the original signal in IMFs and a residue.

Each IMF has a higher frequency than the next extracted one. The residue which is the lowest frequency component represents the trend of the signal.

The spectral analysis of the IMFs allows to separate them into three groups. The first contains the IMFs located in a

lower frequency band than the stimulus frequency. This frequency and its neighbors constitute the frequency band of the second group. The last one contains the IMFs characterized by several scattered frequencies higher than the stimulation frequency.

Figures 5 (A)-(B) illustrate the Normalized Amplitude Spectrum (NAS) of different IMFs, defined in equation (10).

$$NSA(x) = \frac{FFT(x)}{\sum_{i=1}^M FFT(x_i)} \quad (10)$$

Where: x is an IMF time series, $FFT(x)$ is the fast Fourier transform of x , and M is the number of FFT points.

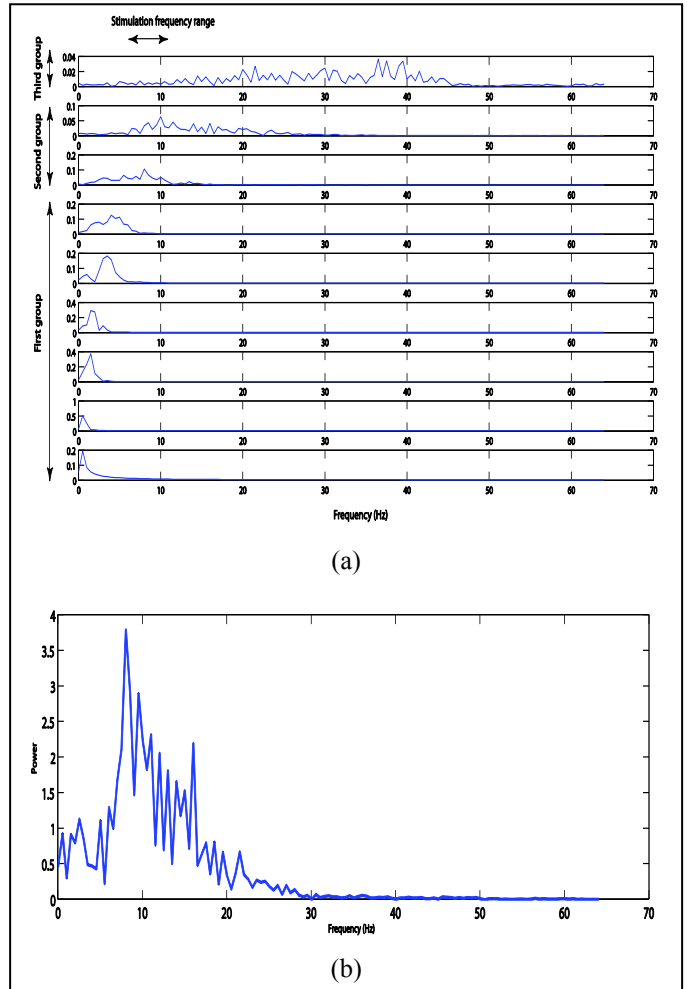


Fig. 5. Spectral analysis of the IMFs: (A) separation between the three groups, (B) spectrum of the stimulation frequency group.

The IMFs located in the first group are considered as noise containing neither stimulation frequencies nor harmonics. To discriminate between the first and the second group, a decision-making criterion is required. R. Sharma et al. [21] have used the sample entropy (SampEn) as a complexity measure of IMFs extracted from EEG signal. This experiment

shows that the SampEn decreases from one IMF to the next. Therefore, SampEn can be a good criterion to separate the IMFs of the first group from the IMFs of the second. Likewise, IMFs which have a SampEn inferior to a predetermined threshold are considered as noise components.

After subtracting the IMFs of the first group from the original signal, the remainder can be considered as interesting or a mixture between the noise and the interesting signal according to the level of noise. The central frequency of the second IMF is in the stimulus frequencies band. Its NAS indicates the noise level. If the NAS is superior to a predetermined threshold, the signal to noise ratio is high enough and the first step of the MEC method is ended. However, in the second case, a filtering is needed to exclude the rest of noise while keeping the sharpest EEG signal response. The moving average filter is optimal for this kind of issue. In spite of its simplicity, it ensures a low curve shape change to keep valid the previous decompositions. To produce each point, some input points are averaging according to the equation (11).

$$y_i = \frac{1}{M} \sum_{j=0}^{M-1} x_{i+j} \quad (11)$$

Where x_i is an input point, y_i is an output point, and M is the number of points in the average.

When M increases the noise decreases but the acute curve angles become obtuse. The best choice of M is about eleven [22].

V. RESULTS AND DISCUSSION

For the comparison purpose, the previously presented methods were implemented as well as the MEC proposed amelioration. The experiment results allow to validate the amelioration.

A. Separation between noise and interesting signal

The SampEn calculates the probability that epochs of window length m that are similar within a tolerance r remain similar at the next point [23]. A study of the different possible combinations shows that the best values of m and r are 6 and 0.2 respectively. Also, the best choices of the thresholds are 0.1 for the SampEn and 0.08 for the NAS.

Figure 6 reports an example of the extracted noise and interesting signal from the EEG recording. In the interesting signal, the stimulus frequency dominates the harmonics frequencies as it has a higher amplitude response than them [8]. Moreover, it follows the shape of the EEG signal i.e. the original features are maintained. The Nuisance signal frequency seems to be higher than the interesting signal as it includes the artifacts.

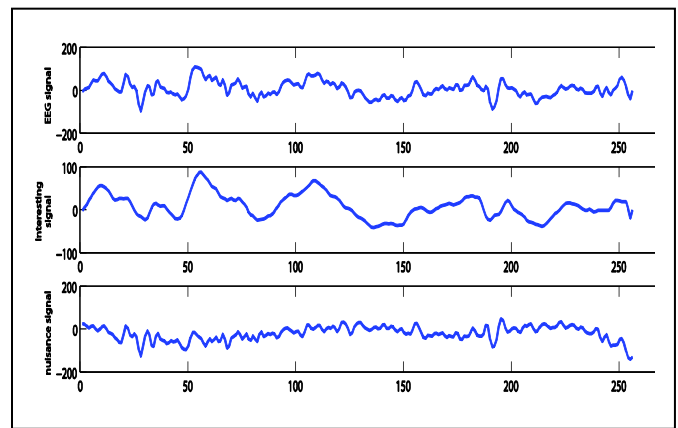


Fig. 6. Extracted noise and interesting signal using the proposed improvement.

B. Stimuli color effect on different recognition methods

Overlapping over several recording intervals can lead to improve the precision of the system. In this experiment, five intervals of a length of 3s each and with a gap of 0.25s were used. The idle period is fixed to 4s. The experiment consists of changing the colors of the three used LEDs. The colors are white, green, red, blue, and yellow. Table 1 illustrates the impact of the colors on the system accuracy rates using the processing methods mentioned before. The results show that the responses of white and yellow colors evoked an accuracy exceeding the 80%. However, the precisions with red, green, or blue stimuli are lowest.

The white and the yellow color are the brightest which explains this result. In fact, D. G. Albrecht et al. [25] examined the effect of the contrast intensity of the visual stimulation on neurons from the Visual area one V1 or the striate cortex. These neurons are sensitive to the object features at which the subject gazes as the color and the direction. The results show that the response of a striate cell increases as the contrast intensifies. The Accuracy rate using a white color is most stable and high. This consequence justifies the choice of the white color for the next experiments.

TABLE I. SSVEP RESPONSES TO STIMULI OF DIFFERENT COLORS

Color	Method	Method		
		MEC	MSI	CCA
Yellow		44%	89%	72%
Blue		50%	55%	55%
Green		67%	78%	78%
Red		50%	67%	67%
White		72%	83%	83%

C. Noise level effects on different recognition methods

As explained before, the nature of the light source has a great effect on the noise level. In the next experiment, two kinds of white light LEDs are used in order to verify the noise sensitivity of each method. The first is the universal LEDs with 3mm and the second is the chips on board (COB) LEDs. COB LEDs embedded ten LEDs in a circular surface with a diameter of 35mm. Therefore they can produce a sufficient light intensity without causing fatigue or losing the subject concentration. The subjects were asked to follow the scenario

of the previous experiment. Figure 7 presents the accuracy rates of MEC, CCA, and MSI methods for the five subjects with the 3mm LEDs. It is clear that the high level of noise leads to decrease the performances of all processing methods.

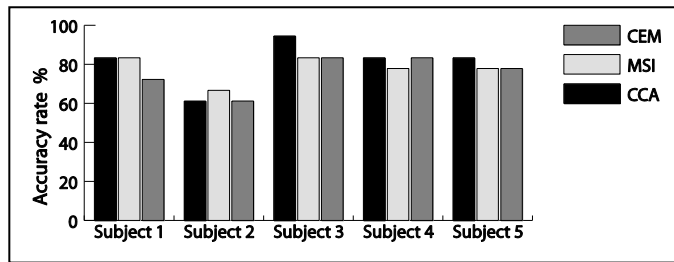


Fig. 7. SSVEP responses to 3mm LEDs stimuli with different processing methods.

Table 2 illustrates the percentage of correct recognitions for each method using 3mm LEDs.

TABLE II. AVERAGE ACCURACY OF EACH METHOD WITH 3MM LEDs

Method	MEC	MSI	CCA
Accuracy rate	75.55%	77.77%	81.1%

The MEC responses remain the lowest even if the accuracy values are close.

In the next section, The COB LEDs replace the universal ones. Figure 8 illustrates the results of different methods.

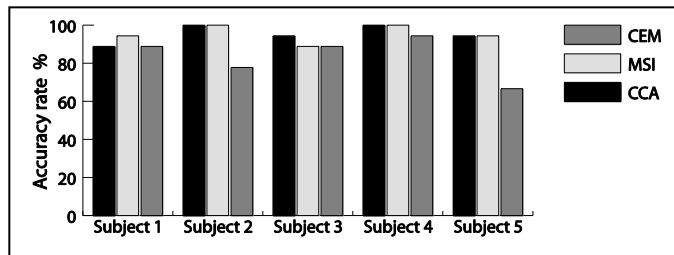


Fig. 8. The accuracy values with COB LEDs stimuli.

The experimental results show that all methods performed better using COB LEDs than using 3mm LEDs. Thus, the level of noise decreases using COB LEDs.

Table 3 illustrates the percentage of correct recognitions for each method using COB LEDs stimuli.

TABLE III. AVERAGE ACCURACY OF EACH METHOD WITH COB LEDs

Method	MEC	MSI	CCA
Accuracy rate (COB LED)	86%	95.55%	95.55%
Amelioration value (%)	10.28%	22.85%	17.8%

Although the three methods are noise sensitive, MEC remains largely poorer. This method is influenced even with little noise levels.

D. Application of the proposed improvement

The goal of this experiment is to validate the proposed improvement procedure of MEC method for the SSVEP features recognition. The COB LEDs are the light sources. The study of the average accuracy rates of MEC, CCA, and MSI methods allows to compare their noise sensitivities as well as the improved MEC for the five subjects. Three data intervals of 2s with a gap of 0.25s were used to identify the target.

Table 4 summarizes the performances of the different methods.

TABLE IV. RECOGNITION METHODS PERFORMANCE

Method	CCA	MSI	MEC	Improved MEC
Accuracy rate	96.66%	96.66%	83.33%	98.88%

The proposed improvement produces more precise results in selecting the stimulation frequency. It ameliorates the accuracy of the MEC method by about 13%. Thus, MEC becomes efficiently resistant to the noise. This method becomes even more performant than other methods and reaches 99% with a total recording interval of 2.5s. This investigation confirms the hypothesis that is using the EMD and the moving average filter allows to separate noise and artifact from the interesting signal. EMD decomposes a nonlinear and nonstationary signal into a sum of IMFs without the need of prior knowledge. In fact, it is an adaptive technique depending on the local characteristic of the signal which explains its compliance with the moving average filter. The results confirm the validity of the improvement.

E. Information-transfer rate

One of the most used metrics to evaluate the performance of BCI systems is the information-transfer rate (ITR). Wolpaw et al. [26] have proposed the most popular method for ITR calculation as defined in equation (12).

$$B = \log_2 N + P \log_2 P + (1 - P) \log_2 [(1 - P)/(N - 1)] \quad (12)$$

Where B is the ITR (bits/symbol), N is the number of possible commands and P is the classification accuracy. In order to make this quantity easier to understand, another ITR definition B_t in bits/min which is derived from B is generally used.

$$B_t = B * (60 / T) \quad (2)$$

Where T is the average time needed to convert a brain feature activity into a command.

A higher ITR leads to a better and more natural use of the system. In fact, this criterion reflects the time during which the subject has to gaze at the target and the number of commands needed to reach the destination. In order to foster this criterion, another light source was added to the stimulation system. Also,

the idle period and the trial length have gradually been reduced. Moreover, in order to reduce the number of needed time intervals, a decision is made as soon as the same target is identified three times by the system in which the total number of intervals is equal or less than five. Otherwise, if five-time intervals are processing without recognizing the target, the trial is considered as erroneous. Table 5 illustrates a comparison of the CCA method and the proposed method where the idle period is equal to 0.5s and the interval is equal to 2s. A lower interval decreases enormously the accuracy of the system and obviously increases the time cost needed to fix the wrong choices.

TABLE V. MEAN ITR AND ACCURACY AS A FUNCTION OF NUMBER OF TARGETS FOR CCA AND IMPROVED MEC METHODS

	Method			
	Improved MEC		CCA	
	3 targets	4 targets	3 targets	4 targets
Average accuracy rate	95.5%	91.78%	91.11%	88.14%
Average trial length	3.108s	3.222s	3.046s	3.133s
Average information-transfer rate (bits/min)	24.617	27.18	20.944	24.64

The results of this experiment show that the proposed method remains better even with the new scenario. The system can reach acceptable performances with four commands.

VI. CONCLUSION

Steady State Visual Evoked Potential is the most effective solution for BCIs in everyday use. Its low required training and high accuracy rate make its use close to the ordinary one. Three LEDs with different frequencies (8, 9 and 10 Hz) were used during the first experiments. Later an additional LED flickers at 11 Hz was added. Each LED represents a possible direction to control the navigation of an electric wheelchair. The EPOC EMOTIVE headset was used to acquire data from the five volunteers. Only the closest six electrodes to the occipital area were used. In this study, a new amelioration to improve the robustness against the noise of the Minimum Energy Combination method was proposed. Results prove that the stimulus characteristics have a great impact on the noise level in the SSVEP signal. The use of COB LEDs allows to increase the SNR. Also the white color increases the excitation of neurons from the visual cortex of the brain and allows to reach the best accuracy values. The Canonical Correlation Analysis and the Multivariate Synchronization Index based methods serve as references for comparing their performances with the performances of the proposed improvement. The results indicated that the improved MEC method performed better than the widely used CCA and MSI. The average accuracy rate reaches 99% and increases by about 13% compared to the original MEC with a data length of 2.5s using three targets. Also using four targets, the system reached an average information-transfer rate of about 27.5 bits/min. This

makes the system more suitable for the wheelchair navigation command.

REFERENCES

- [1] A. L. S. Ferreira, L. C. de Miranda, E. E. C. de Miranda and S. G. Sakamoto "A Survey of Interactive Systems based on Brain-Computer Interfaces", SBC Journal on 3D Interactive Systems, VOL. 4, NO. 1, 2013, pp 3-13.
- [2] M. Middendorf, G. McMillan, G. Calhoun and K. S. Jones "Brain-Computer Interfaces Based on the Steady-State Visual-Evoked Response", IEEE Transactions on Rehabilitation Engineering, VOL. 8, NO. 2, 2000, pp 211-214.
- [3] W. Yijun, W. Ruiping, G. Xiaorong and G. Shangkai "Brain-computer Interface based on the High-frequency Steady-state Visual Evoked Potential", First International Conference on Neural Interface and Control Proceedings, Wuhan, China, 26-28 May 2005, pp 37-39.
- [4] Y. Wang, R. Wang, X. Gao, B. Hong and S. Gao "A Practical VEP-Based Brain-Computer Interface", IEEE Transactions on neural systems and rehabilitation engineering, VOL. 14, NO. 2, 2006, pp 234-239.
- [5] F. Gembler, P. Stawicki and I. Volosyak "A Comparison of SSVEP-Based BCI-Performance Between Different Age Groups", 13th International Work-Conference on Artificial Neural Networks, Palma de Mallorca, Spain, June 10-12 2015, pp 71-77.
- [6] Q. Liu, K. Chen, Q. Ai and S. Q. Xie "Review: Recent Development of Signal Processing Algorithms for SSVEP-based Brain Computer Interfaces", Journal of Medical and Biological Engineering, VOL. 34, NO. 4, 2013, pp 299-309.
- [7] O. Friman, I. Volosyak and A. Gräser "Multiple Channel Detection of Steady-State Visual Evoked Potentials for Brain-Computer Interfaces", IEEE transactions on biomedical engineering, VOL. 54, NO. 4, APRIL 2007, pp 742-750.
- [8] N. Chumerin, N. V. Manyakov, A. Combaz, Ar. Robben, M. vanVliet and M. M. Van Hulle "Steady State Visual Evoked Potential Based Computer Gaming - The Maze", Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, INTETAIN 2011, LNICST VOL. 78, 2012, pp. 28-37.
- [9] Z. Lin, C. Zhang, W. Wu and X. Gao "Frequency Recognition Based on Canonical Correlation Analysis for SSVEP-Based BCIs", IEEE transactions on biomedical engineering, VOL. 54, NO. 6, JUNE 2007, pp 1172-1176.
- [10] G. Hakvoort, B. Reuderink and M. Obbink "Comparison of PSDA and CCA detection methods in a SSVEP-based BCI-system", technical report TR-CTIT-11-03, center for telematics and information technology, university of twente (2011).
- [11] N. Mora, V. Bianchi, I. De Munari and P. Ciampolini "A BCI Platform Supporting AAL Applications", 8th International Conference Universal Access in Human-Computer Interaction, Heraklion, Crete, Greece, June 22-27 2014, pp. 515-526.
- [12] W. Nan, C. M. Wong, B. Wang, F. Wan, P. U. Mak, P. I. Mak and M. Vai "A Comparison of Minimum Energy Combination and Canonical Correlation Analysis for SSVEP Detection", the 5th International IEEE EMBS Conference on Neural Engineering Cancun, Mexico, April 27 - May 1 2011, pp 469-472.
- [13] Y. Zhang, P. Xu, K. Cheng and D. Yao "Multivariate synchronization index for frequency recognition of SSVEP-based brain-computer interface", Journal of Neuroscience Methods, VOL. 221, 15 January 2014, pp 32- 40.
- [14] C. S. Herrmann "Human EEG responses to 1-100 Hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena", Experimental Brain Research, NO. 137, Springer-Verlag, 2001, pp 346-353.
- [15] F.C. Lin, J. K. Zao, K.C. Tu, Y. Wang, Y.P. Huang, C.W. Chuang, H.Y. Kuo, Y.Y. Chien, C.-C. Chou and T.P. Jung "SNR Analysis of High-Frequency Steady-State Visual Evoked Potentials from the Foveal and Extrafoveal Regions of Human Retina", 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society

- EMBS, San Diego, California USA, 28 August - 1 September 2012, pp 1810-1814.
- [16] E. E. Sutter “The brain response interface: communication through visually-induced electrical brain responses”, *Journal of Microcomputer Applications* VOL. 15, issue 1, 1992, pp 31-45.
- [17] A. González-Mendoza, J. L. Pérez-Benítez, J. A. Pérez-Benítez and J.H. Espina-Hernández “Brain Computer Interface based on SSVEP for controlling a remote control car”, *International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Cholula, 25-27 Feb. 2015, pp 93-97.
- [18] A. Duszyk, M. Bierzynska, Z. Radzikowska, P. Milanowski, R. Kus, P. Suffczynski, M. Michalska, M. Labecki, P. Zwolinski and P. Durka “Towards an Optimization of Stimulus Parameters for Brain-Computer Interfaces Based on Steady State Visual Evoked Potentials,” *Plos One*, VOL. 9, NO. 11, 2014, pp.11.
- [19] Y. Liu, X. Jiang, T. Cao, F. Wan, P. U. Mak, P.-I. Mak and M. I Vai “Implementation of SSVEP Based BCI with Emotiv EPOC”, *IEEE International Conference on Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS)*, Tianjin, 2-4 July 2012, pp 34-37.
- [20] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche and B. Arnaldi. “A review of classification algorithms for EEG-based brain-computer interfaces,” *Journal of Neural Engineering*, IOP Publishing, 2007, 4, pp.24.
- [21] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.C. Yen, C. C. Tung and H. H. Liu “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis”, *Proc. The Royal Society, Lond. A* (1998) 454, pp 903–995, Printed in Great Britain.
- [22] R. Sharma, R. BilasPachori and U. R. Acharya “Application of Entropy Measures on Intrinsic Mode Functions for the Automated Identification of Focal Electroencephalogram Signals”, *Entropy*, NO. 17, 2015, pp 669-691.
- [23] S. Smith “Digital Signal Processing: A Practical Guide for Engineers and Scientists, 1stEdition” 07 Nov. 2002 pp. 279-280.
- [24] D. E. Lake, J. S. Richman, M. P. Griffin and J. Randall Moorman “Sample entropy analysis of neonatal heart rate variability”, *American Journal of Physiology Regulatory, Integrative and Comparative Physiology*, 1 September 2002 Vol. 283 no. 3, R789-R797.
- [25] D. G. ALBRECHT and D. B. HAMILTON “Striate Cortex of Monkey and Cat: Contrast Response Function” *journal of neurophysiology*, Vol.48, No.1, 1982.
- [26] J. R. Wolpaw, H. Ramoser, D. J. McFarland, and G. Pfurtscheller “EEG-Based communication: improved accuracy by response verification” *IEEE TRANSACTIONS ON REHABILITATION ENGINEERING*, VOL. 6, NO. 3, SEPTEMBER 1998, pp. 326–33.

An Analysis on Natural Image Small Patches

Shengxiang Xia
School of Science,
Shandong Jianzhu University,
Jinan 250101 P.R.China

Wen Wang
School of Science,
Shandong Jianzhu University,
Jinan 250101 P.R.China

Di Liang
School of Science,
Shandong Jianzhu University,
Jinan 250101 P.R.China

Abstract—The method of computational homology is used to analyze natural image 8×8 and 9×9 -patches locally. Our experimental results show that there exist subspaces of the spaces of 8×8 and 9×9 -patches that are topologically equivalent to a circle and a Klein bottle respectively. These extend the results of the paper "on the local behavior of spaces of natural images." To the larger patches. The Klein bottle feature of natural image patches can be used in image compression.

Keywords—natural image analysis; persistent homology; high-contrast patches; Klein bottle; barcode

I. INTRODUCTION

Many results on statistics of images were obtained in the recent years [1], [2], [3]. Lee, Pedersen, Mumford [3] discuss the distributions of 3×3 image patches, they found that the majority of high-contrast 3×3 patches concentrate near A circle. Carlsson, Ishkanov, de Silva, and Zomorodian [4] analyze 3×3 natural image patches; they find a high density the subset is called the primary circle and prove that there exists the large 2-dimensional subset with the topology of a Klein bottle Which includes the primary circle. In [5], we showed that 4×4 , 5×5 , 6×6 and 7×7 natural image patches have the circle behavior.

In this paper, we utilize the methods of the paper [4] to study the structure of $n \times n$ high-contrast natural image patches for the cases $n=8$ and $n=9$. In particular, we find the largest 2-Dimensional subspace of each case, whose topology is that of a Klein bottle. The results of the paper enlarge the results of [4] To 8×8 and 9×9 patches. The Klein bottle feature of image patches can be used in techniques of image compression [4], [6]. The data sets used in this paper were chosen from INRIA Holidays dataset [7], which are different from that of the paper [4].

II. THE DATA SETS OF NATURAL IMAGE PATCHES

As the dimensional problem of the data, it is very difficult To directly analyze the pixel distribution of images. We divide each natural image into small $n \times n$ -patches, and consider each patch as an n^2 -dimension vector, we study the topology of the space of $n \times n$ -patches for sufficiently small n , here we study the cases of $n=8, 9$.

We sample data sets of high-contrast 8×8 and 9×9 Patches from 550 sampled natural images in INRIA Holidays dataset [7]. Each data set consists of about $55 \cdot 10^5$ high-Contrast log patches. INRIA Holidays dataset is available at <http://lear.inrialpes.fr/%7ejegou/data.php>. Fig.1 has two sam-

Our main spaces X_8 and X_9 are sets of 8×8 and 9×9 Patches of high contrast created by the following steps. The routine handled here is similar to [3], [4], [8].

Step 1. Sample 550 images from INRIA Holidays dataset.

Step 2. Using MATLAB function `rgb2gray` to compute the intensity at each pixel for each image.

Step 3. We randomly select 5000 8×8 and 9×9 patches from each image.

Step 4. We treat each patch as an n^2 -dimensional vector, and take the logarithm of each coordinate.

Step 5. For any vector $\mathbf{x}=(x_1, x_2, \dots, x_n)$, we calculate the D -norm: $\|\mathbf{x}\|_D$. Two coordinates of \mathbf{x} are neighbors, expressed by $i \sim j$, if the corresponding pixels in the $n \times n$ patch are adjacent. The formula of D -norm is: $\|\mathbf{x}\|_D = \sqrt{\sum_{i \sim j} (x_i - x_j)^2}$.

Step 6. We select the patches which have a D -norm in the top t percent in each image. We take $t = 20\%$, as done in [3], [4], [8].

Step 7. Subtract an average of all coordinates from each coordinate.

Step 8. We map X_8 (X_9) into a unit sphere by dividing each vector with its Euclidean norm. We do not translate to the DCT basis for convenience.

Step 9. We randomly select 50,000 points from X_8 and X_9 for computational convenience, the subspaces of X_8 and X_9 are indicated by \bar{X}_8 and \bar{X}_9 respectively.

III. COMPUTATIONAL METHOD

For determining topological features of an underlying space by sampled finite points, the computing method used in this paper is persistent homology, which is set up by Edelsbrunner, Letscher, and Zomorodian [9] and distilled by Carlsson and Zomorodian [10]. To apply persistent homology, we firstly build lazy witness complexes for a sampled point set P from underlying space X .

For a point cloud P , a landmark subset L , for all $p \in P$. Let $t(p)$ be the distance p to the closest landmark point. The lazy witness complex $LW(P, L, \epsilon)$ is formulated as follows: (i) the vertex set is L ; (ii) for vertices a and b , edge $[ab]$ is in $LW(P, L, \epsilon)$ if there is a witness point $p \in P$ such that

$$\max\{d(a, p), d(b, p)\} \leq \epsilon + t(p);$$

(iii) a higher dimensional simplex is in $LW(P, L, \epsilon)$ if all of its edges are.



Fig. 1. Samples from INRIA Holidays dataset

The most important parameter in a sequence of lazy witness complexes is ϵ , but there is no an optimal value of ϵ without prior information of the underlying space, we do not know how To pick the value of ϵ . However, using the Javaplex package developed by Adam and Tausz [12], we can compute the Betti numbers in an interval of ϵ and explain the result by a Betti barcode. The instinctive explanation is that long intervals accord to actual topological features of the underlying space while short ones are explained as noise.

To uncover the topological features of our spaces X_8 and X_9 , we use the different core subsets of X_8 and X_9 . We evaluate the local density of the space at a point by its nearest neighbor. For $y \in X$ and $k > 0$, let $\rho_k(y) = |y - y_k|$, here y_k is the k th nearest neighbor of y . The larger k -Values contribute more global estimations, while small k -Values result in local density estimates. For a given k , we arrange the points of X by descending density; we pick the points with densities in the top p percent, written as $X(k, p)$. The core subset $X(k, p)$ possibly give important topological information, which may be disappeared for all the points of X .

Here we examine core subsets $\bar{X}_n(k, p)$ of \bar{X}_n for $n = 8, 9$. Core subsets have two parameters k and p , they demonstrate some topological features of their underlying space for k and p with suitable values.

IV. RESULTS FOR $\bar{X}_8(k, p)$ AND $\bar{X}_9(k, p)$

The authors of the paper [4] applied persistent homology to detect the topologies of high-density subsets of natural image patches. They discovered that the topology of the core sets vary from a circle to a 3-circle model as decreasing of Density estimator k . In this paper, we use INRIA Holidays data set, which is other than that of the paper [4], to prove experimentally that some core sets of X_8 and X_9 possess the similar results as above.

We take the core subsets $\bar{X}_8(300, 20)$, $\bar{X}_9(300, 20)$, and calculate the barcodes by Javaplex software, their sample barcode plots are displayed in Fig.2, Fig.2 separately. In Fig.2, Fig.3, there exist a single long line of $Betti_0$ and a single long $Betti_1$ line (i.e. $\beta_0 = 1$ and $\beta_1 = 1$), which means that

they have the topology of a circle. When we project core sets onto some plane, their circle feature is visible in Fig.4 and Fig.5. Choosing different landmark points, we run many times experiments on $\bar{X}_8(300, 20)$, $\bar{X}_9(300, 20)$, and the results are very steady.

For 8×8 and 9×9 patches, there are much different cores subsets of \bar{X}_8 and \bar{X}_9 , if we take proper values of parameters k and p , the core sets also have the topology of a circle. For $\bar{X}_8(100, 20)$, $\bar{X}_8(200, 20)$, $\bar{X}_9(100, 20)$ and $\bar{X}_9(200, 20)$, We ran many trials and found that they have the topology of a circle and the results to be robust.

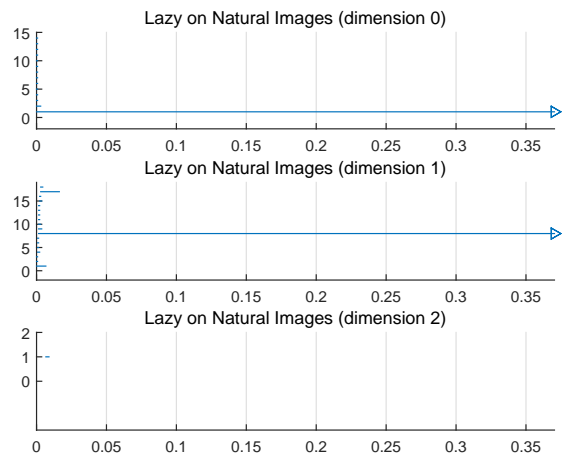


Fig. 2. PLEX results for $\bar{X}_8(300, 20)$

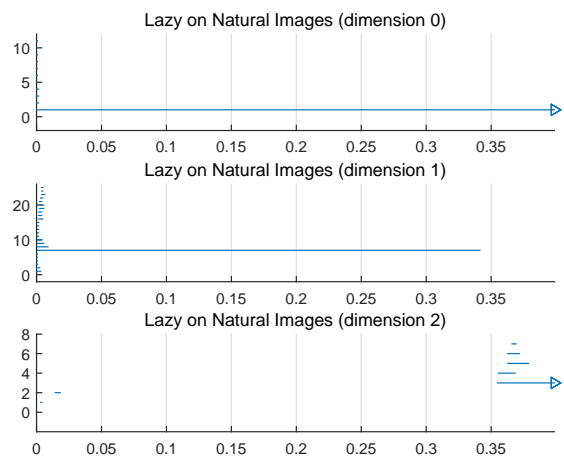


Fig. 3. PLEX results for $\bar{X}_9(300, 20)$

When we consider the core subsets $\bar{X}_8(15, 20)$, $\bar{X}_9(15, 20)$, and calculate the barcodes, their sample barcode plots are shown in Fig.6, Fig.7 separately. In Fig.6 (Fig.7), there are a single long line of $Betti_0$ and five long $Betti_1$ line for ϵ from 0.06 to 0.18 (from 0.05 to 0.19), which shows that they have the topology of three circle model [4] (Fig.8), that is, Betti numbers $\beta_0 = 1$ and $\beta_1 = 5$.

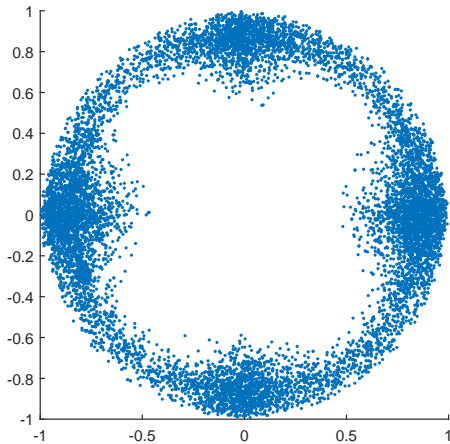


Fig. 4. Projection of $\bar{X}_8(300, 20)$ onto a plane

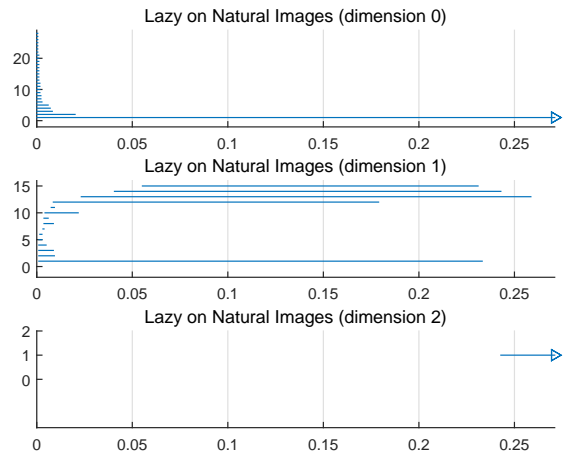


Fig. 6. PLEX results for $\bar{X}_8(15, 20)$

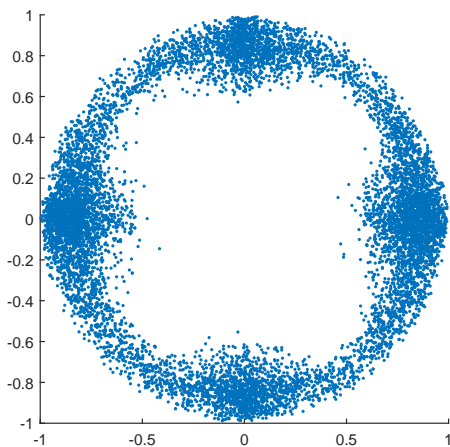


Fig. 5. Projection of $\bar{X}_9(300, 20)$ onto a plane

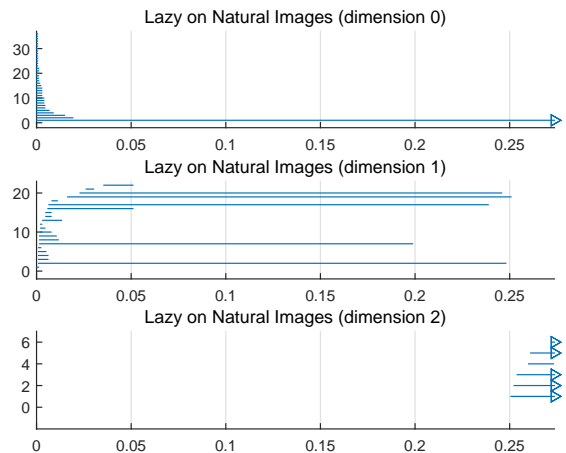


Fig. 7. PLEX results for $\bar{X}_9(15, 20)$

V. EMBEDDING OF THE KLEIN BOTTLE INTO S^{63} AND S^{80}

The Klein bottle is a very important non-orientable surface, it can be sketched by the quotient space of the square $[0, 1] \times [0, 1]$ with sides glued by the relations $(0, y) \sim (1, y)$ for $y \in [0, 1]$ and $(x, 0) \sim (1 - x, 1)$ for $x \in [0, 1]$. To identify the Klein bottle features of subspaces of X_8 and X_9 , we embed the Klein bottle into S^{63} and S^{80} , and get another theoretical model of the Klein bottle.

We define the map $g : S^1 \times S^1 \mapsto \mathcal{P}$ by $(\cos \alpha, \sin \alpha, \cos \beta, \sin \beta) \mapsto \cos \beta(x \cos \alpha + y \sin \alpha)^2 + \sin \beta(x \cos \alpha + y \sin \alpha)$ ([4]), where \mathcal{P} consists of all functions with the form $\cos \beta(x \cos \alpha + y \sin \alpha)^2 + \sin \beta(x \cos \alpha + y \sin \alpha)$, $\alpha, \beta \in [0, 2\pi]$, it is obvious that g is onto, but not one to one, since the points $(\cos \alpha, \sin \alpha, \cos \beta, \sin \beta)$ and $(-\cos \alpha, -\sin \alpha, \cos \beta, -\sin \beta)$ are mapped to the same function, that is, $(\cos \alpha, \sin \alpha, \cos \beta, \sin \beta) \sim (-\cos \alpha, -\sin \alpha, \cos \beta, -\sin \beta)$ is an equivalent relation, the relation can be rewritten as $(\alpha, \beta) \sim (\pi + \alpha, 2\pi - \beta)$. The space $\mathcal{P} = \text{im}(g)$ is homeomorphic to $S^1 \times S^1 / (\alpha, \beta) \sim (\pi + \alpha, 2\pi - \beta)$, as no other identifications produced by g .

A torus has a similar representation to that of the Klein bottle as glued a square with the opposite edges (Fig.9). The effect of the map g on a torus is displayed in Fig.10. Each half is a representation of the Klein bottle, thus the image of g is homeomorphic to the Klein bottle and so is \mathcal{P} ([4]).

We define a map $h_8 : \mathcal{P} \mapsto S^{63}$ by a composite of evaluating a polynomial at each point of the plane grid $G_8 = \{-3, -2, -1, 0, 1, 2, 3, 4\} \times \{-3, -2, -1, 0, 1, 2, 3, 4\}$ subtracting the mean and normalizing. In a similar way, we define $h_9 : \mathcal{P} \mapsto S^{80}$ on the grid $G_9 = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\} \times \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$. Because continuous 1-1 map on a compact space is a homeomorphism onto its image, as Proved in [4], the images $\text{im}(h_8)$ and $\text{im}(h_9)$ are homeomorphic to the Klein bottle.

To embed the Klein bottle into into S^{63} and S^{80} , primarily, we uniformly take 200 points $(\{x_1, \dots, x_{200}\})$ from the unit circle, all possible tuples (x_i, x_j) produce a point set on the torus $S^1 \times S^1$. Secondly, we map each of the 40000 points into S^{63} and S^{80} through compositions of $h_8 \circ g$ and $h_9 \circ g$ separately,

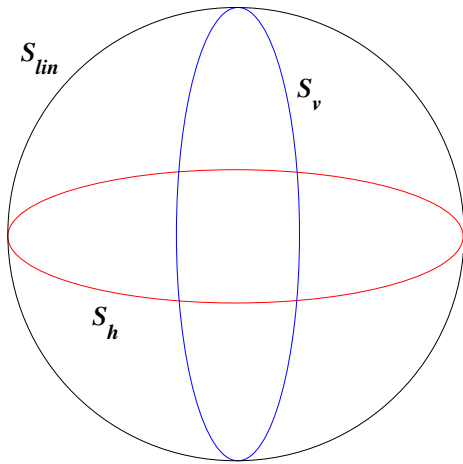


Fig. 8. Three circle model

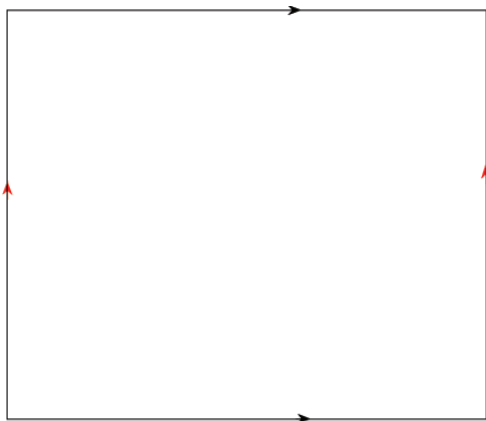


Fig. 9. Denotation of a torus as a quotient space

and the image of the composition is presented by $K_8(200)$ and $K_9(200)$ respectively. Fig.11, Fig.12 display the PLEX results of the spaces $K_8(200)$ and $K_9(200)$ respectively, they provide the Betti numbers $\beta_0 = 1, \beta_1 = 2$ and $\beta_2 = 1$, these are the mod 2 Betti numbers of the Klein bottle. Therefore, $K_8(200)$ ($K_9(200)$) is an appropriate approximation of the Klein bottle in S^{63} (S^{80}).

VI. RESULTS FOR X_8 AND X_9

We have embedded the Klein bottle into S^{63} and S^{80} , and the subspaces $K_8(200)$ and $K_9(200)$ are a proper approximation of the Klein bottle in S^{63} and S^{80} respectively. Applying $K_8(200)$ and $K_9(200)$, we can find subspaces of X_8 and X_9 , whose topology is that of the Klein bottle. The constructing process of the subspaces of X_8 and X_9 are as following.

For each point of X_8 , we compute the Euclidean distance from the point to point set $K_8(200)$, then we resort points of X_8 according to increasing of their Euclidean distances to $K_8(200)$, then we take the top t percent of the closest distances, and represent the subspace of X_8 as $XP_8(200, t)$. The subspace $XP_9(200, t)$ of X_9 is constructed by a similar way.

To find subspaces of X_8 and X_9 having the topology of the Klein bottle, we take the parameter $t=20$ we do

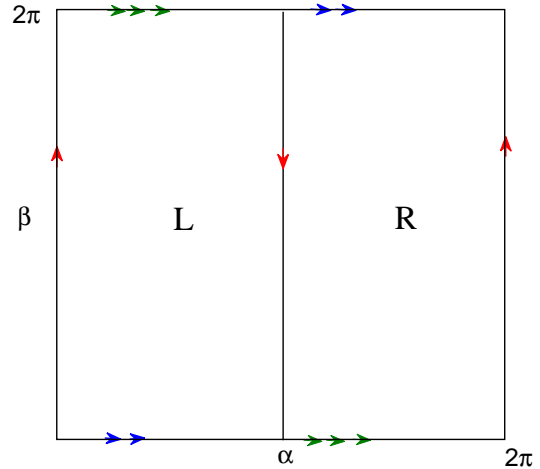


Fig. 10. Klein bottle, the image of the map g

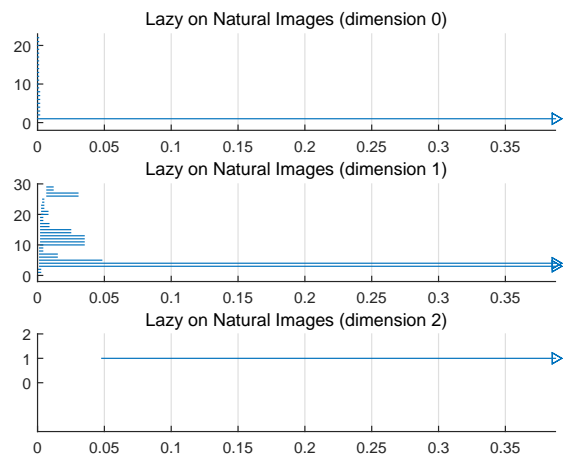


Fig. 11. PLEX result for $K_8(200)$

many experiments on $XP_8(200, 20)$ for the parameter num-landmark-points from 80 to 100 and the result is very stable. Fig.13 displays one PLEX result for $XP_8(200, 20)$, which gives Betti numbers $\beta_0 = 1, \beta_1 = 2$ and $\beta_2 = 1$ for ϵ from 0.019 to 0.059. When taking $t = 25$, the space $XP_8(200, 25)$ experience a topological change. Indeed, we do 50 trials on $XP_8(200, 25)$ for different parameters, where there exist 23 trials whose PLEX results producing the topology of the Klein bottle and most barcode intervals with the homology of the Klein bottle is in very small ranges, the other 27 trials give no the homology of the Klein bottle. Fig.14 gives the Betti numbers of $XP_8(200, 25)$: $\beta_0 = 1, \beta_1 = 2$ and $\beta_2 = 1$ for ϵ from 0.025 to 0.035. The PLEX result Fig.15 of $XP_8(200, 25)$ shows that it has no the Klein bottle's homology. Similarly, we do many experiments on $XP_9(200, 18)$ and $XP_9(200, 23)$ respectively, we discover that the largest subspace of X_9 having the homology of the Klein bottle is about $XP_9(200, 18)$, and the subspace $XP_9(200, 23)$ experiences a topological change. Fig.16 displays one PLEX result for $XP_9(200, 18)$, which gives $\beta_0 = 1, \beta_1 = 2$ and $\beta_2 = 1$ for ϵ in $[0.032, 0.086]$. Fig.17 shows $XP_9(200, 23)$ having the Klein bottle feature

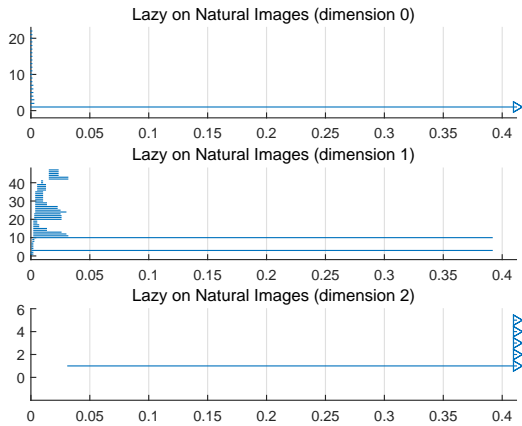


Fig. 12. PLEX result for $K_9(200)$

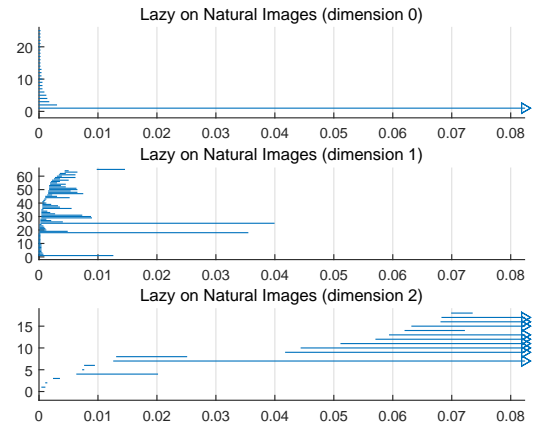


Fig. 14. PLEX result for $XP_8(200, 25)$

in a very small range of ϵ values (from 0.014 to 0.033). The PLEX result for $XP_9(200, 23)$ in Fig.18 gives no the Klein bottle feature.

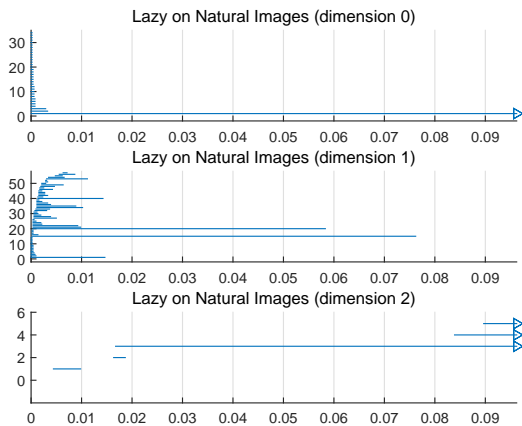


Fig. 13. PLEX result for $XP_8(200, 20)$

From the results of [5], we knew that the largest subspaces with the homology of the Klein bottle of 3×3 , 4×4 , 5×5 , 6×6 and 7×7 patches are about 40%, 35%, 30% and 25% of points of X_3 , X_4 , X_5 , X_6 and X_7 respectively. Combining the current results, we may conclude that the size of the largest subspace having the Klein bottle's homology of $n \times n$ patches depends on the patch size n , and the larger of patch size the smaller the size of the largest subspace. Hence it is necessary to discuss different sizes patches in natural images.

VII. CONCLUSION

In this paper we apply persistent homology to study natural image 8×8 and 9×9 patches, and obtain similar results to the papers [4], [5], the results of in this paper enlarge image analysis to larger patches. We find the largest subspaces of X_8 and X_9 with the Klein bottle's homology, and the size of the largest subspace of $n \times n$ natural image patches having the Klein bottle's homology is decreases as increasing of n . Thus we need only study $n \times n$ natural image patches for sufficiently

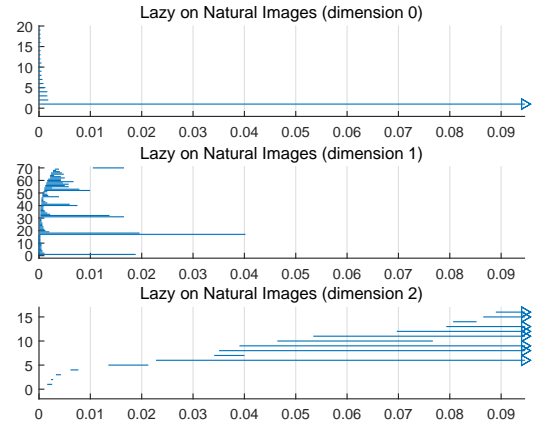


Fig. 15. PLEX result for $XP_8(200, 25)$

small n . The Klein bottle's feature of natural image patches may improve techniques of image compression [4], [6]. But it is worth to study that for how big of n , the $n \times n$ natural image patches have no the Klein bottle feature. As increasing of n , the computing for $n \times n$ patches becomes more difficult.

ACKNOWLEDGMENT

The authors are very grateful to the reviewers for their valuable comments and corrections.

The project is supported by the National Natural Science Foundation of China (Grant No.61471409).

REFERENCES

- [1] J. Huang, and D. Mumford, *Statistics of natural images and models*. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, (1999), pp. 541–547.
- [2] B.A. Olshausen, and D.J. Field, *Natural image statistics and efficient coding*. Network: Computation in Neural Systems, 7 2(1996), pp. 333–339.
- [3] A. B. Lee, K. S. Pedersen, and D. Mumford, *The non-linear statistics of high-contrast patches in natural images*. Internat. J. Computer Vision, 54 1-3(2003), pp. 83–103.

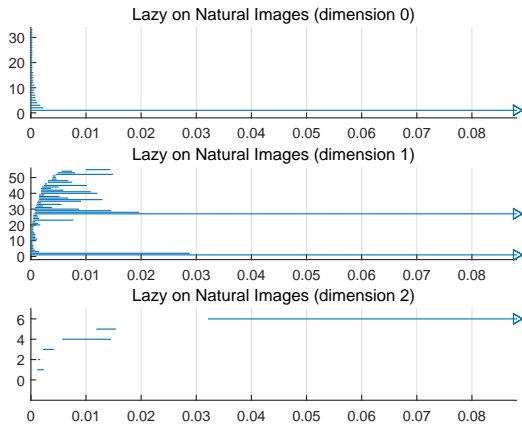


Fig. 16. PLEX result for $XP_9(200, 18)$

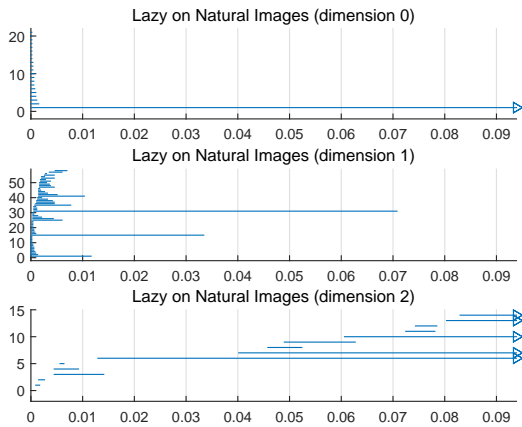


Fig. 17. PLEX result for $XP_9(200, 23)$

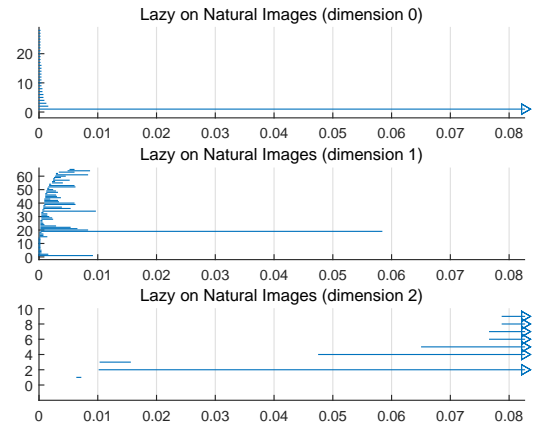


Fig. 18. PLEX result for $XP_9(200, 23)$

[4] G. Carlsson, T. Ishkhanov, V. de Silva, A. Zomorodian, *On the local behavior of spaces of natural images*. Internat. J. Computer Vision, 76 (2008), pp. 1–12.

[5] S. Xia, *A topological analysis of high-contrast patches in natural images*. J. Nonlinear Sci. Appl., 9 (2016), pp. 126–138.

[6] J. Perea and G. Carlsson *A Klein-Bottle-Based Dictionary for Texture Representation*. Internat. J. Computer Vision, 107 (2014), pp. 75–97.

[7] H. Jegou, M. Douze, and C. Schmid, *Hamming Embedding and Weak geometry consistency for large scale image search*. Proceedings of the 10th European conference on Computer vision, October, (2008), pp. 304–317.

[8] H. Adams and G. Carlsson, *On the nonlinear statistics of range image patches*. SIAM J. Imag. Sci., 2 (2009), pp. 110–117.

[9] H. Edelsbrunner, D. Letscher, and A. Zomorodian, *Topological persistence and simplification*. Discrete Comput. Geom., 28 4(2002), pp. 511–533.

[10] A. Zomorodian and G. Carlsson, *Computing Persistent Homology*. Discrete Comput. Geom., 33 (2005), pp. 249–274.

[11] V. de Silva and G. Carlsson, *Topological estimation using witness complexes*. Proc. Sympos. Point-Based Graphics, (2004), pp. 157–166.

[12] H. Adams and A. Tausz, *Javaplex tutorial*. http://javaplex.googlecode.com/svn/trunk/reports/javaplex_tutorial/javaplex_tutorial.pdf.

[13] G. Carlsson, *Topology and data*. Bulletin (New Series) of the American Mathematical Society, 46 2(2009), pp. 255–308.

Intelligent Pedestrian Detection using Optical Flow and HOG

Huma Ramzan, Bahjat Fatima, Ahmad R. Shahid, Sheikh Ziauddin and Asad Ali Safi
Department of Computer Science
COMSATS Institute of Information Technology
Park Road, Islamabad

Abstract—Pedestrian detection is an important aspect of autonomous vehicle driving as recognizing pedestrians helps in reducing accidents between the vehicles and the pedestrians. In literature, feature based approaches have been mostly used for pedestrian detection. Features from different body portions are extracted and analyzed for interpreting the presence or absence of a person in a particular region in front of car. But these approaches alone are not enough to differentiate humans from non-humans in dynamic environments, where background is continuously changing. We present an automated pedestrian detection system by finding pedestrians' motion patterns and combing them with HOG features. The proposed scheme achieved 17.7% and 14.22% average miss rate on ETHZ and Caltech datasets, respectively.

Keywords—Pedestrian detection, pedestrian protection system, HOG descriptor, optical flow, motion vectors, FPPI, miss-rate

I. INTRODUCTION

According to National Highway Traffic Safety Administration (NHTSA), traffic fatality rate has been increased by 6% in 2012 and on an average nearly 4,743 pedestrians were killed which accounted for 14% of the total traffic related fatalities along with 76, 000 ended up injured in USA [1]. In countries of Asia and Europe due to high population, the rate of road user deaths is much higher. This rate of pedestrian deaths and injuries could be reduced by employing intelligent frameworks for detecting people on road. But to give viable safety, such frameworks need to recognize people on foot in changing ecological conditions, as well as anticipate the probability of impact. Road users particularly pedestrians are more susceptible to serious injuries in contrast with drivers in such collisions. Pedestrian Safety has therefore gain the attention of many researchers now a days. Safety components are intended to avoid the impacts and resulting casualties and injuries by offering advancements that alert the driver to potential issues before time. These safety advancements might lit up the car light, automatic braking, consolidate GPS traffic flow notifications, interface with cell phones, alert driver about other cars and dangers, keep the driver in the right lane, or show what is in blind corners. For effective road user protection, frameworks like Advanced Driver Assistance System (ADAS) and Intelligent Vehicles (IV) are built up that lessened the auto collisions by giving knowledge to drivers [2][3]. Pedestrian recognition is a critical and all the more difficult problem in the field of machine vision. The essential target of road user detection based vision frameworks is to avoid impact of vehicles with people while driving.

Although, several pedestrian detection algorithms have

been proposed so far, but still there is a need for an automatic system that can detect the human in urban environments that are more challenging as compared to highway traffic. Therefore, the need is to develop a system capable of discriminating humans from a captured image or video with high accuracy, precision, and low miss rate. In this paper, we have presented a motion vector based pedestrian detection system. We will discuss and analyze the experimental results of our proposed technique on two publicly available ETHZ and Caltech pedestrian benchmarks. Novel contributions of this work include:

- Development of a representation of human motion which is extremely efficient for detecting people.
- Implementation of proposed technique on two pedestrian benchmarks.
- Accurate human detection with smaller FPPI (false positive rate per image).
- A system for automatic pedestrian detection with Low average miss rate.

The rest of the paper is organized as follows: A comprehensive survey of related schemes can be found in Section II., Section III presents the proposed methodology employed for effective pedestrian detection, Section IV includes performance evaluation of our technique and Section V concludes the paper.

II. LITERATURE REVIEW

In spite of advancements in pedestrian protection, numerous road accidents still happen all around the globe because of poor driving conditions (e.g. low light or mist) or a transient diversion of either the driver or pedestrian. A programmed framework to identify people in the surroundings of a vehicle is exceedingly desirable and is one of the fundamental concerns for both auto makers and researchers today. Related literature for pedestrian detection and recognition framework are reviewed from both the application and also, from hardware point of view.

From the application's point of view, pedestrians location can be utilized by intelligent vehicles and for surveillance videos and aided driving systems [4][5]. From technology perspective, detection and recognition of on foot people in front of the vehicle can be done by utilizing the perceptible and imperceptible light range, for example, visible, ultrasonic, infrared, sensors and radars [6][7].

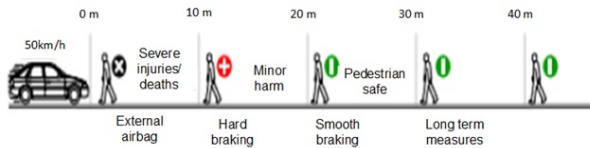


Fig. 1. PPS driving scenario [8]



Fig. 2. Difference in pedestrian appearances [9]

A. Pedestrian Protection Systems (PPS)

PPS are a special kind of intelligent frameworks concerned to pedestrian safety. It is a safety framework that typically recognizes moving as well as stationary pedestrians in front of the vehicle in order to perform braking actions or to provide knowledge to driver [2]. Any decline in the speed of the vehicle can intensely lower the fatality rate due to reduced kinetic energy of the approaching vehicle in case of impact. Pedestrians are having a 90% likelihood of lasting the accident due to vehicle impact coming at small speeds like 30km/h or below, however if the vehicle is approaching at the speed of 45km/h or above then in such cases there is less than half likelihood of surviving [2] as shown in Fig. 1.

B. Challenges to the Task of Pedestrian Recognition

These are the principle elements and difficulties that influence the performance of pedestrian detection frameworks:

1) *Appearance changeability*: Pedestrian constantly changes in appearances because of the color, shading, texture, size of the garments, and they move with different items and articles (boxes, bags, umbrellas) as appeared in Fig. 2.

2) *Inconsistency of the surroundings in which people appear*: Different type of environments include urban and congested city zones which are more complex to handle as compared to highways under different climate conditions and changes in light add irregularity in the information. Fig. 3 indicates distinctive surrounding environmental conditions.

3) *Variability in pedestrian shapes and postures*: Pedestrians may have diverse weights, postures, poses and tallness figures. Fig.4 shows such differences.

4) *Variability of the activities*: Different kind of actions which they may perform positions which they may have (stand,



Fig. 3. Inconsistency of the surroundings in which people on foot appear [2]



Fig. 4. Variable Body Postures [2]

run, walk, sit, and hands shake etc.) Besides, they can show up under different observing angles (longitudinal or sidelong positions).

5) *Motion of camera and Road user*: When both (walker and camera) are in moving condition, this marks detection and tracking more troublesome. The vast majority of the frameworks are focused on the high-hazard zone i.e. separation from 5m-25m to the camera [9]. However, 50m identification range speaks to a generally safe area that demonstrates an extraordinary support for PPSs in the long period crashes.

The problem of pedestrian detection has been approached both from hardware and software perspective, utilizing different sensors and creating numerous different algorithms for detection.

C. Hardware Based Solutions

Finding which sensor is most appropriate for pedestrian identification and detection is a major question especially for cluttered urban traffic environments. Today the most appealing sensors include Radar and Laser. Sensors that use visible light are efficient during day but do not work at night. In addition, there is an emerging attention towards infrared frameworks that are guaranteed to be extremely useful being less expensive as compared to other sensing mediums like Laser, Lidars and Radars and are not affected by the weather conditions or time of the day. From the hardware aspect, there are passive or active sensors [10].

Passive sensors capture light by using scan chips like CMOS or CCD cameras and can perform in infrared or visible spectra, consisting of all camera-based systems. Active sensors that recently have provided the good results are the laser sensors, particularly LIDARs, but these are too expensive for intelligent systems [11][12]. For that reason, mostly passive sensors are used in intelligent vehicles to detect pedestrians because they are cheaper and provide useful information from different clues like color and texture information [2].

D. Software Based Machine vision for Pedestrian Detection

Based on humans owned knowledge, vision-based pedestrian detection is a preferred choice. It can acquire much richer

information about the environment than laser scanner or radar [7]. For this purpose, various types of cameras have been used for the detection of pedestrians. Camera may be either still or moving that is installed on the vehicle. According to their working range, cameras can be divided in the electromagnetic spectrum. The range of visible spectrum (VS) is in 0.4-0.74 (μm), near infrared is in 0.75-1.4 (μm) and thermal infrared (TIR) covers 6-15 (μm) [6], [7], [13]. Visible cameras are more commonly used because pedestrian detection is mainly focused on day time as compared to thermal infrared cameras that are used on night time [8].

The problem of pedestrian detection is to determine whether a local image area/section represents people or not and thus it is a typical two class classification problem. The detection process can be applied in two steps: feature extraction and classification. In the learning-based discriminative framework, various feature descriptors and classification approaches have been proposed for use in visible images [2], [3]. Some of the computer vision based approaches for pedestrian detection are:

1) *Feature Based Detection*: Detectors are developed that use feature information by extracting gradient features from the image to detect pedestrian in front of the vehicle within an unsafe range. The two popular gradient based feature detection methods are Covariance matrix (COV) and HOG descriptors [4], [14], [15], [16].

Such local gradient feature based detector was first presented by Dalal and Triggs [17]. They represented a human in the image as a thick web of Histogram of oriented Gradients then these feature maps were given to an SVM classifier to detect human. The image is first segmented into equal sized 16x16 blocks and each block further consists of 8x8 cells. These cells combined together to form a dense grid of with each cell representing edge orientation. Orientation gradients for each individual cell is quantized into 9 equal sized bins defined for 0-180 unsigned orientation each comprising of 20. The combined histograms represent a feature vector of normalized 1-D histograms from each block it is then given to an SVM classifier which classifies it as either a pedestrian or not. The dense overlapping grids, normalized histograms and non-smoothing gradients made HOG descriptor a better detector by decreasing the count of false positives. However HOG-feature extraction performs poorly for the cluttered images where it becomes difficult to detect images to create histograms.

Costea et al. [4] proposed a novel method for image recognition using word channels because of their high discriminating power. This approach uses image with a single size and a single classifier for each pedestrian sliding window scale. For extracting features this approach uses high level word channels inspired from codebook based techniques instead of low level pixel gradients used in HOG techniques. At each pixel level three descriptors HOG, LRP and LUV color channel values are computed. The authors use LUV color channel as the descriptor for computing feature vector. The computed results are matched to the visual codebooks. Three different word maps are generated with multiple word channels one channel per descriptor. Approach was found to have promising results on both INRIA and Caltech pedestrian datasets.

Richer image representation provides a good chance for improving the detector performance in image analysis. But

for richer image representations detectors require more computational time to process an image that is the improvement in recognition comes at the cost of more computational time. Dollar et al. [18] proposed that for richer image representation such finely sampled pyramids could be obtained by estimation. Experiments indicate that by extrapolating the features for coarsely sampled pyramids, we could estimate the features at any given scale and hence get a rich representation of the image inexpensively. The effectiveness of the proposed scheme is demonstrated with different detection architectures including integral channel features, aggregate channel features and deformable part models. Results indicate that the proposed method has the same detection rate as the current state-of-the-art but has decreased computational cost. However, this strategy could not be applied in cases where the image contains texture or white noise.

Dollar et al. [19] proposed a multi scale pedestrian detection technique. This technique uses feature approximation for the features at nearby scales by computing the feature at one scale. This approximation is shown accurate within entire scale octave. Algorithm thus computes features in the image once every half octaves and approximate features for the rest of scales resulting an overall speed up in the detection process with a little loss in detection accuracy.

Dollar et al. [20] proposed a novel method for pedestrian detection using Integral channel features. Using linear and non-linear transformations multiple image channels are computed. Features from these multiple channels are computed by summing the local rectangles. Using integral images features such as Haar wavelets, local histograms are computed. A feature is defined as the weighted sum of the integral channels. Much time of this algorithm is spent on constructing these channel features, making it a fast detector. The authors combined Histogram of Oriented gradients with LUV color channels coupled with a boosting classifier. Integral channel features when combined with a boosting classifier are proved as a fastest object detector. On Cal Tech dataset the detection rate of integral channel features was found to be 60% while that of its competent HOG is 50%. Results indicate that channel filters outperform other feature extraction techniques including HOG.

[21], [22], [23], [24] are the other feature based methods proposed by the researchers for detecting pedestrian in order to avoid or anticipate the likely hood of collision.

2) *Texture Based Detection*: Various texture based approaches like Local binary patterns (LBP) [25], [26] also provide good outcomes to the Pedestrian detection. The LBP advantages are its less computational complexity and multi-scaling [25]. Multi-block local binary patterns (MBLBP) [27], have been presented as efficient applications of traditional LBP for detection. Zhang et al. [28] propose that for accurate and fast detection of a human in an image scene or a video sequence, a robust detector is needed which can compute promising features in least computational time. They presented a set of effective features that can be computed easily and are robust to external noise. This set includes dense center-symmetric local binary patterns (CS-LBP) which captures the gradient information combined with texture details and pyramid center-symmetric local binary/ternary patterns (CS-LBP/LTP) which is more descriptive and computationally

efficient for real life applications. Experiments on INRIA pedestrian dataset indicate that the proposed features of CS-LBP when coupled with linear SVM give comparable results as HOG/SVM and pyramid CS-LBP when coupled with HIKSVMS outperforms the previous PHOG. The authors also suggest that by combining pyramid C-LBP with PHOG feature, detection performance could further be improved.

For the detection of a pedestrian in a still image Jiu Xu et al. [29] proposes a novel feature named Bidirectional Local Template Patterns (B-LTP). B-LTP is inherited from CS-LBP and HOT and thus combines their desirable properties. It takes texture properties from CS-LBP feature and gradient based properties from HOT. Moreover, B-LTP is a short length feature and thus it is cheaper to implement as well as cost less memory making it suitable for real time applications. This technique proposed a two directional template in which for each pixel, four templates are defined containing the pixel itself and its two center-symmetric neighbors. Results on INRIA dataset shows that B-LTP performs better than its competent features like HOG, HOT and COV in both speed and detection rate.

3) *Deep Learning for effective pedestrian Detection:* In the recent past, Deep learning has been applied to the domain of pedestrian detection problem which learns features in a supervised or unsupervised fashion and has shown very promising results. The input data moves from lower layers and is gradually transformed into higher level representations. The output features from the top layer is then given to classifier and the network is fine tuned with back propagation algorithm.

In his paper, Luo et al. [30] proposed a deep learning architecture named "Switchable Deep Network (SDN)" for pedestrian detection. His work focuses on using deep networks to model hierarchical features, stressed locations from multiple feature maps called saliency map and a mixed representation of body parts. SDN is an extension of the traditional convolutional neural network with the addition of multiple switchable layers. In order to model the complex visual postures the paper introduces a Switchable Restricted Boltzmann Machine (SRBM) that explicitly develops saliency maps at each level indicating if the pixel belongs to the background or a pedestrian and hence suppressing background clutters from discriminative regions containing pedestrian. Results indicate the state-of-the-art performance on the public pedestrian detection datasets.

4) *Template Based Detection:* This technique is used to find the features in a particular image region and then compare it with a standard template. Image is scanned to find a set of features representing pedestrian and then are compared with a template image. But these methods fail to handle articulations and occlusions in the scene and are computationally expensive [31]. The disadvantage of template based detection is that object occlusion is difficult to compute and high computational complexity.

5) *Deformable part based Detection:* In his paper, Yan et al. [32] presented an extension of the prevalent deformable part model [33] (DPM) called Multi task DPM (MT-DPM) which aims to explore the relation among multiple resolutions by combining an optimal DPM detector and resolution aware transformations. It takes the pedestrians from multiple resolutions and determines their commonness and differences

jointly. To map the pedestrians from different resolutions the model transforms them into a common space where a detector separates the pedestrians from the background. The global spatial assembly for example part configurations remains the same while the differences exist in the local features. The differences among these local features are reduced by mapping them from multiple resolutions to a common subspace and a detector is learned on these locally mapped features. The authors further develops a context model depending on the vehicle-pedestrian relation to improve pedestrian detection by reducing the false positive rates especially in crowded scenes. Results indicate a reduction of miss rate to 60% for Caltech dataset which outperforms the recent state of the art.

6) *Infrared Thermal Imaging for Pedestrian Detection:* Effective pedestrian detection and tracking algorithms in visible spectrum have found many important applications from video surveillance to intelligent vehicles. However, under certain circumstances (e.g., in nights or bad weathers), sensing in visible spectrum becomes infeasible or severely impaired, which calls for the imaging modalities beyond visible spectrum. In particular, the cost of thermal sensors has reduced dramatically in the past decades.

Infrared imaging is here to rescue the environments in which we have little or no light. Dai et al. [11] proposed a Generalized expected maximization (EM) algorithm using IR imagery. The image is first segmented into a layered structure consisting of foreground and background layer. In the second pass using the shape and appearance details a pedestrian is traced from the foreground layer. Shape based classification is performed by SVM and appearance based localization is done through principal component analysis(PCA). Similarly for a video , the sequence is first divided into segments called shots and pedestrian detection through EM algorithm is then applied within each shot. The pedestrians present in the same shot are identified through a graph matching technique. The algorithm performs quiet well in case of crowded scenes and does not require prior assumptions about the motion trajectory. Experimental results showed the overall accuracy of 88%.

7) *Machine Learning Based Methods:* The accuracy of the detection system usually comes at the cost of high false positive rate. In order to reduce this false detection rate Z. Wang et al. [34] presented a two stage machine learning algorithms based approach for efficient and accurate pedestrian detection. This approach is based on highly efficient combination of cascade AdaBoost detector and vector function link net derived from machine learning domain. By using multi-scale sliding window detectors; all sub windows extracted from a still image and are normalized and resized. Then the two detectors cascade AdaBoost [35] deetctor and random vector functional-link net [36][37] are applied simultaneously on this candidate feature set to check if it is a pedestrian or not. Experiments with four datasets have shown that this technique outperforms other methods in terms of detection accuracy and false positive rates.

Behera et al. [38] presented a real time vision based image segmentation algorithm for accurate pedestrian detection in day time. The image is scanned in all directions for finding the edges. Before segmentation, the edges are first linked by an edge linking algorithm. The correctness of the edge map is required for accurate segmentation. After segmentation the image is divided into foreground and background segments. In

order to boost the probability to find the presence of pedestrian, a combination of head and leg edges are used. Using these head and leg edge patterns the whole pedestrian image is reconstructed. Accuracy could be further improved by applying a classifier on the extracted segments. Results indicate that this algorithm performs well on real images for accurate pedestrian detection. In order to improve the accuracy of a pedestrian detector much work has already been done in the current state-of-the-art, the paper by Smedt et al. [39] in this regard presented a generic framework to combine multiple pedestrian detectors in an optimal and efficient manner. Each pedestrian detection technique uses a different set of candidate features. Highly accurate results could be achieved by an intelligent combination of these features from multiple detectors. The authors used the simple AND OR combinations of multiple detectors and by using performance measures determine the best combination which has the highest yield in terms of detection accuracy. However combining multiple detectors; results into long computational time but results showed that an improved accuracy is obtained by hiring this optimal combination approach as compared to the current state-of-the-art detection methods.

Image feature description can be improved significantly by using HOG features based on variant scale blocks. This idea was presented by Hoang et al. [40] who suggested that without restricting HOG blocks a comprehensive feature space is obtained with the help of which highly distinguished features can be obtained for classification in the next step. Image is first segmented into grid windows and affine transformations from each window are obtained after which optical flow from each transformed window are extracted. After morphological processing, correlated features of human shape are obtained as candidate regions within each window. In the next step HOG features from each segmented window are obtained in order to detect a human by using SVM classification. Experiments showed that the proposed detector gives 5% improved results as compared to standard HOG using SVM.

Hongyan Li [41] proposed a new method of segmentation and detection of objects in a video. Their algorithm uses mean, variance and standard deviation as features calculated from gray scale multi-frame images. These features are used to train SVM that views the categorization of pixels as a binary classification task. Trained SVM classifies a pixel given to it as either a static background pixel or a foreground moving object pixel. Accuracy of this SVM classifier could be significantly improved by customizing its kernel function and other parameter values.

8) *Motion Based Detection:* Use of the object motion provides extra information but the change of position make the process of detection more complex and problematic [31]. One of the popular method in this regard is Background subtraction in which the image is segmented into foreground and background layers but this is only possible in the cases in which video is captured by a fixed camera. However, this has an apparent disadvantage for pedestrian detection from the automobile because the moving vehicle provides a continually dynamic background. Therefore, motion-based pedestrian detection algorithms do not work as a primary detection approach for a camera that is mounted on a moving vehicle [9]. Sparse scene flow information is used to detect

objects using stereo cameras and optical flow information in. [42] has extracted interest points from consecutive video frames; using flow information the complete scene flow is constructed to model the movement of the background. Scene elements whose motion pattern varies from this background flow model are considered as distinct objects. Interest points belonging to adjacent segments represent a single rigid object. The proposed method employs a class independent approach for object detection using stereo cameras and optic flow. Experiments indicate that the proposed method outperforms the previously known techniques that only use optical flow information. Moreover, these solutions can work only for those classes on which the detector was trained during training phase.

Hariyono et al. [43] presented a novel method for moving pedestrian detection through moving camera using motion information and HOG features. After segmenting the regions that represent same motion vectors different moving objects are extracted. In order to differentiate a pedestrian from non-pedestrian HOG features are extracted from candidate segments. This feature vector is given as an input to the Linear SVM classifier that classifies the given segment in an image or video frame as pedestrian or non-pedestrian. Experiments reveal an outstanding performance on ETHZ pedestrian dataset as compared to the original HOG approach. Detection rate obtained was 99.3% with 0.09 false positive rate. Another method is CodeBook method [31]. A codebook collects a series of code words or color values for every pixel of background. After that, codewords will found out what color each pixel have and from that it determine the pixels of background. Advantage of this method is that it can handle dynamic scenes [31].

From the cited literature, the issues pertaining to the real-time pedestrian detection problem are as follows:

- Occluded and continuously changing backgrounds put a limit on the detection rate.
- Feature extraction techniques alone perform poorly for the cluttered images.
- Deep learning based methods have long learning and detection times.
- Template based approaches are affected by occluded environments.
- Relative motion between camera and pedestrian marks detection and tracking more troublesome.

A system with the following properties is therefore desired:

- Detect pedestrian in real time accurately and reliably with lower false positives.
- Detect people regardless of the differences in their postures and appearances.
- Improved system accuracy with small miss rates.
- Detect pedestrians in all type of environmental conditions.
- Detection of people even if there is relative motion between pedestrian and camera.

This research will combine the motion vectors method with HOG feature extraction technique for improved detection results in images and videos. Datasets used are ETHZ pedestrian dataset and Caltech pedestrian dataset.

III. PROPOSED METHODOLOGY

Motion is a natural property of the world and is a rich foundation of data that supports a wide range of visual responsibilities. Identification of moving objects in videos is vital for numerous computer vision based applications, including action acknowledgment, activity recognition, and car safety. The issue of motion based object detection can be separated into two sections:

- 1) Identifying moving subjects in each video frame.
- 2) Associating the recognitions relating to the same subject over time.

A. Work Flow

The proposed methodology to detect the pedestrian in videos as indicated in systems block diagram below in Fig. 5 is composed of the following sequences of steps.

1) *Acquire images and videos:* The first step to create a successful system for pedestrian detection is the selection of right dataset. There are many standard publicly available datasets for solving the problem of moving person detection. We have used ETHZ and Caltech datasets for our research. The challenging part of this research is to compensate the motion of both camera and pedestrian while reducing the miss rate in order to achieve accurate detections with small false positives.

2) *Compute motion vectors to extract moving objects in a scene :* In order to make the task of detection easy we have opt for detection through motion vectors because it separates the dynamic objects in the foreground from static background in the very first step. Here in our research, first of all we have computed motion vectors from the each video frame and by associating these detections same object over time is identified. This gives a set of bounding boxes around each of the moving object in the scene as the output.

3) *Extract features from the computed flow vector:* Feature extraction is the process of selecting relevant attributes in the data that are used for the construction of the model. In the next step, by computing HOG features from each bounding box we are able to learn a single feature vector which is the combination of motion information from the previous step and HOG features. This feature set is given to the classifier in the next step that will learn the features of Pedestrians with the help of the input feature vector.

4) *Feed this feature set to a classification module:* This feature set is given to SVM/AdaBoost classifier one by one to train the classifier and learn the features of a human so as to discriminate human from non-human objects in the test videos.

5) *Detect pedestrians in the test set:* Detection is performed on test videos and performance of the detector for each classifier in terms of accuracy and average miss rate is computed and compared with state-of-the-art methods of pedestrian detection.

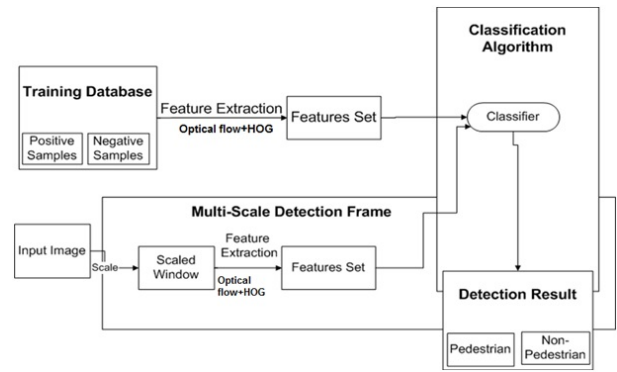


Fig. 5. Block diagram of proposed system



Fig. 6. Images from ETHZ and Caltech Datasets

IV. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

We ran the experiments on Intel core i3 with 4GB of memory using MATLAB Release R2014 b as integrated development environment and runtime platform. We have evaluated our proposed methodology on two benchmark pedestrian datasets namely ETHZ and Caltech pedestrian benchmarks.

ETHZ is a popular and most commonly used pedestrian dataset. It consists of videos made from AVT Marlins F033C camera, it contains video frames of the size 640x480 each with a frame rate of approximately thirteen to fourteen frames per second. The dataset contains three setups consisting of three set of videos each. Second dataset is the Caltech which is the largest dataset amongst other datasets. It consists of 10 hours of 640 x 480 videos in an urban environment. These videos are taken from a CCD camera mounted on a vehicle. All these datasets are publicly available and some of the images of these datasets are shown in Fig. 6.

A. Evaluation Metrics

Evaluation of our work is done based upon the following evaluation metrics. The terms are defined as follows:

We trained our detector by using positive and negative samples once from ETHZ and next time with samples from Caltech training sets. First the training of SVM is performed and a model is created, one for each dataset. Similarly, the procedure is repeated to create AdaBoost training models. For the rest of our experiments, we test our pedestrian detectors on the reasonable subset of videos from both ETHZ and Caltech test sets.

TABLE I. METRICS FOR EVALUATING CLASSIFIERS PERFORMANCE

Metrics	Definition/Formula
Miss Rate	= 1- Recall = No. of FP in positive samples/ Total no. of positive samples
FPPI	= False positives/ Total images
Detection Rate	= TPR= Number of correctly classified pedestrians
Precision	= Fraction of true positives/ All positives
F1 Measure	= $2 * TP / (2*TP + FP + FN)$
False Detection Rate	= 1- Precision

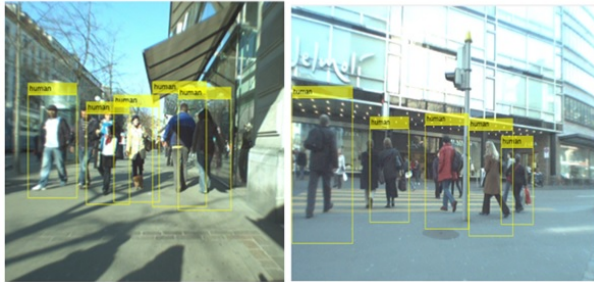


Fig. 7. Detectors Detection Results

B. Testing the Detector Performance

When training process is completed, the next step is to test the performance of classifiers by giving any unknown images to the detector. The input image can be either a single image or a cell array of images. For detection, we slide a window over the whole image and consider the multiple window strides. For each video from the test set the experiment is performed with 8x8 (default) and 6x6 window stride. Detector uses this measure to slide the window over the image, smaller strides produce better detection rate. After this, the detector returns different bounding boxes in the form of [x, y, w, h, score]. Confidence of detection is measured by its score. Only detections with the score above threshold are returned. A few detection results are shown below in Fig. 7:

Confidence is high for the pedestrians in front of the camera or within a suitable range. Those already standing in the safe range are given small scores by the detector as shown in Fig 8.

C. Experimental Results

1) *Detection:* Detection results on both datasets for each classifier are shown below. Detector creates a bounding box



Fig. 8. Testing the detector (a)Input test image (b)Detection scores

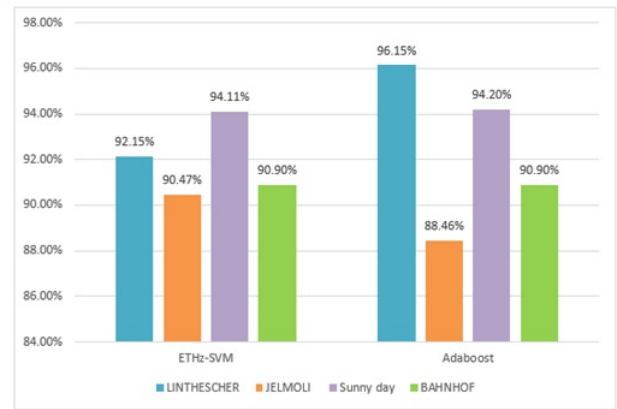


Fig. 9. Detection Results on ETHZ Pedestrian benchmark

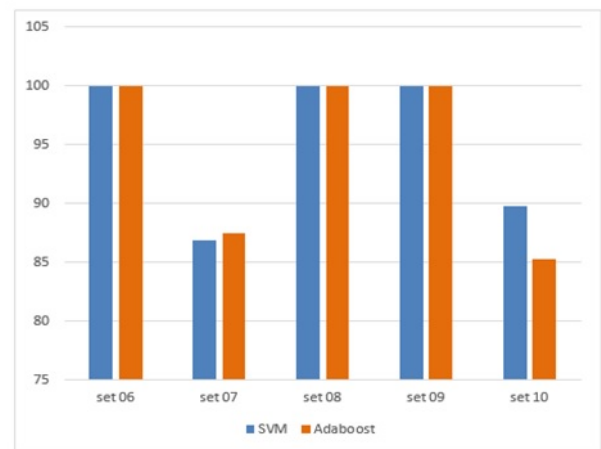


Fig. 10. Detection Results on Caltech Pedestrian benchmark

around a human. HOG and flow features from these bounding boxes are computed. Detector marks each bounding box with a score that represent the detection confidence. Confidence is high for the pedestrians in front of the camera or within a suitable range. Those already standing in the safe range are given small scores by the detector.

Detection results for each classifier on ETHZ benchmark are shown in Fig. 9. On average we have achieved 91.82% detection accuracy with SVM while AdaBoost provided 92.96% accuracy.

Detection results for each classifier on Caltech benchmark are shown in Fig. 10. On average we have achieved SVM provided 95.33% and AdaBoost provided 94.5% accuracy.

2) *Evaluation of Results:* In the second phase the detector performance is evaluated for both datasets by comparing the number of detections with the ground truth annotation which is provided with each test video. These annotation files contain information in the form of bounding boxes coordinates for each person present in the scene for each frame. However, for ETHZ, ground truth annotation is provided for every 4th frame in the video. Miss rate and the detection rates are computed here with the possible number of false positives per image and overall accuracy, recall and precision are computed for both ETHZ and Caltech datasets.

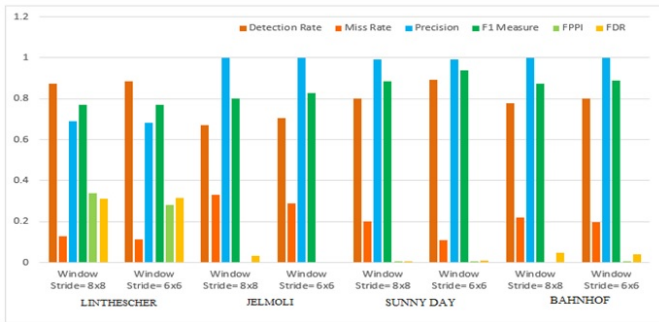


Fig. 11. Evaluation Results on ETHZ Pedestrian benchmark

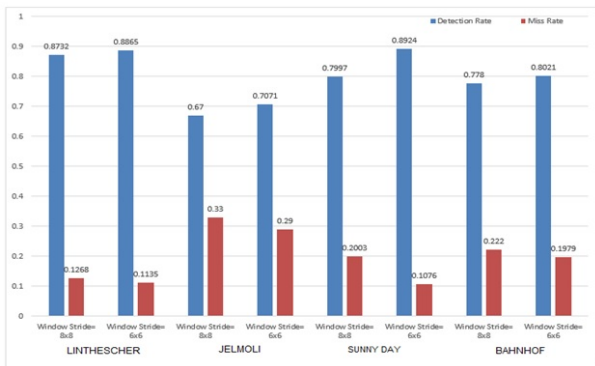


Fig. 12. Detection Rate VS Miss Rate on ETHZ Dataset

valuation with respect to the ground truth data is performed and results of evaluation are shown in Fig. 11. We have achieved an average miss rate of 17.7% on ETHZ Test set. Miss rate is calculated for each video with respect to detection rate as shown in Fig. 12.

Results of evaluation on Caltech are shown in Fig. 13. We have achieved an average miss rate of 14.22% on Caltech Test set. Miss rate is calculated for each video with respect to detection rate for each test set from Caltech benchmark as shown in Fig. 14.

V. PERFORMANCE COMPARISON

In our proposed work we have developed a pedestrian detection system by combining two features that is HOG and motion vectors. The two features are concatenated to get a

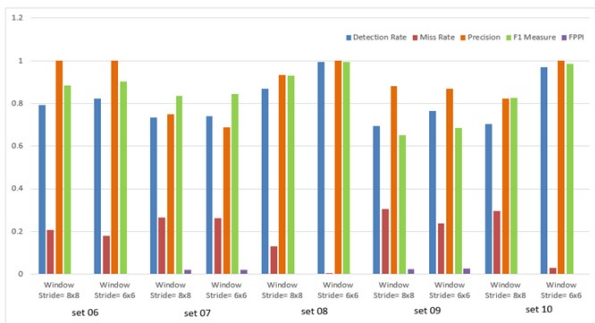


Fig. 13. Evaluation Results on Caltech Pedestrian benchmark

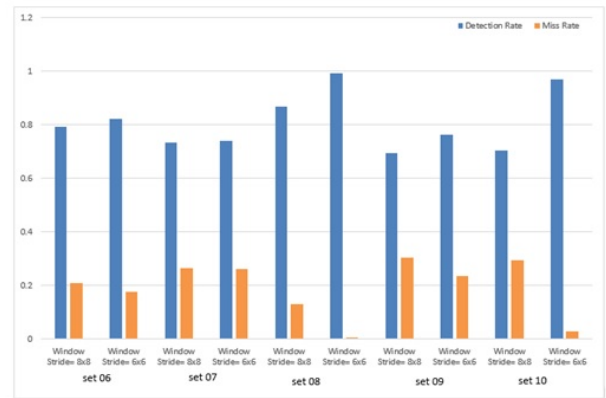


Fig. 14. Detection Rate VS Miss Rate on Caltech Dataset

single feature vector for the training of the classifier in the next step. Two classification models created during training phase for each classifier (SVM and AdaBoost) are used to discriminate human from non-human in the test videos and the detection results and accuracy is computed. Average miss rate computed from our proposed methodology is compared with those present in existing literature as indicated in Table 4.2. Our detector yields significant performance improvement as compared to the baseline HOG detector.

VI. CONCLUSION

The problem of efficient pedestrian detection is studied in this research. This work has presented a novel technique for underlying problem of pedestrian detection by incorporation of motion information with feature extraction technique of HOG. Based on our experiments, we observe that the performance of pedestrian detection yields significant improvement with the use of motion vectors. Furthermore, implementation parameters also play an important role to achieve the best detection performance. We have achieved 17.7% miss detections on ETHZ and 14.22% miss detections on Caltech with a window stride of 6x6. Our future research objective include pedestrian tracking over multiple frames by using the optical flow based motion estimation. Currently, we have used datasets made in different lightening conditions during day time only and the pedestrian detection at night time is not included in the current scope of work. In future, we will also include the datasets that contain videos made at night. Moreover, the proposed scheme can also be extended to work in different weather conditions including rain, snow etc. Similarly, by combining more features with currently used features can help further improve the detector performance.

REFERENCES

- [1] National Highway Transportation Safety Administration, "TRAFFIC SAFETY FACTS 2012 Data," <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812018>, 2014, online; accessed 27 July 2016.
- [2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.
- [3] Y. Yang, R. Mingwu, and Y. Jingyu, "Obstacles and pedestrian detection on a moving vehicle," *International Journal of Advanced Robotic Systems*, vol. 11, 2014.

TABLE II. PERFORMANCE COMPARISON WITH EXISTING LITERATURE

	Paper	Year	Techniques used	Average Miss Rate		
				INRIA	CalTech	ETHZ
1	Dalal [17]	2005	HOG, SVM	46%	68.46%	–
2	Dollár [20]	2009	HOG + LUV Channels, AdaBoost	22.18%	56.34%	–
3	Wang [26]	2009	HOG + LBP, SVM	39.10%	67.77%	–
4	Dollar [19]	2010	Channels, AdaBoost	21%	–	–
5	Yan [34]	2013	MT-DPM, HOG, Latent SVM	–	37.64%	–
6	Benenson [35]	2013	Channels, SVM	13.53%	48.35%	–
7	J. J. Lim [36]	2013	Channels, Random Forest	13.32%	–	–
8	P. Luo [30]	2014	SDN, HOG, SVM	–	37.87%	–
9	Zhang [24]	2014	HoG + LUV, AdaBoost	14.30%	34.60%	–
10	Katamari [37]	2014	Square Chns + DCT + Optical Flow, AdaBoost	–	22%	–
11	Proposed Scheme		HOG + Optical Flow, SVM, AdaBoost	–	14.22%	17.70%

- [4] A. Daniel Costea and S. Nedeveschi, "Word channel based multiscale pedestrian detection without image resizing and using only one classifier," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2393–2400.
- [5] Y. Wei, Q. Tian, and T. Guo, "An improved pedestrian detection algorithm integrating haar-like features and hog descriptors," *Advances in Mechanical Engineering*, vol. 5, p. 546206, 2013.
- [6] C.-Y. Chan and F. Bu, "Vehicle-infrastructure integrated approach for pedestrian detection: feasibility study based on experimental transit vehicle platforms," in *Transportation Research Board 85th Annual Meeting*, no. 06-0118, 2006.
- [7] C.-Y. Chan, F. Bu, and S. Shladover, *Experimental vehicle platform for pedestrian detection*. California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2006.
- [8] D. G. Gomez, *A global approach to vision-based pedestrian detection for advanced driver assistance systems*. Universitat Autònoma de Barcelona, 2010.
- [9] D. Gerónimo and A. M. López, *Vision-based pedestrian protection systems for intelligent vehicles*. Springer, 2014.
- [10] A. Miron, A. Bensrhair, R. I. Fedriga, and A. Broggi, "Swir images evaluation for pedestrian detection in clear visibility conditions," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 2013, pp. 354–359.
- [11] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 288–299, 2007.
- [12] C. Premebida, O. Ludwig, and U. Nunes, "Lidar and vision-based pedestrian detection system," *Journal of Field Robotics*, vol. 26, no. 9, pp. 696–711, 2009.
- [13] A. Bartsch, F. Fitzek, and R. Rasshofer, "Pedestrian recognition using automotive radar sensors," *Advances in Radio Science*, vol. 10, no. B. 2, pp. 45–55, 2012.
- [14] B. Besbes, A. Rogozan, A.-M. Rus, A. Bensrhair, and A. Broggi, "Pedestrian detection in far-infrared daytime images using a hierarchical codebook of surf," *Sensors*, vol. 15, no. 4, pp. 8570–8594, 2015.
- [15] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*. Springer, 2006, pp. 428–441.
- [16] R. Ranjan, V. M. Patel, and R. Chellappa, "A deep pyramid deformable part model for face detection," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 2015, pp. 1–8.
- [17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.
- [18] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.
- [19] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *BMVC*, vol. 2, no. 3. Citeseer, 2010, p. 7.
- [20] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.
- [21] S. Wang, Z. Han, L. Zhu, and Q. Chen, "A novel approach to design the fast pedestrian detection for video surveillance system," *International Journal of Security and Its Applications*, vol. 8, no. 1, pp. 93–102, 2014.
- [22] C.-H. Chuang, S.-S. Huang, L.-C. Fu, and P.-Y. Hsiao, "Monocular multi-human detection using augmented histograms of oriented gradients," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.
- [23] W. Xing, Y. Zhao, R. Cheng, J. Xu, S. Lv, and X. Wang, "Fast pedestrian detection based on haar pre-detection," *International Journal of Computer and Communication Engineering*, vol. 1, no. 3, p. 207, 2012.
- [24] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 947–954.
- [25] S. Yang, X. Liao, and U. Borasy, "A pedestrian detection method based on the hog-lbp feature and gentle adaboost," *International Journal of Advancements in Computing Technology*, vol. 4, no. 19, 2012.
- [26] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 32–39.
- [27] A. Halidou, X. You, and B. Bogno, "Pedestrian detection based on multi-block local binary pattern and biologically inspired feature," *Computer and Information Science*, vol. 7, no. 1, p. 125, 2014.
- [28] Y. Zheng, C. Shen, R. Hartley, and X. Huang, "Effective pedestrian

- detection using center-symmetric local binary/trinary patterns,” *arXiv preprint arXiv:1009.0892*, 2010.
- [29] J. Xu, N. Jiang, and S. Goto, “Pedestrian detection based on bidirectional local template patterns,” in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 400–404.
- [30] P. Luo, Y. Tian, X. Wang, and X. Tang, “Switchable deep network for pedestrian detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 899–906.
- [31] V. Bandarupalli, “Evaluation of video based pedestrian and vehicle detection algorithms,” 2010.
- [32] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, “Robust multi-resolution pedestrian detection in traffic scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3033–3040.
- [33] H. Cho, P. E. Rybski, A. Bar-Hillel, and W. Zhang, “Real-time pedestrian detection with deformable part models,” in *Intelligent Vehicles Symposium (IV), 2012 IEEE*. IEEE, 2012, pp. 1035–1042.
- [34] Z. Wang, S. Yoon, S. J. Xie, Y. Lu, and D. S. Park, “A high accuracy pedestrian detection system combining a cascade adaboost detector and random vector functional-link net,” *The Scientific World Journal*, vol. 2014, 2014.
- [35] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. 1–511.
- [36] Y.-H. Pao, G.-H. Park, and D. J. Sobajic, “Learning and generalization characteristics of the random vector functional-link net,” *Neurocomputing*, vol. 6, no. 2, pp. 163–180, 1994.
- [37] G. H. Park and Y. H. Pao, “Unconstrained word-based approach for off-line script recognition using density-based random-vector functional-link net,” *Neurocomputing*, vol. 31, no. 1, pp. 45–65, 2000.
- [38] J. Xu, L. Yao, and L. Li, “Argumentation based joint learning: A novel ensemble learning approach,” *PloS one*, vol. 10, no. 5, p. e0127281, 2015.
- [39] F. De Smedt, K. Van Beeck, T. Tuytelaars, and T. Goedemé, “The combinator: Optimal combination of multiple pedestrian detectors,” in *ICPR, 2014*, pp. 3522–3527.
- [40] V.-D. Hoang, M.-H. Le, and K.-H. Jo, “Hybrid cascade boosting machine using variant scale blocks based hog features for pedestrian detection,” *Neurocomputing*, vol. 135, pp. 357–366, 2014.
- [41] H. Li and J. Cao, “Detection and segmentation of moving objects based on support vector machine,” in *Information Processing (ISIP), 2010 Third International Symposium on*. IEEE, 2010, pp. 193–197.
- [42] P. Lenz, J. Ziegler, A. Geiger, and M. Roser, “Sparse scene flow segmentation for moving object detection in urban environments,” in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 926–932.
- [43] J. Hariyono, V.-D. Hoang, and K.-H. Jo, “Moving object localization using optical flow for pedestrian detection from a moving vehicle,” *The Scientific World Journal*, vol. 2014, 2014.

Modeling and Analyzing Anycast and Geocast Routing in Wireless Mesh Networks

Fazle Hadi*, Sheeraz Ahmed[†], Abid Ali Minhas[‡], Atif Naseer[§]

*Preston University Kohat, Peshawar Campus, Pakistan

[†]Preston University Kohat, Peshawar Campus, Pakistan

[‡]Al Yamamah University Riyadh, Saudi Arabia

[§]Science and Technology Unit, Umm Al Qura University, Saudi Arabia

Abstract—Wireless technology has become an essential part of this era's human life and has the capability of connecting virtually to any place within the universe. A mesh network is a self healing wireless network, built through a number of distributed and redundant nodes to support variety of applications and provide reliability. Similarly, anycasting is an important service that might be used for a variety of applications. In this paper we have studied anycast routing in the wireless mesh networks and the anycast traffic from the gateway to the mesh network having multiple anycast groups. We have also studied the geocast traffic in which the packets reach to the group head via unicast traffic and then are broadcasted inside the group. Moreover, we have studied the intergroup communication between different anycast groups. The review of the related literature shows that no one has considered anycasting and geocasting from gateway to the mesh network while considering the multiple anycast groups and intergroup communication. The network is modeled, simulated and analyzed for its various parameters using OMNET++ simulator.

Keywords—Mesh Network; Anycast; Geocast; Routing; Unicast

I. INTRODUCTION

The basic aim of the wireless mesh networks (WMNs) is to guarantee the connectivity. WMNs are gaining popularity for its wide range of applications. The networks have gained substantial consideration as an unconventional solution to applications such as community networks, enterprise networks, and last mile access networks to the Internet [1]. WMNs are citywide multi-hop networks. They have a fixed infrastructure in the form of gateways and either mobile or fixed wireless mesh clients. The gateways have neither the mobility nor the power issues. There might be a series of fixed points- they relay the traffic to the sparsely distributed nodes. Gateways provide the internet access to the wireless nodes. This is the most common application of the wireless mesh networks. Video on demand and IP-TV are other interesting applications of a high speed wireless mesh networks. Among all these applications group communication is an important paradigm to study. Considering the citywide mesh network there might be many groups like the group of educational institutes, industries, vehicular network clusters etc. Contemporary studies mostly focus on optimal reaching to a gateway for the internet access.

Routing in the wireless mesh networks always attracts the researchers. Field base routing (FBR) is recently introduced for WMNs [2]. FBR relies on a routing field, and it is exchanged among the participating nodes. The data travels along the path having a relatively larger value of heat (the value computed

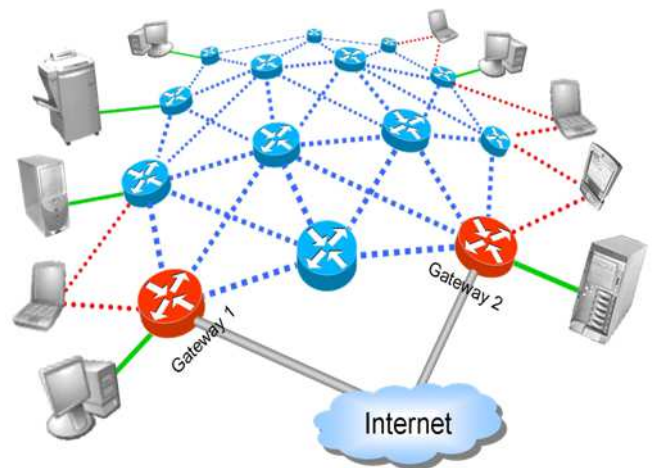


Fig. 1. Basic mesh architecture [21]

for every node considering the gateway as a source of heat) [2].

Anycast is an important service that always applies the greedy approach to deliver the packets to the nearest destination. If there are various groups of the same category then the anycast traffic will be forwarded to the next hop towards the group head having larger calculated parameter (temperature field). It considers the group head as the heat source.

The major contribution of this study is the proposal of an anycast model for the traffic from gateway to the mesh nodes using various anycast groups. Moreover, the study also analyzed the geocast and unicast communication. To the best of our knowledge the anycast and geocast communication have not been studied for the traffic forwarded by mesh gateways to the mesh clients. Though in [3] Tracy Camp et al. studied Geocast Adaptive Mesh Environment for Routing (GAMER) and the presented technique is about the geocast communication in ad hoc network. But the geocast technique presented in this paper is for WMNs and specifically for the traffic generated by gateway to mesh clients based on the field base routing. The geocasting is achieved by delivering the packets using unicast to a group head and then is broadcasted within the group. In addition to geocasting the main contribution of this study is the presentation of anycast model and group communication.

The existing literature does not handle multiple groups. So the inter-group communication is not yet covered. These issues

are of serious apprehension, that's why it is the main focus of this study. The research work will also consider multiple anycast groups, in the result the inter-group communication becomes possible.

The rest of this paper is organized as follows. Related

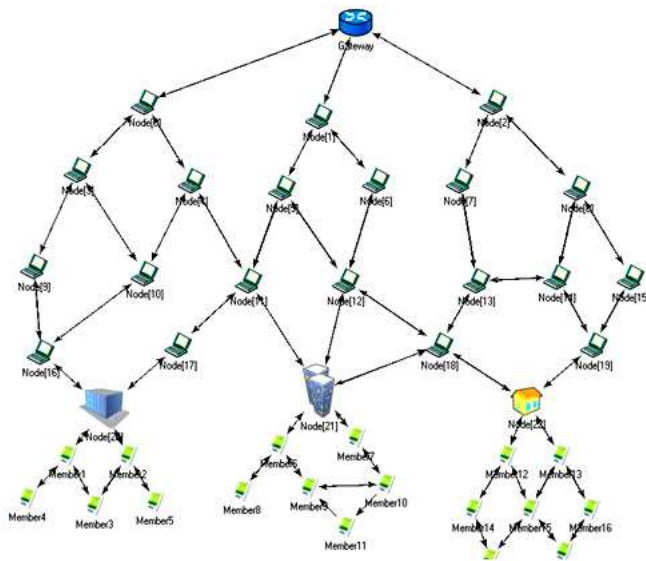


Fig. 2. Architecture of mesh network for anycast routing

work is underscored in section II. Section III describes the various challenges encountered in the design and analysis of wireless network structures like WMNs. Section IV depicts our proposed model for anycasting in WMNs, and the section V spells out the simulation environment and discussion on various traffic results. Finally, section VI concludes the paper with future directions.

II. RELATED WORK

Due to its demanding structure and application oriented architecture, there is a rich literature available about the wireless mesh networks. But after discussing the basic architecture we will converge our attention to the routing and group communication in the WMNs.

Akyildiz et al. [1] states that the wireless mesh network is a class of network where some nodes are fixed. These nodes serve as the gateway for the Internet connectivity. Others nodes are mobile, give access to the mobile nodes in a multihop fashion. Due to the redundant links, connectivity is not an issue in these networks. Most of the recent studies regarding the routing in the wireless mesh networks focus on the traffic flows from mesh nodes to the mesh gateways, for example, AODV [4] or OLSR [5]. Various unicast routing algorithms have been proposed in different studies like in [4], [5], [6] and [7]. Because of its scalability and robustness many researchers like Lenders et al. [8] and V. Park et al. [9] presented the field based routing algorithms. Recently, Baumann et al. [2] presented a field based routing algorithm for routing the packets from the mesh nodes to the gateway in anycast fashion. The authors give a field based routing algorithm, HEAT, it computes the temperature field keeping the gateway as the source of heat. In their later work Baumann et al. [10] presented the gateway source routing (GSR) algorithm for routing the packets through

the wireless mesh network. The authors use the routing path in backward direction, the path which is build up by the mesh clients by sending the packets to the gateway. In order to route the packets from gateway to the mesh nodes it is necessary that the mesh clients first send the data to the gateway. it seems to be very anomalous limitation of the proposed scheme.

The concept of the field based routing algorithms is very straight forward. In these algorithms the data moves along the steepest path towards its destination. In [11], Bahr introduced a hybrid wireless mesh protocol (HWMP). It is the merging of two seemingly opposite technologies i.e. flexibility of on-demand route discovery and enabling efficient proactive routing as well.

Field base routing algorithms are ingeniously used for various applications like load balancing in wide area networks [12], data gathering in sensor networks [13], placement of sensor nodes [14] and routing in MANETs [2], [9]. This study also using the gradient based routing algorithm for routing unicast, anycast and geocast traffic in wireless mesh networks. We are presenting the anycast model considering the traffic from the gateway to the mesh clients, having different anycast groups. Geocasting is phenomenon in which packets are delivered to a particular group belonging to a specific geographical location. There are various geocast algorithms available in the literature e.g. [15] and [16] exclusively depends upon the exact geographical information of the source and the destination.

The exact geographical information need specialized devices and is very hard to obtain [17]. The authors also present a geocast model in which the traffic moves from the gateway till the group head in unicast fashion following the gradient base routing and then group head broadcasts it inside the group. In [18], the authors propose a joint traffic splitting, routing, rate control and scheduling algorithm called CLC-DGS, which splits and distributes network traffic into multiple gateways in an optimal way. The authors prove by extensive simulations that CLC-DGS can achieve maximum network utility and improves the performance of WMNs under various network environments including link and gateway heterogeneities and various interference models.

In paper [19], the authors propose a jamming technique which targets the periodic nature of the routing protocol residing in the network layer. The technique is based on the concept of null-frequency jamming which refers to periodic attacks targeting specific protocol period/frequency of operation. The effects of this jamming technique are investigated in stack, half-diamond, full-diamond, full-mesh and random topologies employing the optimised link state routing protocol. In order to fully utilize spectrum resource in WMNs, [20] proposes a combination of some communication techniques, including link scheduling, spatial reuse, power and rate adaptation and Network Coding (NC). This was done to activate as many transmission links as possible during one scheduling period, so that the total scheduling length can be minimized and network throughput can be maximized. They consider interplay among these techniques and present an optimal NC-aware link scheduling mechanism in multi-rate WMNs, which relies on the enumeration of all possible schedules. Due to the high computational complexity of proposed model, they utilize a column generation (CG)-based method to resolve the optimization problem. Also, they present a distributed power control algorithm, by which the computational complexity of the CG-

based scheme can be largely reduced.

III. ROUTING CHALLENGES IN WMNS

There are various design and data transfer challenges in WMNs because of their applications and network topologies. Also, these challenges play a vital role in the design of routing algorithms for these networks as well the performances. Following factors are very important to be considered in the design of routing algorithms and analysis for WMNs.

1. Mobility of the nodes:

All the nodes represent an autonomous route for peer-to-peer connection.

2. Network topology:

A non-systematic mobility of nodes with varying speeds make the topology of the network vary randomly.

3. Link between the nodes:

The network nodes are connected by an air medium and do not have any fixed infrastructure.

4. Battery lifetime:

To retain the residual power in the nodes is a major trouble; so WMNs rely on batteries.

5. Network attacks:

Wireless networks have more chances of security attacks comparative to wired networks.

6. Quality of Service(QoS):

The QoS depends on various parameters in delivery of data, resulting in lower performance.

7. Consumption of power:

The energy conservation plays an important part in network evaluation.

8. Bandwidth constraint:

The unfixed infrastructure based network has less throughput metrics than the fixed ones.

IV. NETWORK MODEL

In contrast to ad hoc networks, WMNs do not impose the strict infrastructure less property. The basis architecture of the wireless mesh network is shown in figure 1 [19]. It has some fixed nodes (gateways) which provide access to the Internet acting as the backbone. The other nodes may be fixed access points, provide connectivity to the wireless mobile nodes in multihop fashion. Every node in the network acts as a router.

In this paper, we have developed a scenario for the anycast routing in wireless mesh networks. Figure 2, shows the architecture of wireless mesh network for anycast routing. There are three types of nodes in this architecture.

1) *Gateway*: The gateway provides the Internet connectivity to the mesh clients. As in this paper we are considering the traffic from the gateway to the mesh clients so the gateway is the main source of traffic in this scenario.

2) *Routing Nodes*: These nodes have the capability to route the packet towards its destinations. The routing is based upon the Gradient-Based Routing. Routing field is calculated while keeping the destination as the ultimate source

(maximum field value).

3) *Groups*: These are different groups having their own group members. Group members are registered with the group head. The group heads are registered with the gateway.

Presence of either anycast traffic at the gateway will be

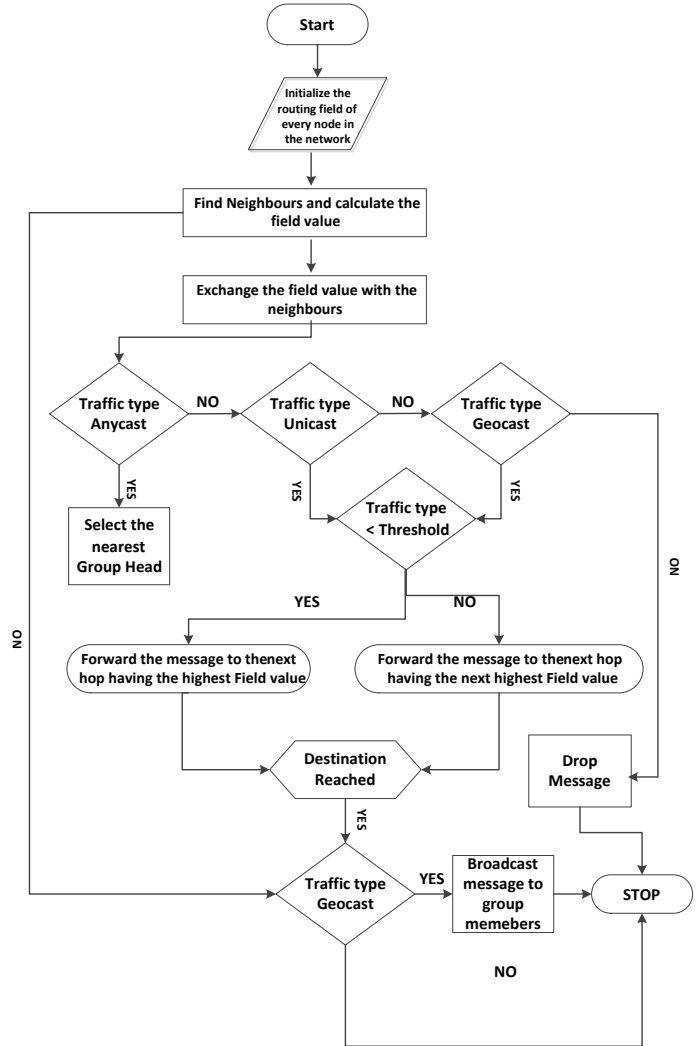


Fig. 3. Flow chart of WMN proposed routing

delivered to nearest group head via gradient based routing mechanism. If, on other hand, there is geocast traffic for any group, it will be routed to the group head following the gradient based routing in unicast fashion. It will then be broadcasted inside the group. Any routing node or group head may be regarded for the unicast traffic trending the same routing mechanism. There can be different interest groups suggested for various types of traffic. In such a way, intergroup communication is also made possible.

The routing field value is computed in the following equation. Let y_1, \dots, y_m be the link paths of nodes in the network from their respective destinations and d is the total distance of every node to its destination. Then:

$$W_i = f(d_i) = \sum_{i=1}^m |y_i| = d \quad (1)$$

which reflects the field value fv of each sensor. Flow-chart in figure 3 depicts anycast, unicast and geocast routing with load balancing.

A. Load Balancing

Forwarding nodes make their decision on the basis of fv of their neighbors. If proper load balancing technique is not followed then the entire traffic will follow the same route. This may not suit to energy constrained nodes. In our scheme, the load balancing is ensured by putting a threshold on the traffic flow. As traffic exceeds that of threshold value, the forwarding algorithm selects the next better route as shown in figure 3.

B. Algorithm

The algorithm for anycast, unicast and geocast routing trending the gradient base routing technique with load balancing is explained below. Every sensor of the network computes its fv considering the fv of its neighbors and then advertising the value to update routing table of the neighbors. The packet is relayed to the next hop with the highest fv . This process continues till the traffic flow exceeds the threshold considered for load balancing. When the traffic flow exceeds the threshold, our algorithm selects next hop with the next highest fv . For geocast traffic it carries the data in unicast fashion till the group head and then broadcasts it inside the group. The detailed algorithm is given below:

- Gm : Group member
- Tt : Traffic type
- Fv : Field value
- Nn : Neighbour node
- Th : Threshold value
- Nh : Nearest head

```

do
{
initialize Fv of every Nn
calculate Fv for every Nn
check the Tt of every message
if (Tt == Anycast)
{
select any nearest head Nh
if (traffic flow of Nn ≤ Th)
{
select Nn having highest Fv
forward the message to Nn
}
}
else
{
select next hop having next highest value
forward the message to the next hop
}
}
else if (Tt == Unicast)
{
if (traffic flow of Nn ≤ Th) {
select Nn having highest Fv
forward the message to Nn
}
}
else if
{

```

```

select next hop having next highest value
. forward the message to the next hop
}
}
else if (Tt == Geocast)
{
if (traffic flow of Nn ≤ Th)
{
select Nn having highest Fv
forward the message to Nn
}
}
else
{
select next hop having next highest value
forward the message to the next hop
}
if (Node == Group Head) Broadcast the message to Gm
}
end if
until (Destination Reached)
}

```

V. SIMULATION RESULTS

For the performance analysis we have used OMNet++ [22] simulator. The scenario developed to simulate the anycast, unicast and geocast traffic contains twenty four nodes. It includes one gateway that relays the data from and to the Internet. The group heads recognize three anycast groups in this research, as we study the traffic from the Internet towards the mesh clients. The gateway is the ultimate source of the traffic. It relays the Internet traffic towards the mesh clients and groups.

Figure No. 4 depicts the comparison of packet delays

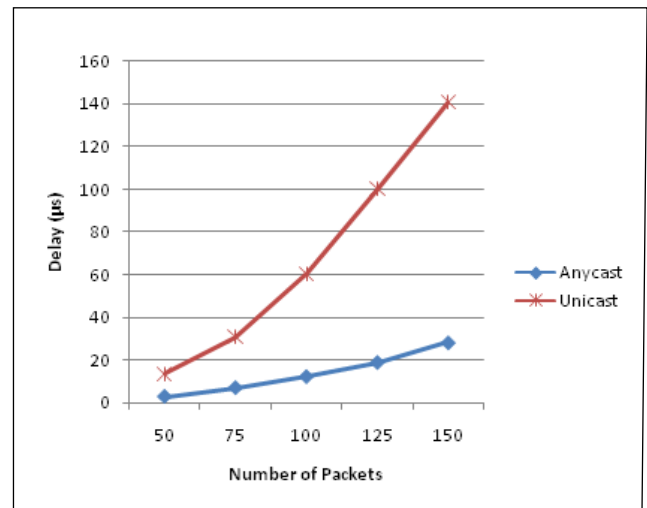


Fig. 4. Packet Delay comparison of Anycast and Unicast traffic

experienced by unicast and anycast traffic. As the idea of anycast group communication is proposed for the first time by considering the traffic from the gateway towards the mesh clients. We analyze the delay experienced by unicast and anycast traffic. Before the idea of anycast the only type of

traffic flow was unicast. The analyses show that when the volume of data increases the delay experienced by unicast traffic is considerably increased as compared to the anycast traffic. This is because the anycast traffic always chooses the best possible path and delivers the packet to the nearest group head. As a precaution we have implemented the load balancing mechanism that diverts some traffic. The idea has been already explained earlier.

The packet delivery ratio is always an important parameter

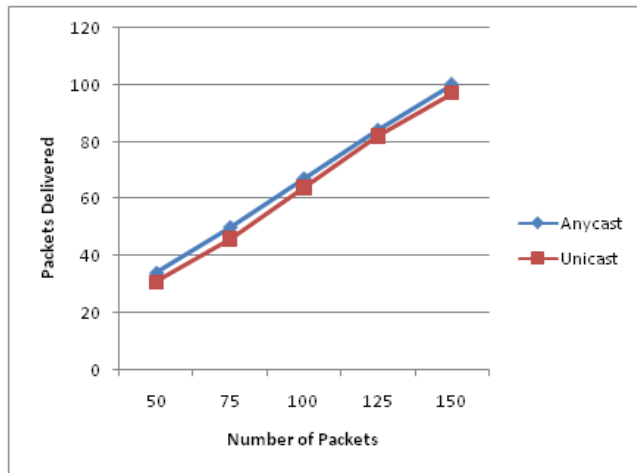


Fig. 5. Packet Delivery Ratio comparison of Anycast and Unicast traffic

to study. Figure 5 shows the comparison of the packet delivery ratio between the unicast and anycast type of traffic. Packets may be dropped uniformly in routing due to buffer overflow at the intermediate devices, link failure or any unseen problem. It has been observed that anycast type of traffic always gives better packet delivery ratio as compared to the unicast traffic. This is because that the unicast traffic will have to follow a predefined path towards a predefined destination and the anycast communication is adaptable and selects the best choice while forwarding the packets to the next hop towards any optimal destination.

Geocasting is the phenomenon of delivering the data packets to a particular geographical location. We consider the entire group members to be the part of a particular geographical location. Now, one way is to deliver the data packets to all group members in unicast fashion. We name it unicast based geocasting (UG). The other way is to deliver a packet to the group head (possible in the proposed group communication) and then broadcasts it within the group. We name it anycast based geocasting (AG). The figure 6 portrays the packet delay analysis experienced by UG and AG. The traffic delay in case of UG is directly proportional to the number of group members and in case of AG it is a delay to reach to the group head plus the delay involved in the broadcast phase. The intra-group broadcasting delay has been considered in AG. Group members are normally in the direct range of the group head.

VI. CONCLUSION

In this paper we have studied the anycast and geocast routing in wireless mesh networks. The concept of the

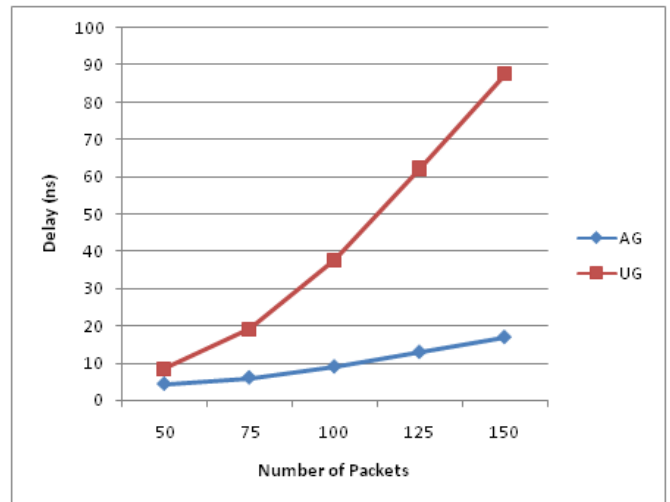


Fig. 6. Packet Delay comparison of Anycast and Unicast base Geocasting

anycast routing by considering the traffic from gateway towards the mesh clients has been proposed. The proposed technique has been studied against the unicast traffic by varying the volume of traffic from gateway towards the anycast groups. The anycast communication outperforms other type of communication in terms of packet delays and packet delivery ratio. Intergroup communication has also been made possible. Moreover geocast communication technique has been proposed in which the data travels in unicast fashion till the group head and then broadcasted inside the group. It lessens the delay time experienced by the geocasting based upon unicasting or geographical location. In future work we will create more realistic scenarios that will contain mobile nodes and adaptive group formation and head selection.

REFERENCES

- [1] F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks*, vol. 47, pp. 445-487, 15 March 2005.
- [2] Rainer Baumann and Vincent Lenders and Simon Heimlicher and Martin May, "HEAT: Scalable Routing in Wireless Mesh Networks using Temperature Fields," in *Proceedings of the IEEE WoWMoM, 2007*.
- [3] Tracy Camp, Yu Liu, "An adaptive mesh-based protocol for geocast routing," *Journal of Parallel and Distributed Computing archive* Vol.63, No.2, 2003.
- [4] Sung-Ju Lee, Elizabeth M. Belding-Royer, Charles E. Perkins, "Scalability study of the ad hoc on-demand distance vector routing protocol," *International Journal of Network Management*, vol. 13, No. 2, pp.97-114, 2003.
- [5] T. Clausen and P. Jacquet, "Optimized Link State Routing Protocol," *IETF Internet Draft*, draft-ietf-manet-olsr-11.txt, July 2003.
- [6] D. B. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks," in *Mobile Computing*, Imielinski and Korth, Eds. Kluwer Academic Publishers, 1996, vol. 353.
- [7] C. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," in *ACM SIGCOMM 94 Conference on Communications Architectures, Protocols and Applications*, 1994, pp. 234-244.
- [8] V. Lenders, M. May, and B. Plattner, "Density-based vs. Proximity-based Anycast Routing for Mobile Networks," in *IEEE INFOCOM, Barcelona, Spain, April 2006*.

- [9] V. Park and S. Corson, "A Performance Comparison of the Temporally-Ordered Routing Algorithm and Ideal Link-State Routing," Third IEEE Symposium on Computers and Communications, Athens, Greece, 2002.
- [10] Rainer Baumann, Simon Heimlicher, Vincent Lenders, Martin May, "Routing Packets into Wireless Mesh Networks," in Proceedings of the IEEE WoWMoM, 2007.
- [11] Michael Bahr, "Update on the Hybrid Wireless Mesh Protocol of IEEE 802.11s," IEEE International Conference on Mobile Ad hoc and Sensor Systems, Italy 2008.
- [12] A. Basu, A. Lin, and S. Ramanathan, "Routing Using Potentials: A Dynamic Traffic-Aware Routing Algorithm," in Proceedings of the ACM-SIGCOMM03, 2003.
- [13] J. Faruque and A. Helmy, "RUGGED: RoUting on finGerprint Gradients in seEnsor Networks," in Proceedings of the IEEE ICPS, 2004.
- [14] S. Toumpis and L. Tassiulas, "Packetostatics: Deployment of Massively Dense Sensor Networks as an Electrostatic Problem," in IEEE INFOCOM, 2005.
- [15] Q. Fang, J. Gao, and L. J. Guibas, "Locating and bypassing holes in sensor networks," Mobile Networks and Applications, Vol 11, No. 2 pp. 187-200, 2006.
- [16] F. Kuhn, R. Wattenhofer, Y. Zhang, and A. Zollinger, "Geometric Ad-Hoc Routing: of Theory and Practice," In twenty-second annual symposium on Principles of distributed computing, pages 6372, 2003.
- [17] M. ODell, R. ODell, M. Wattenhofer, and R. Wattenhofer, "Lost in Space Or Positioning in Sensor Networks," In Proceedings of the Workshop on Real-World Wireless Sensor Networks (RE-ALWSN'05), June 2005.
- [18] Zhou, Anfu, Min Liu, Zhongcheng Li, and Eryk Dutkiewicz. "Joint Traffic Splitting, Rate Control, Routing and Scheduling Algorithm for Maximizing Network Utility in Wireless Mesh Networks." 2015.
- [19] Lall, Shruti, B. T. J. Maharaj, and PA Jansen van Vuuren. "Null-frequency jamming of a proactive routing protocol in wireless mesh networks." Journal of Network and Computer Applications 61 (2016): 133-141.
- [20] Ning, Zhaolong, Qingyang Song, Lei Guo, Zhikui Chen, and Abbas Jamalipour. "Integration of scheduling and network coding in multi-rate wireless mesh networks: Optimization models and algorithms." Ad Hoc Networks 36 (2016): 386-397.
- [21] Sichitiu, Mihail L. "Wireless mesh networks challenges and opportunities." Electrical and Computer Engineering Department, NC State University, Raleigh, NC, USA (2006).
- [22] M. Kozlovsky, A. Balasko, A. Varga, "Enabling OMNeT++-based simulations on grid systems," Proceedings of the 2nd International Conference on Simulation Tools and Techniques, Rome, Italy, 2009.

MOSIC: Mobility-Aware Single-Hop Clustering Scheme for Vehicular Ad hoc Networks on Highways

Amin Ziagham Ahwazi

Department of Computer Science and Engineering
Islamic Azad University, Ahvaz Branch
Ahvaz, Iran

MohammadReza NooriMehr

Department of Computer Science and Engineering
Islamic Azad University, Ahvaz Branch
Ahvaz, Iran

Abstract—As a new branch of Mobile ad hoc networks, Vehicular ad hoc networks (VANETs) have significant attention in academic and industry researches. Because of high dynamic nature of VANET, the topology will be changed frequently and quickly, and this condition is causing some difficulties in maintaining topology of these kinds of networks. Clustering is one of the controlling mechanism that able to grouping vehicles in same categories based upon some predefined metrics such as density, geographical locations, direction and velocity of vehicles. Using of clustering can make network's global topology less dynamic and improve the scalability of it. Many of the VANET clustering algorithms are taken from MANET that has been shown that these algorithms are not suitable for VANET. Hence, in this paper we proposed a new clustering scheme that use Gauss Markov mobility (GMM) model for mobility predication that make vehicle able to prognosticate its mobility relative to its neighbors. The proposed clustering scheme's goal is forming stable clusters by increasing the cluster head lifetime and less cluster head changes number. Simulation results show that the proposed scheme has better performance than existing clustering approach, in terms of cluster head duration, cluster member duration, cluster head changes rate and control overhead.

Keywords—Vehicular Ad hoc Networks; Mobile ad hoc Networks; Network Topology Control; Clustering Scheme

I. INTRODUCTION

Vehicular ad hoc networks (VANETs) makes a new vision in the field of Intelligent Transportation Systems (ITS). Recently, VANET becomes a most important area of research both in academic and industry field, because it has the potential to create numerous applications such as dissemination of safety, routing plans, traffic condition message, entertainment (e.g. information sharing, gaming), e-commerce and control of vehicle flow formations [1], [21], [22]. In principle, VANET is a special form of MANETs, with the difference that there are mobile nodes (Vehicles) have high dynamic mobility. In VANET vehicles equipped with an on-board unit (OBU) which make them able to communicate with each other (vehicle-to-vehicle, V2V) and via roadside units (vehicle-to-roadside, V2R) also called as RSUs. The communication standard that vehicles used to communicate with each other is Wireless Access for Vehicular Environments (WAVE), which it is an approved amendment to the IEEE 802.11 standard. WAVE is also known as IEEE 802.11p [13].

Due to high mobility, VANET topology changes rapidly, so establishing new control topology cause to introducing high communication overhead for exchanging information. There are several control schemes for media access and topology management have been proposed. One of these schemes is clustering structure. In clustering structure, the mobile nodes are divided into a number of virtual groups based on certain metrics. These virtual groups are called clusters [2].

Some cluster-based approaches have been proposed and applied in Ad-hoc Networks, because the clustering have more advantages such as reduce the delay, overhead and solving the scalability problem in large scale networks. However, in dynamic environments the clusters usually are unstable and frequently disjointed. Hence the clustering schemes which are proposed for Mobile ad hoc networks (MANET) and Wireless Sensor networks (WSN) are not suitable for VANET. On the other word, in VANET, vehicle move with high and variable speeds which causing to frequent changes in the network topology, and it can significantly reduce the cluster stability and efficiency. CH duration is one of the reasons that can be caused to this reduction. It means that whatever CH duration increased respectably cluster stability will be increased. On the other hand, an efficient cluster maintenance has directly impact on CH lifetime. Hence, this parameters should be considered in the design of new cluster scheme. The aim of this work is proposed a scheme to construct a stable single-hop clusters with more CH lifetime, more CM duration and less cluster change rate. In this scheme CH selection conducted base on relative mobility, which calculated as the average relative distance and relative velocity.

The rest of this paper is organized as follows. In Section II previous work related to cluster formation and maintenance will be described. Section III explain preliminaries of proposed scheme. Section IV present our proposed algorithm processing. Section V includes simulation description with comparative results. The paper is concluded in Section VI.

II. RELATED WORKS

As a well-known organizing and controlling networks, node clustering is widely used in MANET and Wireless Sensore Networks (WSN). Clustering technique can be used for diverse purpose such as broadcasting, routing and QoS.

There are many clustering solution based on topology, energy, neighbor have been proposed [16], [7], [8], [9], [10], [11], [12]. However, these clustering algorithms significantly are not suitable for dynamic environments such as VANET. One of the well-known clustering scheme which frequently used for comparison with other VANET clustering algorithms, is MOBIC [4]. Indeed, this algorithm is based on the lowest-ID algorithm [16]. In MOBIC, cluster head selection is based on the signal power which received at any node from its neighbors derived from successive receptions. The performance of MOBIC is medium and not effective for dynamic scenarios.

The aggregate local mobility (ALM) algorithm is proposed in [5]. This algorithm used a relative mobility which calculate based on distance between a node and its neighbors. ALM algorithm aims to extend cluster lifetime using ALM.

Another known clustering algorithm which was proposed is affinity propagation (AP) algorithm [14]. AP algorithm is a distance-based clustering scheme which vehicles exchange the availability and responsibility information with their neighbors and based on this information, CH is selected. The drawback of AP is that frequent changing of CH increased when vehicle's speed increased. it is because of that the AP does not take the speed difference of vehicles into consideration.

Adaptable mobility-aware clustering algorithm based on destination positions (AMACAD) [17] is clustering scheme which is proposed for VANETs. This algorithm used set of parameters, including position, speed and distance as a metric for CH selection. DMMAC is a novel clustering algorithm which proposed by Hafeez et al [15]. DMMAC used velocity as main factor to form clusters, meanwhile it utilized fuzzy system to processing vehicle's velocity to enhance stability of cluster. Beside aforementioned aspects, DMMAC algorithm used a temporary cluster head concept which will be used when the main CH are unavailable. But this algorithm suffers from CHs frequently change when the vehicle's speed increased.

At the end of review the previous works, we will refer to laned-based clustering (LBC) scheme [19]. This scheme is designed specifically for the urban environments, which the number of lanes in its traffic flow considered as metric for CH selection process. However, this scheme does not consider the exact number of vehicles for each flow.

III. PRELIMINARIES

The proposed clustering scheme uses Gauss-Markov Mobility (GMM) model [3] to calculate the future vehicles position and based on that predicted position and other metrics (e.g. Relative velocity, relative distance), the proposed scheme try to form a stable single-hop cluster. We call this, MObility-aware and SIngle-hop Clustering scheme (MOSIC). The term of single-hop cluster refer to a cluster architecture which cluster-member can communicate with cluster-head directly. The MOSIC focuses only on V2V (Vehicle-to-Vehicle) communication and the main objective of proposed scheme is to make a large network with highly dynamic nodes appear smaller and able to sustaining clusters for long period by increasing the cluster-head and cluster-member duration. So in following some essential assumptions and definitions which MOSIC used will be described.

A. Assumptions and Definitions

The proposed clustering scheme assumes that all vehicles traveling in the same direction (one way) on highway and all of them are equipped with Global Position System (GPS) receivers an On Board Units (OBU). Location information of all vehicles needed for clustering scheme is collected with the help of GPS receivers. Also The roads in highway have a maximum allowed velocity (V_{max}). Each vehicle have the same transmitting capability since they have equal chance to be elected as CHs. In this paper we use some definitions that we'll explain them in the following.

In addition, Table II provides the notations that utilize in this paper.

Definition 1: (Vehicle State): In proposed scheme, vehicles have four kinds of state which listed in Table I. Where

TABLE I: Vehicle States

State	Description
<i>NC</i>	Non-Clustered
<i>CM</i>	Cluster-Member
<i>CH</i>	Cluster-Head
<i>TCM</i>	Temporary Cluster-Member

NC indicate as a vehicle is standalone and doesn't belong to any cluster, *CM* as a vehicle which belong to a cluster, *CH* as a vehicle that has task of coordination among cluster members and responsible for management of cluster structure [20] and *TCM* represent as a vehicle which doesn't receive the information broadcasted by the CH for ΔT interval.

Figure 1 show the vehicle's state in a highway environment.

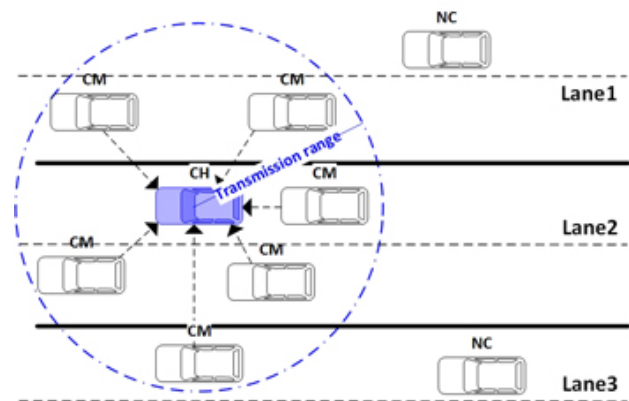


Fig. 1: Vehicle's States

Definition 2: (r -neighbors): We considered two vehicles are r -neighbors if the distance between them is less than r . Consequently, the neighborhood N_i of a vehicle i is defined as follows:

$$N_i = \{j, D_{i,j} < r\} \quad (1)$$

Where $D_{i,j}$ is the average distance between vehicles i and j .

TABLE II: Notations and Description

Notation	Description
ID_i	Unique ID of vehicle i
r	Communication Range of vehicle i
N_i	Neighborhoods of vehicle i
Deg_i	The nodal degree of vehicle i
$d_{i,j}$	Relative distance between vehicle i and j
$d'_{i,j}$	Predicted relative distance between vehicle i and j
x_i	Position of vehicle in axe x
y_i	Position of vehicle in axe y
x'_i	Predicted position of vehicle in axe x
y'_i	Predicted position of vehicle in axe y
\bar{R}	Relative Distance
V_i	Velocity of vehicle i
\bar{V}_{m_i}	Mean relative velocity of vehicle i among their neighbors
\bar{V}_{rel_i}	Relative Velocity of vehicle i
V_{max}	Maximum speed of road
\bar{M}_i	Relative Mobility metric of vehicle i
M_i	Mobility metric of vehicle i
ΔV_{th}	The threshold used to classify neighboring nodes as stable or non stable neighbors
CFV	Cluster forming vehicle

Definition 3: (Nodal degree): The total number of r -neighbors of a vehicle is called the nodal degree of the vehicle i which calculated as follow:

$$Deg_i = |N_i| \quad (2)$$

The nodal degree of a node i can be concluded as the cardinality of the set N_i .

Definition 4: (Stable r -neighbors): Two vehicles are considered as a stable r -neighbors if the difference speed between them is less than $\pm\Delta V_{th}$. Where ΔV_{th} is a predefined threshold.

B. Gauss-Markov Mobility (GMM) Model

The Gauss-Markov Mobility (GMM) Model [3] is a memory-based mobility model which able to calculate next position of mobile node based on its current mobility metric. In this model, each mobile node is assigned to the initial speed and direction. The GMM model used alpha α , $0 \leq \alpha \leq 1$, parameter which determines variability in mobile node movement. In this model, at each fixed interval of time, n , the mobile node update it current speed and direction which the new speed and direction are calculated as follows:

$$s_n = \alpha s_{n-1} + (1 - \alpha)\bar{s} + \sqrt{(1 - \alpha)^2} s_{x_{n-1}} \quad (3)$$

$$d_n = \alpha d_{n-1} + (1 - \alpha)\bar{d} + \sqrt{(1 - \alpha)^2} d_{x_{n-1}} \quad (4)$$

where s_n and d_n are the new speed and direction of the mobile node at time interval n ; \bar{s} and \bar{d} are representing the mean value of speed and direction and $s_{x_{n-1}}$ and $d_{x_{n-1}}$ are random variables from a Gaussian (normal) distribution. At each time interval the next location is calculated based on the current location, speed, and direction of movement.

Specifically, at time interval n , an Mobile node's position is given by the equations 5 and 6:

$$x_n = x_{n-1} + s_n \cos(d_{n-1}) \quad (5)$$

$$y_n = y_{n-1} + s_n \sin(d_{n-1}) \quad (6)$$

where (x_n, y_n) and (x_{n-1}, y_{n-1}) are the x and y coordinates of the mobile node's position at the (n^{th}) and $(n - 1)^{\text{st}}$ time intervals, respectively, and s_n and d_n are the speed and direction of the mobile node, at the $(n)^{\text{st}}$ time interval which achieved based on equations 3, 4.

C. Message passing format

As previously mentioned, the VANET is running under WAVE (Wireless Access for Vehicular Environments) architecture (IEEE 802.11p) and messages are encapsulated in UDP packets in the network layer. Each vehicles exchange their status message with their neighbors in its communication range, r , periodically. The status message contains information about the vehicle's ID, vehicle state, current speed V , communication range r , CH's ID (CHID) and position POS , as shown in Fig. 2.

ID	$State$	V	r	$CHID$	POS
------	---------	-----	-----	--------	-------

Fig. 2: Status message packet format

In addition, POS consist two parts, Geographical and Predicted position, which both of them are based on Cartesian coordinates. The geographical location include (x, y) and predicted location consist (x', y') .

D. Cluster Metrics

In this section the cluster metrics, which plays an important role in cluster formation and cluster maintenance, will be described.

1) *Average relative velocity:* In every time interval, each vehicle will be aware about all r -neighbor vehicles, using exchange status message, and based on that information, average relative velocity \bar{V}_{rel_i} will calculated as follow:

$$\bar{V}_{rel_i} = max \left\{ \frac{\bar{V}_{m_i}}{V_{max}}, 0 \right\} \quad (7)$$

where V_{max} is the maximum allowed velocity on the road, and \bar{V}_{m_i} is the average velocity of vehicle i against their r -neighbors which defined as follow:

$$\bar{V}_{m_i} = \frac{1}{|Deg_i|} \sum_{j=1}^{Deg_i} (V_i - V_j) \quad (8)$$

where j is a potential neighboring vehicle, and V_i, V_j are the velocity of vehicle i and j respectively in m/s and Deg_i is nodal degree of vehicle i .

2) *Average relative distance*: Each vehicle will collect its mobility information such as its location at every time interval ΔT and send this information to all its r -neighbors via Control Channel. So each vehicle able to calculate its average relative distance among its r -neighbors. Relative distance is one of the measure that play a key role to elect CH.

Consequently relative distance defined and calculated as follow:

$$\bar{R}_i = \frac{1}{|Deg_i|} \sum_{j=1}^{Deg_i} R_{i,j} \quad (9)$$

where $R_{i,j}$ is obtained from the below equation:

$$R_{i,j} = 10 \times \log_{10} \left(\frac{d_{i,j}}{d'_{i,j}} \right) \quad (10)$$

We use the metric proposed by Basu [4] to calculate average relative distance (Equation 10), but whit difference that we used distance and predicted distance between two nodes instead of Packet Delay, which is used in [4]. In formula 10, $d_{i,j}$ is distance between vehicle i and j which can achieved and calculate via Euclidean distance.

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (11)$$

also $d'_{i,j}$ represents the distance between vehicle i and j which is predicted with mobility model, and similar to $d_{i,j}$ calculated as follow:

$$d'_{i,j} = \sqrt{(x'_i - x'_j)^2 + (y'_i - y'_j)^2} \quad (12)$$

where x'_i and y'_i is future position of vehicle i that calculated and obtained using Gauss-Markov Mobility model (see Sect. III-B).

3) *Average relative mobility*: Average relative mobility is an important measurement that vehicles can be informed about their r -neighbors and based on this parameter, vehicles decides which vehicle is more suitable to selected as CH. \bar{M}_i is defined as follow:

$$\bar{M}_i = \bar{R}_i + \bar{V}_{rel_i} \quad (13)$$

where \bar{R}_i is average relative distance and \bar{V}_{rel_i} is average relative velocity which described in previous subsections. As you can see, whatever the nodal degree of a vehicle is increased, then correspondingly, the value of \bar{R}_i and \bar{V}_{rel_i} , will be decreased. Because according to Formula 8 and 9, the relative distance and relative velocity are inversely proportional with nodal degree (Deg_i). So a vehicle with lower value of \bar{M}_i is more considerable.

IV. MOSIC PROCESSING

This section contains the description of the procedures that form part of the proposed clustering scheme. In brief, the proposed clustering scheme is formed by the four phase (Initialization, CH Selection, Cluster Formation and Cluster maintenance), described in the following subsections. When a node is not belong to any cluster (in Non-Clustered state), it executes the initialization phase. after that, depending

on whether the cluster head can be found in nearby or not, the node can launches the join procedure or the cluster formation phase. Hence, after the cluster formation phase or after joining a cluster, the maintenance procedure will be executed and it checks the validity of the cluster periodically.

A. Initialization phase

This phase is executed by any vehicle which its state is NC (Non_Clustered) and also receives a status message from its r -neighbors. In any interval time, ΔT , a vehicle which its state is NC broadcast its status message to discover weather a Cluster Head exist in vicinity or not. If there is at least one Cluster Head can be found, then the vehicle launches the join procedure. Otherwise, it execute the cluster formation phase. The pseudo code of the Initialization phase shown in Algorithm 1.

Algorithm 1 Node Initialization

```

1:  $State_i$ : state of vehicle  $i$ ;
2:  $N_i$ :  $r$ -neighbors set of vehicle  $i$ ;
3: if ( $State_i == NC$ ) then
4:   Broadcast its status message;
5: end if
6: if (Receive messages after  $\Delta T$  interval) then
7:   Update its  $N_i$  sets;
8:   if ( $State_i == NC$ ) then
9:     if ( $r$ -neighbors indicate CH exist) then
10:      Call  $JoinCluster()$ ;
11:     else
12:      Call  $ClusterFormation()$ ;
13:     end if
14:   end if
15: else
16:   After  $\Delta T$  interval, broadcast its status message again;
17: end if

```

B. Cluster Head selection phase

In principle, CH is a coordinator with the task of coordination among cluster members and also responsible for management of cluster structure [20]. One of the most frequently used technique to increase cluster stability is CH duration. CH duration has impact direct relative with cluster stability. It means that select a more stable CH can be beneficial to keep cluster structure for long periods and stable cluster can reduce packet loss probability. Consequently, select a CH that can be stable for long period, is an important factor in the design of MOSIC. In proposed scheme, we defined a mobility measure, M_i , that each vehicle calculated it based on status messages which received in interval time ΔT from r -neighbors and each vehicle has greatest value of M_i , will be selected as CH. Mobility measure calculated as follow:

$$M_i = \begin{cases} |\frac{1}{\bar{M}_i}| & , \bar{M}_i \neq 0 \\ 0 & , \bar{M}_i = 0 \end{cases} \quad (14)$$

where \bar{M}_i is average relative mobility. As mentioned in sec: III-D3, a vehicle with lowest value of \bar{M}_i is more considered to be CH, So, for simplicity calculations, the value

of \bar{M}_i will be reversed, because the lowest value becomes to the greatest value, And it's exactly what Formula 14 shows.

Once status message are received, the vehicle with highest M_i among its r -neighbors will elect itself as CH. Vehicle with highest M_i will set its $CHID$ field to its own ID and send the status message to r -neighbors and subsequently all r -neighbors will join cluster (All r -neighbors sets their $CHID$ field to vehicle's ID which selected as CH). It should be noted that nodes with Non-Clustered state, can't participate in the election process and they must commence the initialization phase.

The pseudo code of the CH selection shown in Algorithm 2.

Algorithm 2 Cluster-Head Selection

```
1:  $State_i$ : state of vehicle  $i$ ;  
2:  $N_i$ :  $r$ -neighbors set of vehicle  $i$ ;  
3:  $M_i$ : mobility measure of vehicle  $i$ ;  
4:  $CHID$ : cluster head;  
5:  $ID_i$ : ID of vehicle  $i$ ;  
6: Receive status message from  $r$ -neighbors in  $\Delta T$ ;  
7: Update its  $N_i$  sets;  
8: Calculate the  $M_i$  based on received status messages;  
9: if ( $N_i > 0$  and  $State_i \neq NC$ ) then  
10:   if ( $M_i = \max(M_j | j \in N_i)$ ) then  
11:     if  $State_i = CH$  then  
12:       DoNothing();  
13:     else  
14:        $CHID = ID_i$ ; //select itself as a CH  
15:        $State_i = CH$ ;  
16:       Broadcast head message and  $r$ -neighbors will join  
        cluster;  
17:     end if  
18:   end if  
19: else  
20:   Call  $Initialization()$ ;  
21: end if
```

C. Cluster formation phase

The cluster formation phase is executed every time interval, ΔT , with nodes in NC state that already before run the initialization phase and discover that there is no CH in vicinity. However, after initialization phase (which all NC state nodes broadcast its status message and receive reply messages), a vehicle whose speed is the slowest among all its NC r -neighbors, start the cluster formation process. This vehicle is called cluster forming vehicle (CFV). At the beginning of the process, CFV select itself as a CH and broadcast status message to r -neighbors. Thus vehicles upon receipt the status message, set its $CHID$ field to CFV 's ID and also update its state to CM.

The pseudo code of the Cluster formation shown in Algorithm 3.

D. Cluster maintenances phase

The main aims of clusters maintenance phase is to maintain the cluster structure as stable as possible. Because of the dynamic nature of the VANET, joining and leaving the

cluster happen frequently. However, there are three events that can affect on stability of a cluster include: Joining Cluster, Leaving Cluster and Cluster merging. In the following cluster maintenance procedure will be described.

1) *Joining Cluster*: When a NC (Non-Clustered) state vehicle approach a CH (comes within CH transmission range), then the vehicle and CH compare and check their relative velocity, \bar{V}_{rel_i} , and if the velocity difference is within $\pm\Delta V_{th}$, then the vehicle will join to the cluster and subsequently, CH add it to its members list. In some cases, a NC state's vehicle maybe comes in multiple CHs transmission range, r , then in this condition, vehicle join to cluster which has more nodal degree.

2) *Leave Cluster*: When a cluster-member moves out of the CH's transmission range, r , it is not removed from the cluster members list maintained by the cluster-head immediately. In the other hand, if a CM does not receive the information broadcasted by the CH every ΔT interval, the state of this node changes from CM to TCM (Temporary Cluster Member). It does not leave the cluster immediately, because this disconnection maybe due to the weak quality of the wireless signals. If the temporary member receives the information broadcasted by the CH again in the coming m interval, the state of this node changes to CM again. But when a temporary member does not receive the CH information consecutively for m times, it means that the node moves out of the cluster range. Thus the state of that node changes to the NC. Meanwhile, the CH will delete this member from the members list. Then, the node can either join another cluster or form a new cluster.

3) *Cluster Merging*: Whenever two CH approach and come in each other transmission ranges, and they stay connected over a time period and also their relatively velocity is within the $\pm\Delta V_{th}$, then the cluster merging process will commence. In this process, the CH with less nodal degree abandon their CH's role and joins to the cluster with more nodal degree. The other members of the merged CH according to its condition can join another cluster or become a standalone member (NC).

Algorithm 3 Cluster Formation

```
1:  $State_i$ : state of vehicle  $i$ ;  
2:  $N_i$ :  $r$ -neighbors set of vehicle  $i$ ;  
3:  $M_i$ : mobility measure of vehicle  $i$ ;  
4:  $CHID$ : cluster head;  
5:  $CFV$ : cluster forming vehicle;  
6:  $ID_i$ : ID of vehicle  $i$ ;  
7: if ( $V_i = \min(V_j | j \in N_i)$ ) then  
8:    $CFV = i$ ;  
9:    $State_{\{cfv\}} = CH$ ;  
10:   $CHID_{\{cfv\}} = ID_i$ ;  
11:  Broadcast(ClusterFormation( $CFV_i$ ));  
12: end if  
13: if (Receive temporary cluster formation message from  
     $CFV_i$ ) then  
14:    $CHID_j = ID_{\{cfv\}}$ ;  
15:    $State_j = CM | \{j \in N_i\}$ ;  
16: end if
```

TABLE III: Simulation Parameters

Parameter	Value
Simulation Time	300 s
Area Size	1000 m × 1000 m
Number of lanes for each direction	3
Maximum Vehicle Velocity	10 - 35 m/s
Maximum Allowed Velocity	55 m/s
The threshold for stable r -neighbors	10
Number of vehicles	100
Interval Time	1 s
Packet Type	UDP
Packet Size	100 Bytes
Transmission Range	200 m
Chanel	IEEE-802.11
Tuning parameter (α)	0.85

V. SIMULATION AND PERFORMANCE EVALUATION

The aim of the simulation is to compare the performance of the our proposed mobility-aware single hop clustering scheme (MOSIC) to the previously proposed VANET clustering schemes. However, the performance of the clustering scheme is evaluated by using the metrics of cluster head duration, cluster member duration, cluster head change rate, the number of cluster and control overhead, which these performance metrics can demonstrate the stability of our clustering scheme [14], [1].

The MOSIC is implemented in NS-3 simulator at version 3.24.1 [18]. The simulation scenario is based on one directional highway segment of 1000 m in length and three lanes. The vehicles are injected into the road randomly. Maximum Velocity varies from 10 to 35 m/s and the transmission range is 200 m. The total simulation time is 600 s. The clustering process start at the 300th second where all the vehicle are on the road. All of the performance metrics are evaluated for the remaining 300 s. Also we consider that the maximum allowed velocity in the road is 55 m/s. The general and important simulation parameters are listed in Table III. Also we used Gauss-Markov mobility model, as temporary hybrid model beside vehicles mobility. In other words, we used Gauss-Markov mobility (GMM) model as a prediction model for calculated next position of vehicles, which used in equation 10. We set α , Tuning parameter, to 0.85, as shown in Table III.

A. Cluster-Head Duration

Cluster-Head duration refers to the interval during which the vehicle' state is in CH and remain in this state until its state changed into CM or NC. The average CH duration is calculated by dividing the total CH duration with the total number of state changes from CH to CM or NC. Figure 3 illustrate the average CH duration of MOSIC and other clustering schemes for different maximum vehicle velocities. In Figure 3, the average CH duration decreases when the vehicle velocity increases. The reason for this is that when the vehicle velocity increase, the topology of network becomes more dynamic and eventually this makes it difficult for CHs to maintain a relatively stable condition with their neighbor vehicles for a long period. As

shown in Figure 3, the MOSIC has better performance in term of CH duration against N-Hop [1], AMACAD [7], ASPIRE [23] and Lowest-ID [16] respectively.

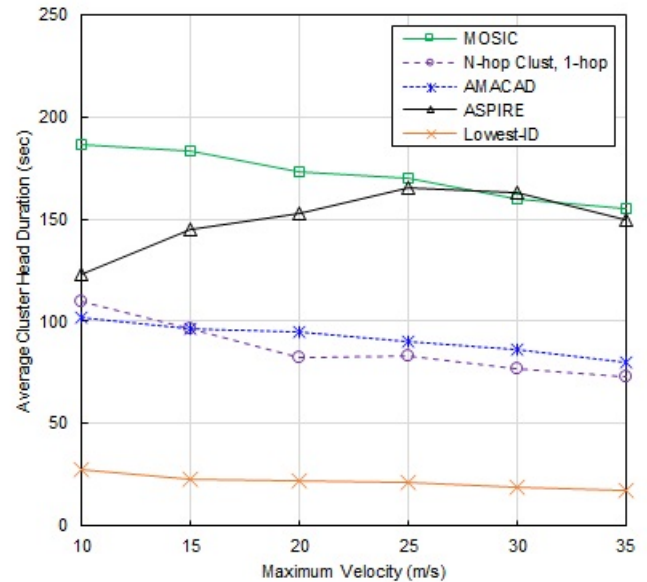


Fig. 3: Average CH Duration

B. Cluster-Member Duration

Cluster-Member duration is the interval from the time during which a vehicle joins a specified cluster to the time when it leaves the cluster. By dividing the total cluster member duration into total cluster member changes, average cluster member duration is calculated. Figure 4, shows the CM duration of MOSIC and other approaches for different maximum vehicle velocities. As shown in Figure 4, CM duration increases when vehicle velocity increases and it's because of the efficient cluster maintenance mechanism. The result which shown in Figure 4 indicate that the MOSIC CM duration is higher than N-Hop, AMACAD, ASPIRE and Lowest-ID respectively in most cases.

C. Cluster-Head Change Rate

Cluster-Head change number is the number of vehicles whose state changes from CH to CM or NC during a simulation process, and the rate of CH Change is defined as the changing per unit time. Figure 5, shows the CH change rate of MOSIC and other clustering schemes for different maximum vehicle velocities. A low CH change rate leads to a stable cluster structure. As shown in Figure 5, CH change rate increases when vehicle velocity increases. This is because of the dynamic nature of network. It means that with increasing velocity, it will be difficult for CH to keep efficient relatively stable with their CMs for a long period and maybe CH exited from cluster or in another condition, maybe CH into range of other CHs and merged with it and this situation effect of CH changes rate.

D. Number of Clusters

In VANET, because of highly dynamic movement of vehicles, clusters are created and vanished frequently over time and it cause to increase clusters number and consequently, increase maintenance cost. A Few clusters can enhance efficiency and performance of VANETs. A clustering algorithm is suitable, if it could reduce the number of clusters in system. This suitability achieved by create a relatively metric which sustain the current cluster scheme stable as much as possible. Figure 6 demonstrate the number of clusters under different transmission ranges and velocity scenarios. As shown in Figure 6, With increasing velocity the changes in the number of clusters is minimally and it because of good relative mobility metric which utilized in our scheme.

E. Average Control Message Overhead

All clustering schemes incur some additional control overhead to form and maintain their cluster structures and most of this overhead related to cluster formation and CH selection. So, in this paper we consider the overhead of the cluster formation and cluster head selection as the control message overhead. However, the average control message overhead is the count of total control message received by each vehicle in the network at cluster formation and CH selection procedures. Figure 7 shows the average control message overhead of MOSIC, N-Hop, AMACAD and ASPIRE at different velocities. Compared with above-mentioned clustering algorithm, MOSIC performs better in terms of control overhead. In MOSIC, each vehicle creates a control message during every channel interval and broadcast it to its single-hop neighbors to calculate the relative mobility between vehicle and its neighbors. This condition is equal to all above-mentioned clustering algorithms. But because of high stability of cluster structure in MOSIC, with more CH duration and low CH change rate, the control message to reestablish the clusters structure and CH selection will be reduced and consequently the control overhead will be decreased.

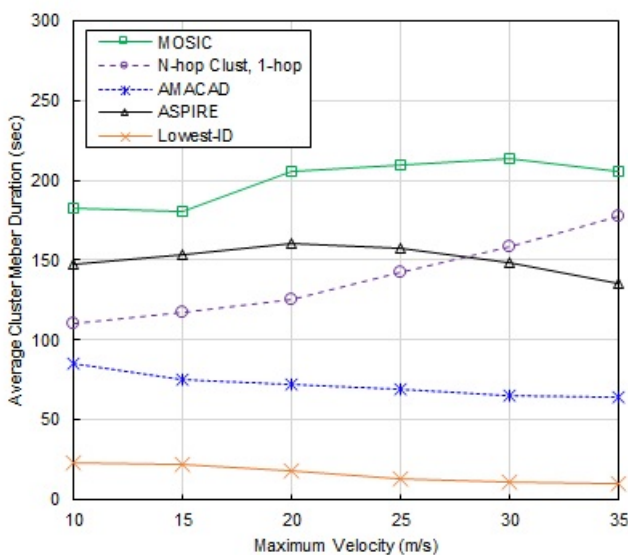


Fig. 4: Average CM Duration

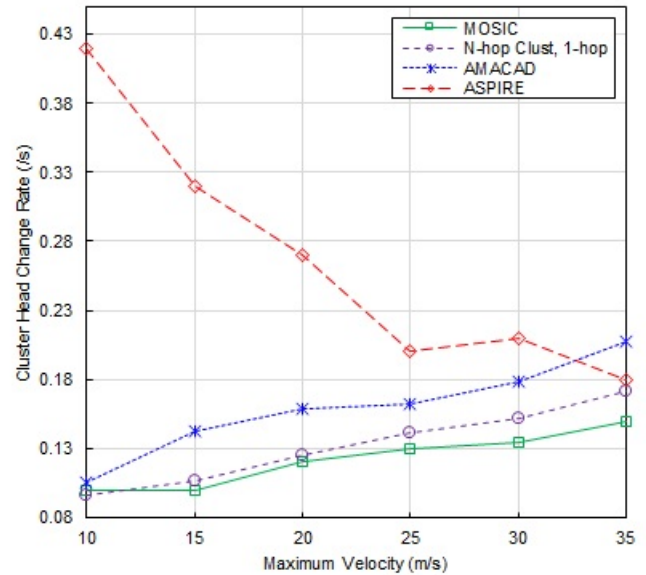


Fig. 5: Average CH Changes Rate

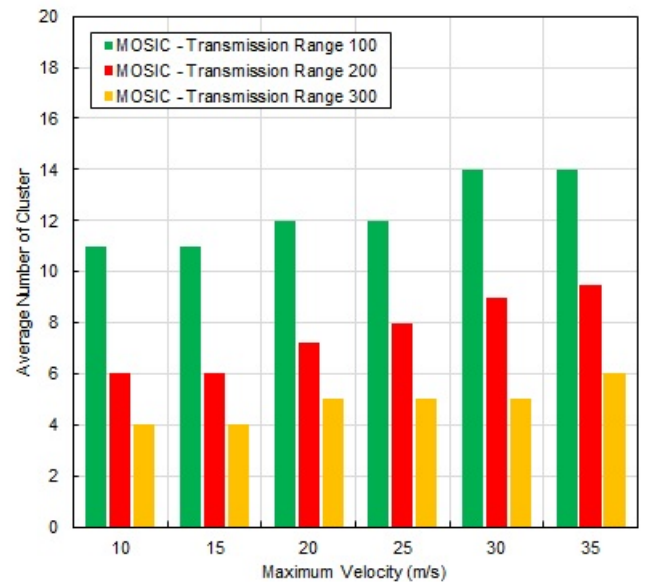


Fig. 6: Average Number of clusters

VI. CONCLUSION

Clustering mechanism is one of existence organizing mechanism which designed to adapt to the VANET environment. In this study, a mobility-aware and single-hop clustering scheme (MOSIC) was proposed. The MOSIC is based on the changes in the relative mobility of the vehicles, which is calculated by finding the average of the relative velocity, the nodal degree and relative distance of all the same direction neighbors. It used Gauss-Markov mobility model to predict the vehicle next location and based on the vehicle's location and its predicted location, relative distance will be calculated and consequently

relative mobility can be obtained. The MOSIC simulated on NS-3 and its performance compare to some clustering approach. Simulation indicate that the clustering of MOSIC outperforms than N-Hop, AMACAD, ASPIRE and Lowest-ID clustering in terms of CH duration, CM duration, CH change rate metrics and Control Message Overhead at various vehicle velocity scenarios. As future work, we aim to investigate the use of MOSIC in urban traffic scenarios and design the efficient routing protocol based on this scheme.

REFERENCES

- [1] Zhang, Z., Boukerche, A. and Pazzi, R.: "A novel multi-hop clustering scheme for vehicular ad-hoc networks", Proceedings of the 9th ACM international symposium on Mobility management and wireless access, New York (2011), 19-26.
- [2] Bali, R., Kumar, N. and Rodrigues, J.: "Clustering in vehicular ad hoc networks: Taxonomy, challenges and solutions", Vehicular Communications, 1, 3 (2014), 134-152.
- [3] Liang, B. and Haas, Z.: "Predictive distance-based mobility management for PCS networks", IEEE INFOCOM '99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. New York (1999), 1377-1384.
- [4] Basu, P., Khan, N. and Little, T. (n.d.): "A mobility based metric for clustering in mobile ad hoc networks". Proceedings 21st International Conference on Distributed Computing Systems Workshops, Mesa, AZ (2001), 413-418.
- [5] Souza, E., Nikolaidis, I. and Gburzynski, P.: "A New Aggregate Local Mobility (ALM) Clustering Algorithm for VANETs". 2010 IEEE International Conference on Communications, Cape Town, South Africa (2010), 1-5.
- [6] Lin, C. and Gerla, M.: "Adaptive clustering for mobile wireless networks". IEEE J. Select. Areas Commun., 15, 7 (97), 1265-1275.
- [7] Selvam, R. P., Palanisamy, V.: "Stable and flexible weight based clustering algorithm in mobile ad hoc networks", international Journal on Computer Science and information Technology, 2, 2 (2011), 824-828.
- [8] Yu, J. and Chong, P.: "3hBAC (3-hop between adjacent clusterheads): a novel non-overlapping clustering algorithm for mobile ad hoc networks", IEEE Pacific Rim Conference on Communications Computers and Signal Processing, PACRIM (2003), 318-321.
- [9] Yadav, N.S., Deosarkar, B.P., Yadov, R.P.: "A low control overhead cluster maintenance scheme for mobile ad hoc networks", International Journal of recent trends in engineering, 1, 1 (2009), 1-9.
- [10] Bentalab A., Boubetra A., Harous S.: "Survey of clustering schemes in mobile ad hoc networks", Communication Networks, 5, 2 (2013), 8-14.
- [11] Ni, M., Zhong, Z. and Zhao, D.: "MPBC: A Mobility Prediction-Based Clustering Scheme for Ad Hoc Networks". IEEE Trans. Veh. Technol., 60, 9 (2011), 4549-4559.
- [12] Xu, Y., Bien, S., Mori, Y., Heidemann, J., Estrin, D.: "Topology control protocols to conserve energy in wireless ad hoc networks", center for Embedded Network Sensing, 2003, 1-18.
- [13] Eichler, S.: "Performance Evaluation of the IEEE 802.11p WAVE Communication Standard", IEEE 66th Vehicular Technology Conference, Baltimore (2007), 2199-2203.
- [14] Shea, C., Hassanabadim B., Valaee, S.: "Mobility-Based Clustering in VANETs Using Affinity Propagation", IEEE Global Telecommunications Conference, Honolulu (2009), 1-6.
- [15] Hafeez, K.A., Zhao, L., Mark J.W., Shen, X., Niu, Z.: "Distributed Multichannel and Mobility-Aware Cluster-Based MAC Protocol for Vehicular Ad Hoc Networks", IEEE Transactions on Vehicular Technology, 62, 8 (2013), 3886-3902.
- [16] Lin, C. R., Gerla, M.: "Adaptive clustering for mobile wireless networks". IEEE Journal on Selected Areas in Communications, 15, 7 (1997), 1265-1275.
- [17] Morales, M., Hong, C. and Bang, Y.: "An Adaptable Mobility-Aware Clustering Algorithm in vehicular networks". 13th Asia-Pacific Network Operations and Management Symposium, Taipei (2011), 1-6.
- [18] Network simulator 3 (ns-3), Discrete Event Network Simulator, Sep. 2015. [Online]. Available: <http://www.nsnam.org>.
- [19] Mohammad, S. and Michele, C.: "Using traffic flow for cluster formation in vehicular ad-hoc networks". IEEE Local Computer Network Conference, Denver (2010), 631-636.
- [20] Su, H. and Zhang, X.: "Clustering-Based Multichannel MAC Protocols for QoS Provisionings Over Vehicular Ad Hoc Networks". IEEE Transactions on Vehicular Technology, 56, 6 (2007), 3309-3323.
- [21] Gupta, N., Prakash, A., Tripathi, R.: "Medium access control protocols for safety applications in Vehicular Ad-Hoc Network: A classification and comprehensive survey". Vehicular Communications, 2, 4 (2007), 223237.
- [22] Cheng, H. T., Shan, H., Zhuang, W.: "Infotainment and road safety service support in vehicular networking: From a communication perspective". Mechanical Systems and Signal Processing, 25, 6 (2011), 20202038.
- [23] A. Koulakezian. (2011, August). ASPIRE: Adaptive Service Provider Infrastructure for VANETs. Master thesis at The University of Toronto. [Online]. Available: <http://hdl.handle.net/1807/29581>.

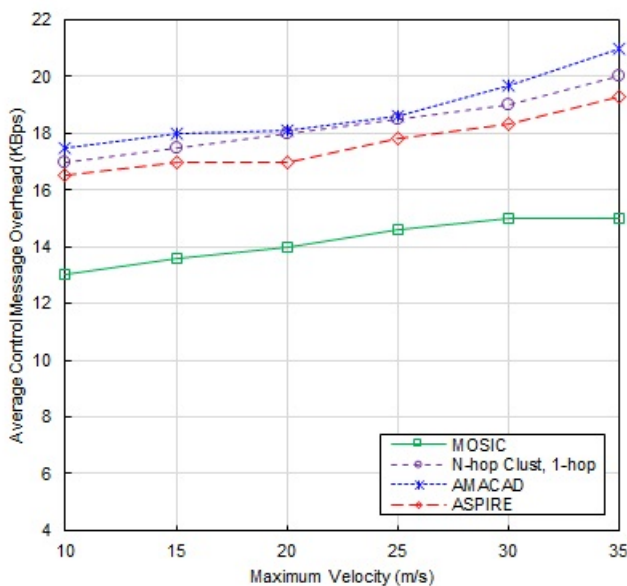


Fig. 7: Average Control Message Overhead

Peak-to-Average Power Ratio Reduction based Varied Phase for MIMO-OFDM Systems

Lahcen Amhaimar¹, Saida Ahyoud²
and Adel Asselman¹

¹Optics and Photonics Team

²Information Technology and Systems Modeling Team

Abdelmalek Essaadi University,

Faculty of Science

Tetuan, Morocco

Elkhaldi Said

Electronic and Instrumentation Team,

Abdelmalek Essaadi University,

Faculty of Science

Tetuan, Morocco

Abstract—One of the severe drawbacks of orthogonal frequency division multiplexing (OFDM) is high Peak-to-Average Power Ratio (PAPR) of transmitted OFDM signals. During modulation the sub-carriers are added together with same phase which increases the value of PAPR, leading to more interference and limits power efficiency of High Power Amplifier (HPA), it's requires power amplifier's (PAs) with large linear operating ranges but such PAs are difficult to design and costly to manufacture. Therefore, to reduce PAPR various methods have been proposed. As a promising scheme, partial transmit sequences (PTS) provides an effective solution for PAPR reduction of OFDM signals. In this paper, we propose a PAPR reduction method for an OFDM system with variation of phases based on PTS schemes and Solid State Power Amplifiers (SSPA) of Saleh model in conjunction with digital predistortion (DPD), in order to improve the performance in terms of PAPR, the HPA linearity and for the sake of mitigating the in-band distortion and the spectrum regrowth. The simulation results show that the proposed algorithm can not only reduces the PAPR significantly, but also improves the out-of-band radiation and decreases the computational complexity.

Keywords—OFDM; MIMO; PAPR; PTS; HPA; GA

I. INTRODUCTION

A combination of multiple-input multiple-output (MIMO) with orthogonal frequency division multiplexing (OFDM) has lead to significant advancement in wireless communication systems. It has been receiving a great deal of attention as a solution of high-quality service for next generation. OFDM systems has been adopted by several emerging wireless applications, such as WiMAX and digital video broadcasting/digital audio broadcasting (DVB/DAB) [1], thanks to its robustness over frequency selective fading channels and high bandwidth efficiency [2]. Still, some challenging issues remain unresolved in the design of the OFDM systems such as Peak to Average Power Ratio (PAPR) in the transmission system. The high PAPR reduces the efficiency of OFDM systems by introducing the intermodulation distortion and undesired out-of-band radiation, due to, the nonlinearity of the high power amplifier (HPA). Thus, it is highly desirable to improve PAPR performance of the signal. In the literature, many PAPR reduction techniques have been proposed and each of them has their own advantages and disadvantages [3], such as interleaving [4], clipping, companding [5], selective mapping

(SLM) [6], [7], partial transmit sequence (PTS) [8]–[10], tone reservation [11], coding technique [12], adaptive mode with low complexity [13], signal set expansion [14], and active constellation expansion [15].

This paper investigates the performance of PAPR reduction in OFDM and MIMO-OFDM system. An optimization with the same phase weighting will be proposed, and throughput results are presented for IEEE 802.11a and IEEE 802.16 standard.

The remainder of this paper is organized as follows. Section 2, present the basic concept of OFDM system, such as OFDM signals, definition and measurement of PAPR of OFDM signal, and PTS algorithm is going to be showed in this study. Section 3 introduces the principles of the proposed system. In Section 4, the performances of proposed method are discussed, and the simulated results are going to be stated. Finally, in section 5, a conclusion is drawn.

II. CHARACTERISTICS OF OFDM SIGNALS

A. PAPR in OFDM Signals

In OFDM system, the baseband operations at the transmitter include mapping the information data bit stream to symbols according to a certain modulation scheme, such as M-PSK or M-QAM, to create a vector of complex-valued symbols, $X = [X_0, X_1, \dots, X_{N-1}]^T$. The data streams transmitted simultaneously by subcarriers. Each of the sub-carriers is independently modulated and multiplexed. Then, an OFDM signal is formed by summing all the N modulated independent subcarriers which are of equal bandwidth. The IFFT generates the ready-to-transmit OFDM signal. The sub-carriers are chosen to be orthogonal so that the adjacent sub-carriers can be separated. In the discrete time domain, the mathematical representation of baseband OFDM signal $x[n]$ can be expressed as

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi nk}{LN}}, \quad 0 \leq n \leq LN - 1 \quad (1)$$

where X_k is the complex symbol transmitted on the kth subcarrier, N is the size of the IFFT and L is the oversampling factor.

The PAPR of an OFDM signal is defined as the ratio of the maximum to the average power of the signal, as follows

$$PAPR\{x[n]\} = \frac{\max\{|x[n]|^2\}}{E\{|x[n]|^2\}}, 0 \leq n \leq LN - 1 \quad (2)$$

where $E\{\cdot\}$ denotes the expected value operation.

The PAPR for the discrete-time baseband signal $x[n]$ may not be the same as that for the continuous time baseband signal $x(t)$. In fact, the PAPR for $x[n]$ is lower than that for $x(t)$, simply because $x[n]$ may not have all the peaks of $x(t)$. Fortunately, to avoid this problem, the oversampling is usually employed. It is shown in [3], [16], that an oversampling factor $L=4$ is sufficient to approximate the real PAPR results.

The complementary cumulative distribution function (CCDF) of PAPR provides information about the percentage of OFDM signals that have PAPR above a particular level. It denotes the probability that the PAPR of an OFDM symbol exceeds the given threshold $PAPR_0$, which can be expressed as

$$CCDF(N, PAPR_0) = Pr\{PAPR > PAPR_0\} \quad (3)$$

B. PAPR in MIMO-OFDM

We define the PAPR of a MIMO-OFDM signal as the maximum of the PAPRs among all the parallel transmit antenna branches. PAPR at the one transmit antenna is defined as the ratio of the peak power to the average power of an OFDM signal in that branch. The MIMO-OFDM PAPR system can be expressed as [17]

$$PAPR_{MIMO} = \max PAPR_i, i = 1, \dots, N_T \quad (4)$$

where N_T is the number of transmission antennas.

C. PTS for PAPR Reduction

The block diagram of OFDM transmitter with partial transmit sequence (PTS) technique is shown in Fig.1 [8]. The PTS approach partitions an input data block of N complex symbols into V disjoint sub-blocks as follows

$$X = [X^0, X^1, X^2, \dots, X^{V-1}]^T \quad (5)$$

where the sub-carriers in each sub-block are weighted by phase rotations, and all the subcarriers which are occupied by the other sub-blocks are set to zero. The various partitioning methods to divide complex symbols into disjoint sub-blocks are proposed in [18], these include pseudorandom, adjacent and interleaved partitioning schemes. In PTS, shown in Fig. 1 each partitioned sub-block is multiplied by a corresponding complex phase factor to produce the sequences

$$X = \sum_{v=1}^V X^v b^v, b^v = e^{j\varphi^v}, v = 1, 2, \dots, V \quad (6)$$

Afterward taking its IFFT to yield [8], [10]

$$x = IFFT \left\{ \sum_{v=1}^V b^v X^v \right\} = \sum_{v=1}^V b^v \cdot IFFT\{X^v\} = \sum_{v=1}^V b^v x^v \quad (7)$$

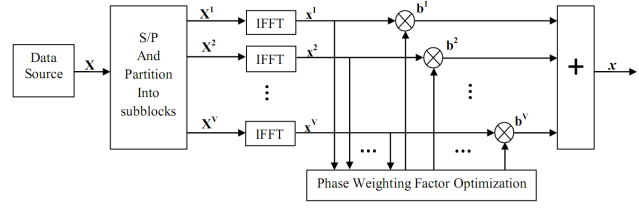


Fig. 1: Block diagram of PTS technique

where x denotes the candidate sequence.

The phase vector is chosen so that the PAPR can be minimized [8], which is shown as

$$[b^1, \dots, b^V] = \arg \min_{[b^1, \dots, b^V]} \left(\max_{n=0,1,\dots,N-1} \left| \sum_{v=1}^V b^v x^v[n] \right| \right) \quad (8)$$

In the practical application of PTS, a set of phase weighting factors is usually selected for generating phase weighting sequences. To assume that there are W allowed phases weighting factors in this set. Without any loss of performance, we can set phase weighting factor for the first sub-block to one and observe that there are $(V-1)$ sub-blocks to be optimized. To match the optimal phase weighting sequence for each input data sequence, W^{V-1} possible combinations should be checked, and then the candidate sequence with the minimum PAPR is selected for transmitting [8], [19].

III. PROPOSED APPROACH FOR PAPR REDUCTION

In this Section we are going to describe the proposed method, it is based on combining signal sub-blocks which are phased shifted by the same phase factors to generate multiple candidate signals, so as to select the ideal PAPR signal. The steps involved in the generation and process of phase weighting factors are summarized as follows:

- As a first step, we find the PAPR of original signal and set it as $PAPR_{min}$
- Then, the input data sequence is partitioned into V sub-blocks $X_{(v)}$ as in equation (5).
- Next, Then, we set the search space of phase weight factor ($angle[0, 2\pi]$) as $(0^\circ, 10^\circ, 20^\circ, \dots, 360^\circ)$ with an increment of 10° , and start the optimization of each sub-block with the same phase factor φ .

$$X_{(v)} \cdot \varphi, \text{ and } \varphi = e^{j\theta} \quad (9)$$

where $v = 1, \dots, V$. and $\theta = 0^\circ, 10^\circ, 20^\circ, \dots, 360^\circ$

- After that, compute the PAPR of the combined signal x (Eq. 10). If $PAPR > PAPR_{min}$, switch φ to the next phase. Otherwise, update $PAPR_{min} = PAPR$.

$$x = IFFT \left\{ \sum_{v=1}^V X_{(v)} \cdot \varphi \right\} \quad (10)$$

- The algorithm continues in this fashion until all the 36 phase factor are searched. Then, we retain the signal

with the PAPR minimum and the set of optimal phase factors.

The proposed approach is used to find a suitable phase factor set that minimizes the PAPR in a transmitted signal without exhaustive search. It decreases the computational load of the PTS technique by searching a small piece of a set of possibilities instead of the whole set as in the classical technique. Finally, The transmitted signal $x(t)$ can be linearly amplified by virtue of the predistorter, a technique that corrects the nonlinearity by compensate nonlinear distortion of the power amplifier.

IV. SIMULATION RESULTS

In this section, we numerically evaluate the performance of the PAPR reduction scheme by extensive simulations were performed using the CCDF. The systems is carried out based on IEEE 802.11a and IEEE 802.16 standard with $N=64$ and $N=256$ subcarriers successively. 10^4 and 10^5 OFDM symbols are randomly chosen for the simulation with QPSK modulation and set the oversampling factor $L=4$ for selecting and estimating the signal with minimum PAPR. The simulation parameters are also documented in the Table 1

TABLE I: Simulation Parameters

Parameter	Standards	
	IEEE 802.11a	IEEE 802.16
FFT size	64	256
User carriers	52	200
Pilot carriers	4	8
Number of null/guard band subcarriers	12	56
Cyclic prefix or guard time	1/4, 1/8, 1/16, 1/32	
Modulation	QPSK, 3/4	
Oversampling Factor	$L=4$	
Number of subblocks	$V = 2, 4, 6, 8$	
Number of phase factors	$W = 4 \{1, -1, j, -j\}$	
Channel bandwidth	3.5MHz	
Noise Channel	AWGN	
SNR	30dB	
Number of generations	$G = 10$	
Population	$P=5$	
Crossover rate	$CR=1.0$	
Mutation rate	$MR=0.05$	

A. PAPR Reduction

In Fig. 2, the CCDF of PAPR are obtained by simulation, in which 10^5 OFDM symbols are randomly generated, four sub-blocks ($V=4$) are used and the sets of phase weighting factors are varied between $[0, 2\pi]$. The simulation results show that the proposed approach can reduce the PAPR significantly. It is shown in Fig. 2 that the PAPR reduction achieved with the new algorithm was 6.6 dB compared with original OFDM (10.7 dB) when $CCDF = 10^{-3}$.

Figure 3 illustrate the effectiveness of the proposed scheme as the number of sub block varies. The results of the simulation are based on the transmission of randomly generated OFDM symbols. It is seen that the PAPR performance improves as the number of sub blocks increases with $V= 2, 4, 6,$ and 8 . So, by increasing the number of sub-blocks, the proposed system requires low transmitted power.

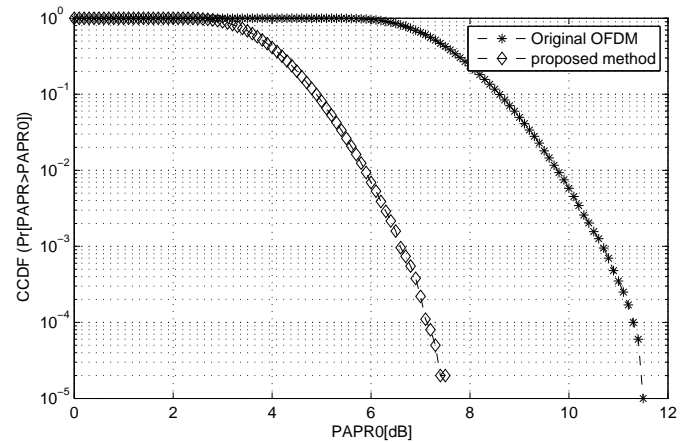


Fig. 2: PAPR reduction with proposed method

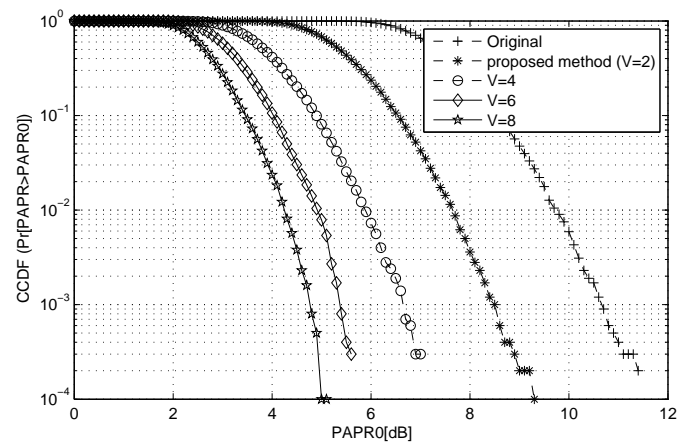


Fig. 3: PAPR performance with new technique when the number of sub-blocks varies

Next, figure 4 show the performance comparison between the proposed scheme and existing PTS and SLM in terms of CCDF. The curves of the simulations results are given for 10^4 OFDM symbols, in which four sub-blocks ($V=4$) are used and the sets of phase weighting factors for PTS are $\{1, -1, j, -j\}$ ($W=4$).

From Fig.4, it is very clear that all schemes can reduce the PAPR largely in OFDM system. However, their performances of the PAPR reduction are different.

For example, when $CCDF = 10^{-3}$, the PAPRs are 5.23 dB, 6.6 dB, 7 dB and 10.7dB for the PTS, proposed scheme, SLM scheme and original OFDM signals, respectively. Although the PTS has better PAPR than the proposed technique, the computational load of the PTS is larger than our approach.

Finally, we consider PAPR reduction performance in MIMO-OFDM system. As shown in Fig. 5, the CCDF of the PAPR of original and optimized signals have been given for randomly generated QPSK symbols. MIMO-OFDM system uses two antennas based on Alamouti scheme with $N=256$

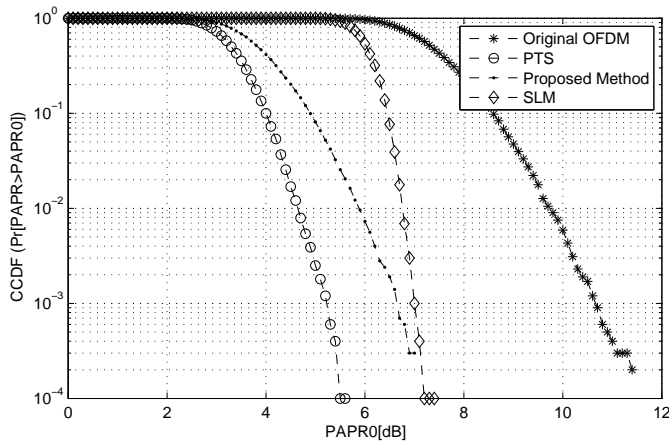


Fig. 4: Comparisons of CCDF based on different PAPR reductions schemes.

carriers per antenna, among those, 56 unused free carriers (as in IEEE802.16 WiMAX standard, Table 1). We set the oversampling factor $L = 4$. This approach achieves significant reduction in PAPR for MIMO OFDM systems. As we can see, the PAPR in transmitted signal is 7.5 dB for the proposed method compared with original MIMO-OFDM PAPR 11.25 dB at the 10^{-3} probability level.

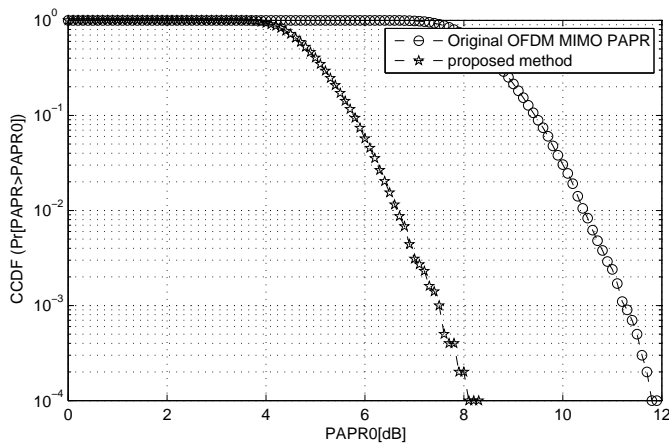


Fig. 5: The PAPR reduction performances in MIMO-OFDM systems

B. Complexity Performances

Figure 6 shows the CCDF curves of OFDM system with different PAPR reduction techniques which are the proposed method, the original PTS technique, and the Genetic Algorithm with PTS technique (GA-PTS). The basic system parameters for the simulations are summarized in Table I. Although the PTS and the GA-PTS techniques have better PAPR than the proposed approach, the computational loads of these methods are larger than our method.

The GA is an optimization method serves as a solution to find a suitable phase factor set that minimizes the PAPR in

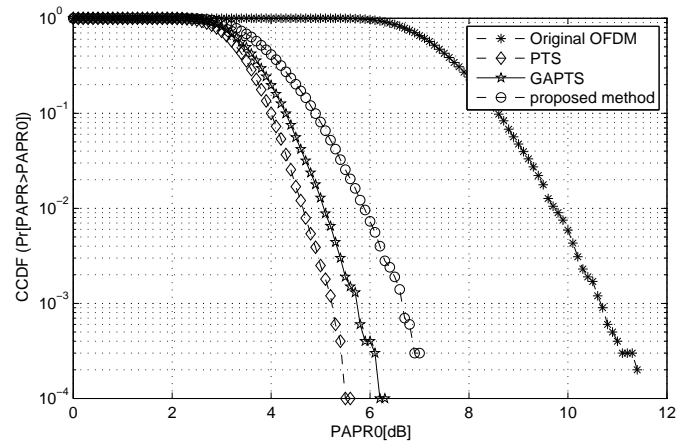


Fig. 6: Comparison of the $PAPR_0$ (dB) versus CCDF in OFDM systems for proposed method, GA-PTS and original PTS

a transmitted signal [20], [21]. It decreases the computational load of the proposed technique by searching a small piece of a set of possibilities instead of the whole set as in the classical technique (PTS). The GA has a good convergence and high explorative ability. However that will result in higher complexity and slower convergence rate, usually we need a decent sized population and a lot of generations to guarantee the accuracy and explore the entire space, which takes more time for convergence.

Table II shows the number of search values for OFDM, using different schemes to find the phase factors. It is shown that when $Pr(PAPR > PAPR_0) = 10^{-3}$, the $PAPR_0$ of the PTS is 5.23dB with exhaustive searching number $W^V = 4^4 = 256$ (number of possible phase factor combinations). Values close to PTS scheme can be obtained by GA-PTS with $P \times G = 5 \times 10 = 50$ searches, where P is the maximum size of the population and G is the generation. while proposed approach is 6.64dB with a small computational load, 36 iteration.

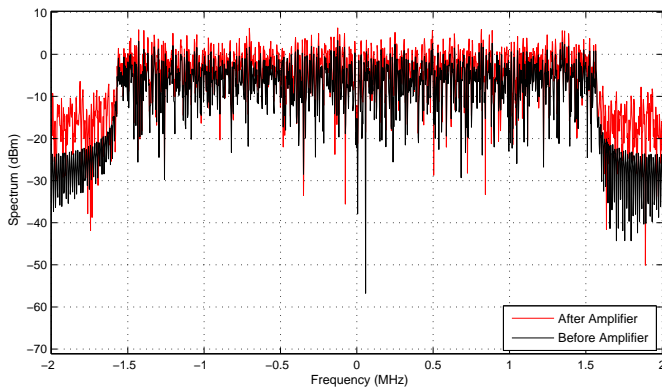
TABLE II: Computational Complexity of the Different Methods for $CCDF = 10^{-3}$

Method ($V=4, W=4$)	Number of search	PAPR (dB)
Original OFDM	0	10.7
PTS	$W^V = 4^4 = 256$	5.23
GA-PTS	$P \times G = 50$	5.7
Proposed method	36	6.64

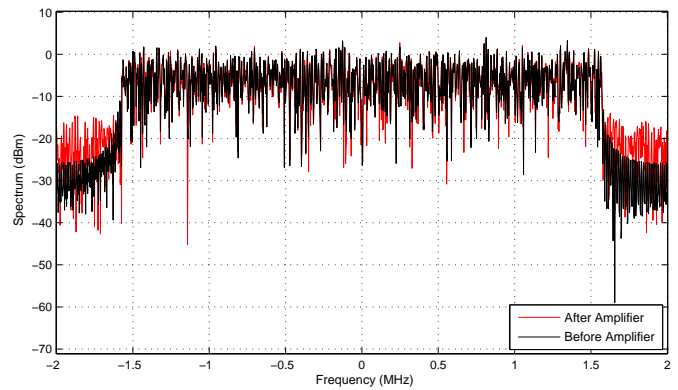
As a result, the proposed method with 36 searches was only 1.4 dB and 0.9dB higher than optimum PTS method and GA-PTS respectively, in OFDM systems. On the contrary, our approach has a low search complexity when compared with each method alone.

C. Spectrum Performances

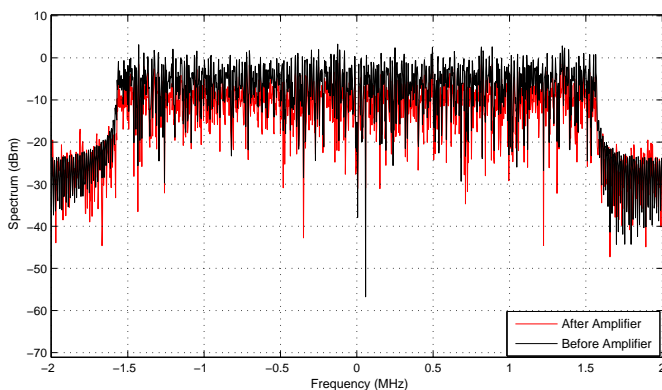
This analysis is carried out in the presence of two models of high power amplifier (HPA), are Solid State Power Amplifiers (SSPA) of Saleh model [22] and Rapp model [23] in conjunction with digital predistortion (DPD) as linearization technique.



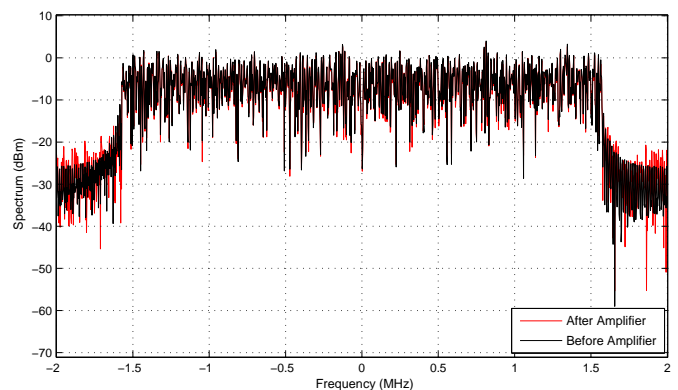
(a) Without PAPR reduction



(b) PAPR reduction with DPD



(c) Without PAPR reduction



(d) PAPR reduction with DPD

Fig. 7: Spectrum performance comparison of an OFDM signal with and without proposed method with severe nonlinearity. (a,b) saleh model and (c,d) rapp model.

Fig.7 show the signal of spectrum performance before and after passing through HPA model for conventional OFDM system and OFDM system with PAPR reduction method and DPD. The results are obtained for 1dB below the input power that causes amplifier saturation.

Figure 7.a and fig.7.b shows the frequency spectra when the HPA is working near the saturation region based saleh model. It is clearly seen that the proposed method with DPD has the ability to reduce the spectral regrowth around each of the two bands, the transmitted spectrum are kept below -15 dBm.

Next, fig.7.c and fig.7.d shows the spectrum with and without proposed technique with Rapps solid state power amplifier model, Comparing the two spectra, we can view a few change on spectrum due to characteristics of rapp model, it does not apply a phase change to the input signal. Therefore, the HPA power efficiency will be much higher in this case compared with saleh model.

V. CONCLUSION

In this paper, a phase weighting method based PTS scheme with low computational complexity is proposed followed by a

predistortion technique in order to reduce the PAPR, improves the spectrum efficiency and the PA's linearity simultaneously. The sets of phase weighting factors are varied between $[0, 2\pi]$ to search for a good set of phase factors, and the optimization was by one phase to obtain the desirable PAPR reduction. The simulation results show that the performances given by our approach, the PAPR reduction is better than original OFDM and maintained close to PAPR values of the PTS and G-PTS technique while providing a low computational load. We have also shown that our proposed method provides a good PAPR reduction in MIMO-OFDM at the transmitter. Moreover, the joint scheme using our approach with DPD decreased the transmitted spectrum regrowth at 1dB HPA backoff below amplifier saturation point. That's means when the HPA is working near or in the saturation region, the proposed PAPR reduction helps to improve the HPA efficiency and linearization.

REFERENCES

- [1] Amhaimar, L., Ahyoud, S. and Asselman, A. "PAPR reduction performance for WiMAX OFDM systems;" In Proceeding IEEE of The Third International Workshop on RFID And Adaptive Wireless Sensor Networks, DOI: 10.1109/RAWSN.2015.7173274, 2015.

- [2] Wu, Y., and Zou, W. Y. "Orthogonal frequency division multiplexing: A multi-carrier modulation scheme," IEEE Transaction on Consumer Electronics, Vol. 41, 392-399, 1995
- [3] Jiang, T. and Wu, Y. "An overview: Peak-to-average power ratio reduction techniques for OFDM signals," IEEE Transaction on Broadcasting, Vol. 54, 257-268, 2008.
- [4] Sakran, H., Shokair, M. and Elazm, A. A., "An efficient technique for reducing PAPR of OFDM system in the presence of nonlinear high power amplifier," Progress In Electromagnetics Research C, Vol. 2, 233-241, 2008.
- [5] Sakran, H., Shokair, M. and Elazm, A. A. "Combined interleaving and companding for PAPR reduction in OFDM systems," Progress In Electromagnetics Research C, Vol. 6, 67-78, 2009.
- [6] Bauml, R.W, Fischer, R.F.H, and Huber, J.B. "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," Electron. Lett., Vol. 32, No. 22, 2056-2057, 1996
- [7] Ying, Liang, H. "Integrating CE and modified SLM to reduce the PAPR of OFDM systems," Wireless Personal Communication. doi:10.1007/s11277-014-2036-0, 2015
- [8] Muller, S.H. and Huber, J. B., "OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," IEE Electronics Letters, Vol. 33, No. 5, 368-69, 1997.
- [9] Cimini, L.J., and Sollenberger, N. R., "Peak-to-Average Power Ratio Reduction of an OFDM Signal Using Partial Transmit Sequences," IEEE Commun. Lett, Vol. 4, No. 3, 86-88, 2000.
- [10] Varahram, P., Mohammady, S., and Ali, B. M. "A robust peak-to-average power ratio reduction scheme by inserting dummy signals with enhanced partial transmit sequence in OFDM systems," Wireless Personal Communications, 72(2), 1125-1137, 2013.
- [11] Yang, M., and Shin, Y., "PAPR reduction using tone reservation for NC-OFDM transmissions," Ubiquitous and Future Network (ICUFN). In 4th International Conference, 454-455, 2012.
- [12] Jones, AE., Wilkinson, TA. and Barton, SK., "Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes," Electron Lett, Vol. 30, No. 25, 2098-2099, 1994.
- [13] Rajbanshi, R., Wyglinski, A. M., & Minden, G. J. "Adaptive-mode peak-to-average power ratio reduction algorithm for OFDM-based cognitive radio." In Proceeding IEEE 64th Vehicle Technology Conference, Vol. 6, pp. 1350-1354, 2006.
- [14] Han, S. H. and Lee, J. H., "Peak-to-average power ratio reduction of an OFDM signal by signal set expansion," IEEE International Conference Communication, Vol. 2, 867-871, 2004.
- [15] Zhou, Y., and Jiang, T., "A novel clipping integrated into ACE for PAPR reduction in OFDM systems," In International Conference on Wireless Communications and Signal Processing, 1-4, 2009.
- [16] Tellado J, "Peak to average power reduction for multicarrier modulation," Ph.D. Thesis, Stanford University, Calif, USA, September 1999.
- [17] Fischer, R.F.H., Hoch, M., "Peak-to-average power ratio reduction in MIMO OFDM," IEEE International Conference on Communications (ICC 2007), Glasgow, United Kingdom, pp. 762-767, June 2007.
- [18] Kang, S.G., Kim, J.G. and Joo, E.K., "A novel sub-block partition scheme for partial transmits sequence OFDM," IEEE Trans. Broadcast. 45, pp. 333-338, 1999.
- [19] S.H. Han and J.H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission, IEEE Wireless. Communications, vol. 12, no. 2, pp. 56-65, April 2005.
- [20] Lixia, M, Murrioni, M, Popescu, V, "PAPR reduction in multicarrier modulations using genetic algorithms," 12th International Conference on Optimization of Electrical and Electronic Equipment; Brasov, Romania. New York, NY, USA: IEEE. pp. 938-942, May 2010.
- [21] Liang, H, Chen, Y, Huang, Y, Cheng, C, "A modified genetic algorithm PTS technique for PAPR reduction in OFDM systems," In: 15th Asia-Pacific Conference on Communications; pp. 182-185, 10 October 2009, Shanghai, China. New York, NY, USA: IEEE.
- [22] ADEL, A, Saleh, M, "Frequency-independent and frequency-dependent non-linear models of TWT amplifiers," IEEE Trans. Commun., vol. 29, No. 11, 1715-1720, 1981.
- [23] Rapp, C., "Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadcasting System," in Proceedings of the Second European Conference on Satellite Communications, Liege, Belgium, pp. 179-184, Oct. 22-24 .1991

Solving Nonlinear Eigenvalue Problems using an Improved Newton Method

S.A Shahzadeh Fazeli
Parallel Processing Laboratory
Faculty of Mathematics
Yazd University
Yazd, P. O. Box 89195-741, Iran

F. Rabiei
Parallel Processing Laboratory
Faculty of Mathematics
Yazd University
Yazd, P. O. Box 89195-741, Iran

Abstract—Finding approximations to the eigenvalues of nonlinear eigenvalue problems is a common problem which arises from many complex applications. In this paper, iterative algorithms for finding approximations to the eigenvalues of nonlinear eigenvalue problems are verified. These algorithms use an efficient numerical approach for calculating the first and second derivatives of the determinant of the problem. Here we present and examine a technique for solving nonlinear eigenvalue problems using Newton method. Computational aspects of this approach for a nonlinear eigenvalue problem are analyzed. The efficiency of the algorithm is demonstrated using an example.

Keywords—nonlinear eigenvalue problems; Newton method; LU-decomposition; refined eigenvalues

I. INTRODUCTION

A method is presented in [5] for obtaining lower and upper bounds on eigenvalues and eigenfunctions for linear integral equations.

Another method is described in [6] for the calculation of the eigenvalues of general integral operators. Several classical results from functions of a complex variable and the theory of integral equations are combined with a recent technique for converting Fredholm integral equations into an initial-valued system of differential equations. The algorithm, which is based on a Cauchy system for the Fredholm determinants, is related to the Nyström method and results of Anselone and Atkinson become applicable.

Another method is presented in [7] for the calculation of the eigenvalues and eigenfunctions of complex-valued symmetric kernels which occur in laser theory. The method combines some classical results of integral equations and complex variables with a recent technique for transforming Fredholm integral equations into a Cauchy system of differential equations.

In this paper, a numerical method for solving the following eigenvalue problem is proposed. Here the method and the notations presented in [4] is used to simplify our method.

Let $D(\lambda)$ be a given n -by- n matrix that is a nonlinear function of the spectral parameter λ . It is required to find the values $\lambda \in C$ (called the eigenvalues) such that the equations

$$x^*D(\lambda) = 0, \quad D(\lambda)y = 0, \quad (1)$$

have nontrivial solutions $x, y \in C^n$.

Here, the asterisk in the superscript indicates the Hermitian adjoint operation. Both problems in (1) have the same desired values of λ that solve the equation

$$f(\lambda) \equiv \det D(\lambda) = 0. \quad (2)$$

In what follows, it is assumed that the entries of $D(\lambda)$ are sufficiently smooth functions of λ varying in a certain domain. This process is an improved Newton method as applied to finding a simple real eigenvalue considered as a root of the corresponding nonlinear scalar equation (2); however, in equation (2), the left-hand side is not expressed in an explicit form. Instead, it is proposed an algorithm for calculating the values of $f(\lambda), f'(\lambda)$ at a fixed λ ; to this end, the LU-decomposition of $D(\lambda)$ is used.

Moreover, the proposed algorithm, combined with the argument principle for analytic functions, makes it possible to find the number of eigenvalues belonging to a given domain G in the complex λ -plane, as well as to find initial approximations to all of these eigenvalues. The approximations found can then be refined using any of the available iterative methods; in particular, an improved Newton method can be applied.

II. CALCULATING $f(\lambda)$ AND $f'(\lambda)$

It is well known that, for any fixed λ , the matrix $D(\lambda)$ can be represented in the form

$$D(\lambda) = L(\lambda)U(\lambda), \quad (3)$$

where $L(\lambda)$ is a lower triangular matrix with unit diagonal and $U(\lambda)$ is an upper triangular matrix. It follows that

$$f(\lambda) = \det L(\lambda)\det U(\lambda) = \prod_{i=1}^n u_{ii}(\lambda),$$

Since the entries of the square matrix $D(\lambda)$ (and, hence, those of $U(\lambda)$)are differentiable functions of λ , that is

$$f'(\lambda) = \sum_{r=1}^n v_{rr}(\lambda) \prod_{i=1, i \neq r}^n u_{ii}(\lambda),$$

for any λ , here $v_{ii}(\lambda) = u'_{ii}(\lambda)$.

To find $v_{ii}(\lambda)$ equation (3) is differentiable with respect to λ . This yields

$$B(\lambda) = M(\lambda)U(\lambda) + L(\lambda)V(\lambda), \quad (4)$$

where

$$B(\lambda) = D'(\lambda), M(\lambda) = L'(\lambda), V(\lambda) = U'(\lambda),$$

and $v_{ii}(\lambda)$ are the entries in the matrix $V(\lambda)$. Thus, to calculate $f(\lambda), f'(\lambda)$, at a fixed $\lambda = \lambda_m$, it is necessary to find the matrices

$$\begin{aligned} D &= LU, \\ B &= MU + LV. \end{aligned} \quad (5)$$

This yields

$$\begin{aligned} f(\lambda_m) &= \prod_{i=1}^n u_{ii}(\lambda_m), \\ f'(\lambda_m) &= \sum_{r=1}^n v_{rr}(\lambda_m) \prod_{i=1, i \neq r}^n u_{ii}(\lambda_m). \end{aligned} \quad (6)$$

The entries of the matrices appearing in decomposition (5) can be calculated using the recursions

$$\begin{aligned} r &= 1, 2, \dots, n \\ u_{rk} &= d_{rk} - \sum_{j=1}^{r-1} l_{rj} u_{jk}, \\ l_{ir} &= [d_{ir} - \sum_{j=1}^{r-1} l_{ij} u_{jr}] u_{rr}^{-1}, \quad i = r + 1, \dots, n, \\ v_{rk} &= d_{rk} - \sum_{j=1}^{r-1} (m_{rj} u_{jk} + l_{rj} v_{jk}), \quad i = r + 1, \dots, n, \\ m_{ir} &= [d_{ir} - \sum_{j=1}^{r-1} (m_{ij} u_{jr} + l_{ij} v_{jr}) - l_{ir} v_{rr}] u_{rr}^{-1}, \\ i &= r + 1, \dots, n. \end{aligned}$$

This algorithm may be unstable and even incorrect if $u_{rr} = 0$ for some r . To avoid such occurrences, one uses permutations of the rows (and/or columns) of D in the process of its LU -decomposition; simultaneously, a pivot is chosen similarly to the Gaussian elimination. In this case, decomposition of (5) can be written as

$$\begin{aligned} PD &= LU, \\ PB &= MU + LV, \end{aligned}$$

where P is a permutation matrix; and $\det P = (-1)^q$, where q is the number of permutations. Thus, relations (6) take the form

$$\begin{aligned} f(\lambda_m) &= (-1)^q \prod_{i=1}^n u_{ii}(\lambda_m), \\ f'(\lambda_m) &= (-1)^q \sum_{r=1}^n v_{rr}(\lambda_m) \prod_{i=1, i \neq r}^n u_{ii}(\lambda_m). \end{aligned}$$

After that, this algorithm is used for calculating the derivatives on the basis of the LU -decomposition of $D(\lambda)$.

III. COMPUTATIONAL ASPECTS OF THE ALGORITHM

The argument principle was repeatedly used to solve various problems in which the number of eigenvalues belonging to a given domain must be determined. By assumption, the characteristic function (2) is analytic. Suppose that f has m zeros $\lambda_1, \dots, \lambda_m$ in G (with the multiplicities counted) and has no zeros on the boundary Γ of G .

It is well known from the argument principle that the number m is determined by the formula [1]

$$m = S_0 = \frac{1}{2\pi i} \int_{\Gamma} \frac{f'(\lambda)}{f(\lambda)} d\lambda. \quad (7)$$

Define the quantities

$$S_k = \frac{1}{2\pi i} \int_{\Gamma} \lambda^k \frac{f'(\lambda)}{f(\lambda)} d\lambda, \quad k = 0, 1, 2, \dots, \quad (8)$$

then, it can be shown that

$$\sum_{j=1}^m \lambda_j^k = S_k, \quad k = 0, 1, 2, \dots \quad (9)$$

Thus, if m and $S_k, (k = 1, 2, \dots, m)$ are known, then system (9) determines the zeros of function (2), that is, all the eigenvalues of problem (1) belonging to the given domain G .

Let us dwell on the computational aspects of this algorithm, namely, on the stage at which the quantities $S_k (k = 1, 2, \dots, m)$ are calculated. Thus, it is proposed to use the LU -decomposition of $D(\lambda)$ for calculating both $f(\lambda)$ and $f'(\lambda)$.

Without loss of generality, take the circle $G(\lambda^*, \rho)$ of radius ρ centered at λ^* as the domain G bounded by the contour Γ . The change of the variable

$$\lambda(t) = \lambda^* + \rho \exp(2\pi i t)$$

transforms integral (8) to the form

$$S_k = \int_0^1 \lambda(t)^k \rho \exp(2\pi i t) \frac{f'(\lambda(t))}{f(\lambda(t))} dt. \quad (10)$$

Partition the interval $[0, 1]$ into N equal subintervals and replace integral (10) by a quadrature (for instance, following [2], the rectangle rule can be used).

This yields

$$S_k = \frac{1}{N} \sum_{j=1}^N \lambda_j^k \rho \exp(i \frac{2\pi j}{N}) \frac{f'(\lambda_j)}{f(\lambda_j)}, \quad (11)$$

where

$$\lambda_j = \lambda^* + \rho \exp(i \frac{2\pi j}{N}).$$

Thus, formula (11) requires that only the values of $f(\lambda)$ and of its derivative on the boundary of G is calculated. This can be done by using decomposition (5). Then, using

representations (6), it can be rewritten the ratio $\frac{f'(\lambda)}{f(\lambda)}$ in the form

$$\frac{f'(\lambda_j)}{f(\lambda_j)} = \sum_{r=1}^n \frac{v_{rr}}{u_{rr}}. \quad (12)$$

In view of (12), the following formulas for calculating $S_k (k = 0, 1, \dots, m)$ is obtained:

$$S_k = \frac{1}{N} \sum_{j=1}^N (\lambda_j^k \rho \exp(i \frac{2\pi j}{N} \sum_{r=1}^n \frac{v_{rr}}{u_{rr}})). \quad (13)$$

Hence, using relations (13), the number $m = S_0$ of the eigenvalues belonging to G , as well as the right-hand side of equation (9) can be found. Following [4], the Newtons method for solving this system, may be applied which results in certain (in general, rather rough) approximations to all the eigenvalues are belonging to G .

IV. AN IMPROVED NEWTON METHOD(IM_NEWTON)

The use of iterative method for refining the rough approximations to the eigenvalues that were obtained by the algorithm described above is proposed.

Theorem 1: Let $p(x)$ be a real polynomial of degree $n \geq 2$, all zeros of which are real, $\xi_1 \geq \xi_2 \geq \dots \geq \xi_n$. Let α_1 be the largest zero of $p'(x): \xi_1 \geq \alpha_1 \geq \xi_2$. For $n = 2$, it is also required that $\xi_1 > \xi_2$. Then for every $z > \xi_1$, the numbers

$$z' = z - \frac{P(z)}{P'(z)},$$

$$y = z - 2 \frac{P(z)}{P'(z)},$$

$$y' = y - \frac{P(y)}{P'(y)}$$

are well defined and satisfy $\alpha_1 < y$, $\xi_1 \leq y' \leq z'$. It is readily verified that $n = 2$ and $\xi_1 = \xi_2$ imply $y = \xi_1$ for any $z > \xi_1$.

Suppose that an approximation λ_0 to the eigenvalue λ_* is given such that the Im_Newton method described in [3] can be initiated from λ_0 :

$$y_m = \lambda_m - 2 \frac{f(\lambda_m)}{f'(\lambda_m)}, \quad (14)$$

$$y'_{m+1} = y_m - \frac{f(y_m)}{f'(y_m)}. \quad (15)$$

At each step of iterative process (14), the values of $f(\lambda)$ and its derivatives at a specific λ are used. Therefore, to calculate these values, use of decomposition (5) and of relations (6) is made. As a result, process (14) and (15) takes the form

$$y_m = \lambda_m - 2 \left(\sum_{k=1}^n \frac{v_{kk}}{u_{kk}} \right)^{-1}, \quad (16)$$

$$y'_{m+1} = y_m - \left(\sum_{k=1}^n \frac{v_{kk}}{u_{kk}} \right)^{-1}. \quad (17)$$

Thus, the following algorithm for solving the nonlinear eigenvalue problem (1) is proposed.

Algorithm 1: Iterative process for refining the rough approximations

Step 1. Choose an initial approximation λ_* to the s th eigenvalue of problem (1).

Step 2. for $m = 0, 1, \dots$ until the required accuracy is attained do.

Step 3. Determine the entries u_{kk}, v_{kk} in decomposition (5).

Step 4. Calculate y_m and y'_{m+1} using formula (16) and (17).

Step 5. Let $\lambda_{m+1} = y'_{m+1}$.

Step 6. end for.

V. NUMERICAL EXAMPLE

The algorithm for calculating rough approximates of eigenvalues presented in section II was tested for the quadratic eigenvalue problem with the matrix

$$D(\lambda) = \lambda^2 A_0 + \lambda A_1 + A_2,$$

where

$$A_0 = \begin{bmatrix} 1 & 0.17 & -0.25 & 0.54 \\ 0.47 & 1 & 0.67 & -0.32 \\ -0.11 & 0.35 & 1 & -0.74 \\ 0.55 & 0.43 & 0.36 & 1 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0.22 & 0.02 & 0.12 & 0.14 \\ 0.02 & 0.14 & 0.04 & -0.06 \\ 0.12 & 0.04 & 0.28 & 0.08 \\ 0.14 & -0.06 & 0.08 & 0.26 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} -3.0475 & -2.1879 & -1.9449 & -2.8242 \\ -2.6500 & -2.4724 & -2.3515 & -2.1053 \\ -0.7456 & -0.6423 & -1.3117 & -0.1852 \\ -4.0500 & -3.0631 & -2.8121 & -3.7794 \end{bmatrix}.$$

First, the number of eigenvalues belonging to a given domain G was determined. In presented calculations, G was a circle $G(\lambda^*, \rho)$ centered at $\lambda^* = 0$.

The following radii were used: $\rho = 0.3, 1.0, 1.3$, and 3.0 . The results are presented in TABLE I. For each ρ , the number of eigenvalues belonging to G , the eigenvalues themselves (λ), and the number k of iteration steps required to calculate the eigenvalues to an accuracy of $\epsilon = 10^{-4}$ are shown.

The approximate eigenvalues found at this stage may not have the required accuracy. To refine these approximations, the

TABLE I. THE NUMBER OF EIGENVALUES

$\rho = 0.3$ $k = 2$	$\rho = 1.0$ $k = 68$	$\rho = 1.3$ $k = 25$	$\rho = 3.0$ $k = 77$
λ	λ	λ	λ
0.2423	0.6383	0.2383	2.3229
	0.2423	-0.8499	0.7967
	-0.3778	-1.2261	0.6383
	0.7967	0.3878	0.2423
	-0.8394	0.7956	-0.3778
		0.6408	-0.8394
			-1.2234
			-2.6354

[7] M. R. Scott and R. E. Kalaba, "An Initial Value Method for Integral Operators: IV -Complex-Valued Kernels of Laser Theory," J. Quant. Spec. Rad. Trans., 13, pp. 509-515 (1973).

algorithms presented in above can be used.

The results of refine are presented in TABLE II. The exact eigenvalues, and initial approximations, are respectively shown in columns 2 and 3. The next column provides numbers of iterations steps required to calculate the eigenvalues to an accuracy of $\epsilon = 10^{-9}$ and the refined eigenvalues are presented in the last column.

TABLE II. REFINED EIGENVALUES

k	<i>exact eigen.</i>	<i>initial approx.</i>	m	<i>refined eigen.</i>
1	0.2422606951	0.2423	2	0.2422606954
2	0.6382838292	0.6383	3	0.6382838295
3	0.7967066727	0.6383	3	0.7967066725
4	2.322748800	2.3229	2	2.322748800
5	-0.777442689	-0.3778	2	-0.777442685
6	-0.8393977662	-0.8394	1	-0.8393977663
7	-1.223471197	-1.2234	2	-1.223471198
8	-2.635389128	-2.6354	2	-2.635389128

VI. CONCLUSION

In this paper, an improved Newton method for solving non-linear eigenvalue problems was presented. The results showed that the improved Newton method is an efficient method. The presented method is very efficient and competitive with other methods used to solve nonlinear eigenvalue problems. Here MATLAB 2012 software was used to implement the algorithm.

REFERENCES

[1] A. V. Bitsadze, Fundamentals of the Theory of Analytical Functions of a Complex Variable (Nauka, Moscow, 1972) [in Russian].

[2] S. V. Kartyshev, Numerical Method for Solving the Eigenvalue Problem for Sparse Matrices Depending Nonlinearly on a Spectral Parameter, Zh. Vychisl. Mat. Mat. Fiz. 29, pp. 1898-1903 (1989).

[3] A. Melman, The double-step Newton method for polynomials with all real zeros, Department of Applied Mathematics, School of Engineering, Santa Clara University, Santa Clara, CA 95053, United States.

[4] B. M. Podlevskii, On Certain Two-Sided Analogues of Newtons Method for Solving Nonlinear Eigenvalue Problems equations, Institute of Applied Mathematics and Mechanics, National Academy of Sciences of Ukraine, ul. Nauchnaya 3-b, Lviv, 79000 Ukraine.

[5] M. R. Scott and J. W. Burgmeier, "A Method for Obtaining Bounds on Eigenvalues and Eigenfunctions by Solving Non-Homogeneous Integral Equations," Computing, 10, pp. 3-22 (1972).

[6] M. R. Scott and R. E. Kalaba, "An Initial Value Method for Integral Operators: I -Complex Eigenvalues," J. Comp. Phys., 12, pp. 364-3648 (1973).

Automatic generation of model for building energy management

Quoc-Dung Ngo
People's security academy
Hanoi, Viet Nam

Yanis Hadj-Said
G-SCOP lab, CNRS UMR 5272
Grenoble Institute of Technology, France

Stéphane Ploix
G-SCOP lab, CNRS UMR 5272
Grenoble Institute of Technology, France

Ujjwal Maulik
Jadavpur University, Kolkata-700032
West Bengal, India

Abstract—This paper proposes a model transformation approach for model-based energy management in buildings. Indeed, energy management is a large area that covers a wide range of applications such as simulation, mixed integer linear programming optimization, simulated annealing optimization, model parameter estimation, diagnostic analysis, . . . Each application requires a model but in a specific formalism with specific additional information. Up to now, application models are rewritten for each application. In building energy management, because the optimization problems may be dynamically generated, model transformation should be done dynamically, depending on the problem to solve. For this purpose, a model driven engineering approach combined with the use of a computer algebra system is proposed. This paper presents the core specifications of the transformation of a so-called high level pivot model into application specific models. As an example, transformations of a pivot model into both an acausal linear model for mixed integer linear programming optimization and a causal non-linear model for simulated annealing optimization are presented. These models are used for energy management of a smart building platform named Monitoring and Habitat Intelligent located at PREDIS/ENSE3 in Grenoble, France.

Keywords—building energy management system, model transformation, model driven engineering, optimization, mixed integer linear programming, simulated annealing

I. INTRODUCTION

Nowadays, the building sector represents about 38% of the total energy consumption in Europe and 63% in France [1, 10]. Therefore, energy consumption reduction in building has become an important challenge for researchers. A lot of Building Energy Management Systems (BEMS) have been proposed aiming at minimizing the daily energy consumption while maintaining a satisfactory level of comfort for occupants using models of the building systems. Modern building systems may be complex in terms of number of appliances, including production and storage means but also in terms of applications, which may cover functionalities like monitoring and state estimation, model parameter estimation, simulation synchronized with measurements for replays but also model based energy management using optimization algorithms, . . . Therefore, tools to handle and transform models are required. In the last decade, sophisticated methods, formalism and tools have been developed for different applications in order to better master dwelling energy consumption and production such as:

- global optimization for anticipative energy management [3, 7, 11, 12]. Actually, a day ahead energy management plan proposes to occupants the best configurations for building envelope and appliances for the next 24 hours in order to optimize a cost/comfort compromise. Optimization problem is dynamically generated according to the appliances and occupant activities impacting the management time horizon. To deal with thousands of variables and constraints in a acceptable computation time, a *mixed integer linear programming* (MILP) solver is used. Therefore, an acausal linear problem is required for this application.
- fast optimization for interactions with occupants [43]. In residential sector, building energy management cannot be fully automatized: it results from an interactive process where occupants shape the anticipative plan by modifying or adding constraints. It is often not necessary to re-perform a global optimization: a local optimization is often sufficient and more interesting because it is less time consuming when using the global solution as initialization.
- simulation for analyzing impacts of actions [46, 47]. Simulation approach can be causal such as with Matlab or acausal with Modelica. There is no optimization process but depending on the kind of simulation, the nature of the required models may differ.
- parameter estimation to learn the building intrinsic behavior thanks to recorded datasets. [44, 45]. The variables that were previously considered as parameters become yet the optimization variables while other variables are set to the values belonging to datasets.

These applications require each a dedicated formalism, whose nature may be very specific. Generally speaking, this problem is not a new one but in building energy management, where models are dynamically generated depending on the used appliances and on the question the energy manager has to solved, it cannot be handled manually by an expert.

Automatic model generation is a promising approach to avoid the above issue. gPROMS [8] or General Algebraic Modeling System (GAMS) [9] have been developed for the user to focus on the modeling problem by forgetting the application formalism requirements. Once the core model is defined, these systems manage the time-consuming transformation required for most common optimization solvers (GLPK,

CPLEX, GUROBI, . . .). Although these approaches are based on a superset high level language, they are not able to change the nature of a model: a causal model cannot become acausal, a nonlinear model cannot be linearized, a causal model for simulation cannot be transformed into a causal model for parameter estimation. To handle such transformations, models have to be deeply modified using a computer algebra system to reformulate and transform constraints.

Because the model construction of a whole dwelling is not a trivial task, due to the system complexity and dynamicity, it is not a good solution to build a whole dwelling model at once but it is preferable to compose step by step element models before generating sub-systems and finally the overall system. When there is a change, it is just needed to add or remove some element models.

In order to fit the building energy management system (BEMS) needs for automatic transformation between application models, a Model Driven Engineering (MDE) [4] approach combined with a computer algebra system (CAS) is proposed. The MDE main objective is to reduce software production cost by using standardized models and increasing their flexibility to deal with computer technology evolution. This methodology is largely implemented in object oriented modeling. This paper proposes to adapt this approach to the transformation of composed pivot model into application specific models.

The paper is composed of 5 main sections. The next section aims at formulating different key concepts using an illustrative example. The third section presents the transformation process principles. An application to the *PREDIS Monitoring and Habitat Intelligent* platform is presented in the fourth section and the last section is dedicated to analysis of two model transformation results dedicated to model based energy management.

II. PROBLEM STATEMENT

In this section, different key concepts aiming at composing an overall dwelling model based on Model Driven Engineering approach are presented. An example is used as an illustration.

A. Concept of model transformation

Let's consider resistor modeled by: $C_0 : U_1 = R_1 \times I_1$.

This simple model may be used by a designer into different optimization problem, adding information like lower and upper bounds of the possible value domains of variables or an objective function. Consider for instance the following (unrealistic) optimization problem:

$$\begin{aligned} C_0 & : U_1 = R_1 \times I_1 \\ C_1 & : R_1 \in [0, 5] \\ C_2 & : U_1 \in [0, 4] \\ \text{Objective} & : \max_{R_1, U_1} I_1 \end{aligned}$$

In spite of its simplicity, if another resistor R_2 is added in parallel, the whole system model has to be rewritten:

$$\begin{aligned} C_0 & : U_1 = R_1 \times I_1 \\ C_1 & : U_2 = R_2 \times I_2 \\ C_2 & : I_{total} = I_1 + I_2 \\ C_3 & : R_1 \in [0, 5] \\ C_4 & : R_2 \in [0, 3] \\ C_5 & : U_1 \in [0, 4] \\ C_6 & : U_2 \in [0, 4] \\ C_7 & : U_1 = U_2 \end{aligned}$$

$$\text{Objective} : \max_{R_1, U_1, R_2, U_2} I_{total}$$

Although the rewriting process is not that time consuming in this example, it becomes a tough work for complex systems which contain hundreds of variables and constraints. In addition, model may also have to be rewritten depending on the target application. For instance, some optimization algorithms require a causal ordering (Simulated Annealing, for example), some others require linearization (Mixed Integer Linear Programming, for example). Therefore, two difficulties must be dealt with:

- a model must be composed of model elements that can be reused
- a model has to be transformable.

To handle model transformation in optimization problems, the concept of pivot model is introduced. Actually, a pivot model is a high level application-independent description that can be transformed into target application formalisms, which may require a reformulation of some constraints including equations i.e. equality constraints.

In the computer science literature, model rewriting processes are usually managed using the concepts of Model Driven Engineering.

B. Concept of Model Driven Engineering

Basically, the Model Driven Engineering approach aims at separating models based on company know-how and those related to software implementations in order to maintain the sustainability of the company know-how in spite of the changes of development environment [4]. To do this, it is necessary to firstly define Platform Independent Models (PIM) i.e. pivot model, technically independent from any execution platform. It enables the generation of a set of Platform Specific Models (PSM) afterwards. Based on the MDE approach, the problem can be decomposed into 2 abstraction levels. The two concepts of PSM and PIM are corresponding respectively to the level M0 and M1. Shortly, the signification of each level is:

- level M0 (PSM) is the real system that contains executable object
- level M1 (PIM) is the model that represents the system

The main objective of MDE is to perform transformations between PIM and different PSMs. There are two types of transformations of models:

- transformation model-to-code (PIM to PSM)
- transformation model-to-model (PIM to PIM)

Generally speaking, the model-to-code transformation can be seen as a special case of model-to-model transformation. A classification of model transformation approaches is presented in [5]. Basically, a model-to-model transformation is performed with the help of transformation rules that consists in transforming a set of input models into a set of targeted models.

C. Concept of pivot model

Thank to this architecture and according to MDE, a pivot model can be considered as a PIM (level M1) and the PSM can be associated to an application specific model such as optimization model formalism. Basically, a PIM is supposed to be available initially, then PIM to PSM or PIM to PIM transformations have to be computed by applying transformation processes. Generally speaking, the PIM construction is built from elementary models, denoted EM , that describe element parts of the system. An elementary model EM , in the field of optimization, is associated with a subspace of a vector space defined on \mathbb{R}^n . It is considered that the integer set \mathbb{N} is a specialization of the real number space \mathbb{R} : $\mathbb{N} \subset \mathbb{R}$ and that the assertions *True* and *False* are modeled with binary values, respectively 1 and 0, i.e. a specialization of \mathbb{N} . An element model representing an element in a given mode is defined as:

Definition 1: $EM : \text{mode}(EM) \leftrightarrow \mathcal{V}_S \in \mathcal{E}; \mathcal{E} \subset \text{dom}(S) \subset \mathbb{R}^n$
with:

- $\mathcal{V}_S = \{v_0, \dots, v_{n-1}\}$ is a set of variables respectively related to the tuple of symbols $\mathcal{S} = \{\text{symbol}_0, \dots, \text{symbol}_{n-1}\}$
- $\text{dom}(S) = \text{dom}(\text{symbol}_0) \times \dots \times \text{dom}(\text{symbol}_{n-1})$ is the set of value domain of symbols corresponding to variables.
- mode is generally and implicitly ok (except in diagnosis analysis) for normal behavior
- the subspace \mathcal{E} is defined by a set of n_j constraints \mathcal{K} defined over \mathbb{R}^n .

$$\mathcal{K} = \{\mathcal{K}_j(S) \diamond 0; \forall j\} \quad (1)$$

where \diamond stands for a comparison operator.

The notion of element has to be clarified. Let's consider a dual flow ventilation system [6] composed of two speed variation control devices and two electric drives associated with the extraction of indoor air. The electric drive and the control system models are parts of the overall ventilation system model but at a lower level of consideration. The more the elements are decomposed, the more they can be reused. Actually, a pivot model is composed step by step by adding required element models.

Definition 2: A pivot system model $PM = (\mathcal{K}_\Sigma(S_\Sigma, \text{dom}(S_\Sigma)))$ is an union of elementary models EM_i plus connection constraints K_j

Let's consider the transformation of a PIM acausal model into a PSM causal i.e. simulable, model.

Consider the pivot system model $PM = (\mathcal{K}_\Sigma(S_\Sigma, \text{dom}(S_\Sigma)))$. Constraints can be decomposed into

equality constraints, denoted $\mathcal{K}_\Sigma^=$, and inequality constraints, denoted \mathcal{K}_Σ^\leq . A model is said *simulable* if it exists a function $\varphi: \mathcal{S}_\Sigma^{in} \rightarrow \mathcal{S}_\Sigma^{out}$ such as $\mathcal{K}_\Sigma^=(S_\Sigma) \leftrightarrow \varphi(S_\Sigma^{in}) = S_\Sigma^{out}$ on $\text{dom}(S_\Sigma)$ with $(S_\Sigma^{in}, S_\Sigma^{out})$ is a partition of \mathcal{S}_Σ .

Transforming PM into a simulation model for simulated annealing optimization for example, denoted PM_{SA} , consists in selecting and projecting $\mathcal{K}_\Sigma^=(S_\Sigma)$ into $\varphi(S_\Sigma^{in}) = S_\Sigma^{out}$, a causal ordering has to be performed. It requires usually to set values of some variables that will become parameters and input variables. The Dulmage-Mendelsohn algorithm [29] is generally used for this purpose.

A MILP model is defined as $PM_{MILP} = (\mathcal{K}_\Sigma^{MILP}(S_\Sigma^{MILP}), \text{dom}(S_\Sigma^{MILP}))$ with $\mathcal{K}_\Sigma^{MILP}(S_\Sigma^{MILP})$ linear. Transforming PM into PM_{MILP} relies on linearization transformation patterns.

Different transformation processes for composing a pivot model are detailed in section III, then the pivot model is projected into causal simulable model and MILP formalism.

III. TRANSFORMATION PROCESSES

This section gives an overview on different transformation principles using an illustrative example. Transformation process is composed of two main steps. The first one aims at manipulating element models to get a system pivot model. The second step consists in applying different projections to get target application models. The processes leading to either a simulable model or a MILP model are shown in this section.

A. Composition process

This sub-section focuses on how a pivot model is built. According to the definition 2, the most important step to build a system pivot model is the composition of different element models. The objective of the composition is to help the reusability of element models and to make the pivot model more modular. A composition may concern a set of element models, a set of compositions of element models or a set of compositions of compositions and so on. Moreover, recursive compositions can be performed unlimitedly to get bigger compositions. To illustrate this point, consider now an electric circuit as presented by figure 1.

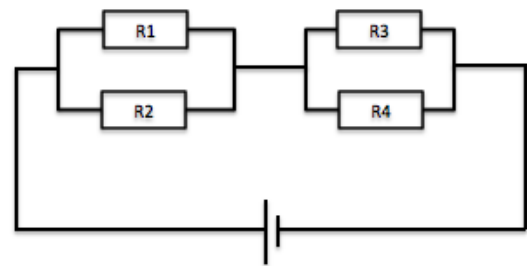


Fig. 1: Example of electric circuit

The system presented in figure 1 is composed of two blocks of 4 resistors R_1, R_2, R_3 and R_4 . Independently of any formalism, the construction of such a pivot model can be

done by composing firstly a bloc of 2 parallel resistors. Then, the pivot model is build by duplicating this bloc and connecting the whole system.

When composing step by step the pivot model, there are two remaining problems that have to be considered. The first one consists in specializing all resistors with the corresponding values, and the second one consists in establishing the different connections between element models.

To deal with the first problem, each element models (*EM*) is necessary specialized before being used in a composition. The specialization concept presented in [34] is well suited for this problem. It makes an element more specific by adding some additional information like a prefix or a type. According to the definition 1, the specialization of an *EM* consists only in adding a distinct prefix to symbol representing a variable each time it is used. For instance, $R_1.U$ is not like $R_2.U$ and so on. An *EM* could be specialized as many times as desired. The more specialized an *EM* is during a composition process, the more specific it is. For instance, $bloc_1.R_1.U$ is not like $bloc_2.R_1.U$. Nevertheless, a set of specialized *EM* cannot form a composition without connections between them. Indeed, two specialized *EM*, for instance resistors R_1 and R_2 require explicitly the following connecting-equations, which is a common concept with [34]:

$$\begin{aligned} R_1.U &= R_2.U \\ I_{total} &= \frac{R_1.U}{R_1}.R + \frac{R_2.U}{R_2}.R \end{aligned}$$

These connecting equations are added into the compositions. A pivot model for this system is given by:

- the parallel bloc composition with the dots '.' represent suffixes of prefixes to symbols standing for variables.

$$C_0 : R_1.U = R.R \times R_1.I \quad (2)$$

$$C_1 : R_2.U = R_2.R \times R_2.I \quad (3)$$

$$C_2 : R_1.U = R_2.U \quad (4)$$

$$C_3 : I_{total} = \frac{R_1.U}{R_1.R} + \frac{R_2.U}{R_2.R} \quad (5)$$

- By duplicating the parallel bloc composition above twice and by adding connecting-equations for establishing the final circuit. The system pivot model is thus built:

$$C_0 : bloc_1.R_1.U = bloc_1.R_1.R \times bloc_1.R_1.I \quad (6)$$

$$C_1 : bloc_1.R_2.U = bloc_1.R_2.R \times bloc_1.R_2.I \quad (7)$$

$$C_2 : bloc_1.R_1.U = bloc_1.R_2.U \quad (8)$$

$$C_3 : bloc_1.I_{total} = \frac{bloc_1.R_1.U}{bloc_1.R_1.R} + \frac{bloc_1.R_2.U}{bloc_1.R_2.R} \quad (9)$$

$$C_4 : bloc_2.R_1.U = bloc_2.R_1.R \times bloc_2.R_1.I \quad (10)$$

$$C_5 : bloc_2.R_2.U = bloc_2.R_2.R \times bloc_2.R_2.I \quad (11)$$

$$C_6 : bloc_2.R_1.U = bloc_2.R_2.U \quad (12)$$

$$C_7 : bloc_2.I_{total} = \frac{bloc_2.R_1.U}{bloc_2.R_1.R} + \frac{bloc_2.R_2.U}{bloc_2.R_2.R} \quad (13)$$

$$C_8 : bloc_1.I_{total} = bloc_2.I_{total} \quad (14)$$

$$C_9 : U_{total} = bloc_1.R_1.U + bloc_2.R_1.U \quad (15)$$

To summarize the above pivot model construction, the resistor model is firstly specialized twice to create two different resistors. Then, a bloc of two parallel resistors is created by adding connecting equations. Finally, the pivot model is built by duplicating this parallel bloc and adding new connecting equations. This pivot model can automatically be generated if these three steps are defined in a recipe. The concept of recipe is an important tool for the systematic generation and transformation processes.

Each generation or transformation step is considered as a transformation rule, which is implemented and put into a common rule-set. Then, recipes have then to be developed: they trigger rules from a rule-set in a relevant order until a desired formalism is obtained for a given application. However, the equation manipulation to create such a pivot model is not a trivial task. To automatize the addition of prefixes, the combinations with connecting constraints, the constraint time duplication, an symbolic calculation engine is required.

In the recent decades, symbolic computation or computer algebra [35, 37] have become an important research area of mathematics and computer sciences aiming at developing tools for solving symbolic equations. The capabilities of major general purpose Computer Algebra Systems (CAS) is presented in [36, 38]. Moreover, among the mathematical features of a CAS, there are transformations allowing to manipulate and optimize symbolic computations in order to automatically generate optimization code [39].

For instance, the GIAC/XCAS CAS [40] has been developed to solve a wide variety of symbolic problems and has been awarded with the 3rd price at the Trophées du Libre 2007 in the scientific software category [41]. This CAS has been used for manipulation of symbols in all the constraints of the pivot model. With GIAC/XCAS, each constraint is considered as a n-ary equation tree as presented in figure 2.

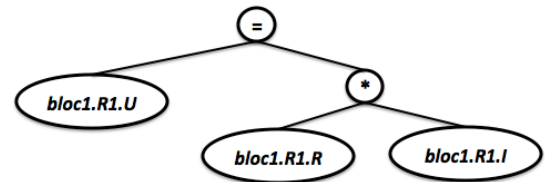


Fig. 2: n-ary tree representation for equation (6)

Finally, the set of required manipulations for composing a pivot model is respectively summarized as follows:

- specialization of *EM* by adding prefixes
- addition of connecting-equations

B. Projection process

Once a pivot model is composed, the next step consists in applying different projection processes to get desired formalism. These projection processes can always be detailed in recipes to automatize the transformation between models. This sub-section shows different steps to get causal simulable model and energy management model formalism. These processes are summarized in figure 3.

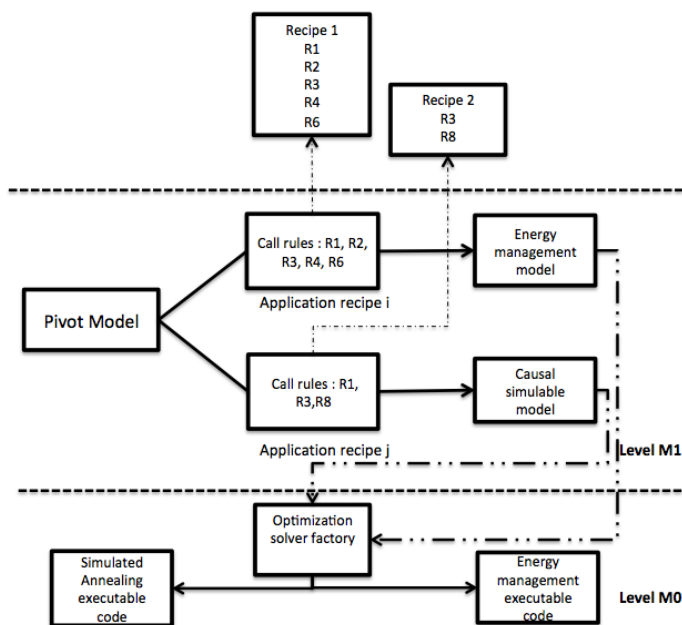


Fig. 3: Example of two projection processes

Energy management application using MILP formalism requires an acausal linear model containing 3 kinds of variables : binary, continuous, integer and 2 kinds of constraints: equality and inequality. It means that MILP transformation process aims at transforming and linearizing all constraints into equality and inequality constraints.

Based on the platform PREDIS/MHI model detailed in section IV, the first point is necessary performed to transform all ODE and logical constraints into equality and inequality constraints. In this study case, an approximation of ODE time discretization is shown instead of the exact transformation solution. This approximation consists in developing the derivative variable into:

$$\frac{dv_i}{dt} = \frac{v_i(t+1) - v_i(t)}{t} \quad (16)$$

with $v_i \in \mathcal{V}_S$ and t is the pre-defined time step. According to the obtained results, it is considered as precise enough for the 1 hour time step of the energy management.

This time discretization transformation of all ODE, $\frac{dv_i}{dt} = f(\mathcal{V}_S)$, is performed symbolically as presented by figure 4.

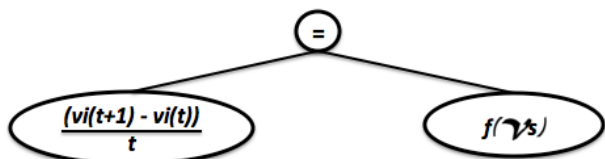


Fig. 4: Time discretization pattern

The main idea of this transformation is the same for logical constraint transformation and it can be found in [12]. Once this step is completed, the pivot model contains equality and inequality constraints. The next step to do consists in searching and linearizing all non-linear terms.

The difference between non-linear terms is based on the nature of variables and/or the nature of functions that contain variables. Indeed, product of two discrete variables cannot be linearized in the same way as a product of two continuous variables or a cosine function for example. To linearize the pivot model, it is preferable to sort out all the non-linear terms in different kinds of non-linearity first. Then each kind of non-linearity is linearized by corresponding rules. It means that recipes, rules and rule-set have to be easily extended to cover all possible changes. The whole linearization process is summarized in figure 5.

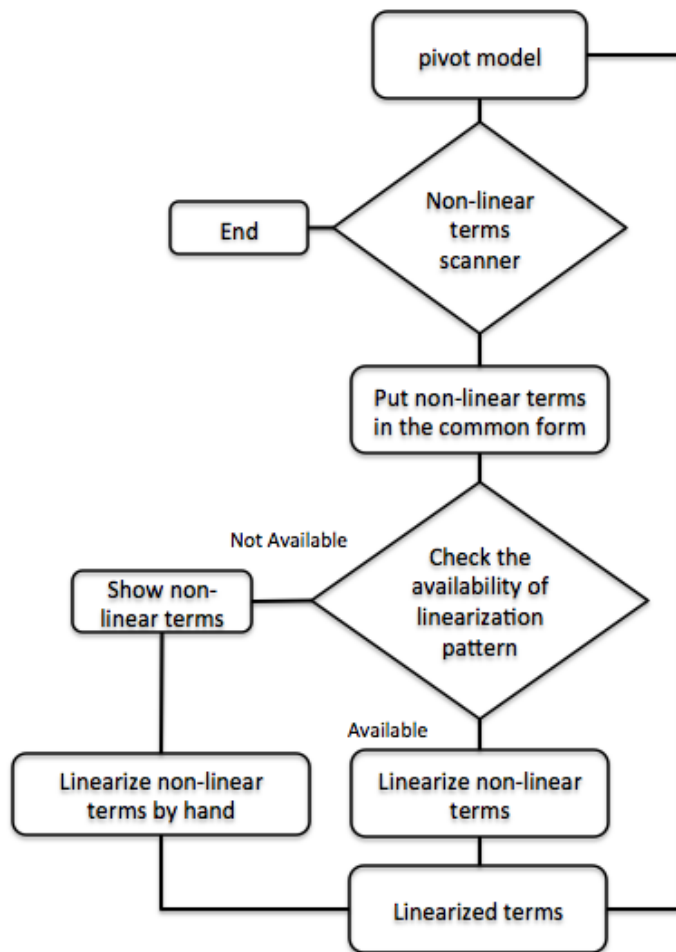


Fig. 5: Linearization process

This schema shows how the linearization process can be automatized using different patterns that were presented in [12, 42]. This process deals with non-linear terms as follows:

- product of m binary variables with $m \geq 1$
- product of l discrete variables with $l \geq 1$
- product of m binary variables and l discrete variables
- product of m binary variables and 1 continuous variable
- product of l discrete variables and 1 continuous variable
- product of m discrete variables and l discrete variables and 1 continuous variable

However, there are some terms for which the linearization process cannot be automatized and where a human intervention

is required, for instance the product of n ($n \geq 1$) continuous variables. Indeed, it does not exist a linearization pattern for this type of non-linearity to be performed directly. Linearizing a such of non-linearity requires a preliminary step consisting in discretizing the domain of $n - 1$ continuous variables into sets of discrete values. Then, the pattern of discrete and continuous product can be used to get a linear term. Discretization also means approximation to realistic values, therefore the choice of discrete values impacts strongly on final results and this step can not be automatically performed by system. Only expert who masters his dwelling system can take good values for linearization process afterwards.

Let's linearize the circuit system (1) by discretizing for instance the resistor into $R = \{3, 4\}$. Then the discrete and continuous linearization pattern can be used by introducing a new variable, denoted Z , representing the product $R \times I$ with:

$$Z = R \times I = (\delta_1 \times v_1 + \delta_2 \times v_2) \times I \quad (17)$$

with δ_i is a binary variable that takes value in $\{0, 1\}$. Actually, the goal is to select the best value among those of R to maximize or minimize the objective function. Equation (17) can be factorized as:

$$Z = \delta_1 \times v_1 \times I + \delta_2 \times v_2 \times I \quad (18)$$

with v_1 and v_2 standing for parameters. There are two binary and continuous products to be linearized. Let's linearize for instance the first binary and continuous product term: $\delta_1 \times v_1 \times I$. The corresponding pattern implies to create a new continuous variables, denoted Z' with 4 new constraints delimiting the bounds of Z' given by:

$$Z' \leq \delta_1 \times v_1 \times \bar{I} \quad (19)$$

$$Z' \geq \delta_1 \times v_1 \times \underline{I} \quad (20)$$

$$Z' \leq (I - \underline{I} \times (1 - \delta_1)) \times v_1 \quad (21)$$

$$Z' \geq (I - \bar{I} \times (1 - \delta_1)) \times v_1 \quad (22)$$

with \underline{I} and \bar{I} respectively are lower and upper bound of the continuous variable I . The second binary and continuous product $\delta_2 \times v_2 \times I$ is linearized in the same way. Once all the non-linear terms are linearized, the MILP model formalism is obtained.

Regarding the generation of a causal simulable model for local optimization, the required projection aims at transforming the pivot model into a simulable model. Firstly, a model is simulable if only if it is a structurally just-determined model [29] i.e. the number of variables is equal to the number of equality constraints, therefore one solution value can be assigned to each variable. Dulmage-Mendelsohn algorithm [29] has been used to compute just-determined blocks of constraints and variables which are represented by a reorganized structural matrix. However, some pivot models can also be:

- structurally under-determined i.e. they contain less equality constraints than variables. In this case, variables will have an infinite number of possible solutions yielding non-simulable models. Nevertheless, non-simulable models are frequent in energy management optimization applications.

- structurally over-determined i.e. there are more equality constraints than variables. In this case, generally speaking, no one value can be assigned to some variables i.e. no one solution can be computed. This kind of models are nevertheless common in diagnosis application where no solution means failure.

Normally, a correct simulable model gives only a just-determined set while other sets are empty. The equality constraints can be reorganized according to the upper-triangular just-determined part of the incidence matrix of equality constraints. The presence of an under-determined set or of an over-determined part means that the whole model can't be simulated and it is necessary to recheck the element models.

In order to automatize the whole transformation process, a software architecture is proposed in figure 6.

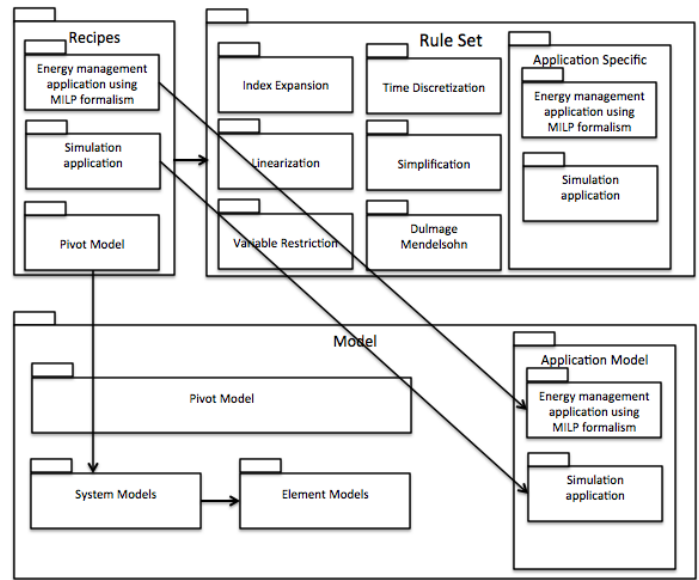


Fig. 6: software architecture

The key steps to transform a pivot model into energy management using MILP formalism and simulation are shown. The next section illustrates the application of the proposed method to the model transformation of the PREDIS/ENSE3 platform.

IV. APPLICATION EXAMPLE

This section presents the platform PREDIS/MHI located in Grenoble, France, that will be used as a "fil-conducteur" to explain the proposed approach. The Monitoring and Habitat Intelligent PREDIS platform is a research platform dedicated to research about smart-building for company, academic researchers and students.

This platform is a low consumption office building highly instrumented where most of the energy flows are measured using different sensor technologies. The structure of this platform is given by figure 7. For the sake of clarity, this section focuses on the classroom zone that is equipped with computers for students and a heating and dual flow ventilation system containing:

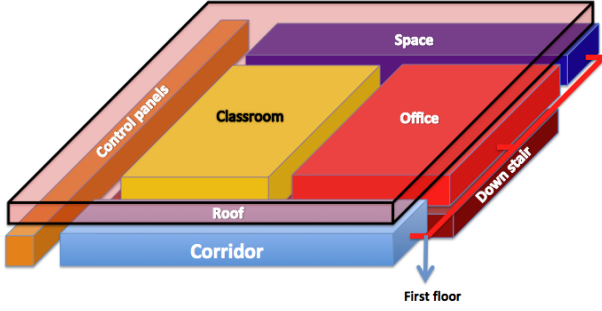


Fig. 7: Overview of PREDIS/ENSE3 platform

- an air treatment unit model:

$$AirFlow = coef \times Q_{Air} \quad (23)$$

$$P_{airTreatmentUnit} = P_{ventilation} + P_{heating} \quad (24)$$

- a thermal balance model:

$$Phi_{Total} = Phi_{Sun} + P_{heating} + Phi_{Occup} \quad (25)$$

- a thermal comfort model depending on whether there is someone or not in the classroom:

$$If\ presence = 1, \quad (26)$$

$$T_{felt} < T_{pref} \Rightarrow sigma_{incomfort} = 1/(T_{pref} - T_{max}) \times T_{felt} - T_{pref}/(T_{pref} - T_{max})$$

$$T_{felt} \geq T_{pref} \Rightarrow sigma_{incomfort} = 1/(T_{max} - T_{pref}) \times T_{felt} - T_{pref}/(T_{max} - T_{pref}) \quad (27)$$

$$If\ presence = 0, sigma_{incomfort} = 0 \quad (28)$$

$$T_{felt} \leq T_{max,absence}$$

$$T_{felt} \geq T_{min,absence}$$

- a Thermal Zone model:

$$R_{Ventilation} = 1/((1 - efficiency) \times Cp_{Air} \times rho_{Air} \times AirFlow) \quad (29)$$

$$R_{Eq} = 1/(1/(R_{Ventilation} + R_w) + \sum(1/R[neighborhood])) \quad (30)$$

$$\frac{d}{dt}T_w = -1/(R_{Eq} \times C_w) \times T_w + 1/((R_{Ventilation} + R_w) \times C_w) \times T_{out} + \sum(T[neighborhood]/(R[neighborhood] \times C_w)) + R_{Ventilation} \times Phi_{total}/(C_w \times (R_{Ventilation} + R_w)) \quad (31)$$

$$T_{In} = R_{Ventilation} \times T_w / (R_{Ventilation} + R_w) + R_w / (R_{Ventilation} + R_w) \times T_{out} + R_{Ventilation} \times R_{Eq} \times Phi_{total} / (R_{Ventilation} + R_w) \quad (32)$$

- a CO₂ Comfort model:

$$sigma_{CO_2} = (C_{CO_2} - C_{fav}) / (C_{max} - C_{fav}) \quad (33)$$

- a CO₂ Zone model:

$$\frac{d}{dt}C_{InCO_2} = Q_{Breath} \times occupancy \times (C_{Breath} - C_{InCO_2}) / Vol_{Zone} + AirFlow \times (C_{OutCO_2} - C_{InCO_2}) / Vol_{Zone} \quad (34)$$

- and finally, the total power consumption model:

$$P_{total} = P_{airTreatmentUnit} + P_{lighting} + P_{computer} \quad (35)$$

$$Total_{cost} = P_{total} \times PricePerKwh \quad (36)$$

These models describe only the physical phenomena of PREDIS/ENSE3. This section illustrates how the pivot model of the classroom is constructed, how it is projected it into MILP formalism and into a causal simulable model. Different required steps to get these application models are summarized in the figure 8.

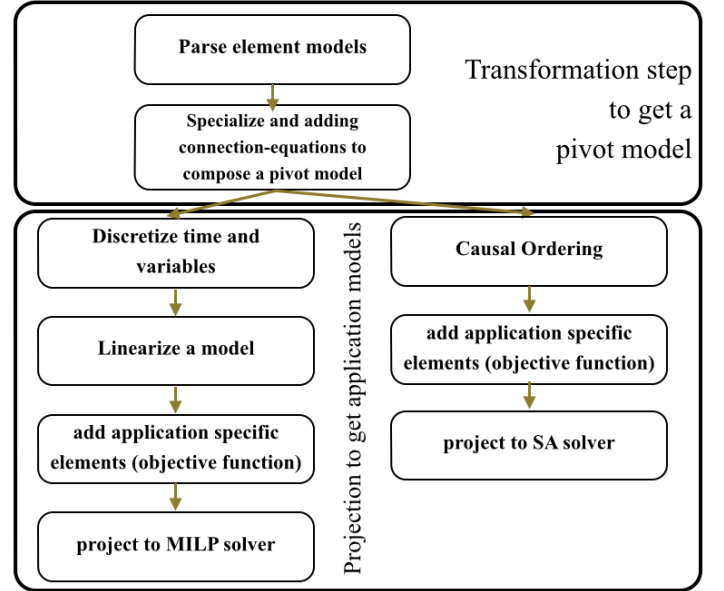


Fig. 8: Transformation processes to get target models

V. FROM PIVOT TO A SIMULABLE MODEL

To generate PREDIS/ENSE3's pivot model, the composition recipe is realized in 3 steps:

- Compose the CO₂ system:
 - specialize : CO₂ comfort with prefix : CO₂Comfort.
 - specialize : CO₂ comfort with prefix : CO₂Zone.
 - connect : CO₂Comfort.C_{CO₂} = CO₂Zone.C_{InCO₂}
- Compose the Thermal system:
 - specialize : thermal comfort with prefix : thermalComfort.
 - specialize : thermal zone with prefix : thermalZone.
 - connect : thermalComfort.T_{felt} = thermalZone.T_{In}
- Compose the final pivot model:
 - specialize : CO₂ system with prefix : CO₂System.
 - specialize : thermal system with prefix : thermalSystem.
 - specialize : power consumption with prefix : powerConsumption.
 - specialize : thermal balance with prefix : thermalBalance.
 - specialize : air treatment unit with prefix : airTreatmentUnit.
 - connect : airTreatmentUnit.AirFlow = thermalSystem.AirFlow

- connect : airTreatmentUnit. $P_{airTreatmentUnit}$ = powerConsumption. $P_{airTreatmentUnit}$
- connect : thermalBalance. $P_{heating}$ = airTreatmentUnit. $P_{heating}$

Initially, element models of PREDIS/ENSE3 are represented in textual description files. Thanks to the GIAC/XCAS computer algebraic system [40], each constraint is represented as a n-ary equation tree. The different variables belonging to the different constraints are collected and represented by symbols. After the parsing process, an elementary model EM is represented by a set of n-ary equation trees that facilitates the different manipulations and projection afterwards. Consider now the representation of the CO₂ zone model. It yields a n-ary equation tree given by figure 9.

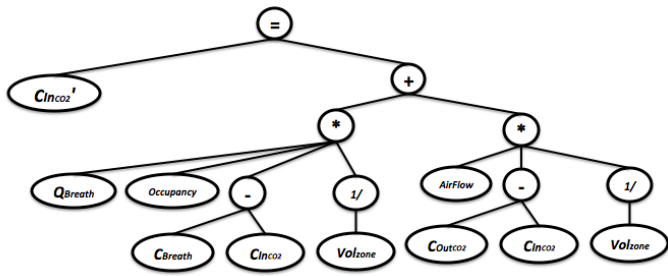


Fig. 9: CO₂ zone model n-ary representation

The name of constraints is then specialized by adding given prefixes and new connection equations are added to compose the pivot model of the system. Once these connection-equations are taken into account, the PREDIS/ENSE3's pivot model generation process is completed.

To transform the PREDIS/ENSE3's pivot model into a simulable model formalism, the key step is to perform Dulmage-Mendelsohn algorithm [29] to verify if the PREDIS/ENSE3's model could be simulable, then the causal ordering of variables if it is simulable.

In building energy management, the reorganized structural matrix is usually upper triangular with no block on the diagonal but sometimes blocks may appear. In this case, the projection cannot be fully automatized because there is no general process to solve implicit sub-systems of nonlinear equations.

It is important to note that only equality constraints are taken into account for the generation of a simulable model. It means that a preliminary step is required: the extraction of equality constraints. If the set of equality constraints is just-determined, the next step consists in making this pivot model simulable. On other words, this pivot model S_{Σ} need to be separated into S_{Σ}^{in} and S_{Σ}^{out} . Therefore, causal ordering process is necessary performed using a Dulmage-Mendelsohn based algorithm.

Let's consider a practical and didactic example consisting in simulating the thermal part of a hybrid panel running under sun:

The model of the hybrid panel is built around Hottel-Whillier equations. This equation describe phenomenon observed in the system of energy caption and transmission.

The system described is a panel of photo voltaic cells which are cooled by liquid in circulation under the layer of PV cells as shown in the scheme

$$C_0 : F_R = \frac{\phi * C_P}{(S_{PV} * U_{loss})} * (1 - e^{-S_{PV} * U_{loss} * F' / (\phi * C_P)})$$

$$C_1 : P_{Ther} = S_{PV} * F_R * \sigma_{abs} * G + U_{loss} * (T_{outdoor} - T_{Input})$$

$$C_2 : P_{Ther} = \phi * C_P * (T_{Output} - T_{Input})$$

with:

F_R : Heat dissipation factor.

ϕ : Flow in the panel.

C_P : Heat capacity.

S_{PV} : Panel Surface Area.

U_{loss} : Heat transfer coefficient.

F' : Thermal resistance between cells.

G : solar radiation.

P_{Ther} : Heat recovery capacity.

T_{Output} : Output temperature of the coolant.

T_{Input} : Input temperature of the coolant.

$T_{outdoor}$: Ambient temperature.

σ_{abs} : Thermal solar panel performance.

The variables presented above contribute to the description of the physical aspects and operational aspects of the hybrid panel. The physical aspects variables are fixed and considered as parameters for the simulation. This parameters are mandatory informed before the simulation process. The causality imposed for the simulation consist to consider the parameters as inputs of simulation and the variables, degrees of freedom, as outputs. The system is simulable if it can supply exact outputs for the inputs chosen we say that is just determined. If the simulation need more information, it is considered as under-determined. If there is one or more degrees of freedom the simulation can be considered as over-determined. The values of the parameters are integrated as equations for dulmage-mendelshon algorithm:

$$C_3 : C_P = value \quad (37)$$

$$C_4 : T_{Outdoor} = value \quad (38)$$

$$C_5 : T_{Input} = value \quad (39)$$

$$C_6 : S_{PV} = value \quad (40)$$

$$C_7 : F_R = value \quad (41)$$

$$C_8 : \sigma_{abs} = value \quad (42)$$

$$C_9 : G = value \quad (43)$$

$$C_{10} : U_{loss} = value \quad (44)$$

$$C_{11} : F_{prim} = value \quad (45)$$

The dulmage-Mendelshon algorithm check the simulability of the system as shown in the matrix I which is triangular. The demonstration of the just determination of the model.

The whole $C_3...C_{11}$ must be informed to get the model simulable . The variables T_{Output} , ϕ , P_{Ther} are considered as

	T_{Output}	ϕ	F_{prim}	P_{Ther}	U_{loss}	G	σ_{abs}	F_R	S_{PV}	T_{Input}	$T_{Outdoor}$	C_P
C_2	1	1	0	1	0	0	0	0	0	1	0	1
C_0	0	1	1	0	1	0	0	1	1	0	0	1
C_{11}	0	0	1	0	0	0	0	0	0	0	0	0
C_1	0	0	0	1	1	1	1	1	1	1	1	0
C_{10}	0	0	0	0	1	0	0	0	0	0	0	0
C_9	0	0	0	0	0	1	0	0	0	0	0	0
C_8	0	0	0	0	0	0	1	0	0	0	0	0
C_7	0	0	0	0	0	0	0	1	0	0	0	0
C_6	0	0	0	0	0	0	0	0	1	0	0	0
C_5	0	0	0	0	0	0	0	0	0	1	0	0
C_4	0	0	0	0	0	0	0	0	0	0	1	0
C_3	0	0	0	0	0	0	0	0	0	0	0	1

TABLE I: Results of Dulmage-Mendelsohn decomposition

the output for simulation. The Dulmage-Mendelsohn algorithm produce the matrix I. This matrix is used to order equations in the execution process, it is the causality ordering. It consist to schedule the resolution of equations to get result from each one. The last equation (row of the matrix) C_3 has to be solved firstly. The resolution of C_3 equation results the value of C_p . The result is put in the inputs set for the next equations. The process run until the whole equations solved. The last equation to be solved is C_2 and give the last output T_{output} . The other outputs are computed before: ϕ from C_0 and P_{Ther} from C_1 .

VI. PROJECTION TO A MILP OPTIMIZATION MODEL

To transform the PREDIS/ENSE3's pivot model into a MILP model formalism, a specific recipe is built as presented by the below part of the figure 8. Let's detail some specific rules to illustrate how the symbolic transformation is performed : time discretization and linearization.

The time discretization consists in discretizing the pivot model into 24 sampling periods standing for one day. Thus, this step multiplies 24 times each constraint of the pivot model with time index ranging from 0 to 23. The ODE implementation of CO_2 zone at 5th time step is given by figure 10.

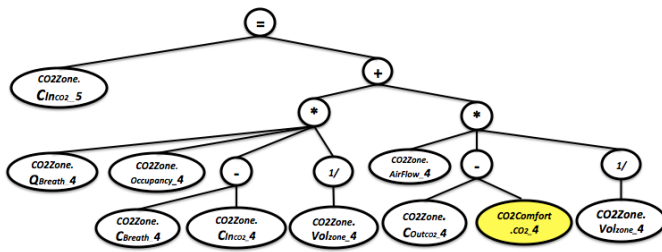


Fig. 10: CO₂ Zone model after the ODE transformation processing

The next important projection consists in linearizing non-linear terms inside the constraints of the pivot model. First, all non-linear terms are detected; then, the nature of each nonlinear term is analyzed before being recursively linearized according to the corresponding pattern according to the linearization process presented by the figure 5. For instance, the binary-continuous product : $CO_2Zone.Q_{Breath} \times CO_2Zone$ is linearized. *occupancy* in the CO_2 zone model where *occupancy* is 0 whether there is nobody or 1 otherwise. In this case, a new variable, denoted z , is used for replacing the considered term in the corresponding constraint as given by figure 11.

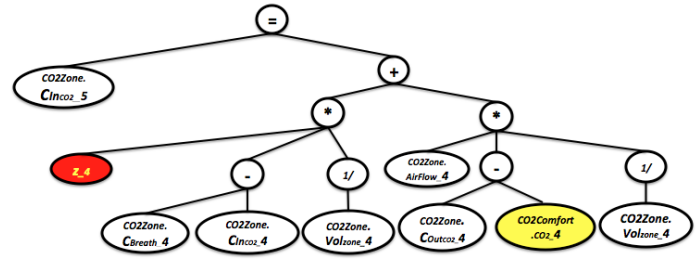


Fig. 11: CO₂ Zone binary model after the first linearization processing

And four new constraints resulting of this binary-continuous product are added into the pivot model :

$$\begin{aligned}
 z_4 &\leq CO_2Zone.occupancy_4 \times \overline{CO_2Zone.Q_{Breath_4}} \\
 z_4 &\geq CO_2Zone.occupancy_4 \times \underline{CO_2Zone.Q_{Breath_4}} \\
 z_4 &\leq \overline{CO_2Zone.Q_{Breath_4}} - (1 - CO_2Zone.occupancy_4) \times \overline{CO_2Zone.Q_{Breath_4}} \\
 z_4 &\geq \underline{CO_2Zone.Q_{Breath_4}} - (1 - CO_2Zone.occupancy_4) \times \underline{CO_2Zone.Q_{Breath_4}}
 \end{aligned}$$

Resulting of this linearization pattern represents exactly the considered binary-continuous product because:

- if $occupancy_4 = 1$:
 $z_4 \leq \sup(CO_2Zone.Q_{Breath_4})$
 $z_4 \geq \inf(CO_2Zone.Q_{Breath_4})$
 $z_4 \leq \overline{CO_2Zone.Q_{Breath_4}}$
 $z_4 \geq \underline{CO_2Zone.Q_{Breath_4}}$
 In this case, the two first constraints are always true so they can be eliminated. The last two constraints make it possible to take into account the real values of $CO_2Zone.Q_{Breath_4}$.
- if $occupancy_4 = 0$:
 $z_4 \leq 0$
 $z_4 \geq 0$
 when there is nobody in the classroom, it means that the Q_{Breath} is equal to 0, too.

After linearizing all non-linear terms, the obtained model is ready to provided to MILP solver to generate energy management plans. This MILP formalism has 1679 constraints/1129 variables whose 1011 new constraints/459 new variables added resulting of linearization process.

VII. PROJECTION TO A SIMULATED ANNEALING OPTIMIZATION MODEL

The simulated annealing process is based on stochastic approach which finds an optimum for single objective optimization problem. In this study, it is used as complement to MILP optimization process to find quickly a better solution in case of minor changes in the optimizing problem.

The aim in the development of simulated annealing optimizer is to complete the offer of MILP optimizing. MILP optimization, thanks to tools like CPLEX or GLPK, offer the guaranty of global optimum when the optimization is achieved. The issue with this approach is the computation time which is considerable in case of complex problems. The long time computation can be acceptable in case of anticipation but in real time situations this running time is not acceptable. For example, when the system interacts with user, recommend, for an ergonomic use, thirty seconds as maximum waiting time during interaction.

The Simulated annealing is not the fastest algorithm in absolute. But the interaction means changes and adaptations of an initial optimization problem. The idea in using the SA Algorithm is to take into account the results of optimizing process for an initial problem solved by Cplex, GLPK or other MILP solvers. The SA optimization takes as initial solution the one that is generated by the MILP solver for the initial problem. The new problem posed by interactions with occupant is quite different. The difference can be in the value of parameters like comfort temperature for example or additional constraints in the problem to describe a limitation like minimum of ventilation air flow because of steam cooking for example.

The simulation models used in SA has to describe the same phenomenon as for MILP. this common base is the guaranty of credibility in initialization point use.

The common description of the phenomenon is done in the pivot model, which is common for both MILP and SA optimization approaches. The recipe to transform the pivot model to simulation model is used. The SA algorithm 12 uses the model in simulation step.

The other important side of the simulated annealing optimization process is the neighborhood definition function. This function determine the direction and the jump of the value for degrees of freedom in each iteration. To illustrate the aims in definition of this function let's get a simple example. A continuous one dimension problem in term of inputs: only one degree of freedom x for the problem.

$$Obj_i = X_i^2 + X_i + 1 \tag{46}$$

A simple and universal neighborhood function is

$$X_n = X_{n-1} + rand(-1, 1) * radius \tag{47}$$

the radius is a parameter fixed for SA optimization algorithm

The optimization problem treated here is to find the minimum for the variable Obj . The results for this simple problem are shown in the convergence curve 13

The optimization algorithm used for the problem 46 is used on four parallel executions. This chose of four executions is

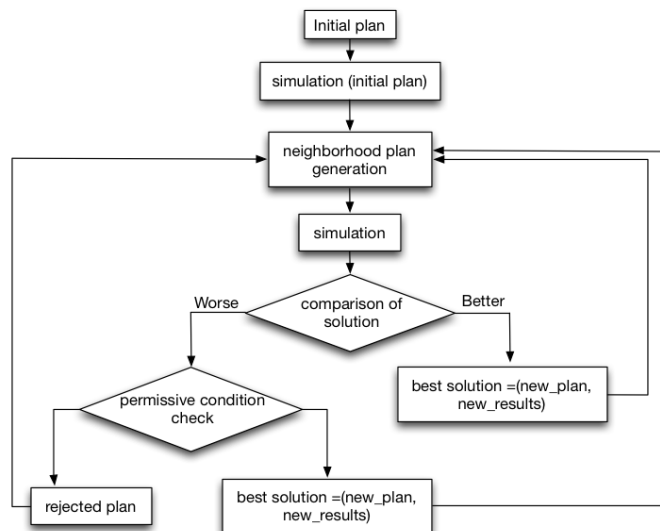


Fig. 12: Simulated Annealing Algorithm

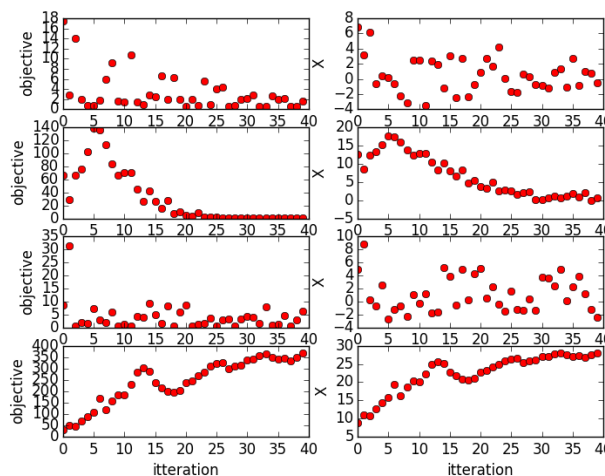


Fig. 13: optimal research evolution on four parallel processes

done because of four core processors in the computer used. It is to maximize the used of the processors during computation and to enhance chances to find best solution in the same time computation as simple execution.

The example shown elapse 00.017753 seconds with 40 iterations per process to find the best solution which is: (best plan: 1.0092122628718672, best performance: 0.5000424328936102).

The four parallel executions show that, some times, the algorithm diverges (the last process). There is no guaranty of convergence, that is why it is better to run parallel instances with limited number of iterations then to run simple instance with large number of iterations, it enhances chances to find solution.

In BEMS, the problem is more complex then this example

but it is in the same spirit. The complexity is added by the diversity of kinds of variables: continuous, discrete and binary. The multi-dimensions of the problem in BEMS are quite different than the single dimension problems treatment. In this paper we will not explain the details of treatment but we will draw major lines of the treatment procedure.

For these problems there is three main solutions:

the sequential treatment of variables The optimization process is done for each variable separately. The neighborhood is defined for the current variable to be optimized with considering the rest of variables fixed.

the global treatment of variables In this case, a new neighborhood is defined for the whole variables at each iteration. It is more random.

the clustering in the treatment of variables The clustering is a way to gather variables of the same type in term of optimization or to select variables which are physically close. The neighborhood of each vector (cluster of variables formed) is generated for each iteration in the SA running. The clusters are treated sequentially.

VIII. ANALYSIS OF OBTAINED RESULTS

After the computation with MILP solver (IBM ILOG CPLEX 12.3), the result obtained within 2 minutes for the classroom temperature and the corresponding cost for next 24 hours is given by figure 14.

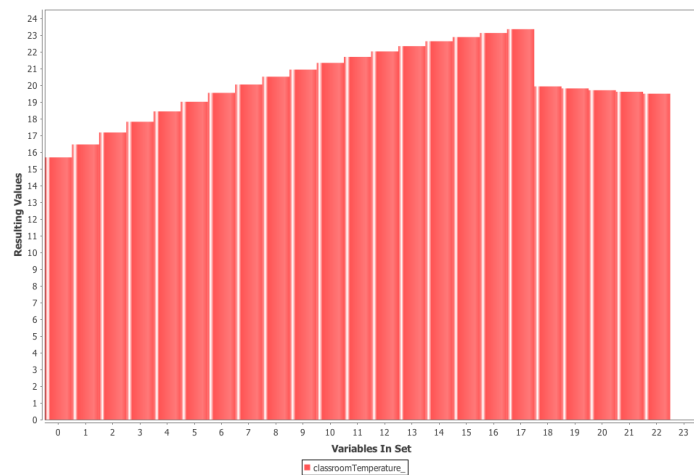


Fig. 14: Classroom temperature for 24 hours generated by MILP solver

Thanks to the pivot model, we kept the results of the MILP problem resolution and use it in the SA algorithm initializing. The problem of optimizing takes into account model of the envelop with its different faces and resistances and the HVAC system. The HVAC system is a couple of air ventilator and complex system of heating. The system of heating is water heated flowing in closed loop. In this description we consider only the power distributed by the exchangers of the heating system. An energy price model is used too. The degrees of freedom are the heating power injected by set-point temperature adjustment in the room and the air flow of the HVAC. These degrees of freedom projected in time constitute the

plan of management. The objective of the management here is quite interesting to develop. To generate the initial plan of management with cplex solver, the problem was considered with twice objectives coupled in optimizing objective. The first one are economic objective, it is to minimize the final bill during 24 hours. The second objective is comfort one, it is to maximize the occupants comfort according to comfort function description shown before. The objective for SA algorithm are exclusively economical. This difference is supposed to represent a will of a user during his interaction with the BEMS.

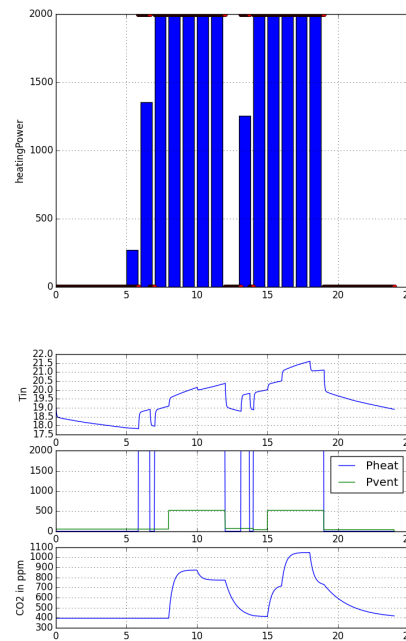


Fig. 15: initial plan with initial orders

The results presented in 15 represent the initial situation, before SA optimizing process. It is the results of CPLEX optimization with the mixed objective (economical and comfort) The results presented in 16, 17 represent two instances of SA runs. It is quite different because of the stochastic nature of the problem. The optimal solution for the problem given go toward zero for the power of heating. The optimum solution are not reached in the two tests but in the twice it was approached. The main idea for the use of SA in the BEMS is to give better solutions then the initial ones in acceptable time. This goal is more or less achieved by the results shown.

To enhance the efficiency of the algorithm and to use the maximum of computing capacities of the material, a parallelism process had been integrated. Four process run in parallel for 40 iterations per process. The elapsed time computation for this problem was 38.583835 seconds.

IX. CONCLUSIONS AND FUTUR WORKS

This paper presents a new method aiming at performing automatically model transformations to energy management system in buildings. The main contribution of this method is to avoid the model rewriting for different energy management applications that represent a significant time-consuming and

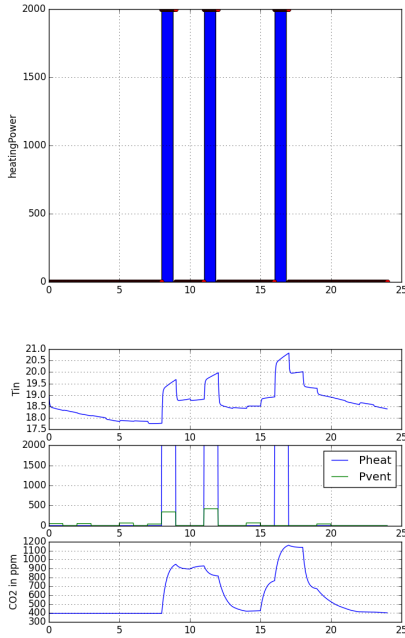


Fig. 16: SA result plan executions

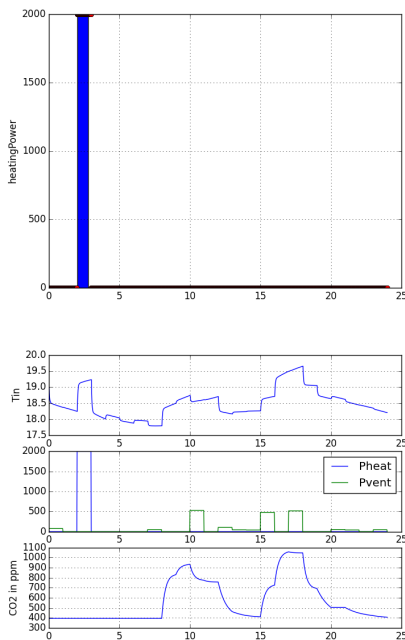


Fig. 17: SA result plan execution

error-prone. The core specifications of the construction a pivot model and the projection this last into target application models are shown via illustrative examples in order to give a better comprehension. The construction modular of pivot model encourages not only the reusability of element models but it allows to deal with equipment changes.

Based on this propose method, a software has been im-

plemented to valid this proposed method. Two kinds of target optimization models: MILP and causal simulable formalisms have been generated for the platform PREDIS/ENSE3. The first one allows generating a It is planed to extend the rule set to be able to generate results based on a global optimization approach while the second one shows acceptable results resulting of a fast heuristic local optimization approach in order to reduce time-consuming. To enhance this proposed method application field, different target applications such as : sizing, diagnosis, parameter estimation will be taken into account.

REFERENCES

REFERENCES

- [1] C. A. Balaras, A. G. Gaglia, E. Georgopoulou, S. Mirasgedis, Y. Sarafidis, and D. P. Lalas, European residential buildings and empirical assessment of the hellenic building stock, energy consumption, emissions and potential energy savings, *Building and Environment*, Volume 42, Issue 3, pp:1298-1314, 2007.
- [2] Europe's Energy Position: Markets and Supply. Market Observatory for Energy, Directorate General for Energy, report 2009
- [3] G. Levermore, *Building Energy Management Systems: An Application to Heating, Natural Ventilation, Lighting and Occupant Satisfaction*, Taylor & Francis, 2002.
- [4] S. Kent, Model driven engineering, in *Integrated Formal Methods*, ser. Lecture Notes in Computer Science, M. Butler, L. Petre, and K. Sere, Eds. Springer Berlin Heidelberg, Volume 2335, pp:286298, 2002.
- [5] P. Harmon, Mda: an idea whose time has come, Technical report, 2003.
- [6] H.A. Dang, B. Delinchant, S. Gaaloul, F. Wurtz, Electrical performance optimization of an HVAC dual flow, *Electrimacs 2011*, Cergy-Pontoise, France, 2011.
- [7] H. Doukas, K. D. Patlitzianas, K. Iatropoulos and J. Psarras, Intelligent building energy management system using rule sets, *Building and Environment*, Volume 42, Issue 10, pp:3562-3569, October 2007.
- [8] PSE. gPROMS. <http://www.psenterprise.com/gproms/>, 2011.
- [9] M. R. Bussieck, A. Meeraus, General algebraic modeling system (GAMS), In *Modeling languages in mathematical optimization*, pp:137-157, 2004.
- [10] Europe's Energy Position: Markets and Supply. Market Observatory for Energy, Directorate General for Energy, report 2009.
- [11] S.A. Ahmadi, I. Shames, F. Scotton, L. Huang; H. Sandberg, K.H. Johansson and B. Wahlberg, Towards more Efficient Building Energy Management Systems, Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on , pp:118-125, 2012.
- [12] D. L. Ha, H. Joumaa, S. Ploix, and M. Jacomino, An optimal approach for electrical management problem in dwellings, *Energy and Buildings*, Volume 45, Issue 0, pp:114, 2012.
- [13] C.A. Floudas, and P.M. Pardalos, *Encyclopedia of Optimization*, Springer Reference, 2008
- [14] S. S. Rao, *Engineering Optimization: Theory and Practice*, Third Edition, Wiley, 1996.

- [15] M. Grotschel and L. Lovász, Combinatorial optimization, Chapter 28 in R.L. Graham, Handbook of Combinatorics, Elsevier, Volume 2, pp:1541-1597, 1995.
- [16] C. H. Papadimitriou and K. Steiglitz, Combinatorial optimization: algorithms and complexity. Courier Dover, 1998.
- [17] C. Blum and A. Roli, Metaheuristics in combinatorial optimization: Overview and conceptual comparison, ACM Computing Surveys, pp:268-308, 2003
- [18] M. R. Garey and D. S. Johnson, Computers and Intractability: a guide to the theory of NP-completeness, W. H. Freeman, 1979.
- [19] I. E. Grossmann and Z. Kravanja, Mixed-integer nonlinear programming techniques for process systems engineering, Computers & Chemical Engineering, Volume 19, Supplement 1, pp:189-204, 1995.
- [20] E. L. Lawler and D. E. Wood, Branch-and-bound methods: A survey, Operations research 14.4, pp:699-719, 1966.
- [21] R. E. Barber and H. C. Lucas, System response time operator productivity, and job satisfaction, Communications of the ACM, pp:972-986, 1983.
- [22] B. Shneiderman, Response time and display rate in human performance with computers, ACM Computing Surveys 16.3, pp:265-285, 1984.
- [23] L. Jan, Impact of system response time on state anxiety, Communications of the ACM, pp:342-347, 1988.
- [24] V. Laarhoven, J. M. Peter and E. H. L. Aarts, Simulated annealing. Springer Netherlands, 1987.
- [25] H. Szu and R. Hartley, Fast simulated annealing, Physics Letters A, Volume 122, Issues 34, pp:157-162, 1987.
- [26] H. L. Minh, S. Ploix, M. Jacomino, D.L. Ha, A mixed integer programming formulation of the home energy management problem, Energy Management, INTECH, 2010.
- [27] S. G. Chouikh, Interopérabilité basée sur les standards Modelica et composant logiciel pour la simulation énergétique des systèmes de bâtiment, PhD Thesis, University of Grenoble, France, 2012.
- [28] M. Blanke, C. W. Frei, F. Kraus, R. J. Patton, M. Staroswiecki, What is fault-tolerant control?, Aalborg University, Department of Control Engineering, 2000.
- [29] A. L. Dulmage and N. S. Mendelsohn, Coverings of bipartite graphs, Canadian Journal of Mathematics, Volume 10, pp:517-534, 1958.
- [30] J. S. Ulmer, J. P. Belaud, and J. M. Le Lann, Towards a pivotal - based approach for business process alignment, International Journal of Computer Integrated Manufacturing, Volume 24, no. 11, pp:1010-1021, 2011.
- [31] J. Bézivin, On the unification power of models, Software & Systems Modeling, Volume 4, no. 2, pp:171-188, 2005.
- [32] Q. Omg, Meta object facility (mof) 2.0 query/view/transformation specification, Final Adopted Specification (November 2005), 2008.
- [33] K. Czarnecki and S. Helsen, Classification of model transformation approaches, Proceeding of the 2nd OOP-SLA Workshop on Generative Techniques in the Context of the Model Driven Architecture, Volume 45, pp:1-17, 2003.
- [34] Modelica, Modelica® - A Unified Object-Oriented Language for Physical Systems Modeling, Technical report, Volume 3.2, 2010.
- [35] J. Von Zur Gathen, J. Gerhard, Modern Computer Algebra, Cambridge University Press, 2003.
- [36] W.S. Brown, A.C. Hearn, Applications of symbolic algebraic computation, Computer Physics Communications, Volume 17, Issues 12, pp:207-215, 1979.
- [37] F. Winkler, Advances and Problems in Algebraic Computation, in Contributions to General Algebra 12 (Proc. AAA'58, Vienna), Verlag Johannes Heyn, Klagenfurt (Austria), 2000.
- [38] M. Wester, A review of CAS mathematical capabilities, Computer Algebra Nederland, Volume 13, pp:41-48, 1994.
- [39] A. Dall'Osso, Computer algebra systems as mathematical optimizing compilers, Science of Computer Programming, Volume 59, Issue 3, pp:250-273, 2006.
- [40] P. Bernard, Giac/Xcas, a free computer algebra system. Technical report, University of Grenoble, 2008.
- [41] <http://www-fourier.ujf-grenoble.fr/parisse/giac.html>
- [42] G. D. Oliveira, M. Jacomino, D. L. Ha and S. Ploix, Optimal power control for smart homes, 18th IFAC World Congress, Elsevier, 2011.
- [43] Y. Hadjsaid, B. Lechat, C. Latremoliere and S. Ploix, Generating global energy management problems: application to the CANOPEA prototype, Building Simulation (BS2013), IBPSA World Conference, Chambéry, France, 2013.
- [44] A. Le Mounier, B. Delinchant and S. Ploix. Determination of relevant model structures for self-learning energy management system. In Building Simulation and Optimization 2014, London, U.K., 23-24 June 2014.
- [45] S. Sarabi, S. Ploix, H. Minh Le, H.-A. Dang and F. Wurtz. Assessing the relevance of reduced order models for building envelop. In Building Simulation (BS 2013), 2013.
- [46] S. Gaaloul, B. Delinchant, F. Wurtz and F. Verdire. Software components for dynamic building simulation. In Proceedings of Building Simulation 2011: 12th Conference of IBPSA 2011, Sydney, Australie, 14-16 November 2011.
- [47] S. Gaaloul, X. Hoa Binh Le, B. Delinchant, F. Wurtz and S. Ploix. Software component architecture for co-simulation applied to the coupling between a building's thermal envelope and its inhabitant behaviour. In the 38th Annual Conference of IEEE Industrial Electronics (IECON), pp: 4620 - 4625, Montréal, Canada, 25-28 October 2012.
- [48] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi Optimization by Simulated Annealing. In Science, New Series, Vol.220, No 4598 May 1983
- [49] A. Basu, LN Frazer Rapid determination of the critical temperature in simulated annealing inversion. In Science 1990