



International Journal of Advanced Computer Science and Applications

Volume 6 Issue 12

December 2015



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 6 Issue 12 December 2015

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

- **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 Karim ABED**
- **Abdur Rashid Khan**
Gomal University
- **Abeer ELkorany**
Faculty of computers and information, Cairo
Univesity
- **ADEMOLA ADESINA**
University of the Western Cape
- **Aderemi A. Atayero**
Covenant University
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**
Ahlia University
- **Ahmed Nabih Zaki Rashed**
Menoufia University
- **Akbar Hossain**
- **Akram Belghith**
University Of California, San Diego
- **Albert S**
Kongu Engineering College
- **Alcinia Zita Sampaio**
Technical University of Lisbon
- **Alexandre Bouënard**
Sensopia
- **Ali Ismail Awad**
Luleå University of Technology
- **Amitava Biswas**
Cisco Systems
- **Anand Nayyar**
KCL Institute of Management and Technology,
Jalandhar
- **Andi Wahju Rahardjo Emanuel**
Maranatha Christian University
- **Andrews Samraj**
Mahendra Engineering College
- **Anirban Sarkar**
National Institute of Technology, Durgapur
- **Antonio Formisano**
University of Naples Federico II
- **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 EL-SAYED**
Computer Science and Eng. Dept., Faculty of
Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Badre Bossoufi**
University of Liege
- **BALAMURUGAN RAJAMANICKAM**
- **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
 - **Brahim Raouyane**
FSAC
 - **Bright Keswani**
Department of Computer Applications, Suresh Gyan Vihar University, Jaipur (Rajasthan) INDIA
 - **Brij Gupta**
University of New Brunswick
 - **C Venkateswarlu Sonagiri**
JNTU
 - **Chandrashekhar Meshram**
Chhattisgarh Swami Vivekananda Technical University
 - **Chao Wang**
 - **Chao-Tung Yang**
Department of Computer Science, Tunghai University
 - **Charlie Obimbo**
University of Guelph
 - **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
 - **Dana PETCU**
West University of Timisoara
 - **Daniel Albuquerque**
 - **Dariusz Jakóbczak**
Technical University of Koszalin
 - **Deepak Garg**
Thapar University
 - **Dheyaa Kadhim**
University of Baghdad
 - **Dong-Han Ham**
Chonnam National University
 - **Dr Kannan**
Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750, Tronoh, Perak, Malaysia
 - **Dr KIRAN POKKULURI**
Professor, Sri Vishnu Engineering College for Women
 - **Dr. Harish Garg**
Thapar University Patiala
 - **Dr. Manpreet Manna**
Director, All India Council for Technical Education, Ministry of HRD, Govt. of India
 - **Dr. Mohammed Hussein**
 - **Dr. Sanskruti Patel**
Charotar Univeristy of Science & Technology, Changa, Gujarat, India
 - **Dr. Santosh Kumar**
Graphic Era University, Dehradun (UK)
 - **Dr. JOHN MANOHAR**
VTU, Belgaum
 - **Dragana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational sciences
 - **Driss EL OUADGHIRI**
 - **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
 - **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 Mastorakis**
Technological Educational Institute of Crete
- **George Pecherle**
University of Oradea
- **Georgios Galatas**
The University of Texas at Arlington
- **Gerard Dumancas**
Oklahoma Baptist University
- **Ghalem Belalem**
University of Oran 1, Ahmed Ben Bella
- **Giacomo Veneri**
University of Siena
- **Giri Babu**
Indian Space Research Organisation
- **Govindarajulu Salendra**
- **Grebenisan Gavril**
University of Oradea
- **Gufran Ahmad Ansari**
Qassim University
- **Gunaseelan Devaraj**
Jazan University, Kingdom of Saudi Arabia
- **GYÖRÖDI ROBERT STEFAN**
University of Oradea
- **Hadj Tadjine**
IAV GmbH
- **Hamid Alinejad-Rokny**
The University of New South Wales
- **Hamid Mukhtar**
National University of Sciences and Technology
- **Hamid AL-Asadi**
Department of Computer Science, Faculty of Education for Pure Science, Basra University
- **Hany Hassan**
EPF
- **Harco Leslie Hendric SPITS WARNARS**
Surya university
- **Hazem I. El Shekh Ahmed**
Pure mathematics
- **Hesham Ibrahim**
Faculty of Marine Resources, Al-Mergheb University
- **Himanshu Aggarwal**
Department of Computer Engineering
- **Hossam Faris**
- **Huda K. AL-Jobori**
Ahlia University
- **Iwan Setyawan**
Satya Wacana Christian University
- **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 Jia**
Assistant Professor
- **Jim Wang**
The State University of New York at Buffalo, Buffalo, NY
- **John Sahlin**
George Washington University
- **JOSE PASTRANA**
University of Malaga
- **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
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Khalid Sattar Abdul**

- Assistant Professor
- **Khin Wee Lai**
Biomedical Engineering Department, University Malaya
 - **KITIMAPORN CHOCHOTE**
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
 - **Lazar Stošić**
Collegefor professional studies educators Aleksinac, Serbia
 - **Leandros 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**
Masters in Cyber Law & Information Security
 - **Manju Kaushik**
 - **Manoharan P.S.**
Associate Professor
 - **Manoj Wadhwa**
Echelon Institute of Technology Faridabad
 - **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. 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
 - **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
 - **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 Azzeh**
Applied Science university
 - **Mohammad Alomari**
Applied Science University
 - **Mohammad Haghighat**
University of Miami
 - **Mohammed Kaiser**
Institute of Information Technology
 - **Mohammed Sadgal**
Cadi Ayyad University
 - **Mohammed Al-shabi**
Associate Professor
 - **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science & Technology
 - **Mohd Helmy Abd Wahab**
Universiti Tun Hussein Onn Malaysia
 - **Mona Elshinawy**
Howard University
 - **Mostafa Ezziyyani**
FSTT

- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**
University Malaysia Pahang
- **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
- **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) India
- **Neeraj Tiwari**
- **Nestor Velasco-Bermeo**
UPFIM, Mexican Society of Artificial Intelligence
- **Nidhi Arora**
M.C.A. Institute, Ganpat University
- **Nilanjan Dey**
- **Ning Cai**
- **Noura Aknin**
University Abdelamlek Essaadi
- **Oliviu Matei**
Technical University of Cluj-Napoca
- **Om Sangwan**
- **Omaima Al-Allaf**
Asesstant Professor
- **Osama Omer**
Aswan University
- **Ousmane THIARE**
Associate Professor University Gaston Berger of
Saint-Louis SENEGAL
- **Paresh V Virparia**
Sardar Patel University
- **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
- **Professor Ajantha Herath**
- **Purwanto Purwanto**
- **Qifeng Qiao**
University of Virginia
- **Rachid Saadane**
EE departement EHTP
- **raed Kanaan**
Amman Arab University
- **Raghuraj Singh**
Harcourt Butler Technological Institute
- **Rahul Malik**
- **Raja Ramachandran**
- **raja boddu**
LENORA COLLEGE OF ENGINEERNG
- **Rajesh Kumar**
National University of Singapore
- **Rakesh Dr.**
Madan Mohan Malviya University of Technology
- **Rakesh Balabantaray**
IIIT Bhubaneswar
- **Rashad Al-Jawfi**
Ibb university
- **Rashad Al-Jawfi**
Ibb university
- **Rashid Sheikh**
Shri Aurobindo Institute of Technology, Indore
- **Ravi Prakash**
University of Mumbai
- **RAVINDRA 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
- **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
- **Sasan Adibi**
Research In Motion (RIM)
- **Satyendra Singh**
Professor
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Seema Shah**
Vidyalankar Institute of Technology Mumbai,
- **Selem Charfi**
University of Pays and Pays de l'Adour
- **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**
- **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**
- **Sohail Jabbar**
Bahria University
- **Sri Devi Ravana**
University of Malaya
- **Sudarson Jena**
GITAM University, Hyderabad
- **Suhas J Manangi**
Microsoft
- **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
- **Süleyman Eken**
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia
- **Sumit Goyal**
National Dairy Research Institute
- **Suresh Sankaranarayanan**
Institut Teknologi Brunei
- **Susarla Sastry**
JNTUK, Kakinada
- **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
- **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
- **Urmila Shrawankar**
GHRCE, Nagpur, India
- **Vaka MOHAN**
TRR COLLEGE OF ENGINEERING
- **VENKATESH JAGANATHAN**
- **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
- **Wei Wei**
Xi'an Univ. of Tech.
- **Wenbin Chen**
360Fly
- **Xiaoqing Xiang**
AT&T Labs
- **Xiaolong Wang**
University of Delaware
- **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
- **Zenzo Ncube**
North West University
- **Zhao Zhang**
Deptment of EE, City University of Hong Kong
- **Zhixin Chen**
ILX Lightwave Corporation
- **Ziyue Xu**
National Institutes of Health, Bethesda, MD
- **Zlatko Stapic**
University of Zagreb, Faculty of Organization and
Informatics Varazdin
- **Zuraini Ismail**
Universiti Teknologi Malaysia

CONTENTS

Paper 1: Introducing a Method for Modeling Knowledge Bases in Expert Systems Using the Example of Large Software Development Projects

Authors: Franz Felix Füssl, Dellef Streifferdt, Weijia Shang, Anne Triebel

PAGE 1 – 7

Paper 2: A Prediction Model for Mild Cognitive Impairment Using Random Forests

Authors: Haewon Byeon

PAGE 8 – 12

Paper 3: Spectrum Sensing Methodologies for Cognitive Radio Systems: A Review

Authors: Ireyuwa E. Igbinosa, Olutayo O. Oyerinde, Viranjay M. Srivastava, Stanley Mneney

PAGE 13 – 22

Paper 4: A Posteriori Pareto Front Diversification Using a Copula-Based Estimation of Distribution Algorithm

Authors: Abdelhakim Cheriet, Foudil Cherif

PAGE 23 – 35

Paper 5: Vitality Aware Cluster Head Election to Alleviate the Wireless Sensor Network for Long Time

Authors: P. Thiruvannamalai Sivasankar, Dr. M. RamaKrishnan

PAGE 36 – 41

Paper 6: Designing an IMS-LD Model for Collaborative Learning

Authors: Fauzi El Moudden, Prof. Mohamed Khaldi, Prof. Aammou Souhaib

PAGE 42 – 48

Paper 7: An Enhanced Steganographic Model Based on DWT Combined with Encryption and Error Correction Techniques

Authors: Dr.Adwan Yasin, Mr.Nizar Shehab, Dr.Muath Sabha, Mariam Yasin

PAGE 49 – 55

Paper 8: A Multimedia System for Breath Regulation and Relaxation

Authors: Wen-Ching Liao, Han-Hong Lin, He-Lin Ruo, Po-Hsiang Hsu

PAGE 56 – 63

Paper 9: A Secure Network Communication Protocol Based on Text to Barcode Encryption Algorithm

Authors: Abusukhon Ahmad, Bilal Hawashin

PAGE 64 – 70

Paper 10: Comparison Contour Extraction Based on Layered Structure and Fourier Descriptor on Image Retrieval

Authors: Cahya Rahmad, Kohei Arai

PAGE 71 – 74

Paper 11: Arabic Sentiment Analysis: A Survey

Authors: Adel Assiri, Ahmed Emam, Hmood Aldossari

PAGE 75 – 85

Paper 12: A Novel Ball on Beam Stabilizing Platform with Inertial Sensors

Authors: Ali Shahbaz Haider, Muhammad Bilal, Samter Ahmed, Saqib Raza, Imran Ahmed

PAGE 86 – 94

Paper 13: A Feature Analysis of Risk Factors for Stroke in the Middle-Aged Adults

Authors: Haewon Byeon, Hyeung Woo Koh

PAGE 95 – 99

Paper 14: Analysis on Existing Basic Slas and Green Slas to Define New Sustainable Green SLA

Authors: Iqbal Ahmed, Hiroshi Okumura, Kohei Arai

PAGE 100 – 108

Paper 15: EMCC: Enhancement of Motion Chain Code for Arabic Sign Language Recognition

Authors: Mahmoud Zaki Abdo, Alaa Mahmoud Hamdy, Sameh Abd El-Rahman Salem, Elsayed Mostafa Saad

PAGE 109 – 117

Paper 16: A Novel Approach for Ranking Images Using User and Content Tags

Authors: Arif Ur Rahman, Muhammad Muzammal, Humayun Zaheer Ahmad, Awais Majeed, Zahoor Jan

PAGE 118 – 123

Paper 17: A Disaster Document Classification Technique Using Domain Specific Ontologies

Authors: Qazi Mudassar Ilyas

PAGE 124 – 130

Paper 18: Intrusion Detection System in Wireless Sensor Networks: A Review

Authors: Anush Ananthakumar, Tanmay Ganediwal, Dr. Ashwini Kunte

PAGE 131 – 139

Paper 19: A Survey on the Internet of Things Software Architecture

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

PAGE 140 – 143

Paper 20: A Carrier Signal Approach for Intermittent Fault Detection and Health Monitoring for Electronics Interconnections System

Authors: Syed Wakil Ahmad, Dr. Suresh Perinpanayagam, Prof. Ian Jennions, Dr. Mohammad Samie

PAGE 144 – 150

Paper 21: A Synchronous Stream Cipher Generator Based on Quadratic Fields (SSCQF)

Authors: Younes ASIMI, Ahmed ASIMI

PAGE 151 – 160

Paper 22: Pneumatic Launcher Based Precise Placement Model for Large-Scale Deployment in Wireless Sensor Networks

Authors: Vikrant Sharma, R B Patel, H S Bhadauria, D Prasad

PAGE 161 – 167

Paper 23: Tree-Combined Trie: A Compressed Data Structure for Fast IP Address Lookup

Authors: Muhammad Tahir, Shakil Ahmed

PAGE 168 – 175

Paper 24: Performance Evaluation of K-Mean and Fuzzy C-Mean Image Segmentation Based Clustering Classifier

Authors: Hind R.M Shaaban, Farah Abbas Obaid, Ali Abdulkarem Habib

PAGE 176 – 183

Paper 25: Identifying Cancer Biomarkers Via Node Classification within a Mapreduce Framework

Authors: Taysir Hassan A. Soliman

PAGE 184 – 189

Paper 26: Intelligent Mobility Management Model for Heterogeneous Wireless Networks

Authors: Sanjeev Prakash, R B Patel, V. K. Jain

PAGE 190 – 196

Paper 27: Development of Adaptive Mobile Learning (AML) on Information System Courses

Authors: I Made Agus Wirawan, Made Santo Gitakarna

PAGE 197 – 202

Paper 28: Ontology-Based Clinical Decision Support System for Predicting High-Risk Pregnant Woman

Authors: Umar Manzoor, Muhammad Usman, Mohammed A. Balubaid, Ahmed Mueen

PAGE 203 – 208

Paper 29: Distributed Optimization Model of Wavelet Neuron for Human Iris Verification

Authors: Elsayed Radwan, Mayada Tarek, Abdullah Baz

PAGE 209 – 218

Paper 30: Composable Modeling Method for Generic Test Platform for Cbtc System Based on the Port Object

Authors: WAN Yongbing, WANG Daqing, MEI Meng

PAGE 219 – 225

Paper 31: JPI UML Software Modeling

Authors: Cristian Vidal Silva, Leopoldo López, Rodolfo Schmal, Rodolfo Villarreal, Miguel Bustamante, Víctor Rea Sanchez

PAGE 226 – 235

Paper 32: Association Rule Hiding Techniques for Privacy Preserving Data Mining: A Study

Authors: Gayathiri P, Dr. B Poorna

PAGE 236 – 242

Paper 33: Improving Video Streams Summarization Using Synthetic Noisy Video Data

Authors: Nada Jasim Al-Musawi, Saad Talib Hasson

PAGE 243 – 249

Paper 34: A New Algorithm for Post-Processing Covering Arrays

Authors: Carlos Lara-Alvarez, Himer Avila-George

PAGE 250 – 254

Paper 35: Database Preservation: The DBPreserve Approach

Authors: Arif Ur Rahman, Muhammad Muzammal, Gabriel David, Cristina Ribeiro

PAGE 255 – 266

Paper 36: Detection of Denial of Service Attack in Wireless Network using Dominance based Rough Set

Authors: N. Syed Siraj Ahmed, D. P. Acharjya

PAGE 267 – 278

Paper 37: Enhanced Version of Multi-algorithm Genetically Adaptive for Multiobjective optimization

Authors: Wali Khan Mashwani, Abdellah Salhi, Muhammad Asif jan, Rashida Adeeb Khanum, Muhammad Sulaiman

PAGE 279 – 287

Paper 38: Extracting Topics from the Holy Quran Using Generative Models

Authors: Mohammad Alhawarat

PAGE 288 – 294

Paper 39: Localisation of Information and Communication Technologies in Cameroonian Languages and Cultures: Experience and Issues

Authors: Mathurin Soh, Jean Romain Kouesso, Laure Pauline Fotso

PAGE 295 – 300

Paper 40: Real-Time Talking Avatar on the Internet Using Kinect and Voice Conversion

Authors: Takashi Nose, Yuki Igarashi

PAGE 301 – 307

Introducing a Method for Modeling Knowledge Bases in Expert Systems Using the Example of Large Software Development Projects

Franz Felix Füssl, Detlef
Streitferdt

Institute for Computer and Systems
Engineering
Technische Universität Ilmenau
Ilmenau, Germany

Weijia Shang

Computer Engineering Department
Santa Clara University
Santa Clara (California), United
States of America

Anne Triebel

Institute for Business Information
Systems Engineering
Technische Universität Ilmenau
Ilmenau, Germany

Abstract—Goal of this paper is to develop a meta-model, which provides the basis for developing highly scalable artificial intelligence systems that should be able to make autonomously decisions based on different dynamic and specific influences. An artificial neural network builds the entry point for developing a multi-layered human readable model that serves as knowledge base and can be used for further investigations in deductive and inductive reasoning. A graph-theoretical consideration gives a detailed view into the model structure. In addition to it the model is introduced using the example of large software development projects. The integration of Constraints and Deductive Reasoning Element Pruning are illustrated, which are required for executing deductive reasoning efficiently.

Keywords—*Knowledge Engineering; Ontology Engineering; Knowledge Modelling; Knowledge Base; Expert System; Artificial Intelligence; Deductive Reasoning Element Pruning*

I. INTRODUCTION

IT technologies are subjects to a fast changeable field of application. Software development teams have to adapt continuously for fitting newest stakeholder needs and finding success in the market. Especially success of large software development projects for example product line developments depends on many different influencing factors, introduced in [1]. These influencing factors determine for example the composition of teams, the choice of software tools or the selection of a suitable software development process.

There are a couple of previous and current projects with the goal of developing an open source expert system (ES) including the ability of machine learning in a specific field. Examples are [2], the “scikit-learn” library [3], the “Mlpy” library [4] or “Orange” library [5]. As opposed to these projects and publications the project behind this paper focuses on large software developments that typically have many various influences and a large set of required or requested software tools and business artifacts.

The most important part of an ES is the basis of decisions. There are a couple of systems, based on a simple decision tree or relational data models [6]. But currently in the domain of software developments there is no appropriated model for illustrating knowledge bases (KB) [6], which are necessary for

automated handling machine learning and deductive reasoning. For this reason an abstract human readable meta-model (defined in [7]) should be developed that deals as architectural basis for further investigations in autonomous decision making having regard to different specific influences as project-specific, personal, economic-driven, product-related or technology-based.

A. Knowledge bases in Expert Systems

According to basic literature as [8], [9] or [10] an ES is a knowledge-based system that is used for intelligent assistance, decision making or problem solving. Main goal of the research project behind this paper is to develop a system that is able to detect any objects that are helpful to bring a software project to success. Thus the system needs to make decisions for solving a specific problem. To do so, it is necessary, to have a profound KB what can be used for inference, in particular deductive and inductive reasoning.

So what exactly is a KB and why it is important to consider? According to [11] or [12] KB are specialized bodies or nets of knowledge and skills. So it is a construct of information, data and associations, where knowledge can be generated and derived by “heuristics or informal ‘rules of thumb’ experts” and “reasoning methods” [13]. In the field of this research project the KB, which is to develop, contains knowledge of software developers, project managers, software architects, software producer or experts and consultant in the field of software development processes.

B. The origins of the model: Ontology, Topic Map and Artificial Neural Network

According to [14] an artificial neural network (ANN) is organized into layers with processing elements, called units. Every unit has its own specific setup, but they are similar in activation events and output functions. Associations can be done among units of the same layer and between elements on different layers. The units are associated by weighted connection paths. As described in [15] “successful training [of ANN] can result [...] in performing tasks such as predicting an output value, classifying an object [...] and completing a known pattern”.

An ANN seems to be the right KB to pursue the goal of this paper. But there are differences to the underlying KB of this paper. ANN work with an unspecified number of layers. They have a known input pattern and a known output pattern. In case of a multi-layer feedforward ANN there is at least one hidden layer in-between these patterns [16], which leads to a worse human readability. Also it is not provided to have different views to the knowledge, for instance a descriptive and a deciding view. The model should be able to perform inferring processes, making decisions and edit knowledge with a focus to human-readability.

With searching a possible solution for solving the human-readability problem, Ontologies have to be mentioned. As described in [6] they serve as method for representing knowledge and it is possible to integrate machine learning by 'Ontology learning'. The problem of Ontologies is the degree of formalization, which is too low for scalable reasoning. Thus generic usable and efficient deductive reasoning algorithms are difficult to integrate and not a goal of Ontologies. Nevertheless the main idea of Ontologies, the descriptive functionality, has been used for developing the model behind this work. [6]

Another possibility to represent knowledge is creating Topic Maps. But here are no approaches for machine learning integration. As with Ontologies the focus of Topic Maps is the presentational view of knowledge. [6]

Taken as a whole, the KB, which builds a foundation of the model in this work, is lightly adopted to multi-layered feedforward ANN with advantages of Ontologies and Topic Maps.

II. MODELING THE KNOWLEDGE BASE

The architecture of the knowledge base can be described as five-layered meta-model. The layers represent different abstraction levels, where information can be stored and processed. Fig. 1 shows these five layers of the meta-model.

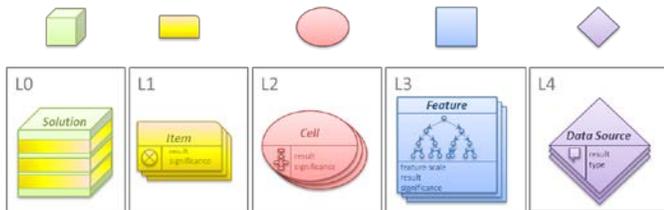


Fig. 1. Multi-layered Meta-Model: Layer Overview

One layer contains at least one *element*, exemplified by repetitive geometric figures of Fig. 1. Elements are associated by different connection types. Thus association rule learning is possible, as it is used for example in artificial intelligence (AI) or Data Mining [17]. With these associations, elements have relations to each other within the same layer or through different layers. The model provides best possibilities for using the principle of loose coupling, which leads to high interchangeability and extensibility of all containing elements [18]. Each element contains implicitly a problem, a specific way for finding the solutions (logical part) and a result for using the element explicitly. The models hierarchical and logical structure helps to break these problems down into

manageable pieces, which is one of the major and well known paradigms in AI [19], [20], [21].

Beginning with the general description of terminology the models characterization is followed below:

A. Layer description and terminology

The Layer **L4** contains the *Data Source* elements of the KB. It serves as application layer and forms the main communication between the AI system and project participants (users), hardware or software interfaces. Elements are for instance questions, measuring tools, sensor technologies or API definitions. This layer collects primary circumstances by a simple request-response method. Thus it is possible to gather actual issues and specific factors of influence.

Each factor of influence is stored as *Feature* in layer **L3**. Examples of these Features are "project budget", "operating system", "personal motivation" or "personal experience". Impacts of each feature correlate closely with research and technological development, which is why layer **L3** is subject to high degree of variability. For simplifying the analysis of results, features are classified in nominal scaled, ordinal scaled and metric scaled. This classification depends on the connected data source element and bases for example on different types of questions as "multiple-choice" and "single-choice" or on different types of input data, as Integer and String. Input data mean typically data generated by measurements. Questions could have this data type too, especially when user input is necessary.

Each Feature is associated with at least one *Cell* that is collected by another layer of the model: **L2**. According to Landauer each Cell could be described as context block [22], which collects data by their associations and interprets it as information by deposited algorithms. Thus they represent the basis for generating knowledge. In addition to analyze specifically adjacent Features, Cells are able to access other Cells by connecting among each other and involving the associated result into the own analyses. They are weighted differently according to their number of associations. Cells can be dependent on each other, for instance "Scrumwise" as software solution and "Scrum" as development process.

The information that is stored in Cells is used by specific *Items*, which build layer **L1**. Each Item contains an abstract component that could be necessary for realizing the project. Items serve as partial solution for reacting to a determined problem. For that reason all Items have to examine carefully, how they conduce to project's success. They can be used in different parts and steps of the project. For instance, one Item symbolizes "requirements engineering", which is necessary for product management and very important for project success. Further examples of Items are "software development process", which has to be ascertained or "software architecture pattern".

For determination Cells and Items in a more understandable manner: **L1** (layer of Items) serves as 'descriptive' view on the information of the KB, while **L2** (layer of Cells) builds a 'deciding' view within the system using the KB. Cells contain something concrete while Items are more a general view into the KB.

The solution layer **L0** represents a complete build package for solving a predefined problem, for example finding the ‘projectalized’ development process or a customized developer environment. The decision of combining project relevant elements bases on a simple suitability test. Each association of every single Item verifies its project suitability. Then the Items determine their linked Cell with the highest suitability value. The corresponding results are abstracted to a project-specific solution.

B. Mathematical consideration as graph

When describing the five-layered meta-model with basic definitions of graph and set theory, for example by [23] or [24], the model is a weighted directed graph $G = (V, E, i)$ without loops and a non-regular property. The layers (L_0 to L_4) are sets of ordered pairs (V, E) with

- $V(G)$ as finite set of all nodes (elements of the layers)
- $E(G)$ as set of all edges (associations) and
- $i = i_G$ as mapping that assigns to each edge $e \in E$ a pair $i(e) = \{x, y\}$ with elements $x, y \in V$.

Set V can be divided into five subsets, as shown in Fig. 2:

$$L_0, L_1, L_2, L_3, L_4 \subset V(G)$$

They contain each element of the different layers of the model. L_2 (Cells layer) and L_1 (Items layer) are exceptional as opposed to the other layers. They are induced subgraphs $C_1, C_2 \subset G$. Within L_1 and L_2 it is possible to build edges between the nodes (on the same layer). According to [25] C_1 and C_2 have maximum associations of

$$\binom{n}{2} = \frac{n(n-1)}{2} \text{ with } n = |L_2| \text{ or}$$

$$\binom{m}{2} = \frac{m(m-1)}{2} \text{ with } m = |L_1|$$

This maximum achievable number of nodes of L_2 and L_1 implies the completeness of the respective subgraphs. It can only be achieved by associating each element of one layer with every other element of the same layer.

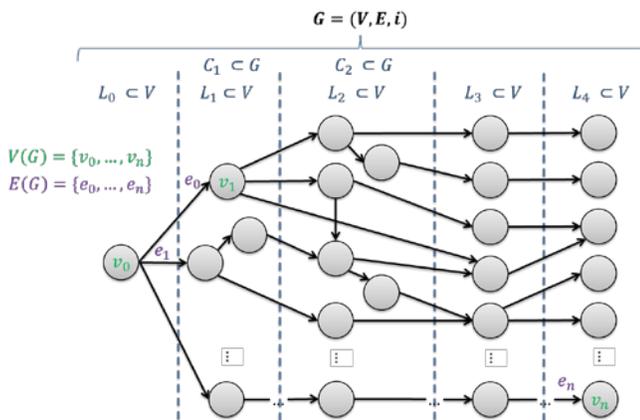


Fig. 2. Model illustration of an example as graph

By integration of weights deductive reasoning can be accelerated. Goal is processing the graph on prioritized paths,

for faster reaching important elements (nodes, information). Here the importance of an element can be concluded from the degree centrality, similar to the PageRank-Algorithm [26]. According to this the significance of a node depends mainly on the number of its edges. [26]

The set $E_G(v)$ is the set of all adjacent edges of a node $v \in V$. Further the set $E_G^+(v)$ is defined as a set containing all edges to successors and $E_G^-(v)$ as a set with all edges to predecessors of one node v .

The weighting is made by assigning an edge $e \in E_G(v)$ a weight $\omega(e)$, which is a real number $\omega : E \rightarrow \mathbb{R}$ [27]. It can be simplified by $\omega : E \rightarrow \mathbb{Z}$. Each node is able to attach its degree $d_G(v)$ as weight to all of its adjacent edges. If an edge has already had a higher weight, this edge keeps the original weight value:

$$\forall v \in V: \omega(e_v) < d_G(v) \rightarrow \omega(e_v) = d_G(v) \text{ with } e_v \in E_G(v)$$

Each node has a specific result that can influence a calculation of a neighbors result. The result of a node $v \in V$ is defined as $r(v)$. Calculation of results can be done in different manner. For example it might be calculated by a linear function, where the variable represents a dependency to an adjacent result:

$$r(v_x) = r(v_y) - 4 \text{ with } v_y \in N_G(v_x)$$

Here $N_G(v_x)$ is the set of all neighbors of v_x . Also it might be possible that results are sets containing other nodes, for instance:

$$r(v_x) = \{v_a, v_b, v_c\} \text{ with } v_a, v_b, v_c \in N_G(v_x)$$

C. Edge types and their usage

The system should be able to distinguish between optional and obligate connections between elements. Reason for this is to handle optional paths while deductive reasoning, for example by asking the user for his needs. An optional path is an edge with a specific successor that might be interesting for decision making, but the actual need of this successor is not sure until the decision making process is executed. The set with optional-edges is defined as followed:

$$E_{opt}(G) = \{e_x, \dots, e_n\}$$

As opposed to optional paths, required paths are those that are included in a decision making process in any case. They represent the usual edge type and are contained in the set of required-edges:

$$E_{req}(G) = \{e_x, \dots, e_n\}$$

The visualization in Fig. 3 illustrates the usage of optional (Fig. 3, a) and required (Fig. 3, b) paths as well as the usage of four other path categories. Required and optional paths can be visually distinguished by the end of each edge. An optional path ends with a non-filled connection. A required (usual) path is represented by a filled arrow head.

In addition to optional and required paths it is necessary to distinguish between more types of paths:

An ‘is-path’ is used for building inheritance relations, for example ‘MS Visio is Software’. It is visualized with a cross-filled circle (Fig. 3, c) on the predecessor element: ‘ v_5 is v_6 ’. This type is always a required path, otherwise the system would not be able to decide, if an element is another or not.

Also paths that represent ‘used-for’-relations between elements are always required edges (Fig. 3, e). An element ‘is used for’ an activity or not; it is not consistent to say ‘perhaps an element is used for an activity’.

For modeling characteristics or attributes the model provides a further type of an edge, the ‘has’-path (Fig. 3, d). These paths serve as instruments to specify elements. For example every software ‘has’ a price and an installation type. Modeling this knowledge means three Items: software, price and installation type. The edges between these elements would be ‘software has price’ and ‘software has installation type’. Now every element, which ‘is’ software, has a price and an installation type, too.

The sixth path type is the ‘part-of’-edge (Fig. 3, f), which is used to build part-whole relations between elements. In opposite to a ‘has’-relation the ‘part-of’ element is not able to exist for its own, which means the whole-unit has to exist.

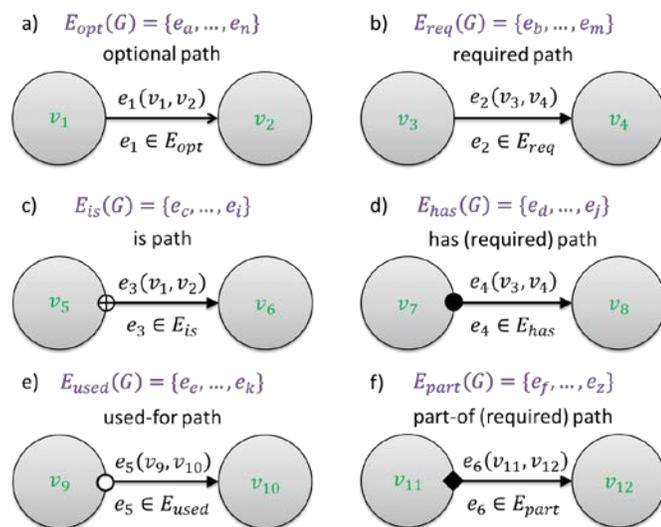


Fig. 3. Visualization of path types

Fig. 4 demonstrates an example with three Items. ‘Requirements documentation’ builds the successor, ‘Documentation of functional requirements’ and ‘Creating Wireframes’.

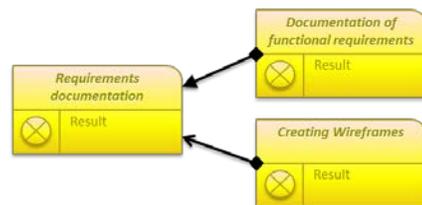


Fig. 4. Example of using optional and required paths

Wireframes’ are predecessors. Both predecessors might be interesting for a deductive reasoning process, for example ‘Find suitable software for requirements documentation!’. In

addition to this both are connected with ‘part-of’ relations, which means the predecessors are part of ‘Requirements documentation’.

The difference between the relations is creating wireframes, which is not a mandatory functionality of requirements documentation, whereas the documentation of functional requirements represents one of the most important topics. Thus in case of decision making ‘Which software would be the most suitable for a specific software development project?’, users has to determine first, if they need the functionality of creating wireframes.

D. Constraints and Deductive Reasoning Element Pruning

Constraints are barriers, which help excluding paths from the set of all paths within the graph. Thus any association between elements can define preconditions or requirements. So the processing of an element is solely necessary if all of its constraints are complied. Goal is:

- more efficient processing through the graph by
- more intelligent operating on elements and thereby
- shorter times of results in deductive reasoning

If at least one constraint of an element is false, the element can be excluded from the entirety of all possible elements of deduction. Using the example of this publication, the integration of constraints leads faster to a list of matching project artifacts. Visualization of constraints is done by a dotted line as illustrated in Fig. 5.

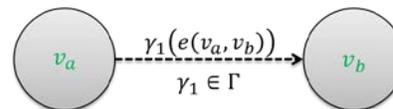


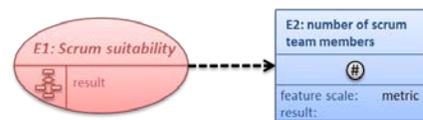
Fig. 5. Illustration and definition of constraints

The set of constraints is defined as:

$$\Gamma(G) = \{\gamma_1, \dots, \gamma_n\}$$

Considering an example of a simple association between Scrum and the number of regarded team members, Fig. 6 shows the functionality of constraints within the graph. A constraint γ is assigned to the edge $e(E1, E2)$ by $\gamma(e(E1, E2)): r(E2) > 5 \wedge r(E2) < 9$ that means the result of $E2$ has to be greater than five, but less than 9.

So γ can be interpreted as follows: Only if the result of node $E2$ (the number of team members) is between five and nine, element $E1$ (Scrum) is relevant for the deductive reasoning. If the result of $E2$ is less than five or higher than nine, $E1$ is irrelevant for the deductive reasoning, because at least one constraint is false.



$$\gamma(e(E1, E2)): r(E2) > 5 \wedge r(E2) < 9 \text{ mit } r(E2) \in \mathbb{R}$$

Fig. 6. Example of using constraints

By integration of constraints the entirety of nodes within a deductive reasoning process can be decreased. This process of reducing nodes can be named as Deductive Reasoning Element Pruning (DREP). Element Pruning is a well-known term in different Selection Algorithms as [28], [29] and [30] or in Clustering Methods for example [31] and [32]. In the domain of this model and publication DREP is an optimization measure during deduction, which is able to reduce the number of elements at the time T to be proceeded until $T + 1$. With DREP parallelism can result in jumping to nodes, which are possibly not reachable anymore, because they are on a pruned path. This case occurs with pruning of bridges (graph theory). Therefore it should be completely dispensed with parallelism or an event-driven system should be used, in case of integrating pruning.

The usage of DREP will be introduced in future work, where an algorithm for deductive reasoning will be shown that is suitable for the model of this paper.

E. Inference

As described above, the structure of the model should serve as KB with the ability of self-learning functionality and deductive reasoning. This is why the architecture of the model is a composition of ANN, which are used in artificial intelligence systems, and ontologies that have their usage in knowledge engineering. Furthermore it has already been expounded that each element within the model can output a specific result, which can be used by any other element. Considering it all together – artificial intelligence components, ontology-based representation of knowledge and the specific result of each element – lead to an architecture that is able to handle inductive reasoning and deductive reasoning in different ways.

By the derived structure of multi-layer feedforward ANN, the model can be used with well-known approaches of machine learning paradigms, as supervised learning, unsupervised learning or reinforcement learning. In addition to these inferring methods, it is possible to make simple decisions by searching the result of a specific element. Example of such a simple decision is ‘Is Scrum suitable for a specific project team?’. Assumed that a Cell ‘Scrum suitability’ exists and that the result of this Cell is the actual value of how suitable scrum is, there would be no need to perform a complicated machine learning algorithm for answering the predefined question. The only need is to output the result. This can be done on two different ways. On the one hand the connected knowledge of the ‘Scrum suitability’ can be interpreted manually, by human reading and reasoning, or on the other hand automated, by a decision-making system. For doing it automatically, one of the next steps of this project is defining an algorithm that handles this behavior under consideration of constraints and DREP.

III. EXAMPLE OF APPLICATION

As outlined above, the model can be used for illustrating information and knowledge. One of the major goals of the research project behind this paper is to develop an automated decision system for identification suitable project tools and required artifacts, especially for large software development projects. Fig. 7 shows an extract of this use case and

demonstrates the complexity of modeling a KB. The figure illustrates four out of five layers of the model: $L1$ Items (rectangular, rounded corner), $L2$ Cells (elliptical), $L3$ Features (rectangular) and $L4$ Questions (rhombic). The lines between the elements serve as connectors and represent the associations with their corresponding types.

There are two major levels of abstraction in the model: a descriptive and a deductive level. The first mentioned descriptive view is represented by Items as ‘Software’ and Features with a connection to Items as ‘Price (amount)’. Descriptive elements are essential for learning and generating information. Items can also represent problems or goals, for example ‘Classification of requirements’, which is part of ‘Requirements Engineering’. For solving this problem ‘Jira’ can be used. Jira is concrete ‘Software’ and so on.

In the example, ‘Software’ and their connected elements can be read as followed:

- ‘Software’ has ‘Price’ and ‘Installation Type’
- ‘Operating System’ is Software
- ‘Gliffy’, ‘Jira’ or ‘Astah’ is concrete Software
- ‘Jira’ can be used for ‘Documentation of non-functional requirements’,
- ‘Documentation of non-functional requirements’ is a part of ‘Requirements Engineering’.

Features can be connected to Cells and Items. In cases of connection with Items, they will be executed with inductive reasoning, which means during learning phases. As opposed to Features that are connected to Items, Cell-Features need to be executed within a concrete deductive reasoning process, which is for example a decision process. Reason for this behavior is distinction between general and specific. The following two examples describe this approach:

Example 1: Sentence to learn: ‘Jira is software.’ The system already learned ‘Software has a price’. Connected question to price: How much is the price? So the system has to ask: ‘How much is the price [for Jira]?’

Example 2: Constraint to learn: ‘Jira requires Windows 8.’ The KB knows: ‘Windows 8’ is an ‘Operating System’. So whenever the system has to make a decision through the Cell Jira, it has to ask: ‘What is the Operating System?’. Obviously the system might ask directly ‘Do you use Windows 8?’, but in this case the operating system of the user would remain unknown and element pruning could not be made in the same efficiency as with the question above.

IV. CONCLUSION

A new method for modeling knowledge, information and data is introduced in this paper. It serves as architectural basis for developing expert systems by building knowledge bases in the field of knowledge engineering.

The abstract meta-model is constructed by five layers. They consist of descriptive and deductive elements and are derived by an artificial neural network. The model is illustrated both in a descriptive way and in a mathematical view by considering it

as graph. In addition to the general description the authors give an insight into a case of application using the example of large software development projects.

By integration of constraints and DREP the proceeding time of the graph can be decreased. Thus a deductive reasoning process is more efficient and the interaction with users can be reduced, when using the model as knowledge foundation. A further advantage of using the model as basis for additional research projects is the ability of extensibility. For instance it is very easy to assimilate different approaches for realizing deductive and inductive reasoning.

When considering deductive reasoning in one of the next steps, it is important to give solutions for the following problems: (1) Detecting the end of the deductive reasoning

under regard to have an arbitrary entry-point and (2) handling conflicts in case of mutually exclusive Constraints.

In addition to use this model for knowledge engineering in the domain of large software development projects, it can also be used in different other domains. With the solution approach it could be possible to model knowledge of study advisers or career counseling, to build a basis for deciding what kind of occupation is the most suitable in dependency of personal characteristics. Further use cases are settled in 'Health and Medical' systems for building a foundation to detected symptoms and give suitable solutions. In case of building end-user systems the model can be used to develop a knowledge base for example of travel agents, fashion advisers or as product adviser.

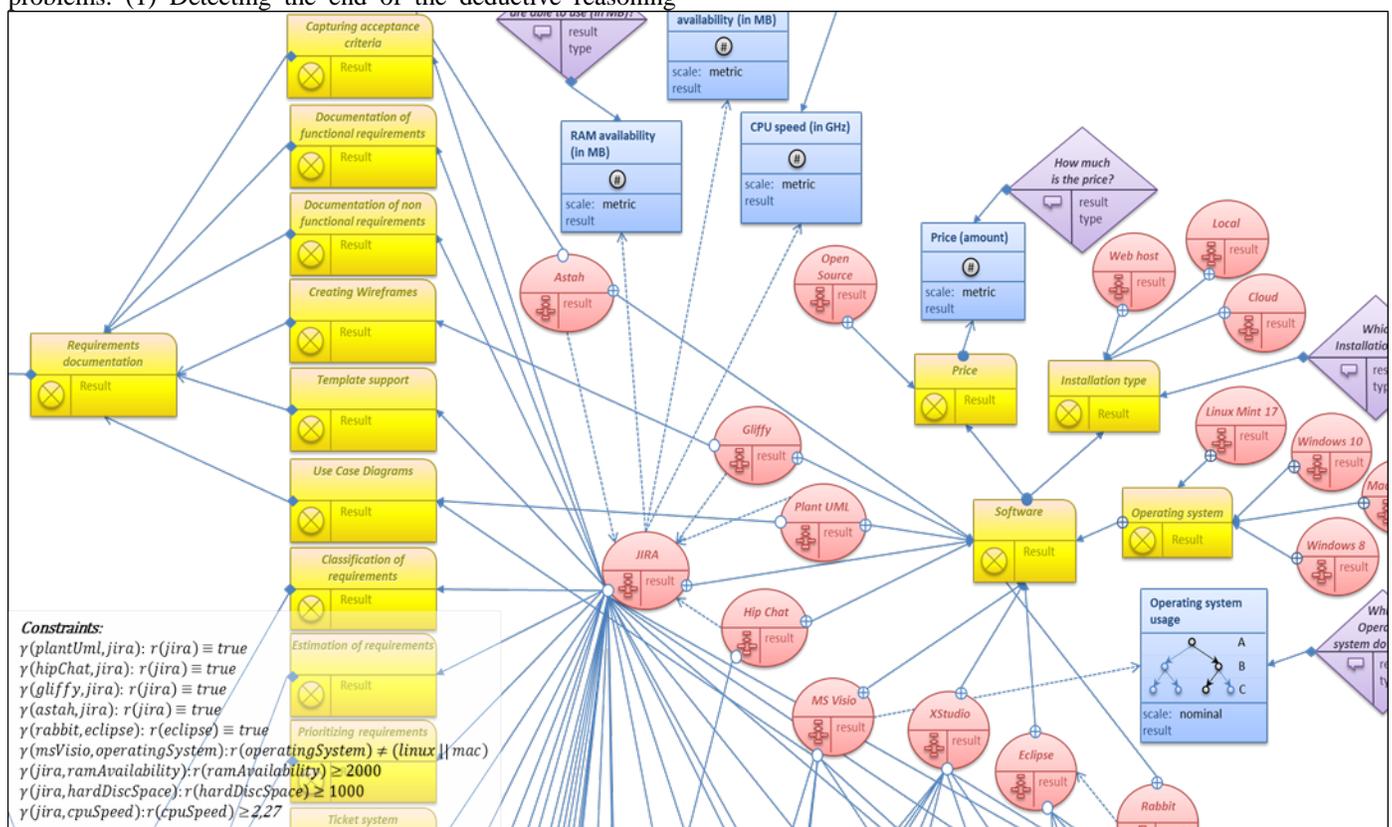


Fig. 7. Extract of using the model for knowledge engineering in large software development projects

REFERENCES

- [1] F.F. Fuessl, J. Ciemala, "Variable Factors of Influence in Product Line Development", Computer Software and Applications Conference Workshops (COMPSACW), 21-25 Jul. 2014, Vasteras, pp. 390-395
- [2] M.S. Mohktar, K. Lin, S.J. Redmond, J. Basilakis, N.H. Lovell, "Design of a decision support system using open source software for a home telehealth application." Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME), 8-9 Nov. 2011, Bandung, pp. 390-395
- [3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel; P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, "Scikit-learn: Machine learning in Python" The Journal of Machine Learning Research 12, 1 Feb. 2011, pp. 2825-2830
- [4] M. H. Nguyen, F. De la Torre, "Optimal feature selection for support vector machines." Pattern recognition 43.3, Mar. 2010, pp. 584-591
- [5] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Stajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, B. Zupan, "Orange: data mining toolbox in Python." The Journal of Machine Learning Research 14.1, Jan. 2013, pp. 2349-2353
- [6] F. F. Fuessl, A. Triebel, D. Streitferdt, "Modeling Knowledge Bases for Automated Decision Making Systems – A Literature Review ", International Journal of Advanced Computer Science and Applications(IJACSA), 6(9), 2015
- [7] K. He, C. Wang, Y. He, Y. Ma, P. Liang, "Theory of Ontology and Meta-Modeling and the Standard" Handbook of Research on Software Engineering and Productivity Technologies: Implications of Globalization, Aug. 2009, pp. 85.
- [8] J.D. Ullman, "Principles of database and knowledge-base systems" Computer Science Press, New York, 1988, p. 24
- [9] R. Akerkar, P. Sajja, "Knowledge-based systems", Jones & Bartlett Publishers, 2010, p. 21

- [10] G.S. Tuthill, S.T. Levym, "Knowledge-based systems: a manager's perspective", TAB Professional and Reference Books, 1991
- [11] R. Donmoyer, M. Imber, J.J. Scheurich, "The knowledge base in educational administration: Multiple perspectives", State University of New York Press, Albany, 1995, pp. 17-18
- [12] L.J. Heinrich, R. Riedl, D. Stelzer, "Informationsmanagement: Grundlagen, Aufgaben, Methoden", Walter de Gruyter GmbH & Co KG, 2014, p. 319
- [13] B. Bouchon, R.R. Yager, "Uncertainty in Knowledge-Based Systems: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems", Springer Science & Business Media, 1987, p. 3
- [14] Bayya Yegnanarayana, "Artificial neural networks" PHI Learning Pvt. Ltd., 2009., p. 29
- [15] J.E. Dayhoff, J.M. DeLeo, "Artificial neural networks", Cancer 91.8, Apr. 2001, pp. 1615-1635
- [16] V.S. Desai, J.N. Crook, G.A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment", European Journal of Operational Research 95.1, Nov. 1996, pp. 24-37
- [17] P.D. McNicholas, Y. Zhao. "Association rules: an overview", Information Science Reference (imprint of IGI Global), 2009, pp. 1-10
- [18] H. Zhu, "Software Design Methodology: From Principles to Architectural Styles", Butterworth-Heinemann (imprint of Elsevier), 2005, pp. 156-157
- [19] P. Norvig, "Paradigms of artificial intelligence programming: case studies in Common LISP", Morgan Kaufmann Publishers Inc., 1992
- [20] E. Bonabeau, M. Dorigo, G. Theraulaz. "Swarm intelligence, From Natural to Artificial Systems", Oxford University Press, 1999
- [21] R.A. Brooks, "Achieving Artificial Intelligence through Building Robots", Massachusetts Institute Of Technology, No. AI-M-899, May 1986.
- [22] C. Landauer, "Data, information, knowledge, understanding: computing up the meaning hierarchy", Systems, Man, and Cybernetics, San Diego (CA), 11-14 Oct.1998, pp. 225-2260 vol.3
- [23] M.C. Golumbic, "Algorithmic graph theory and perfect graphs", Vol. 57. Elsevier B.V., second edition 2004
- [24] M. Foreman, A. Kanamori, "Handbook of set theory", Springer Science & Business Media B.V, 2010.
- [25] P.C. Biswal, "Discrete mathematics and graph theory", Fourth Edition, PHI Learning Pvt. Ltd., Delhi, 2015, P. 101
- [26] C. Rousseau, Y. Saint-Aubin, "Mathematics and technology", Springer Science & Business Media, LLC, 2008, p. 269
- [27] R. Garnier, J. Taylor, "Discrete mathematics for new technology", second edition, IOP Publishing Ltd., 2002, p. 592
- [28] D. Dor, U. Zwick, "Selecting the median", SIAM Journal on Computing 28.5, 1999, pp. 1722-1758.
- [29] D. Dor, "Selection algorithms", Diss., Tel-Aviv University, Sep. 1995
- [30] T. Anand, P. Gupta. "A selection algorithm for $X+Y$ on mesh", Parallel processing letters 08.03, Sep. 1998, pp. 363-370.
- [31] G. Bisson, C. Nédellec, D. Canamero, "Designing Clustering Methods for Ontology Building-The MoK Workbench", ECAI workshop on ontology learning. Vol. 31, 2000
- [32] L.A.F. Fernandes, A.C.B. García, "Association rule visualization and pruning through response-style data organization and clustering", Advances in Artificial Intelligence-IBERAMIA 2012, Lecture Notes in Computer Science Volume 7637, Springer Berlin Heidelberg, 2012, pp. 71-80

A Prediction Model for Mild Cognitive Impairment Using Random Forests

Haewon Byeon

Department of Speech Language Pathology & Audiology
Nambu University
Gwangju, Republic of Korea

Abstract—Dementia is a geriatric disease which has emerged as a serious social and economic problem in an aging society and early diagnosis is very important for it. Especially, early diagnosis and early intervention of Mild Cognitive Impairment (MCI) which is the preliminary stage of dementia can reduce the onset rate of dementia. This study developed MCI prediction model for the Korean elderly in local communities and provides a basic material for the prevention of cognitive impairment. The subjects of this study were 3,240 elderly (1,502 males, 1,738 females) in local communities over the age of 65 who participated in the Korean Longitudinal Survey of Aging (close) conducted in 2012. The outcome was defined as having MCI and set as explanatory variables were gender, age, level of education, level of income, marital status, smoking, drinking habits, regular exercise more than once a week, monthly average hours of participation in social activities, subjective health, diabetes and high blood pressure. The random Forests algorithm was used to develop a prediction model and the result was compared with logistic regression model and decision tree model. As the result of this study, significant predictors of MCI were age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise and high blood pressure. In addition, Random Forests Model was more accurate than the logistic regression model and decision tree model. Based on these results, it is necessary to build monitoring system which can diagnose MCI at an early stage.

Keywords—random forests; data mining; dementia; mild cognitive impairment; risk factors

I. INTRODUCTION

As worldwide aged population increases with the development of science, technology and medicine, number of geriatric diseases increases radically as well. Especially, dementia, a typical geriatric disease, is expected to experience an unprecedentedly rapid increase. According to World Alzheimer Report (2015), worldwide dementia population recorded 44 million in 2013 and will increase more than 3-fold to 135 million in 2050 [1].

In Korea, dementia also increases rapidly due to fast aging. According to a survey on prevalence rate of dementia conducted by Ministry of Health and Welfare, the number of dementia patients in Korea was 540,000 in 2012 and the number is expected rapidly increase to 840,000 in 2020, 1.27 million in 2030 and 2.71 million in 2050 [2]. In particular, as Korea shows the most rapid rate of increase in the world, it is urgent to take measures for geriatric cognitive impairment [3].

Although treatment methods for dementia have been developed globally over the last 20 years, no treatment method developed so far can provide full recovery. It is only possible to postpone cognitive decline of dementia when cognitive function is managed systematically by using drugs such as donepezil [4]. As medication of this kind can produce greater effect with earlier application, early diagnosis and intervention is crucial in dementia.

Especially, as dementia incurs tremendous socio-economical costs, systematic management is required through early intervention. According to a report to Korean National Assembly, social cost for dementia patients is estimated to be US\$ 37.3 billion in 2050, which amounts to 1.5% of GDP [3]. Thus, reduction of prevalence rate through early discovery of dementia can decrease unnecessary social and economic costs [5].

Like this, as early diagnosis of dementia becomes important, Mild Cognitive Impairment (MCI) which is a previous stage of dementia is gaining attention. MCI is defined as intermediate stage between normal aging and dementia with its decline of cognitive function out of normal range but its severity still not reaching the state of dementia [6]. MCI, an earliest stage to discover dementia, is important as a primary target for dementia treatment since its early discovery and treatment can postpone the progress of dementia.

Along with dementia, MCI is also on the rapid increase. MCI in Korea has increased 4.3-fold from 24,000 in 2010 to 105,000 in 2014, attracting attention to its early discovery and prevention [2].

Over the last 20 years, numerous studies have reported that risk factors of MCI were gender, age, smoking, drinking, eating habits, exercise, diabetes and hypertension [5, 7, 8, 9]. And opinions exist that there are limitations to explain the outbreak of MCI with these individual risk factors and studies report different results on affecting risk factors [9]. In addition, necessity to consider mental health such as depression is recently being raised in the search for factors related to MCI [10]. In particular, as there are differences among races in outbreak pattern of cognitive impairment and risk factors, it is necessary to develop MCI prediction model reflecting the living patterns of the Korean elderly.

Meanwhile, as analysis on big data becomes possible with the development of computer, attention is being paid to data mining techniques in developing prediction models. Data

mining is a method of analysis to predict data based on already known attributes by using training data [11, 12]. Especially, Random Forests developed as one analysis method of data mining has high level of prediction capability as it creates multiple decision trees by implementing random sampling in an identical data set, combines them and finally predicts target variables [13, 14]. In addition, Random Forests is known to have an excellent prediction capability in finding out correlation between explanatory variables and a disease and prevent overfit when there are many kinds of explanatory variables applied to the model [15, 16].

This study developed a prediction model of Mild Cognitive Impairment for the elderly in Korean local communities based on random forests algorithm and compared it with logistic regression model and decision tree model to verify its results and accuracy.

Construction of this study is as follows; chapter II explains study subjects and analyzed variables and chapter III defines random forests and explains the procedure of model development. Chapter IV compares the results of developed prediction model with those of logistic regression model and decision tree model. Lastly, chapter V presents conclusion and direction for future studies.

II. METHODS

A. Sources of data

This study analyzed a total of 3,240 elderly people (1,502 males and 1,738 females) over the age of 65 who participated in 2012 Korean Longitudinal Survey of Aging (KLoSA).

KLoSA is supervised by Korea Labor Institute and TNS Korea conducted the survey on commission from July 7, 2012 through December 2012 [17]. Sampling frame was districts of Population and Housing Census 2005 and 261,237 districts were set as sampling units. In 2012 survey, 10,000 people was set as maximum valid sample size and considering that average population over the age of 45 was 1.67 per household in 2000 Population and Housing Census, 1,000 sampling districts were selected. The method of the survey was computer-assisted personal interviews using laptop computers.

B. Measurements

Outcome was defined as prevalence of MCI. Explanatory variables were included as gender, age (65-75, 75+), level of education (middle school and lower, over high school), level of income, marital status (have spouse, divorced or separation, separation by death), smoking (non-smoking, past smoking, current smoking), drinking habits (non-drinker, past drinker, current drinker), regular exercise more than once a week (yes, no), monthly average hours of participation in social activities (less than 1 hour, over 1 hour), subjective health (good, fair, bad), diabetes (yes, no), and hypertension (yes, no).

III. STATISTICAL ANALYSIS

A. Development and evaluation of model

In order to develop MCI prediction model, this study divided data into 70% of training data and 30% of test data. Random forest algorithm was used to develop prediction model

and results of developed prediction model were compared with those of decision tree based on multivariate logistic regression analysis and CART (classification and regression tree). Accuracies of developed models were evaluated with correct classification rate, and importance of variables and major risk factors drawn out were compared respectively.

B. Random Forests

Random forest is a type of ensemble classifier which randomly learns multiple decision trees and is composed of training stage which construct many decision trees and test stage which classifies and predicts incoming input data [18] (Figure1).

Ensemble form of training data can be expressed in Forest $F = \{f_1, \dots, f_n\}$ (Figure2).

Distributions earned from decision trees of each forest are averaged by T , the number of decision trees, and finally classified.

$$L(p) = \frac{1}{T} \sum_{t=1}^T P_t(b|I, p)$$

Figure 3 shows bagging algorithm which creates final model by conducting n times of random sampling on raw data and combining prediction variables coming out of modeling of each sample. For combining method of prediction variables of each sample, average was used when a target variable was a continuous variable while majority vote was used when a target variable was a categorical variable.

Although random forest is similar to bagging in that it enhances stability by combining decision trees created in multiple bootstrap samples with majority principle, it is different from bagging in that it uses only a few explanatory variables randomly chosen from bootstrap samples. In order to adjust correlation of combination model, random forest establishes decision tree by randomly extracting several explanatory variables from boot strap samples and establishes a model with as few pruning as possible.

Random forest has higher prediction capability than decision tree and strength that it can prevent overfitting [19]. This study established random forest model first and then compared it with the results drawn out from multivariate logistic regression analysis and decision tree and accuracy of model respectively.

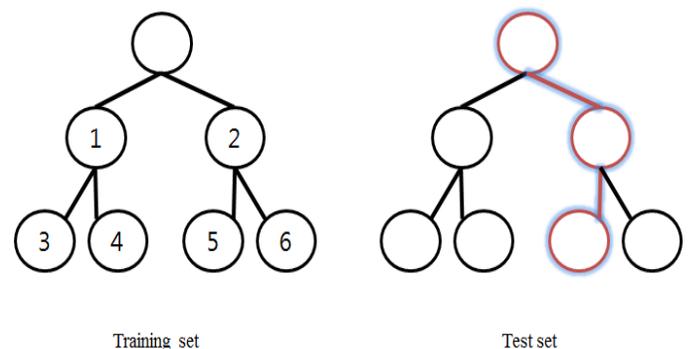


Fig. 1. Training and testing datasets

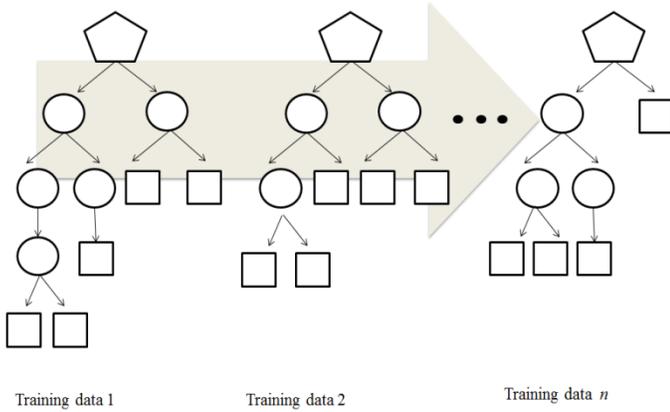


Fig. 2. Random forest: a classifier that combines many single decision trees

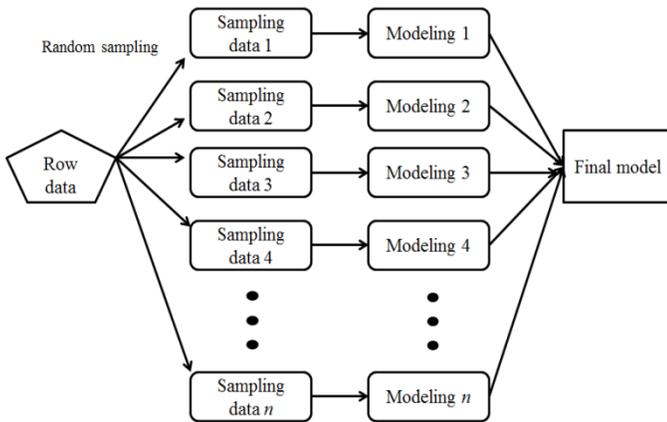


Fig. 3. Bagging algorithm

IV. RESULTS

A. Characteristics of subjects

Among total 4,134 subjects, males were 46.4% and females were 53.6%. Average age was 72 (range=65-99, standard deviation=5.9). As for level of education, high school graduates and over were 24.8% and those living with spouse were 71%. Current smokers were 13.1% and current drinkers were 28.8% while the elderly who exercise regularly more than once a week were 37.3%. 49.8% had hypertension and 20.9% had diabetes. Prevalence rate of MCI was 31.3%.

B. Potential risk factors for Mild Cognitive Impairment (univariate analysis)

Potential risk factors for MCI are presented in Table 1. As the result of cross-tabulation analysis, there were significant differences between the normal elderly and those with MCI in gender, age, level of education, level of income, marital status, smoking, drinking habit, regular exercise of more than once per week, monthly average hour of participation in social activities, subjective health and hypertension ($p < 0.05$).

Prevalence rate of MCI was high in the elderly over the age of 75 (43.7%), females (38.6%), middle school graduates and lower (36.4%), the bereaved of spouses (42.3%), nonsmokers (34.6%), nondrinkers (35.6%), elderly who do not exercise regularly (35.9%), elderly who participate in social activities

less than average 1 hour per month (31.8%), elderly with poor subjective health (43.9%) and elderly with hypertension (34.3%).

TABLE I. GENERAL CHARACTERISTICS OF THE SUBJECTS BASED ON MCI (UNIVARIATE ANALYSIS), N (%)

Characteristics	MCI		p
	Yes (n=1,015)	No (n=2,225)	
Age			<0.001
65-75	475 (23.7)	1,528 (76.3)	
75+	540 (43.7)	697 (56.3)	
Sex			<0.001
Male	345 (23.0)	1,157 (77.0)	
Female	670 (38.6)	1,068 (61.4)	
Level of education			<0.001
Middle school and lower	887 (36.4)	1,549 (63.6)	
Over high school	128 (15.9)	676 (84.1)	
Level of income (quartile)			<0.001
First quartile	478 (40.5)	703 (59.5)	
Second quartile	270 (29.1)	659 (70.9)	
Third quartile	174 (23.1)	579 (76.9)	
Fourth quartile	93 (24.7)	284 (75.3)	
Marital status			<0.001
Have spouse	625 (27.2)	1674 (72.8)	
Divorced/separation	21 (30.9)	47 (69.1)	
Separation by death	369 (42.3)	504 (57.7)	
Smoking			<0.001
Non-smoking	767 (34.6)	1,449 (65.4)	
Past smoking	141 (23.5)	459 (76.5)	
Current smoking	107 (25.2)	317 (74.8)	
Drinking			<0.001
Non- Drinking	628 (35.6)	1,138 (64.4)	
Past Drinking	159 (29.4)	382 (70.6)	
Current Drinking	228 (24.4)	705 (75.6)	
Regular exercise more than once a week			<0.001
Yes	287 (23.7)	923 (76.3)	
No	728 (35.9)	1,302 (64.1)	
Monthly average hours of participation in social activities			0.001
Less than 1 hour	996 (31.8)	2,134 (68.2)	
Over 1 hour	19 (17.3)	91 (82.7)	
Subjective health			<0.001
Good	106 (17.4)	502 (82.6)	
Fair	392 (27.0)	1,061 (73.0)	
Bad	517 (43.9)	662 (56.1)	
Hypertension			<0.001
Yes	554 (34.3)	1,061 (65.7)	
No	461 (28.4)	1,164 (71.6)	
Diabetes			0.185
Yes	226 (33.4)	450 (66.6)	
No	789 (30.8)	1,775 (69.2)	

C. Accuracy comparison between models

Prediction model was developed by using random forests and its accuracy was compared with those of logistic regression model and decision tree (Table 2). As the result of analysis on training data, random forests showed very high accuracy of 72.5% (Figure 4, Figure 5). On the other hand, accuracy of decision tree was 71.2% and accuracy of logistic regression model was the lowest with 68.7%.

In test data, random forests showed the highest accuracy with 72.1% while logistic regression model had the lowest accuracy with 67.5%. Hence, random forests had the highest accuracy in both training data and test data.

TABLE II. ACCURACY COMPARISON BETWEEN MODELS

Data	Model	Accuracy (%)
Training data	Logistic regression	68.7
	Decision tree	71.2
	Random Forests	72.5
Test data	Logistic regression	67.5
	Decision tree	70.8
	Random Forests	72.1

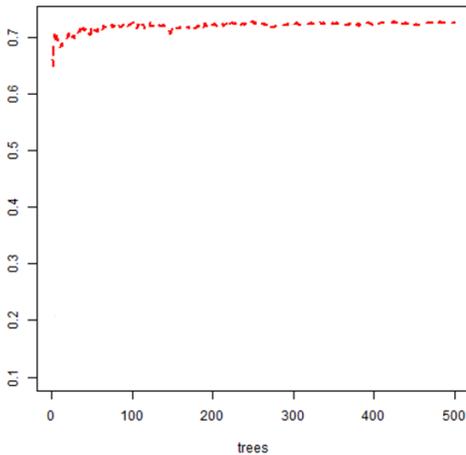


Fig. 4. Accuracy of Random Forests model

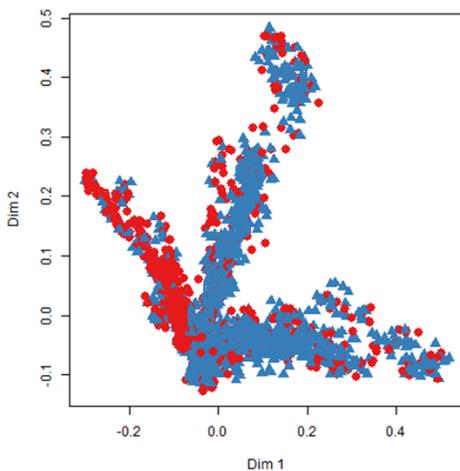


Fig. 5. Multidimensional scaling plot of proximities

D. Comparison of risk factors based on prediction model

The results of prediction models established based on logistic regression model, decision tree and random forest by using a total of 12 explanatory variables to predict MCI are presented in Table 3. Among the prediction models used in this study, major risk factors for random forests model were presumed by using decrease of GINI coefficient.

In logistic regression model, risk factors for MCI were total of 7 variables, which were age, gender, level of income, hour of participation in social activities, subjective health and regular exercise and its accuracy was 67.5%.

Decision tree model predicted 8 variables as risk factors for MCI, which were age, gender, level of education, level of income, subjective health, marital status, smoking, regular exercise and its accuracy was 70.8%.

Random forests model predicted age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise and hypertension as risk factors for MCI, and its accuracy was 72.1%.

TABLE III. COMPARISON OF RISK FACTORS BASED ON PREDICTION MODEL

Model	Number of factors	Risk factors
Logistic regression	7	age, gender, level of education, level of income, social activities, subjective health, regular exercise
Decision tree	8	age, gender, level of education, level of income, subjective health, marital status, smoking, regular exercise
Random Forests	10	age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise, hypertension

V. CONCLUSION

Early discovery of MCI is clinically important as it can postpone cognitive decline. This study developed MCI prediction model for the elderly in local communities by using random forest algorithm.

As the result of developing prediction model based on random forests, major risk factors for MCI were age, gender, level of education, level of income, subjective health, marital status, smoking, drinking, regular exercise and hypertension. Many studies have reported that socio-demographic factors such as old age and level of education and health risk behaviors such as smoking and drinking are risk factors of MCI [3, 5]. In particular, smoking was the most important variable except socio-demographic factors in the MCI prediction model of this study. Although smoking is a modifiable factor compared with socio-demographic factors like age and level of education, quitting smoking is difficult since the elderly have been exposed to smoking for a very long period of time and they do not have strong will to quit. Smoking rate of the elderly over the age of 65 in Korea is still very high of 23.3% [2]. Thus, in order to prevent MCI and maintain healthy cognitive function, quitting smoking is required more than anything else.

As the result of comparison of accuracy among random forests, logistic regression model and decision tree, random forests were the most accurate, which is speculated, because random forests is based on bagging algorithm which creates various decision trees from around 500 bootstrap samples.

As decision tree can compose a node even in the case of outlier, the influence of parameter deciding node is great, which creates risk of overfitting [12]. On the other hand, in the case of random forests which predict target variables through average or probability of each tree, as the bias of trees is maintained and variance decreases, its accuracy is higher than that of decision tree [16]. Therefore, when using data with many variables like disease examination data or establishing prediction model using distributed processing system such as big data, random forests is considered the most proper which extracts training data to create trees and predicts target variables. In order to further enhance accuracy of random forests, future studies are required to develop prediction model using weighted voting.

This study has the strength that it developed MCI prediction model by using examination data representing the whole population. It is necessary to establish a monitoring system which can diagnose old-age cognitive impairment in an early stage based on the MCI prediction model developed by this study.

ACKNOWLEDGMENT

The author wishes to thank the Korea Labor Institute that provided the raw data for analysis.

REFERENCES

- [1] Alzheimer's Disease International, World Alzheimer Report 2015. London, Alzheimer's Disease International, 2015.
- [2] Ministry of Health & Welfare, Nationwide Study on the Prevalence of Dementia in Korean Elders 2012. Sejong, Ministry of Health & Welfare, 2013.
- [3] S. Kim, Analysis on Management Policies for the Dementia. Seoul, National Assembly Budget Office, 2014.
- [4] P. Anand, and B. Singh, A review on cholinesterase inhibitors for Alzheimer's disease. Archives of pharmacal research, vol. 36, no. 4, pp. 375–399, 2013.
- [5] H. Byeon, Y. Lee, S. Y. Lee, K. S. Lee, S. Y. Moon, H. Kim, C. H. Hong, S. J. Son, and S. H. Choi, Association of alcohol drinking with verbal and visuospatial memory impairment in older adults: Clinical Research Center for Dementia of South Korea (CREDOS) study. International Psychogeriatrics, vol. 27, no. 3, pp. 455–461, 2015.
- [6] H. A. Tuokko, and D. F. Hultsch, Mild cognitive impairment: International perspectives. New York, Psychology Press, 2013.
- [7] G. Cheng, C. Huang, H. Deng, and H. Wang, Diabetes as a risk factor for dementia and mild cognitive impairment: a meta-analysis of longitudinal studies. Internal medicine journal, vol. 42, no. 5, pp. 484–491, 2012.
- [8] M. Ganguli, B. Fu, B. E. Snitz, T. F. Hughes, and C. C. H. Chang, Mild cognitive impairment Incidence and vascular risk factors in a population-based cohort. Neurology, vol. 80, no. 23, pp. 2112–2120, 2013.
- [9] T. Etgen, D. Sander, H. Bickel, and H. Förstl, Mild cognitive impairment and dementia: the importance of modifiable risk factors. Deutsches Ärzteblatt International, vol. 108, no. 44, p. 743, 2011.
- [10] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, Mild cognitive impairment: a concept in evolution. Journal of internal medicine, vol. 275, no. 3, pp. 214–228, 2014.
- [11] H. Byeon, Development of prediction model for endocrine disorders in the Korean elderly using CART algorithm: results from a population-based study. International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, pp. 215–219, 2015.
- [12] D. T. Larose, Discovering knowledge in data: an introduction to data mining. New York, John Wiley & Sons, 2014.
- [13] G. Biau, Analysis of a random forests model. The Journal of Machine Learning Research, vol. 13, no. 1, pp. 1063–1095, 2012.
- [14] A. Shameem, and D. Manimeglai, Analysis of significant factors for dengue infection prognosis using the Random Forest Classifier. International Journal of Advanced Computer Science and Applications, vol. 6, no. 2, pp. 240–245, 2015.
- [15] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution. BMC bioinformatics, vol. 8, no. 1, p. 25, 2007.
- [16] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, Screening large-scale association study data: exploiting interactions using random forests. BMC genetics, vol. 5, no. 1, p. 32, 2004.
- [17] Korea Labor Institute, Korean Longitudinal Survey of Ageing 2011. Sejong, Korea Labor Institute, 2014.
- [18] S. N. Devi, and S. P. Rajagopalan, A study on feature selection techniques in bio-informatics. International Journal of Advanced Computer Science and Applications, vol. 2, no. 1, pp. 138–144, 2011.
- [19] S. Hussain, and G. C. Hazarika, Educational data mining model using rattle. International Journal of Advanced Computer Science and Applications, vol. 5, no. 6, pp. 22–27, 2014.

Spectrum Sensing Methodologies for Cognitive Radio Systems: A Review

Ireyuwa E. Igbinsa¹, Olutayo O. Oyerinde², Viranjay M. Srivastava¹, Stanley Mnene¹

¹ School of Electrical, Electronics and Computer Engineering
University of KwaZulu-Natal, Durban, 4041, South Africa

²School of Electrical and Information Engineering,
University of the Witwatersrand, Johannesburg,
2050, South Africa

Abstract—Spectrum sensing is an important functional unit of the cognitive radio networks. The spectrum sensing is one of the main challenges encountered by cognitive radio. This paper presents a survey of spectrum sensing techniques and they are studied from a cognitive radio perspective. The challenges that go with spectrum sensing are reviewed. Two sensing schemes, namely; cooperative sensing and eigenvalue-based sensing are studied. The various advantages and disadvantages are highlighted. Based on this study, the cooperative spectrum sensing is proposed for employment in spectrum sensing in wideband based cognitive radio systems.

Keywords—Cognitive radio; Cooperative sensing; Data Fusion; OFDM; Spectrum Sensing; wideband sensing

I. INTRODUCTION

Spectrum sensing can be said to be the process of performing measurement on a part of the spectrum and making a decision related to spectrum usage based upon measured data [1]. Spectrum sensing is a fundamental operational block of the cognitive radio (CR) which consists of spectrum sensing, management, sharing and spectrum mobility. The growing demand for wireless application has put a lot of strain on the usage of available spectrum. In order to address this situation and improve spectrum efficiency, Mitola proposed a technique that allows secondary users to utilize radio spectrum band allocated to primary users that is not actively used [2]. According to a report from the United States Federal Communication Commission [3], there are larger temporal and geographic variations in the utilization of allocated spectrum. It is also known that allocated spectrum is underutilized because of the static allocation of the spectrum. In order to overcome this, there is need to propose a means of improving utilization of spectrum [4, 5, 6]. The scarcity and underutilization of spectrum has led to the development of cognitive radio (CR) technology, which exploit the existing spectrum opportunistically. Cognitive radio technology was defined in [2] and [7]. In this paper, the definition of the FCC is adopted. It states “Cognitive Radio is a system that senses its operational electromagnetic environment and can dynamically and autonomously adjust its radio operating parameters to modify system operation, such as maximization of throughput, mitigation of interference and facilitation of

inter-operability accessing the secondary markets”. To achieve this goal of cognitive radio, it is a compulsory requirement that a cognitive user (CU) performs spectrum sensing to detect the presence of primary users’ (PU) signal [8]. In the context of cognitive radio, the primary users can be defined as the users who have higher priority or right in the usage of a specific part of the spectrum. The secondary users (SU) on the other hand are the users who have lower priority or lower rights; they use the spectrum in such a way that they do not cause harmful interference to the primary users. However, secondary users need to have cognitive radio capabilities, such as sensing spectrum efficiently to ascertain if it is being occupied by a primary user and also change their radio parameters to exploit the unused part of the spectrum. The three popularly used methods for spectrum sensing are: Energy detection, Matched filtering and Cyclostationary detection [9, 10, 11, 12]. The basic idea of cognitive radio is its ability for spectral reusing or spectrum sharing, which allows secondary users to communicate over licensed spectrum. It also involves determining what type of signals that are occupying the spectrum including modulation, waveform, bandwidth, carrier frequency, etc. This however, requires powerful signal analysis method with additional computational complexity. Wideband spectrum sensing for cognitive radio network has not been sufficiently investigated in literature. Earlier approach uses a tunable narrowband filter and the RF front-end to sense one narrow frequency band at a time [13], in which the existing narrowband sensing techniques can be applied. In order to operate over multiple frequency bands at a time, the RF front-end needs wideband architecture. Spectrum sensing usually involves the estimation of the power spectral density (PSD) [14, 15]. There are so many factors that can cause spectrum sensing to be practically challenging. The rest of the paper is organized as follows. Complexity of spectrum sensing concept is studied in Section II. The challenges associated with spectrum sensing for cognitive radio are discussed in section III. Section IV shows the algorithms for spectrum sensing in cognitive radio. Section V discusses the cooperative spectrum sensing. Section VI discusses the research challenges involved in improving cooperative spectrum sensing and finally section VII concludes this paper.

II. COMPLEXITY OF SPECTRUM SENSING CONCEPT

Spectrum opportunity is conventionally defined in literature as “a *band of frequencies that are not used by a primary user of that band at a particular time and specific geographic location*,” [16]. This definition therefore introduces multi-dimensional spectrum awareness, since a spectrum hole is a function of frequency, time and geo-location. Since noise is present all the time in the entire radio spectrum, then an empty frequency bin doesn't exist. [17] Therefore it is important to be able to differentiate a band occupied by a primary user signal (PU) and the one from a spectrum hole that contains noise only signal. The traditional definition of spectrum sensing only exploits the three dimensions of the spectrum space. These are frequency, time and geo-location. Traditional methods usually relate to sensing the spectrum using these three [18]. However there are other dimensions that can be exploited for further spectrum opportunity. For example the code dimension of the spectrum space has not been extensively explored in details in literature therefore the traditional spectrum sensing algorithms find it

challenging to deal with signals that makes use of spread spectrum, time or frequency hopping codes. As a result, this type of signals causes a lot of challenges in spectrum sensing as discussed in the later part of this paper. Also the angle dimension is another area which is coming up as there are recent advances in multi- antenna technologies such as beam forming, multiple users can be multiplexed into the same channel at the same time in the same geo-location. Hence, in angle dimension, a primary user and a secondary user can be in the same geo-location and share the same channel. Spectrum sensing should include the process of recognizing occupancy in all dimensions of spectrum space and find spectrum holes. For instance, a certain frequency can be occupied for a given time, but may be empty in another time. Hence, a temporal dimension is as important as frequency dimension. The idle periods between bursty transmissions of local area network (WLAN) signals are exploited for opportunistic usage [19]. As a result of this requirement, advanced spectrum sensing algorithms that offers spectrum awareness in multiple dimensions can be developed.

TABLE I. SUMMARY OF DIMENSIONAL SPECTRUM AWARENESS

Dimensions	Sensing Parameters	Observations
Frequency	Frequency domain opportunity	Spectrum opportunity in this dimension means that all bands are not used simultaneously at the same time (some bands may be available for opportunistic usage).
Time	Opportunity of Specific band in time	This involves the availability of a specific part of the spectrum in time. In other words, the band is not continuously used. Hence there would be times where it would be available for opportunistic usage.
Geo-Location	Location and distance of primary users	The spectrum can be available in some parts of the geo-location and occupied in some other part at a given time. This takes advantage of pathloss in space. This measurement can be avoided by simply looking at the interference level. However, one needs to be careful of hidden terminal problem.
Code	Time hopping (TH) or frequency hopping (FH) sequences used by the primary users.	The spectrum over a wideband can be used at a given time through spread spectrum or frequency hopping. This doesn't mean that there is no availability over this band. Hence, simultaneous transmission without interfering with primary users would be possible in code dimension with an orthogonal code with respect to codes that primary users are using. (Not only detecting the usage of the spectrum, but also determining the used codes, and possibly multipath parameters as well).
Angle	Beam (azimuth, elevation angle) and locations of primary users	Along with the knowledge of the location/position or direction of primary users, spectrum opportunity in angle dimension can be created. For instance, if a primary user is transmitting in a specific direction, the secondary user can transmit in other directions without causing interference on the primary user.

III. THE CHALLENGES ASSOCIATED WITH SPECTRUM SENSING FOR COGNITIVE RADIO

Before getting into the details of spectrum sensing techniques, it is advisable to review the challenges. The challenges associated with the spectrum sensing for cognitive radio are discussed in this section.

A. Hardware Requirements

Applications for spectrum sensing in cognitive radio needs high sampling rate, high resolution analog to digital converter (ADC) with large dynamic range and high speed signal processors. Noise variance estimation technique has been popularly used for optimal receiver [14]. Designs such as, channel estimation, soft information etc., as well as improved

handoff, power control and channel allocation techniques [20]. The noise and estimation challenges are easier for these purposes as receivers are tuned to receive signals that are transmitted over a desired bandwidth. However, receivers have the ability to process the narrowband baseband signals with reasonably low complexity and low power processors. Cognitive radio terminals are then required to process transmission over a wideband for utilizing any opportunity [18]. Hence, a cognitive radio should be able to capture and analyze a relatively larger band and utilize any spectrum opportunities. These large operating bandwidths create additional requirements on the radio frequencies (RF) components such as antennas and power amplifiers. Hence high speed processing units are needed for performing computationally demanding signal processing task with

relatively low delay. Spectrum sensing can be performed through two different architectures, such as single radio and dual radio. In the case of the single radio, only one specific time slot is assigned for spectrum sensing. Due to the limitation in sensing time, only certain accuracy can be guaranteed for spectrum sensing result. However, the spectrum sensing efficiency is decreased as some part of the available time slot is used for sensing instead of data transmission. The merit of using the single radio architecture is because of its simplicity and low cost of implementation. However, in the dual radio architecture, one radio chain is allocated for data transmission and reception while the other chain is allocated for spectrum monitoring [21]. The limitations of such approach is that it increases power consumption and hardware cost. In practice, there are already available hardware and software platforms for cognitive radio such as GNU Radio, Universal Software Radio Peripheral and shared spectrum's XG Radio. Energy detector based sensing is mainly used in this platform because of its simplicity.

B. Hidden Primary User Problem

The hidden primary user problem is caused by several factors such as: - severe multipath fading or shadowing observed by secondary users while scanning for primary users. In this condition, the cognitive radio devices causes, unwanted interference to the primary user (receiver) as the primary transmitter, signal cannot be detected due to the location of devices. Cooperative sensing has been proposed as a means of handling hidden primary user problem [22, 11, 23]. Cooperative sensing is discussed in detailed in the later part of this paper.

C. Detecting Spread Spectrum Primary Users

There are two main types of technologies for detecting commercially available devices, they are: fixed frequency and spread spectrum. There are two main spread spectrum technologies available. The frequency hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSS). Fixed frequency devices operate at a single frequency or channel [18]. An example of this kind of system is the IEEE 802.11 a/g based WLAN. FHSS devices have the ability to change their operational frequencies dynamically into multiple narrowband channels. This technique is known as hopping and is implemented according to a sequence that is known to both transmitter and receiver. However, they use a single band to spread their energy. Primary users that use spread spectrum signaling are difficult to detect as the power of the primary user is distributed over a wide frequency range even though the actual information bandwidth is much narrower [11]. This problem can however be avoided partially if the hopping pattern is known and perfect synchronization to the signal can be achieved.

D. Sensing Duration and Frequency

In order to prevent interference to and from primary license owner, cognitive radio should be able to identify the presence of primary users as quickly as possible and should exit the band immediately. Therefore, sensing methods should be able to identify the presence of primary users within certain duration. This requirement poses a limit on the performance of sensing algorithms and creates challenges for cognitive radio

design. Selection of sensing parameters brings about tradeoff between the speeds (sensing time) and reliability of sensing. Sensing frequency is a design parameter that needs to be chosen carefully [23]. In a case when the status of primary users is known to change slowly, sensing frequency requirements can be relaxed. In addition to sensing frequency, the channel detection time, channel move time and some other timing related parameters are also defined in [24]. Another factor that can affect the sensing frequency is the interference tolerance of primary license owners. An example is when a cognitive radio is using opportunities in public safety bands, sensing should be done as frequently as possible in order to prevent any interference. The effect of sensing time on the performance of secondary users is investigated in [25]. The aim is to maximize the average throughput of secondary users while guiding primary users from interference [18]. Similarly, detection time is obtained using numerical optimization in [26]. Channel efficiency is maximized for a given detection probability. Sensing time can be decreased by sensing only changing parts of the spectrum instead of the entire target. A channel that is being used by secondary users cannot be used for sensing. Hence, secondary users have to stop data transmission for spectrum sensing [27]. This however, decreases the spectrum efficiency of the overall system [23]. To solve this problem, a method known as dynamic frequency hopping (DFH) is proposed. The DFH method is based on the assumption of having more than a single channel. This was proposed in [28].

E. Decision Fusion in Cooperative Sensing

Sharing information among cognitive radios and combining results from various measurements is a challenging task [18]. This shared information can either be soft or hard decisions made by each cognitive device [29]. The results presented in [29, 30] shows that soft information combining method performs better than hard information combining method in terms of the probability of missed opportunity. Hard decision is found to perform as good as soft decision when the numbers of cooperating users are high. The optimum fusion rule for combining sensing information is the Chair-Varshney rule with log-likelihood ratio test [31]. Likelihood ratio tests (LRT) are used for making classification using decision from secondary users in [29, 32, 33, 34, 35]. Different techniques for combining sensing result are employed in [12]. The credibility of cognitive radios depends on the channel conditions and their distance from a licensed user. The required number of nodes for satisfying a probability of false alarm rate is investigated in [36].

F. Security

In Cognitive radio, an unauthorized user can change its air interface to look like a primary user. However, this phenomenon misleads the spectrum sensing performed by legitimate primary users. This type of attack is investigated in [37]. The possibilities of primary user emulation (PUE) attacks are realistic due to the facts that CR is highly reconfigurable due to the fact that they are software based air interface [16, 38]. In order to stop such attacks, a robust transmitter verification scheme that can distinguish between legitimate incumbent primary signal transmitters and secondary signal transmitters needs to be designed [16]. The

task of differentiating an incumbent primary signals user from secondary users becomes challenging when the requirement described in FCC's NPRM 03-322, which states that "no modification to the incumbent system should be required to accommodate opportunistic use of the spectrum by secondary user signal". The major technical challenge in spectrum sensing is distinguishing primary signals from secondary user signal. A public key encryption is proposed in [39]. The primary user encrypts its identification with its private key and appends the encrypted value (signature) to its transmission. All secondary users scan for the signature, during the sensing period, the signature from various base stations. The base station then verifies these signatures. Since the primary user knows its signature, a malicious secondary user cannot produce a valid signature.

IV. ALGORITHM FOR SPECTRUM SENSING IN COGNITIVE RADIO

This section presents a study on spectrum sensing techniques that require knowledge of both source signal and noise power information. Some of the most common spectrum sensing techniques in this category is explained in this section.

A. Parametric Method of Spectrum sensing schemes

Three basic parametric method of spectrum sensing are explained as follows:

a) Optimal LRT-Based Sensing

The Neyman-Pearson states that for a given probability of false alarm, the test statistics that maximizes the probability of detection is the likelihood ratio test (LRT) [40, 41, 42] which is defined as:

$$T_{LRT}(x) = \frac{P(x|H_1)}{p(x|H_0)} \quad (1)$$

where $P(\cdot)$ denotes the probability density function (PDF) and (x) denotes the received signal vector that is the aggregation of $x(n)$, $n = 0, 1, \dots, N-1$. Such likelihood ratio test decides \mathcal{H}_1 when $T_{LRT}(x)$ exceeds a threshold γ , otherwise it uses \mathcal{H}_0 . The main challenge in implementing the LRT is the requirement on the distribution given in equation (1). The distribution of random vector x less than \mathcal{H}_1 is related to the source signal distribution, the wireless channels and the noise distribution. The distribution of x under \mathcal{H}_0 is related to the noise distribution [43].

In order to implement the LRT, a prior knowledge of the channel as well as the signal and noise distribution is of paramount importance. This is practically difficult to realize.

Assuming that the channels are flat-fading and the received source signal sample $s_i(n)$ is independent over, the PDF in LRT is decoupled as:-

$$P(x|H_1) = \prod_{n=0}^{N-1} P(x(n)|H_1),$$

$$P(x|H_0) = \prod_{n=0}^{N-1} P(x(n)|H_0), \quad (2)$$

Furthermore assuming that noise and signal samples are both Gaussian distributed, such that $\eta(n) \sim N(0, \sigma^2 \eta I)$ and $s(n) \sim N(0, R_s)$, the LRT becomes the estimator correlator (EC) detector as shown in [54] for which test statistic is given as:

$$T_{EC}(x) = \sum_{n=0}^{N-1} x^T(n) R_s (R_s + \sigma^2 \eta I)^{-1} x(n), \quad (3)$$

From equation (3) it is shown that $R_s (R_s + 2\sigma^2 \eta I)^{-1} x(n)$ is the minimum mean squared error (MMSE) estimation of the source signal $s(n)$. Thus, $T_{EC}(x)$ in (3) can be seen as the correlation of the signal $x(n)$ with the MMSE estimation of $s(n)$. EC detector needs to know the source signal covariance matrix R_s and noise power $\sigma^2 \eta$. Hence when the signal presence is unknown it becomes unrealistic to require signal covariance matrix for detection. It should be noted that if we assume that the noise is Gaussian distributed and the signals source is deterministic and known to the receiver, which in this case is the radar signal processing [44, 45, 46], it would then it can be easy to show that LRT becomes the matched filter based detector and its test statics is [43].

b) Matched Filter

Matched filter (MF) is a linear filter designed to maximize the output signal to noise ratio (SNR) for a given input signal [47]. Matched filtering is also known as optimal method for detection of primary users when transmitted signal is known [48]. Hence, cognitive radio has a prior knowledge of the Primary User Signal at both PHY and MAC layer, such as bandwidth, frequency, modulation type to demodulate received signals [49]. Matched filter detector has a high processing gain, but the sensing devices have to achieve coherency and demodulate primary user signal. This can be achieved since most wireless networks have pilot patterns (or symbols) and preambles that can be used for coherent detection. For examples: TV Signal has narrowband pilot for audio and video carriers; CDMA system have dedicated spreading codes for pilot and packet acquisition. The operation of matched filter detection is expressed as:

$$Y[n] = \sum_{K=-\infty}^{\infty} h[n-k]x[k] \quad (4)$$

where x is the unknown signal (vector) and is convolved with the h . The impulse response of the matched filter is useful only in cases where the information from the primary users is known to the cognitive users.

Matched filter advantage is it requires less detection time because it requires only $O(1/\text{SNR})$ samples to meet a given probability of detection constraint. When the information of the primary user is known to the cognitive radio user, matched detection is optimal [64].

The drawback of matched filter is that it requires prior knowledge of every primary signal. If the information is not accurate, MF would perform poorly. Also the most significant disadvantage of MF is that cognitive radio would need a dedicate receiver for every type of primary user [61].

c) Cyclostationary Based Detection

Cyclostationary based detection is a method that detects primary users by exploiting its Cyclostationary features of the received signals [50, 51]. Modulated signals are in general coupled with sine wave carriers, pulse trains, repeating spreading, hopping sequence or cyclic prefixes; these modulated signals are known as cyclostationary, since they have statistics, mean and autocorrelation. They can also be intentionally induced to assist spectrum sensing [52]. The cyclostationary based detection algorithm can differentiate noise from primary users signal. This is due to the fact that noise is in wider sense stationary with no correlation while modulated signal are cyclostationary with spectral correlation due to the redundancy of signal periodicities [43]. This periodicity trend is used for analyzing various signal processing tasks such as detection, recognition and estimation of the received signals. Even though cyclostationary feature detection have high computational complexity, it performs well satisfyingly well under low SNR due to its robust against unknown level of noise. Free bands in the spectrum are detected following the hypotheses testing problem in received signal $x(t)$ [53].

$$x(t) = s(t)h + w(t) \quad (5)$$

where $s(t)$ is the modulated signal, h is channel coefficient and $w(t)$ AWGN.

- Under H_0 $x(t)$ it is not cyclostationary and thus the band is considered free
- Under H_1 $x(t)$ is cyclostationary and thus the band is considered congested

where H_0 signifies the existence of signals and H_1 the existence of signal. Modulated signal $x(t)$ is considered to be a periodic signal or a cyclostationary signal in wide sense its mean and autocorrelation exhibits periodicity as shown in [54]. Though cyclostationary detection has certain advantages such as its robustness to uncertainty in noise power and propagation channel. It has its own disadvantages as follows:

- It needs a very high sampling rate
- The computation of spectral correlation density (SCD) function would require large number of samples and thus become complex.
- The strength of SCD could be affected by the unknown channel
- Sampling time error and frequency offset could affect the cyclic frequency.

B. Semi Blind Detection Methods

This section shows detection techniques that requires only noise power information. Hence it's called semi-blind detection.

a) Energy Detection

Energy detection is an optimal way to detect primary signals when prior information of the primary signal is unknown to secondary users. It measures the energy of the received waveform over a specified observation time [9, 55]. In addition, as receivers do not require any knowledge on the primary users signal. The signal is detected by comparing the output of the energy detector with a threshold which depends on the noise floor. Energy detector also known as radiometer has been investigated and widely used for signal detection due to its advantage of simple circuitry in practical implementation [56]. Prior to energy detection been proposed, many work have been performed to study energy detection based schemes in radar and security communication areas. Have some advantages that motivate research in this area. These include the following:-

- It is more generic as receivers do not need any knowledge on primary user's signal
- It is very simple to implement

The signals can be detected at low SNR provided the detection interval is adequately long and noise power spectral density is known.

The study of energy detection takes into account the dynamics traffic patterns of primary users, in the form random signal arrival and departure is of theoretical and practical importance. However, some of the existing techniques resort to approximation to certain approximation techniques to characterize the detection performance. In order to improve this technique, we would propose a Bayesian based Energy detection algorithm. There has been recent works which addresses the effect of primary user traffic patterns on the performance of the detection of energy detectors. In [57], they considered the random arrival or departure of the primary user's signal which exploits the distributions of the arrival and departure times. The effect of the primary user traffic on the detection performance is investigated and studied in [58]. However, to improve the robustness of energy detection we would propose a Bayesian –based Energy detection by exploiting the statistical knowledge.

b) Wave form based Sensing

In wireless systems, known patterns such as preambles, midambles, regularly transmitted pilot pattern, spreading sequence etc. [56]. The problems of energy detection which are false detection and difficulty in differentiating modulated signals from interference. Both of these problems are addressed in waveform based sensing. Waveform based sensing is performed in time domain using received signal;

$$y(n) = x(n) + z(n) \quad (6)$$

Where $x(n)$ is the signal to be detected and $z(n)$ is the Additive white Gaussian noise (AWGN). Assuming the known time- domain signal contains N_B signal [56]. We can then consider the following wave forms sensing metric:

$$S = R_e \left[\sum_{n=1}^{N_B} y(n)x^*(n) \right] \quad (7)$$

When there is no primary user signal present, the sensing metric would then be

$$S = S_0 = R_e \left[\sum_{n=1}^{N_B} z(n)x^*(n) \right] \quad (8)$$

When there is presence of primary user's signal present the sensing metric becomes:

$$S = S_1 = \sum_{n=1}^{N_B} |x(n)|^2 + R_e \left[\sum_{n=1}^{N_B} z(n)x^*(n) \right] \quad (9)$$

The decision on the presence of a primary user can be made by comparing the decision metric S against a fixed threshold λ_z . The sensing metrics (7) can then be approximated as a Gaussian random variable when N_B sample is large. [59]. Waveform based sensing outperforms energy detection based sensing in reliability and convergence time. Though waveform based sensing has good advantage, it also has its drawback. Since waveform based sensing requires short measurement time, it is then susceptible to synchronization errors.

C. Totally Blind Detection

This section presents detection techniques of spectrum sensing that requires no information what so ever on source signal or power. These techniques are explained as follow:

a) Eigenvalue based-sensing

This section reviews two sensing algorithms under the totally blind sensing spectrum. The first algorithm is based on the ratio of the maximum eigenvalue to minimum eigenvalue and the other is based on the ratio of average eigenvalue to minimum eigenvalue. There are two major eigenvalue based detection technique that would be studied in this paper, they are:

1) Maximum-minimum eigenvalue detection (MME)

This method generalizes the energy detection because it is used on a basis similar to the energy detection. What makes this unique is that it does not require any prior knowledge of the signal and the channel. It also eliminates the susceptibility of energy detection synchronization error, since it doesn't require synchronization. It is shown that the ratio of the maximum eigenvalue to the minimum can be used to detect signal [59]. This is achieved by some Random matrix theories (RMT), from this we can quantize the ratio and therefore find the threshold. The probability of the false alarm can also be found by using the random matrix theories [60, 4]. This technique overcomes the noise uncertainty difficulty which is peculiar to the energy detection while keeping the advantages of energy detection. It can even perform better than energy detection when the signals to be detected are highly correlated for signal detection as we already know from the beginning of this paper, there are two hypotheses H_0 , signal does not exist and H_1 signal exist. The received signal under the hypothesis is given as follows [13, 40]:-

$$H_0: x(n) = \eta(n), \quad (10)$$

$$H_1: x(n) = s(n) + \eta(n), \quad (11)$$

where $s(n)$ is the transmitted signal sample and $\eta(n)$ is the white noise which is independent and identically distributed (iid). There are two probabilities that are of interest for channel sensing. They are; probability of detection P_d , at hypothesis H_1 and the probability of the sensing algorithm having detected the presence of primary signal. The probability of false alarm P_{FA} which defines the Hypothesis H_1 . [60]. The probability of the presence of the primary signal can be defined by the following vectors. Assuming we consider L consecutive samples and then defines the vectors as follows [61]. The major advantage of the maximum-minimum eigenvalues based detection is that they do not need the noise power for detection. The major similarity with the energy detector is that they both use the received signal for detection and no information on the transmitted signal and channel is needed.

2) Energy with Minimum Eigenvalue based Detection (EME)

In this algorithm, the ratio of the signal energy to the minimum eigenvalue is used for detection of the primary user signal. as discussed in [62]. The difference between the conventional energy detection and EME is:

- Energy detection compares the signal energy to the noise power, which has to be estimated in advance.

While the EME on the other hand compares the signal energy to the minimum eigenvalues of the sample covariance matrix, which is computed from the received signal only. Though they have differences, but are however similar to energy detection. The MME and EME only use the received signal samples for detection and requires no information on the transmitted signal and channel is needed. The major advantage of EME detection over energy detection is:

- Energy detection requires noise power for detection while the EME does not.

The major complexity of EME is the computation of the covariance matrix equations and the eigenvalue decomposition of the covariance matrix. From the work done by Zeng *et al.* [62], the EME is worse than the ideal energy detection but better than energy detection with noise uncertainty 0.5dB [62]. The MME on the other hand performs better than the EME from the experiment done by Zeng *et al.* [60] but there is no theoretical proof yet in literature yet. The eigenvalue based methods can be used for different signal detection application without the knowledge of the signal, channel and noise power such as DTV signal and wireless microphone.

V. COOPERATIVE SPECTRUM SENSING

This scheme was proposed as a solution to the problem of noise uncertainty, fading and shadowing. This scheme decreases the probabilities of false detection and false alarm. Cooperative sensing can also be used to solve the problem of hidden primary user problem and can also reduce sensing time [18]. The major idea of cooperative sensing is that it increases the sensing performance by exploiting the spatial diversity in the observation of spatially located cognitive radio users [19].

By cooperating, cognitive radio users can share their sensing information for making a combined decision more accurate than individual decision [51]. The process of sensing starts with local sensing; this is when spectrum sensing is performed individually at each cognitive radio. The local sensing can be formulated as hypothesis problem as follows [63].

$$x(t) = \begin{cases} n(t) & H_0 \\ h(t), s(t) + n(t), & H_1 \end{cases} \quad (12)$$

Where $x(t)$ is the received signal at the cognitive radio user, $s(t)$ is the transmitted primary signal, $n(t)$ is the zero mean additive white Gaussian noise (AWGN), H_0 and H_1 denote the hypothesis of the absence and presence of signals respectively. The detection performance probability and the probability of false alarm are defined as:

$$P_d = P\{\text{decision} = H_1/H_1\} = P\{Y > \lambda | H_1\} \quad (13)$$

$$P_f = P\{\text{decision} = H_1/H_0\} = P\{Y > \lambda | H_0\}, \quad (14)$$

Where Y is the decision statistics and is λ the decision threshold.

Cooperative spectrum sensing implementation can be categorized into three, they include: Centralized, Distributed and Relay assisted.

a) Centralized Cooperative Sensing

In the centralized cooperative sensing, the central identity also known as fusion center (FC) [62] controls the three steps of cooperative sensing process. The first stage of the process, FC then chooses a channel or frequency band for sensing then delegates all cooperating cognitive radio to individually perform local sensing [64]. In the second process, all cooperating cognitive radio reports the sensing results through the control channel. In the third process the FC combines the received local sensing information, then determines the presence of primary users and then passes the decision to the cooperating cognitive radio users.

All cognitive radio users are tuned to the selected channel or frequency band where a wireless point to point link between the primary user (PU) transmitter and each cooperating radio also known as sensing channel is used for observing the PU and for data reporting all cognitive users are tuned to a control called a reporting channel. From figure 1, the Fusion center (FC) and CR1-CR5 performing local sensing and reporting back to the FC. In centralized system, the cognitive radio base station is the FC. But in cognitive Radio ad hoc network (CRAHNs), the cognitive radio base station is not present; hence any cognitive radio can then act as a FC to coordinate the sensing activities and then combines the sensing information from the cooperating neighbors. In a situation where there are large numbers, the required bandwidth for reporting results becomes huge. The reduction techniques of the sharing bandwidth and local observation are discussed in [65]. Only cognitive radios with reliable information are required to send decision to the fusion center [66].

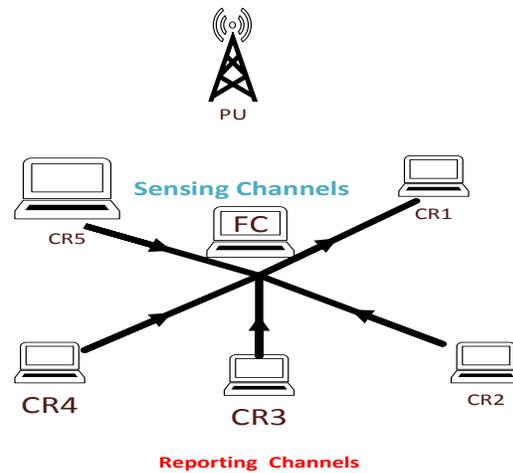


Fig. 1. Illustration of Centralized Cooperative sensing

b) Distributed Cooperative Sensing

In this type of sensing, cognitive node share information among each other. Though, they make their individual decision on the presence or absence of primary users. Fig 3 shows the distributed cooperative sensing. CR1-CR5 shares the locally sensed results with users within the transmission range. Several distribution algorithm have been developed [66, 17, 67], based on these algorithm each cognitive radio user transmits its own data to other users, then combines its results with the received data and then decides if a primary user is present by using local criterion [64]. Distributed sensing is more advantageous than the centralized sensing because it does not require fusion center (FC) for cooperative decision thereby reducing cost.

c) Relay Assisted Spectrum Sensing

Since the centralized and distributed cooperative sensing scheme is not that perfect, it gave birth to the relay assisted scheme. In this scheme, the cognitive radio user observing a weak sensing channel and a strong report channel, a cognitive radio user with a strong sensing channel and a weak report channel [64].

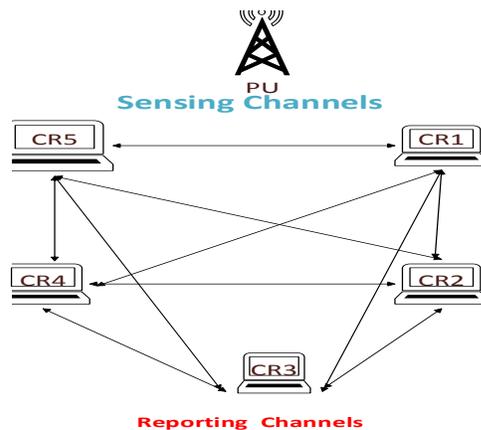


Fig. 2. Illustration of Distributed Cooperative sensing

Figure 2, shows CR1, CR4 and CR5, observes strong primary user signal, which may suffer from a weak reporting channel. CR2 and CR3 which have a strong report channel, serves as a relay to the fusion center (FC). In this situation, the report channels from CR2 and CR3 can also be known as the relay channels. Though, figure 2 shows a centralized structure. The relay assisted cooperative sensing scheme can also exist in distributed scheme. Hence, if the centralized and distributed structures are one hop cooperative sensing, the relay assisted structure can be considered as multi-hop cooperative sensing [17]. Though, the cooperative sensing scheme has some impressive advantages, such as higher accuracy in primary user detection, reduced sensing time and the prevention of shadowing effect and hidden node problem. The disadvantage with the scheme is the complexity of sensor within the cooperation among system cooperation, traffic overhead and the need for a control channel.

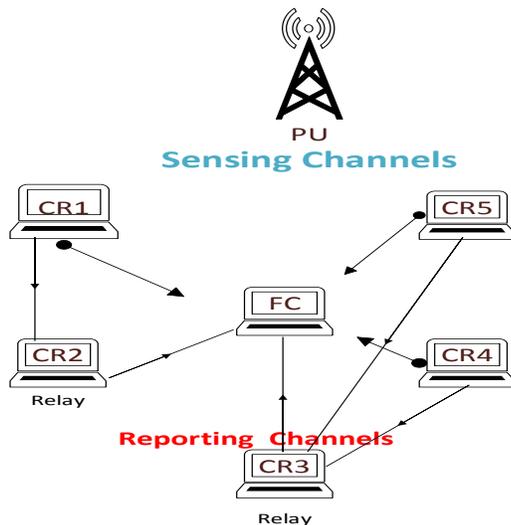


Fig. 3. Illustration of Relay Assisted Cooperative sensing

d) Data Fusion

In cooperative spectrum sensing data fusion is a procedure of combining local sensing data for hypothesis testing that is a constituent of cooperative sensing. This is based on the control channel bandwidth requirement; recorded sensing results can be of different forms, types and sizes [19]. Hence, the sensing results relayed to the FC or shared with cooperating users can be combined in three different ways they include; soft combining, quantized soft local combining and hard local decision

In the case of soft combining, cognitive radio users can either transmit the whole local sensing samples or the total local test statistics for soft decision. The receiver diversity techniques that is utilized for soft combining is the equal gain combining (EGC) and maximal ratio combining (MRC) [68]. The cognitive users can only transmit the quantize local sensing results and send the quantized data for soft combining increase control communication cost. In the case of hard combining, the commonly used fusion rules are AND, OR and Majority Rules. The cognitive radio users make a local decision and transmit the binary decision for hard combining.

VI. RESEARCH CHALLENGES TO IMPROVE EXISTING COOPERATIVE SENSING

The challenges to improve cooperative sensing delay are as follows:

Multiple tradeoffs in cooperative sensing delay: The sensing-throughput tradeoff analysis in cooperative Sensing should consider not only the sensing time and CR throughput, but also the report delay and the delay for synchronization or asynchronous reporting. Thus, the challenge is to balance the tradeoff between the CR throughput and cooperative sensing delay, which consists of multiple delay components depending on the cooperative sensing schemes.

Delay analysis in distributed schemes: Distributed cooperative sensing schemes usually require an iterative process to reach the cooperative decision. The cooperative sensing delay is dominated by the report delays if the number of iterations for convergence is large. As a result, the delay analysis and the convergence of the distributed cooperative algorithm should be jointly considered.

With the above listed factors, we would improve the cooperative spectrum sensing by using an improved energy detection based on second order statistics in a centralized cooperative spectrum sensing scheme.

VII. CONCLUSION

In this paper, the various spectrum sensing schemes have been reviewed. The various aspects of the sensing scheme are explained in details. Based on the different methodologies that were studied, the cooperative sensing scheme was considered as a solution to some specific challenges associated with spectrum sensing such as hidden primary user etc. Cooperative sensing is seen as an effective technique to improve detection performance by exploiting spatial diversity. Special attention was also given to the totally blind sensing methods that do not require prior information on the source signals and the transmitting channel. In conclusion, the review of various sensing techniques would be useful to researchers in developing a novel system for spectrum sensing algorithm. Also we identified some challenges in cooperative spectrum sensing which would be useful to researchers starting their research.

REFERENCES

- [1] "End to End Efficiency (E3)," 2009. [Online]. Available: <http://ict-e3.eu>.
- [2] J.Mitola, "Cognitive Radio: Making software radio more personal," IEEE personal communication, vol. 6, no. 4, pp. 13 - 18, Aug 1999.
- [3] FCC, "Spectrum policy task force report,," Technical report 02-135 Federal communication commission, Nov 2005.
- [4] A.M.Shahzad, M.A.Shah, A.H.Dar,A.Haq, A.U.Khan, T.Javed, S.A.Khan, "Comparative analysis of primary transmitter detection based spectrum sensing technologies in cognitive radio systems," Australian Journal of Basic and applied sciences , vol. 4, no. 9, pp. 4522 - 4532, 2010.
- [5] W.Wang, "Spectrum sensing for cognitive radio," third international symposium on intelligent information technology application workshop, pp. 410 - 412, 2009.
- [6] V.Stoianovici, V.Popescu, M.Murroni, "A Survey on spectrum sensing techniques in cognitive Radio,," Bulletin of the Transilvania University of Brasov, vol. 15, no. 50.

- [7] J.Mitola, "Cognitive radio: An Integrated agent architecture for software defined radio," PHD thesis, Royal Institute of technology (KTH), May 2000.
- [8] C.H.Hwang, S.H.Chen, "Spectrum sensing in wideband OFDM cognitive radio," IEEE transaction in signal processing, vol. 58, no. 2, pp. 709 - 719, Feb 2010.
- [9] H.Urkowitz, "Energy detection of unknown deterministic signals," Proceeding of the IEEE, vol. 55, no. 4, pp. 523 - 531, 1967.
- [10] M.Ghozzi, M.Doiler, F.Marx and J.Palicot, "Cognitive radio:Methods for detection of free bands," Comptes Rendus Physique,Elsevier, vol. 7, pp. 794 - 804 , 2006.
- [11] D.Cabric, A.Tkachenko and R.Brodersen, "Spectrum sensing measurement of pilot, energy and collaborative detection," in IEEE military communication conference, washington DC, USA, 2006.
- [12] F.F.Digham, M.S.Alouini and M.K.Simon, "On Energy detection of unknown signals over fading channels," vol. 55, no. 1, pp. 21 - 24, 2007.
- [13] A.Sahai and D.Cabric, "Spectrum sensing; Fundamental limit and practical challenges," in IEEE international symposium on new frontiers in dynamic spectrum access network (DySPAN '05), Baltimore, MD, USA, 2005.
- [14] H.Sun and A.Nallanathan, "Wideband spectrum sensing for cognitive radio network: A Survey," IEEE wireless communication, vol. 20, no. 2, pp. 74-81, April 2013.
- [15] H.Sun, W.Y.Chiu,J.Jiang, A.Nallanathan and H.V.Poor, "Wideband spectrum sensing with sub-Nyquist samples in cognitive radios," IEEE Transaction on signal processing , vol. 60, no. 11, pp. 6068 - 6073, 2012.
- [16] S.Haykin, "Cognitive radio: Brain-Empowered wireless communication," IEEE journal on selected areas in communication, vol. 25, pp. 201 - 220, Feb 2005.
- [17] J.A.Bazerque and G.B.Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity," IEEE transaction on signal processing , vol. 58, no. 1, pp. 1847 - 1862, Mar 2010.
- [18] T.Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," IEEE communications and survey& tutorial , vol. 11, no. 1, 2009.
- [19] I.F.Akyildiz, F.Brandon, L.Ravikumar, "Cooperative spectrum sensing in cognitive radio networks:A survey," Physical communication, vol. 4, pp. 40 - 62, 2011.
- [20] T.Yucek and H.Arslan, "MMSE Noise plus interference power estimation in adaptive OFDM system," IEEE transaction on vehicular technology, vol. 56, no. 6, pp. 3857 - 3863, Nov 2007.
- [21] Y.Hur,J.Park, W.Woo, K.Lim, C.Lee, H.Kim and J.Lasker, "A wideband analog multi-resolution spectrum sensing (MRSS) techniques for cognitive radio (CR) systems," in Island of KOS, Greece, 2006.
- [22] G.Ganesan and Y.Li, "Agility improvement through cooperative diversity in cognitive radio," in IEEE Global telecommunication conference (GLOBECOM '05), Missouri, USA, Dec. 2005.
- [23] S.D.Jones, E.Jung, X.Liu, N.Merheb and I.J.Wang, "Characterization of spectrum activities in the US public safety band for opportunistic spectrum access," in International symposium on new frontiers in dynamic spectrum access networking (DySPAN '07), Dublin, Ireland, APR. 2007.
- [24] C.Cordeiro, K.Challapali, D.Birnu and S.Shankar, " IEEE 802.22; the first worldwide wireless standard based on cognitive radio," in IEEE international symposium on new frontiers in dynamic spectrum access network (DySPAN'05), Baltimore, MD, USA, Nov. 2005.
- [25] A.Ghasemi and E.S.Sousa, "Capacity of fading channel under spectrum-sharing constraints," in IEEE international conference on communications, Istanbul, Turkey, 2006.
- [26] P.Wang, L.Xiao, S.Zhou and J.Wang, "Optimization of detection time for channel efficiency in cognitive radio systems," in IEEE wireless communication and networking conference, Hong Kong, 2005.
- [27] N.Khambekar, L.Dong and V.Chaudhery, "Utilizing OFDM guard interval for spectrum sensing," in IEEE wireless communication and networking conference, Hong Kong, Mar. 2007.
- [28] W.Hu, D.Wikomm, M.Abusubaih, J.Gross, G.Vlantis, M.Gerla and A.Wolisz, "Dynamic frequency hopping communities for efficient IEEE 802.22 operation," IEEE communication magazine, vol. 45, no. 5, pp. 80 - 87, May 2007.
- [29] E.Visotsky, S.Kuffner and R.Peterson, "On collaborative detection on TV transmission in support of dynamic spectrum sharing," in Proceedings of IEEE International symposium of New Frontiers in Dynamic spectrum access network, Baltimore, Nov 2005.
- [30] T.Weiss, J.Hillenbrand and F.Jondral, "A diversity approach for the detection of Idle spectral resource in spectrum pooling systems," in Proceedings of the 48th International scientific colloquium, Immenau, Germany, 2003.
- [31] Z.Chair and P.K.Varshney, "Optimal data fusion on multiple sensor detection system," IEEE transaction on Aerospace Electronics system, vol. 22, no. 1, p. 98 101 , 1986.
- [32] M.Gandetto, A.F.Catto, C.S.Regazzoni and M.Musso, "Distributed cooperative mode identification for cognitive radio application," in Proceedings of international radio science union (URSI), New Delhi, India, 2005.
- [33] M.Gandetto, A.F.Cattoni, C.S.Regazzoni, "Distributed approach to mode identification and spectrum monitoring for cognitive radio," in Proceedings of SDR forum for technical conference, Orange County, California, USA, Nov. 2005.
- [34] A.F.Cattoni, I.Minetti, M.Gandetto, R.Niu, P.K.Varshney and C.S.Regazzoni, "A Spectrum sensing algorithm based on distributed cognitive model," in Proceeding of SDR forum for technical conference, Orlando, Florida, USA, Nov.2006.
- [35] M.Gandetto andf C.S.Regazzoni, "Spectrum Sensing: A distributed approach for cognitive terminals," IEEE Journal on selected areas of communication, vol. 25, no. 3, pp. 546 - 557, 2007.
- [36] P.Pawelczar, G.J.Janssen and R.V.Prasad, "Performance measure of dynamic spectrum access networks," in Proceedings of IEEE Global telecommunication conference (Globecom), San Francisco, California USA, Nov.2006.
- [37] R.Chen and J.M.Park, "Ensuring trustworthy spectrum sensing in cognitive radio network," in Proceedings of IEEE workshop on networking technologies for software defined radio networks (held in conjunction with IEEE SECON 2006), 2006.
- [38] E.Orumwense, O.Olutayo, S.Mneney, "Impact of primary user emulation attack on cognitive radio network," International journal on communication antenna on propagation, vol. 4, no. 1, pp. 19 - 26, 2014.
- [39] C.N.Mathur and K.P.Subbalakshmi, "Digital signatures for centralized DSA network," in First IEEE workshop on cognitive radio networks, Las Vegas, Nevada, USA, Jan 2007.
- [40] S.M.Kay, Fundamentals of statistical signal processing : Detection Theory, Upper Saddle River, NJ: Prentice Hall, 1998.
- [41] H.V.Poor, An Introduction to signal detection and estimation, Berlin: Springer, 1988.
- [42] H.L. Van-Tress, Detection, Estimation and Modulation Theory, New York: John Wiley & Sons, 2001.
- [43] Y.Zheng, Y.C.Liang, A.T.Hoang and R.Zheng, "A Review on spectrum sensing for cognitive radio:challenges and solutions," EURASIP Journal and advances in signal processing, 2010.
- [44] E.Fishler, A.Haimovich, R.Blum, D.Chizhik,L.Cimini and R.Valenzuela, "MIMO Radar: an idea whose time has come," in Proceedings of the IEEE National rada conference, Philadelphia, USA, Apr.2004.
- [45] A.Sheiki and Zamani, "Coherent detection for MIMO Radars," in Proceedings of IEEE National Radar Conference, Apr. 2007.
- [46] P.Stoica, J.Li and Y.Xie, On Probing signal processes advances in wireless and mobile communications, vol. 1, Prentice Hall, 2001.
- [47] M.Subhedar and G.Birajdar, "Spectrum sensing techniques in cognitive radio: A Survey," International Journal of Next-Generation Networks (IJNGN), vol. 3, no. 2, Jun 2011.
- [48] J.G.Proakis, Digital Communication, 4th ed., McGraw Hill, 2001.
- [49] R.F.Ustok, Spectrum sensing techniques for cognitive radio systems with multiple antennas, 2010.
- [50] S.Shankar, C.Cordeiro and K.Challapali, "Spectrum agile radio: utilization and sensing architectures," in IEEE international symposium

- on new frontiers in dynamic spectrum access network (DySPAN'05), Baltimore, Maryland, USA, Nov. 2005.
- [51] D.Cabric, S.Mishra and Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in Proceedings of Asilomar conference on signal systems and computers, California, USA, No. 2004.
- [52] K.Maeda, A.Benjaebbour, T.Asai, T.Furuno and T.Onya, "Recongnition array OFDM-Based systems utilizing cyclostationarity inducing transmission," in Proceedings of IEEE international symposium on New Frontiers in Dynamic spectrum access networks, Dublin, Ireland, Apr.2007.
- [53] W.A.Gardner, A.Napolitanob and L.Paurac, "Cyclostainoarity:Half a ccentury of research," Elsevier signal processing, vol. 86, pp. 639 - 697, 2006.
- [54] Y.Tengyiz and G.Chi, "Performance of Cyclostationary feature based spectrum sensing methods in multiple antenna cognitive systems," in Wireless communication and networking conference (WCNC), 2009.
- [55] M.Hoyhtya, A.Hekkala, M.Katz and A.Mammela, "Spectrum Awareness: Techniques and challenges for active spectrum sensing," in Cognitive wireless networks, 2007.
- [56] T.Yucek, Channel, Spectrum and wave form Awareness in OFDM-Based Cognitive Radio, 2007.
- [57] N.C.Beaulieu and Y.Chen, "Improved energy detectors for cognitive radios with randomly arriving and departing primaryusers," IEEE signal processing letters, vol. 17, no. 10, pp. 870 - 877, 2010.
- [58] L.Tang, Y.Chen, E.L.Hines and M.S.Alouini, "Effect of primary user traffic on sensing-throughput tradeoff for cognitive radios," IEEE Transaction on wireless communication , vol. 8, no. 4, pp. 1063 -1068, 2011.
- [59] H.Tang, "Some physical layer issues of wiiderband cognitive radio system," in Proceeding of IEEE international sypsymposium of New Frontiers in Dynamic Spectrum access networks, Baltimore, Maryland, USA, Nov.2005.
- [60] Y.Zeng, C.L.Koh and Y.C.Liang, "Maximum Eigenvalue Detection:Theory and Application," in IEEE International conference on communication, 2007.
- [61] A.M.Tulino and S.Verdu, Random Matrix Theory and, Now Publisher Inc, 2004.
- [62] Y.Zeng and Y.C.Liang, "Eigenvalue- Based spectrum algorithm for cognitive radio," IEEE Transaction on communication, vol. 57, no. 6, 2009.
- [63] M.E.Yildizy, T.C.Aysaly and K.E.Barner, "In network cooperative spectrum sensing," in Proceedings of EURASIP European signal processing conference, Glasgow, UK, 2009.
- [64] X.Jing, D.Raychandhuri, "CSCE etiquette protocol," in IEEE DySPAN, 2005.
- [65] C.Sun, W.Zhang and K.B.Letaief, "Cooperative spectrum sensing for cognitive radio under bandwidth constraints," in In proceedings of IEEE communication and networking conference, Hong Kong, Mar 2001.
- [66] Z.Li, F.R.Yu and M.Huang, "A distributed consensus based cooperative spectrum sensing scheme in cognitive radio," IEEE transaction on vehicular technology, vol. 59, no. 1, pp. 383 - 393, 2010.
- [67] W. K.B.Letaief, "cooperative communication for cognitive radio networks," proceeding of IEEE, vol. 97, no. 5, pp. 878 - 893, 2009.
- [68] [68]J.Ma, G. Yeli and B.H. Juang, "Signal processing in cognitive radio," proceeding of IEEE, vol. 97, no. 5, May 2009.
- [69] [69]Y.Zeng , C.L.Koh and Y.C.Liang, "Maximum eigenvalue detection: theory and application," in IEEE International conference on communication, 2007.

A Posteriori Pareto Front Diversification Using a Copula-Based Estimation of Distribution Algorithm

Abdelhakim Cheriet
LESIA Laboratory, Biskra University
Algeria

Foudil Cherif
LESIA Laboratory, Biskra University
Algeria

Abstract—We propose CEDA, a Copula-based Estimation of Distribution Algorithm, to increase the size, achieve high diversity and convergence of optimal solutions for a multiobjective optimization problem. The algorithm exploits the statistical properties of Copulas to produce new solutions from the existing ones through the estimation of their distribution. CEDA starts by taking initial solutions provided by any MOEA (Multi Objective Evolutionary Algorithm), construct Copulas to estimate their distribution, and uses the constructed Copulas to generate new solutions. This design saves CEDA the need of running an MOEA every time alternative solutions are requested by a Decision Maker when the found solutions are not satisfactory. CEDA was tested on a set of benchmark problems traditionally used by the community, namely UF1, UF2, ..., UF10 and CF1, CF2, ..., CF10. CEDA used along with SPEA2 and NSGA2 as two examples of MOEA thus resulting in two variants CEDA-SPEA2 and CEDA-NSGA2 and compare them with SPEA2 and NSGA2. The results of The experiments show that, with both variants of CEDA, new solutions can be generated in a significantly smaller without compromising quality compared to those found SPEA2 and NSGA2.

Keywords—Multiobjective Optimization Problems; Evolutionary Algorithms; Estimation of Distribution Algorithms; Copulas

I. INTRODUCTION

A Multiobjective Optimization Problem (MOP) is an optimization problem that involves multiple functions with objectives that need to be optimized simultaneously. These objectives are usually contradictory so much so improving one objective may degrade many others. Under these circumstances, there does not exist a single solution that optimizes all functions. Instead, there typically are a number of optimal solutions, called Pareto solutions, which are considered equally good and cannot be ordered completely [1].

Although these Pareto solutions are considered equally good, a decision maker involved in working with the Pareto solutions obtained from solving a multiobjective problem may not be satisfied with some of them. In many of these cases, a decision maker may need to solve the multiobjective problem again with the expectation of finding another set of solutions that suit his needs in a better way.

Searching for new solutions by running a multiobjective problem solver each time may not be practical as finding a new solution can be complex and require a significant amount of time and resources, particularly if the solution technique used is not appropriate.

Motivating by the effort to make it more efficient for a decision maker to search for new solutions, this paper target reducing the time needed to generate new solutions without compromising their qualities. This paper propose, a Copula-based Estimation of Distribution Algorithm. CEDA belongs to the class of Estimation of Distribution Algorithms (EDA) [2], which is itself a class of Evolutionary Algorithms (EA) [3] usually used to solve multiobjective problems. In contrast to EA where new solutions are generated using an implicit distribution defined by one or more variation operator (mutation, crossover), EDA uses an explicit probability distribution model to characterize the interactions between the solutions. This feature along with their good global searching ability makes EDA well suited for efficiently generating new solutions.

Although there are many variants of EDA (See Section 4 for details), the work (CEDA) based on Copulas [4] for their ability to provide a scale-free description of how Pareto solutions are distributed. With Copulas, a joint probability distribution function can be constructed which makes is particularly easy to generate new sample solutions according to that joint probability distribution function. This makes CEDA efficient in generating new solutions in quick way with a high degree of quality thereby making it convenient for a decision maker to search for new solutions that would better suit his needs.

Briefly, this is achieved by the way CEDA operates, which starts by selecting the best individual using a MOEA (Multi Objective Evolutionary Algorithm)[5,6,7,8] from a population generated randomly. Then, CEDA uses the selected individuals to estimate their distribution using a Copula. The constructed Copula is used to generate a new population. CEDA continues with generating and selecting the best individuals until the stop condition is met. When CEDA stops, the latest generated individuals are considered Pareto optimal solutions and the last Copulas can be used in later calls of CEDA to generate alternative optimal solutions if those generated do not satisfy the needs of the Decision Maker. This design saves CEDA the need of running an MOEA every time alternative solutions are requested by a Decision Maker when the found solutions are not satisfactory.

The main contributions of this paper are the following:

- Devise a Copula-based EDA to increase the size, deliver high diversity, and achieve a quick convergence of Pareto optimal solutions for a multiobjective optimization problem. We achieve this by exploiting the

statistical properties of Copulas to produce new solutions from the existing ones through the estimation of their distribution.

- Define a new performance metric called solution generation efficiency to measure the speed of generating new Pareto optimal solutions in terms of the number of objective function evaluations.
- Thoroughly test CEDA on a set of benchmark problems traditionally used by the community, namely UF1,...,UF10, CF1,...,CF10, using SPEA2 [6] and NSGA2 [5] algorithms as two example candidates for MOEA selecting methods. Finding that new Pareto optimal solutions can be generated in a significantly smaller time compared to those found by NSGA2 and SPEA2, without compromising the quality (convergence and diversity) of these solutions.

The rest of the paper is organized as follows. In Section 2 a definition of multiobjective optimization problems is given. In Section 3, provide an overview of the work carried out in the area of multiobjective optimization. In Section 4, an overview on EDA is presented and described how they generally operate. In Section 5, provide a mathematical definition of Copula and some of their features used in EDA. We present the contribution CEDA Copula-based EDA in Section 6 and evaluate its performance on various benchmark problems in Section 7. We conclude our paper in Section 8.

II. MULTIOBJECTIVE OPTIMIZATION

A multi-objective optimization problem is an optimization problem that involves multiple objective functions [1]. In mathematical terms, a multi-objective optimization problem can be formulated as follows.

$$\begin{aligned} \min F(x) & \quad \text{where } F = (f_0, f_1, \dots, f_m) \\ \text{subject to } G(x) & \leq 0 \end{aligned} \quad (1)$$

With $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^m$, $\mathbf{G}(\mathbf{x}) \in \mathbb{R}^p$, we have m functions to optimize and p constraints to satisfy. The main goal of optimization methods is to find an optimal solution for the problem described in (1). Note that a multiobjective problem has many objectives to achieve which are usually mutually contradictory. Therefore, a relation for assessing the goodness of a solution compared to another one should be defined. Typically, the Pareto Dominance relation (see Definitions below) is used to achieve this end.

Definition 1 Considering a minimization problem, a decision vector \mathbf{u} **weakly dominates** \mathbf{v} ($\mathbf{u} \preceq \mathbf{v}$) iff

$$\begin{aligned} f_i(u) & \leq f_i(v), \forall i \in \{0, 1, \dots, m\} \text{ and} \\ \exists j & \in \{0, 1, \dots, m\}, f_j(u) < f_j(v). \end{aligned}$$

Definition 2 Considering a minimization problem, a decision vector \mathbf{u} **dominates** \mathbf{v} ($\mathbf{u} < \mathbf{v}$) iff

$$f_i(u) < f_i(v), \forall i \in \{0, 1, \dots, m\}.$$

Definition 3 A solution \mathbf{x}^* is a **Pareto optimal** solution if and only if there is no other admissible solution \mathbf{x} where $f(\mathbf{x})$ dominates $f(\mathbf{x}^*)$. So the solution of a Multiobjective problem is a set of solutions which are not dominated by any other solution, we call this set the **Pareto Solutions PS**. The image of this set in the objective space form the **Pareto Front PF**:

III. RELATED WORK

There are in the literature many methods for solving a multiobjective problem. Those based on EA, referred to as MOEA (Multiobjective Optimization Evolutionary Algorithms) are among the most used ones due to the good quality/cost trade-off of the solutions they provide [9]. MOEA may be classified according to the following aspects: (1) the techniques used to solve the optimization problem and (2) the schemes used for the reproduction of the offspring.

In the MOEA based on decomposition MOEA/D [9], the multiobjective problem (MOP) is decomposed into a number of scalar objective optimizations (SOPs). The objective of each SOP is called sub-problem. The population is composed in every generation with the best solution found for each sub-problem [10].

The Indicator-based MOEA framework is a recent kind of resolution which uses the Quality Indicator of the approximated Pareto Front to guide the search, the Generational Distance and the Hypervolume are two examples of the indicators used in the work of [11, 12, 13].

Another type of the MOEA frameworks is the one that is based on preference. In this class of framework, the Decision Maker (DM) is involved in the choice of preferred solutions, so the MOEA method needs to get a Pareto Front of interest to the DM. Various algorithms exist according to the way of involving the DM, a priori, a posteriori, or interactively.

In many a priori approaches (e.g. [14]), a preference point or region is given to guide the search for solutions process. The preference points are chosen according to the DM demands. After getting the preferred direction, the search process is executed from the begin to the end without involving the DM. Note that the solution obtained after executing the algorithm is usually not the best solution and may not even be close to the most preferred solution.

In a posteriori methods (e.g. NSGAI [5], SPEA2 [6]), optimal solutions are obtained using an evolutionary algorithm ignoring the interaction with the DM. After getting the PS, the DM can choose one of the obtained solutions. A posteriori methods do not provide the DM with the option of guiding the search for new solutions thereby possibly leading to solutions that are not of interest to the DM.

In interactive methods (e.g. [15, 16]), the DM directs the search for new solutions with the aim for finding solution that are of interest to them. Although these methods help the DM find good solutions to their problem, the interaction process significantly slows down the computation of solutions.

MOEA can also be classified according to the method they use for reproduction (e.g. the DE (Differential Evolution)-based algorithms [17], the Immune-based algorithms [18], the

PSO (Particle swarm optimization)-based [19] algorithms, and the probabilistic model-based algorithms).

The probabilistic model-based approaches are considered as a new paradigm in the evolutionary computation. Their principal idea is extracting the statistical information from their previous generations and trying to build a probabilistic distribution model of the best candidate solutions. This distribution is used to sample new individuals (solutions). Examples of probabilistic model-based algorithms include those using Ant Colony Optimization, Cross Entropy [20], and Quantum-inspired Genetic Algorithm [21].

Another very important class of the probabilistic-based models are those based on the estimation of distribution, known as the EDA (Estimation of Distribution Algorithms). This class of algorithms was first introduced by Mühlenbein and Paaß [22]. The rest of this paper mainly deals with this class of algorithms, which will be explained in details in Section 4.

IV. ESTIMATION OF DISTRIBUTION ALGORITHMS

The Estimation of Distribution Algorithms is a class of Evolutionary Algorithms. It is a population based algorithm which starts with an initial population usually a random one, and then tries to select the best solutions using a fitness function (for example in the experimentation, the best solution is the one that is not dominated by any another solution). The statistical properties of the selected solutions (individuals) are used to find a distribution or a kind of function or law representing all the selected solutions. The EDA algorithms try in every generation of the algorithm to estimate the distribution of the best solution in this generation. After finding or estimating the distribution of the best-selected solutions, a number of new individuals are generated using the created function or law. In general, those new individuals have the same properties of the best solutions of the precedent generation. The algorithm runs many generations according to the steps described above until a criterion stop is achieved [2].

The general steps followed by an EDA are described in Algorithm 1:

Algorithm 1 Estimation of Distribution Algorithm

Initialization

While Not termination criteria **do**

 Select best Solutions

 Estimate the best Solutions Distribution

 Generate a candidate Solutions

End While

Most of Estimation of Distribution Algorithms may be classified into two categories: those that deal with discrete variables and those that deal with the real-valued vectors. In the discrete variables class, we find algorithms that use univariate models, which assume that the problem variables are independent. Under this assumption, the probability distribution of any individual variable should not depend on the values of any other variables.

Mathematically, a univariate model decomposes the probability of a candidate solution (X_1, X_2, \dots, X_n) into the product of probabilities of individual variables as

$$p(X_1, X_2, \dots, X_n) = p(X_1)p(X_2) \dots p(X_n)$$

where $p(X_i)$ is the probability of variable X_i , and $p(X_1, X_2, \dots, X_n)$ is the probability of the candidate solution (X_1, X_2, \dots, X_n) . One of simplest algorithms that uses this idea is the Univariate Marginal Distribution Algorithm (UMDA). UMDA works on binary strings and uses the probability vector $p = (p_1, p_2, \dots, p_n)$ as the probabilistic model, where p_i denotes the probability of a "1" at position i of solution strings.

One of the main drawbacks of UMDA is the necessity of keeping the selected individuals to calculate the probability vector. To alleviate this problem, Incremental EDAs propose to update the probability vector incrementally to avoid keeping the list of all individuals. Population-Based Incremental Learning (PBIL) is an example of Incremental EDAs where probability vector elements are calculated according to the following equation:

$$p_i = (p_i * (1.0 - LR)) + (LR * v_i)$$

where p_i is the probability of generating a 1 in bit at position i , v_i is the i th position in the solution string and LR is the Learning Rate specified by the user. Although using univariate models is efficient particularly in saving memory usage, the assumption that problem variables are independent will often prevent efficient convergence to the optimum when problem variables interact strongly.

Tree-based models are another EDAs that deal with discrete variables. This type of EDAs is capable of capturing some pair-wise interactions between variables. In tree-based models, the conditional probability of a variable may only depend on at most one other variable. The Mutual-Information-Maximizing Input Clustering (MIMIC) uses a chain distribution to model interactions between variables. Given a permutation of the n variables in a problem, $\pi = i_1, i_2, \dots, i_n$, MIMIC decomposes the probability distribution of $p(X_1, X_2, \dots, X_n)$ as

$$p_\pi(X) = p(X_{i_1}|X_{i_2})p(X_{i_2}|X_{i_3}) \dots p(X_{i_{n-1}}|X_{i_n})p(X_{i_n})$$

where $p(X_{i_j}|X_{i_{j+1}})$ denotes the conditional probability of X_{i_j} given $X_{i_{j+1}}$.

All EDAs motioned previously are applicable to problems with candidate solutions represented by fixed-length strings over a finite alphabet. However, candidate solutions for many problems are represented using real-valued vectors. In these problems, the variables cover an infinite domain so it is no longer possible to enumerate variables' values and their probabilities. This gives rise to EDAs that deal with the real-coded values. One example of dealing with the real-coded values is to manipulate these through discretization and variation operators based on a discrete representation. Typically, there are three different methods of discretization: fixed-height histograms, fixed-width histograms, and k-means clustering.

The next algorithms are examples of EDAs that work directly with the real-valued variables themselves. The Estimation of Gaussian Networks Algorithm (EGNA) works by creating a Gaussian network to model the interactions between variables in the selected population of solutions in each generation [2].

Recently a new approach to developing EDAs to solve real-valued optimization problem has been developed that is based on Copula theory. The main idea of Copulas is to decompose the multivariate joint distribution into each univariate distribution and a Copula. Copula is a function that embodies the relationship of the variables [23, 24]. The use of Copula-based models in continuous EDAs places these algorithms in an advantageous position in comparison with other EDAs that rely on the assumption of a particular multivariate distribution, such as the multivariate normal distribution [25, 26]. By means of Copulas, any multivariate distribution can be decomposed into the marginal distribution and the Copula that determines the dependence structure between the variables.

The main steps of a Copula-based EDA are resumed in the Algorithm 1.a:

Algorithm 1.a Copula-based EDA

Generate initial population P_0

$t = 1$

While not stop criterion do

P_t^S = select best individual

Use P_t^S to learn (estimate parameters of) a Copula C

P_{t+1}^S = sample individuals from C

End while

Many types of Copula have been used in the literature. In [26], the authors used T-Copula, in [27, 28, 29, 30], the authors used an Archimedean Copula, in [31] the authors used Clayton Copula, and in [32, 33, 34], the authors combined more than one Copula to find the best estimation. This paper, will focus on Archimedean Copulas for their ability to model dependence in high dimensions with only one parameter, which has the good effect of speeding up multiobjective optimization computation time.

V. MATHEMATICAL OVERVIEW ON ARCHIMEDEAN COPULAS

As defined in [4], Copulas are functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions and as distribution functions whose one-dimensional margins are uniform.

Definition 4 A function C is called a Copula if only if is defined:

$$C : [0,1]^d \rightarrow [0,1]$$

It has the following characteristics:

$C(u_1, \dots, u_d) = 0$ If one of its components u_i is equal to zero.

$$C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$$

In addition, C must be d -increasing. Example, for $d = 2$, we have:

$$C(u, v) : [0,1]^2 \rightarrow [0,1]$$

For any $0 \leq u \leq 1$ and $0 \leq v \leq 1$ we have the three following conditions:

$$C(0, v) = C(u, 0) = 0$$

$$C(1, v) = v$$

$$C(u, 1) = u$$

For any u and v , we define the 2-increasing propriety as:

$$C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0$$

Definition 5 According to Sklar's theorem, if C is a Copula, and if F_1, \dots, F_d are a cumulative distribution functions (univariate), then:

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d))$$

is a cumulative distribution function with a dimension d , where the marginals are F_1, \dots, F_d exactly.

The converse is also true: if F is cumulative distribution function with d dimension, there is a C Copula such as:

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d))$$

where all F_i are F marginals' laws.

According to Sklar's theorem, two steps are performed in order to construct the joint probability distribution function of a random vector. The first step is constructing the margins of each random variable separately. The second step is selecting a proper Copula to construct the joint distribution. Therefore, Copulas can be used to study the distribution character of each random variable and their relationship.

There are many families of Copulas. They can be characterized by one parameter or by a vector of parameters. These parameters measure the dependence between the marginals and are called dependence parameters θ . This paper, use Frank Copula, a variant of Archimedean Copulas, because we obtained satisfactory results with it (see Section 6.1.3.1). In general, Archimedean Copulas have one dependence parameter θ that can be calculated using Kendall's τ [4].

Kendall's τ measures the concordance between two continuous random variables X_1 and X_2 . The relation between Kendall's τ and θ in Frank Copula used in this paper is defined as:

$$\tau = 1 - \frac{4}{\theta} [1 - D_1(\theta)] \text{ where } D_1(\theta) = \frac{1}{\theta} \int_0^\theta \frac{t}{e^t - 1} dt$$

The Frank Copula function is defined by:

$$C(u, v; \theta) = -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right) \text{ where } \theta \in (-\infty, \infty)$$

The dependence parameter of a bivariate Copula can be estimated using the maximum likelihood method (MLE). To do so, we need to optimize the log-likelihood function given by:

$$l(\theta) = \sum_{t=1}^T \ln c(F(x_{1t}), F(x_{2t}); \theta)$$

where T is the sample size. The value θ which maximizes the log-likelihood $l(\theta)$ is called maximum likelihood estimator $\hat{\theta}_{MLE}$. Once the value of θ is estimated, the bivariate Copula is well defined. For maximizing the likelihood function, we use the nonparametric estimation of θ given by Kendall's τ as an initial approximation to $\hat{\theta}_{MLE}$.

After the characterization of the Copula, the generation of sample is performed as the following steps:

- 1) Generate two independent uniform (0,1) variables u and t ;
- 2) Set $v = C_u^{(-1)}(t)$, where $C_u^{(-1)}(t)$ denotes a quasi-inverse of $C_u(v)$.
- 3) The desired pair is (u, v) .
- 4) (x_1, x_2) is a sample of the specified joint distribution, where $x_1 = F_1^{(-1)}(u)$, $x_2 = F_2^{(-1)}(v)$

VI. CEDA: COPULA-BASED ESTIMATION OF DISTRIBUTION ALGORITHM

The aim of the proposal is to help the decision maker to get the solutions that are closest to its interest. To achieve this, a two-stage algorithm is proposed, that is composed of the *Optimization* stage which finds a set of the best solutions to a given problem (see Section 6.1) and the *Update* stage which finds another set of the best solutions until the decision maker is satisfied (see Section 6.2). Note that the Update stage runs much faster in finding new solutions compared to the initial Optimization stage.

A. Optimization Stage

Like every evolutionary algorithm the proposed algorithm (Algorithm 2) has two principal steps (i) the *Selection* and (ii) the *Reproduction*. In the Selection step (performed by Function *SelectUsingMOEA*), the proposal use the *NSGA2* [5] or *SPEA2*

[6] to select the best individuals (solutions) that will be used in the Reproduction step where CEDA makes use of Copulas to estimate and regenerate new individuals (performed by Functions *ConstructCopulas* and *GenerateSolutions* respectively).

A pseudo-code of the algorithm that performs the estimation of distribution using a Copula for solving multiobjective problems can be viewed as follows (Algorithm 2).

Algorithm 2 Copula-based EDA

Function CEDA

P₀ = Initialization(m)

P = SelectUsingMOEA(P₀)

While Not termination criteria **do**

C = ConstructCopulas(P)

P' = GenerateSolutions(C)

P'' = SelectUsingMOEA([P'P]^T)

P = P''

End while

Return (P, C)

End function

1) Initialization

Initially CEDA assume that we have a population **P**₀ = [**x**₁, ..., **x**_m]^T where **x**_{*i*}, *i* ∈ [1, m] are the individuals. Each individual **x**_{*i*} = [*x*_{1*i*}, ..., *x*_{*n**i*}] where *x*_{min} ≤ *x*_{*ij*} ≤ *x*_{max}. Both *x*_{min} and *x*_{max} are reals. Where each *x*_{*ij*} is initially picked up according to a uniform distribution in [*x*_{min}, *x*_{max}]. Note that each individual **x**_{*i*}, (*i* ∈ [1, m]) is real-value coded vector, i.e. every *x*_{*ij*}, (*i* ∈ [1, m], *j* ∈ [1, n]) are real values.

2) Selection

In selection step achieved by the function *SelectUsingMOEA*, CEDA use one of the classical algorithms *NSGA2* or *SPEA2* as a MOEA.

The result of the selection is a set of individuals that will be used in the reproduction step. The proposal call **P** the matrix of the individuals resulting from the selection process operated on the precedent population. For the first generation, the algorithm use the initial population **P**₀. **P** is defined as the following:

$$\mathbf{P} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

NSGA2 or *SPEA2* selects the best individuals of the population it operates on according to the dominance relation defined in Section 2. Note that generated solutions (obtained by *GenerateSolutions*) at a given step are not necessarily better than those generated at the step that preceded it. Therefore, the selection of the best solutions operates on the union of the two sets: the solutions obtained from the current step and those resulted from the step that preceded it, as shown in Algorithm 2.

3) Reproduction

To perform the reproduction, CEDA start by calculating the dependency between the best individuals using Copula as shown in Algorithm 3 and then generate new individuals using these Copulas as shown in Algorithm 4.

a) Constructing Copulas

The Copula type used in this paper is Archimedean. The Archimedean Copula deals with two vectors of variables \mathbf{u} and \mathbf{v} ; therefore, CEDA divided each one of the decision variable vectors into two sub vectors to fit into the variables \mathbf{u} and \mathbf{v} . CEDA performed this division into two sub vectors in each generation of the algorithm.

$$\mathbf{P} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = [\mathbf{w}_1, \dots, \mathbf{w}_n]$$

CEDA operate on the transpose of the matrix \mathbf{P} , and take each vector $\mathbf{w}_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$ the proposal take each vector $\mathbf{w}_j (j \in [1, n])$ to construct the sub vectors $\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{v}_1, \dots, \mathbf{v}_n$ according to the following: $\mathbf{u}_1, \mathbf{v}_1$ are extracted from \mathbf{w}_1 where the size of each of \mathbf{u}_1 and \mathbf{v}_1 are equal to $m/2$. The elements of \mathbf{u}_1 are taken randomly from \mathbf{w}_1 , and \mathbf{v}_1 is constructed from the rest of the elements of \mathbf{w}_1 . For the sake of simplicity, CEDA assume that m is an even number. In the case where m is odd, CEDA remove one individual from the initial population to make its size even. The computation of the other sub vectors $\mathbf{u}_2, \dots, \mathbf{u}_n$ and $\mathbf{v}_2, \dots, \mathbf{v}_n$ is performed in a similar way as for \mathbf{u}_1 and \mathbf{v}_1 respectively.

The algorithm create Archimedean Copulas, represented by the vector $\mathbf{C} = [C_1 \dots C_j \dots C_n]$, using the sub vectors $\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{v}_1, \dots, \mathbf{v}_n$, where each Copula $C_j, j \in [1, n]$ is constructed from the sub vectors \mathbf{u}_j and \mathbf{v}_j as shown in Algorithm 3.

Algorithm 3 Construct Copulas

Function ConstructCopulas(P,type)

For all w_j a vector in P do

$\mathbf{u}_j = \text{RandomPick}(w_j)$

$\mathbf{v}_j = \text{Remainder}(w_j, \mathbf{u}_j)$

$C_j = \text{Copula}(\mathbf{u}_j, \mathbf{v}_j, \text{type})$

End for

Return $\mathbf{C} = [C_1 \dots C_j \dots C_n]$

End function

Note that there are many types of Archimedean Copulas. In this paper, CEDA considered three of them namely Gumbel, Clayton and Frank Copula. CEDA have experimented with them on various optimization problems and found that Frank Copulas provides better results in the configurations tested on.

b) Generating New Individuals

The proposal uses the constructed Copulas C_1, \dots, C_n to generate new individuals. The set of the new generated individuals $\mathbf{X}' = [\mathbf{w}'_1, \dots, \mathbf{w}'_n]$ where \mathbf{w}'_j is the concatenation

of \mathbf{u}'_j and \mathbf{v}'_j which are sampled using Copula C_j . Note that the vector \mathbf{w}'_j (resulting from the concatenation of \mathbf{u}'_j and \mathbf{v}'_j) is of size m' that is not necessarily the same of the initial population size m . The new individuals are therefore the vectors $\mathbf{x}'_i, i \in [1, m']$ where $\mathbf{X}' = [\mathbf{x}'_1, \dots, \mathbf{x}'_{m'}]^T$. Algorithm 4 summarizes these steps.

Algorithm 4 Generate Solutions

Function GenerateSolutions(C, m')

For all C_j in C do

$(\mathbf{u}'_j, \mathbf{v}'_j) = \text{GenerateFromCopula}(C_j, m')$

$\mathbf{w}'_j = \text{Concat}(\mathbf{u}'_j, \mathbf{v}'_j)$

End for

return $\mathbf{X}' = [\mathbf{w}'_1 \dots \mathbf{w}'_j \dots \mathbf{w}'_n]$

End function

The function used to generate individuals form the estimated Copula C is performed in the same way defined in Section 5. CEDA start by picking u and t form (0,1) uniform function then we get v by calculating the $C_u^{(-1)}(t)$ the quasi-inverse function of C_u . The generated variables x_1 and x_2 are produced from the quasi-inverse function of each marginal distribution. In every iteration (see Algorithm 4.a), CEDA insert x_1 to the list **ListX1** and x_2 to **ListX2**. Finally, after generating m samples of x_1 and x_2 we return the two lists.

Algorithm 4.a GenerateFromCopula

Function GenerateFromCopula(C, m)

For $i=1, m$ do

$\mathbf{u} = \text{uniform}(0, 1)$;

$\mathbf{t} = \text{uniform}(0, 1)$;

$\mathbf{v} = C_u^{(-1)}(\mathbf{t})$;

$\mathbf{x}_1 = F_1^{(-1)}(\mathbf{u})$;

$\mathbf{x}_2 = F_2^{(-1)}(\mathbf{v})$;

Insert ($\mathbf{x}_1, \text{ListX1}$)

Insert ($\mathbf{x}_2, \text{ListX2}$)

End for

Return (**ListX1**, **ListX2**)

End function

B. Update Stage

The Optimization stage allows us to calculate new solutions as shown in Algorithm 2. These solutions may not suit the needs of the decision maker and thus another stage of new solutions generation is needed. The Update stage that proposed in this paper (as shown in Algorithm 5) makes it possible for the decision maker to find other new solutions quickly by using the Copulas constructed in the Optimization stage. Specifically, CEDA achieves this by calling Function *GenerateSolutions* with arguments \mathbf{C} (the Copulas constructed in the Optimization

phase), and m'' the number of new individuals required. The output of the Update stage is the population P_{update} . If the decision maker is still not satisfied with the obtained solutions, only another round of the Update stage is required thus saving the need for running the Optimization stage another time.

Algorithm 5 Update Solutions

Function UpdateSolutions(C,m'')

$P_{tmp} = \text{GenerateSolutions}(C,m'')$

$P_{update} = \text{SelectUsingMOEA}(P_{tmp})$

Return P_{update}

End function

It is important to note that Copulas used as input in the Update Solution Algorithm have been constructed using a set of the best solutions obtained in the last generation of the algorithm used in the Optimization stage. Therefore, those Copulas inherently characterize the distribution of the best solutions thereby making the new individuals P_{tmp} among the best solutions. The returned solutions at this stage (Update stage) P_{update} are selected from the temporary individuals P_{tmp} according to one of the MOEA to select the best solutions as shown in Algorithm 5.

VII. EXPERIMENTATION

A. Used Benchmark problems

To evaluate the efficiency of the proposed algorithm, we chose to test it on a set of benchmark problems usually used in the literature. Specifically, CEDA uses the benchmark problems UF1, UF2, ..., UF10, CF1, CF2, ..., and CF3 defined in CEC2009 competition [35]. CEDA operates on 100 individuals and set the maximum number of evaluation to 300000. Each algorithm runs independently 30 times for each benchmark problem, as recommended by CEC2009 settings. We vary the number of DM calls and show the results obtained with 5 and 20 DM calls.

B. Used Metrics

In addition to considering the metrics traditionally used to assess the quality of the obtained solutions, namely diversity and convergence, the proposal defines a new metric, **directed regeneration speed**, to measure how quickly new solutions can be obtained. The new metric allows showing the efficiency of the algorithm that enables finding new solutions according to the decision maker needs quickly without compromising their qualities.

Both the diversity and the convergence are calculated from the set of solutions obtained by the used algorithms (CEDA, SPEA2, NSGA2). The diversity of a set of solutions is calculated using the IGD metric defined in [10] to assess the quality of the distribution of the obtained solutions over the PF and the convergence to the PF.

The **solution updating speed** metric, referred to as I_{new} , measures the number of new solutions obtained over a period of time (expressed in terms of the number of function evaluations) as shown in (3).

To show the new aspect guaranteed with the algorithm, which is the ability to get new PS with a very short time (negligible) we have proposed a metric that calculate the number of different PS between two set of PS that can be defined as follows:

$$I_{new} = \frac{\sum_{t=0}^T |PS_t|}{\sum_{t=0}^T FE_t} \quad (3)$$

where $|PS_t|$ represents the number of Pareto Solutions obtained at iteration t , $|FE_t|$ is the number of the function evaluations, and T is the number of iterations (the number of times the decision maker calls the algorithm again to find new solutions).

C. Simulation Results

This section, shows that the proposed CEDA method achieves good diversity and convergence compared to those obtained with state-of-the-art methods such as SPEA2 and NSGA2 by considering benchmark problems (UF1, ..., UF10 and CF1, ..., CF10) taken from CEC2009 [35].

1) Solution Qualities

Figure 1, shows that the Pareto Front solutions obtained when solving the considered benchmark problems with the proposed method CEDA-SPEA2 (CEDA using SPEA2 as selection method) and CEDA-NSGA2 (CEDA using NSGA2 as selection method). We show that both CEDA algorithms find solutions with similar qualities independently of the algorithm used for the selection (SPEA2 or NSGA2), because solutions are generated according to the same Copulas-based technique — SPEA2 or NSGA2 are only used for the selection.

Figure 2 shows that CEDA-NSGA2 and CEDA-SPEA2 provide solutions with different qualities on benchmarks UF2 and UF1 because they use different techniques to find new solutions. Figure 1 and Figure 2 also show that the proposal always provides solutions that are close to the optimal Pareto Front, particularly on benchmarks UF7, UF4, and CF1.

Figure 3 shows that the proposed CEDA algorithm generates more Pareto Solutions compared to those obtained by traditional algorithm NSGA2. Similar results have been obtained with CEDA compared to traditional SPEA2.

We show that both variants of CEDA converge to the optimal Pareto Front in a way that is similar to NSGA2 and SPEA2 (see Figure 1, 2, and 3). This shows that the Copula estimator is very good and comparable to the classical genetic operators (mutation and crossover) in terms of reproduction.

2) Solution Convergence and Diversity

To evaluate the convergence and the diversity of CEDA during, the measure the IGD Indicator obtained with both CEDA variants (CEDA-NSGA2 and CEDA-SPEA2) as well as those obtained with NSGA2 and SPEA2.

Figure 4 and Figure 5 show that CEDA-NSGA2 (resp. CEDA-SPEA2) algorithm achieves mean IGD values that are

close to those obtained with NSGA2 (resp. SPEA2), with both 5 DM and 20 calls. Those IGD values reflect the good convergence and diversity qualities achieved by the method.

3) Solution Update Speed (Update Stage)

Figure 6 shows the IGD of the Pareto Front using the CEDA-NSGA2 and CEDA-SPEA2 in function of the number of function evaluations. Lower IGD values reflect better solution qualities. We show that the speed of generating better solutions achieved by the CEDA algorithms is higher than that achieved by traditional algorithms. For example, to decrease the value of IGD of the approximated Pareto Solutions of the UF4 problem using CEDA-NSGA2, from the beginning of the execution of the algorithm to 0.8, needs 500 function evaluations compared to 10000 required by NSGA2.

For example, in the plot of CEDA-NSGA2 in the subfigure corresponding to the UF1 benchmark, the results corresponding to one call are represented in the leftmost point. That point was found after running 1000 evaluations (as represented in the x-axis). The point next to it on the right corresponds to two calls, which was found after running 2000 evaluations (see corresponding value on the x-axis). The point next to the second point represents the results obtained for three calls, and so on until we reach the rightmost point, which corresponds to the results obtained for 20 calls. The same explanation applies to CEDA-SPEA2, as well as NSGA2 and SPEA2 in the other subfigures.

Note that counting the number of function evaluations in the optimization stage starts from the beginning of finding of new solutions until their convergence to the optimal Pareto Front. However, during the update stage, the CEDA algorithms generate new solutions, which assess the quality by calculating the IGD and compare it with the IGD found in the first phase. If the IGD of the update stage is smaller or equal to the one of the optimization stage which consider that the update stage converged too, which gives DM good alternative solutions.

4) New Solution Count (Update Stage)

Figure 4 and Figure 5, plots the number of new solutions obtained when a Decision Maker wants to generate new ones. CEDA tested the solution with two cases. In the first case, the Decision Maker makes 5 calls to Algorithm 5, and in the second one the Decision Maker 20 calls to the same Algorithm. By considering these two cases, the work aim to reflect various decision making needs requiring different numbers of algorithm calls to obtain Decision Maker satisfaction. We plotted the mean value of the number of new solutions averaged over a 30 simulation runs, as well as the standard deviation, maximum and minimum values. The work shows

that CEDA-based algorithms, both CEDA-SPEA2 and CEDA-NSGA2, generate a significantly greater number of new solutions per objective function evaluation (i.e. new solutions are obtained with fewer objective function evaluations), compared to traditional SPEA2 and NSGA2 algorithms. This is because Copulas based techniques reduce the search space which becomes closer to the optimal Pareto Front thereby making it easier to find new solutions with a smaller number of objective function evaluations compared to SPEA2 and NSGA2.

VIII. CONCLUSIONS

CEDA was presented, a Copula-based Estimation of Distribution Algorithm, to improve the efficiency of solving multiobjective optimization problems. CEDA is based on the statistical properties of Copulas to estimate the distribution of a population and thus its ability to generate new individuals with similar properties. This feature makes CEDA particularly designed to promptly help a Decision Maker find alternative solutions to a multiobjective problem when the solutions obtained by traditional algorithms such as SPEA2 and NSGA2 do not satisfy his/her needs. The production of alternative solutions by CEDA is accelerated by the fact that they are generated according the probabilistic model established by the use of Copulas thereby saving the need for running the costly traditional MOEA algorithms another time to find new solutions.

CEDA was tested on a set of benchmark problems from the CEC2009[35] traditionally used by the community for the evaluation of multiobjective problem solving algorithms and shown that the proposal provides solutions with good convergence and diversity compared to state-of-the-art algorithms such as SPEA2 and NSGA2. This work have also particularly shown that the time needed to generate these solutions is substantially smaller than that needed by state-of-the-art algorithms, which makes the algorithm suitable for prompt alternative solution generation.

CEDA was tested with traditional NSGA2 and SPEA2 as the selection methods. Although the results are encouraging, better results with other methods such as the MOEA/D or a hybrid evolutionary algorithm [36-39] as substitutes for NSGA2 and SPEA2 is expected. CEDA Algorithms may also be used for solving MaOPs (Many-Objective Problems) where there are more than four objectives to optimize. In addition, Copulas creations are independent and thus can be done in parallel, which will even enhance the performance on parallel computers.

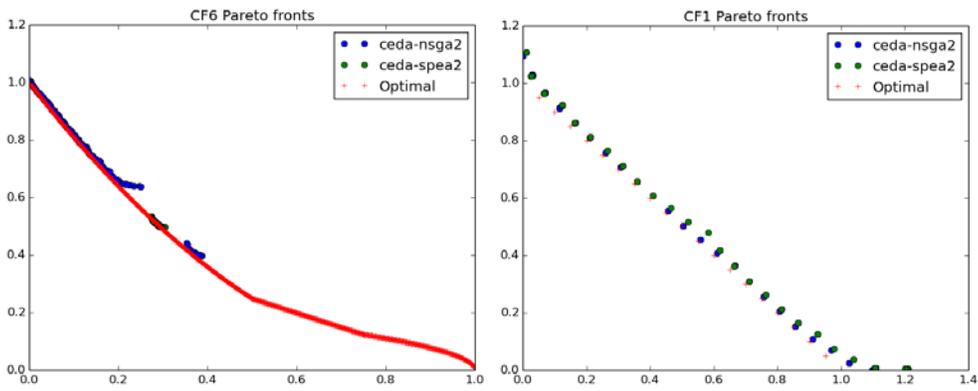


Fig. 1. Pareto front of the CF1 and CF6 Constrained problems

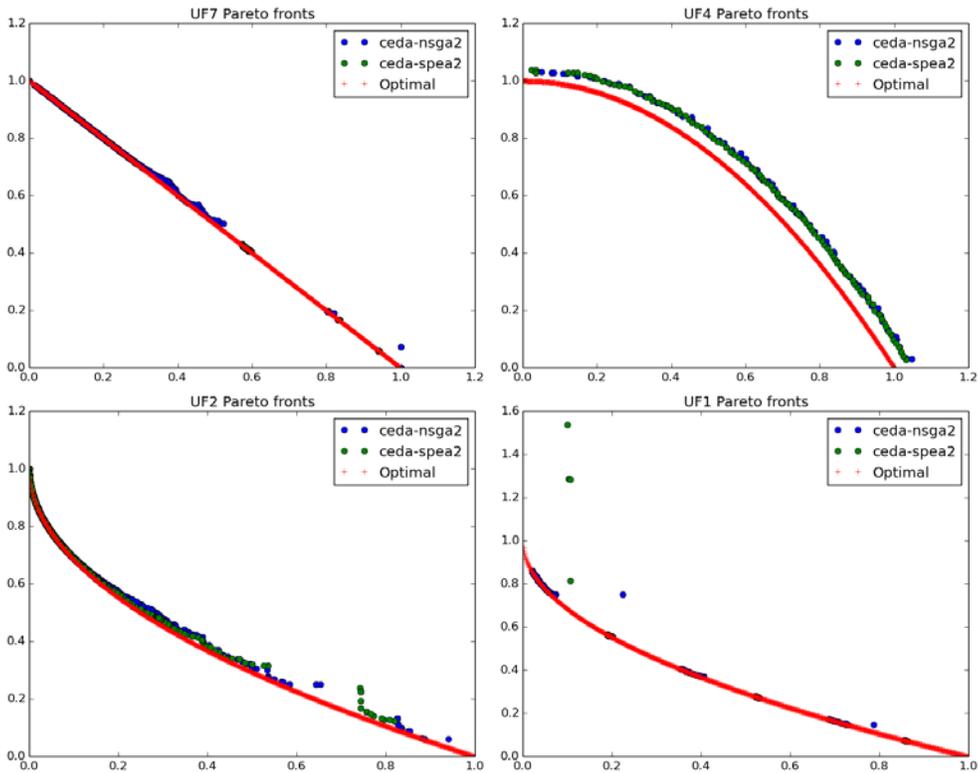


Fig. 2. Pareto front of the UF1,UF2,UF4 and UF7 unconstrained problems

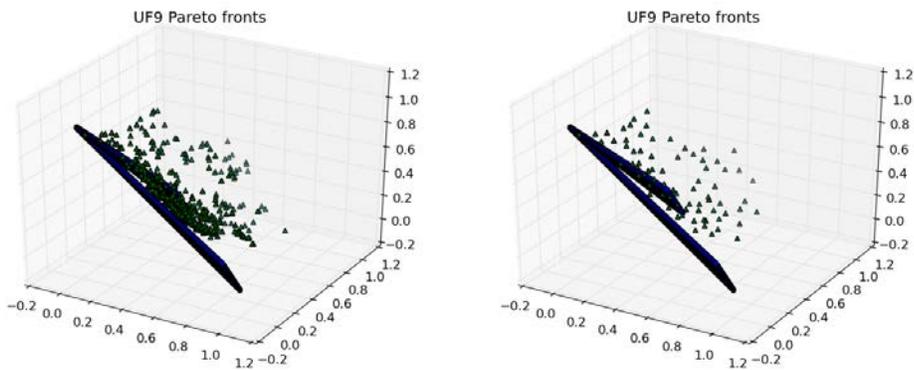


Fig. 3. Pareto Front of the CEDA-NSGA2 and NSGA2 of the UF9 problem

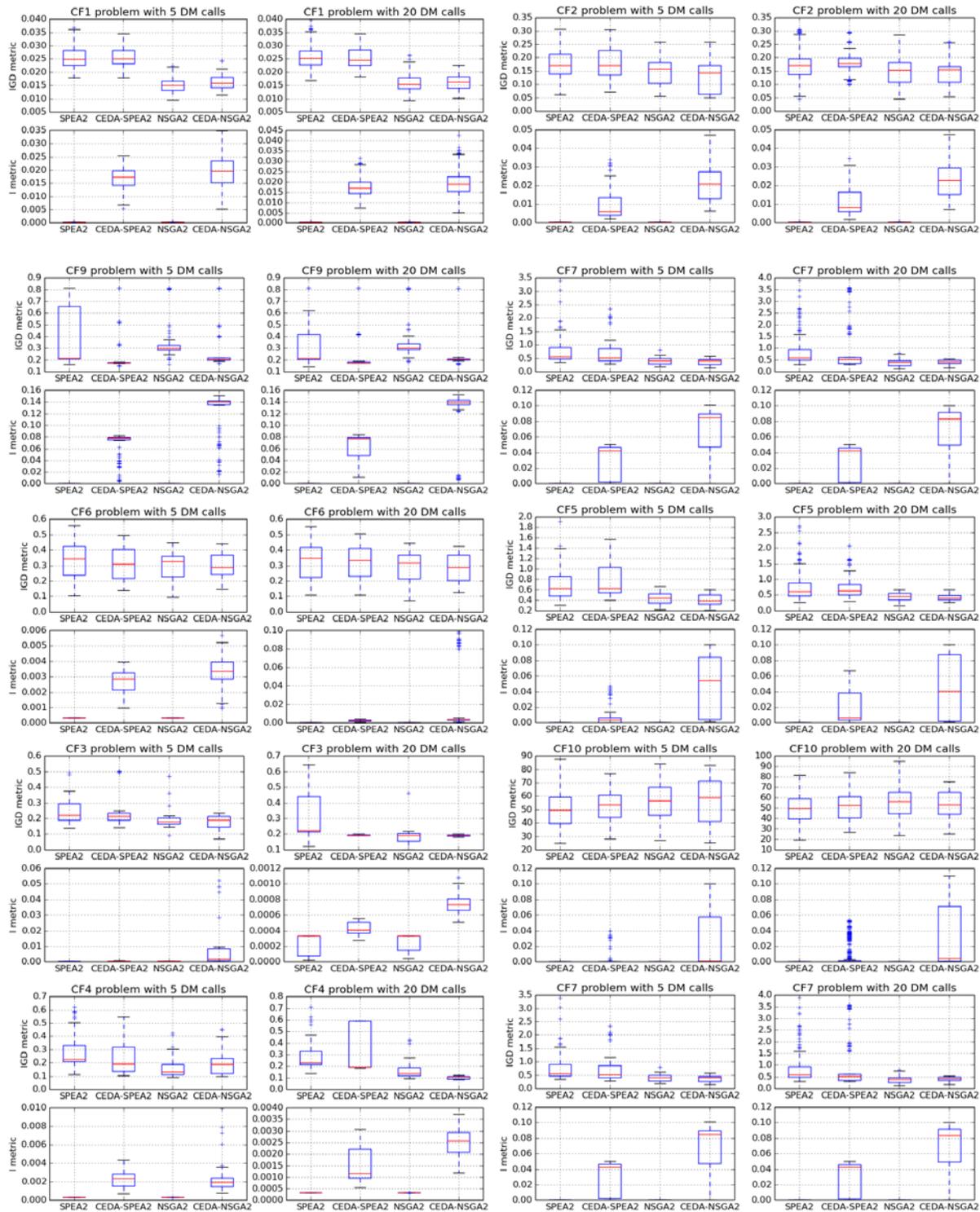


Fig. 4. Mean/Deviation of the *IGD* and *I_{new}* metrics for the constrained problems benchmarks

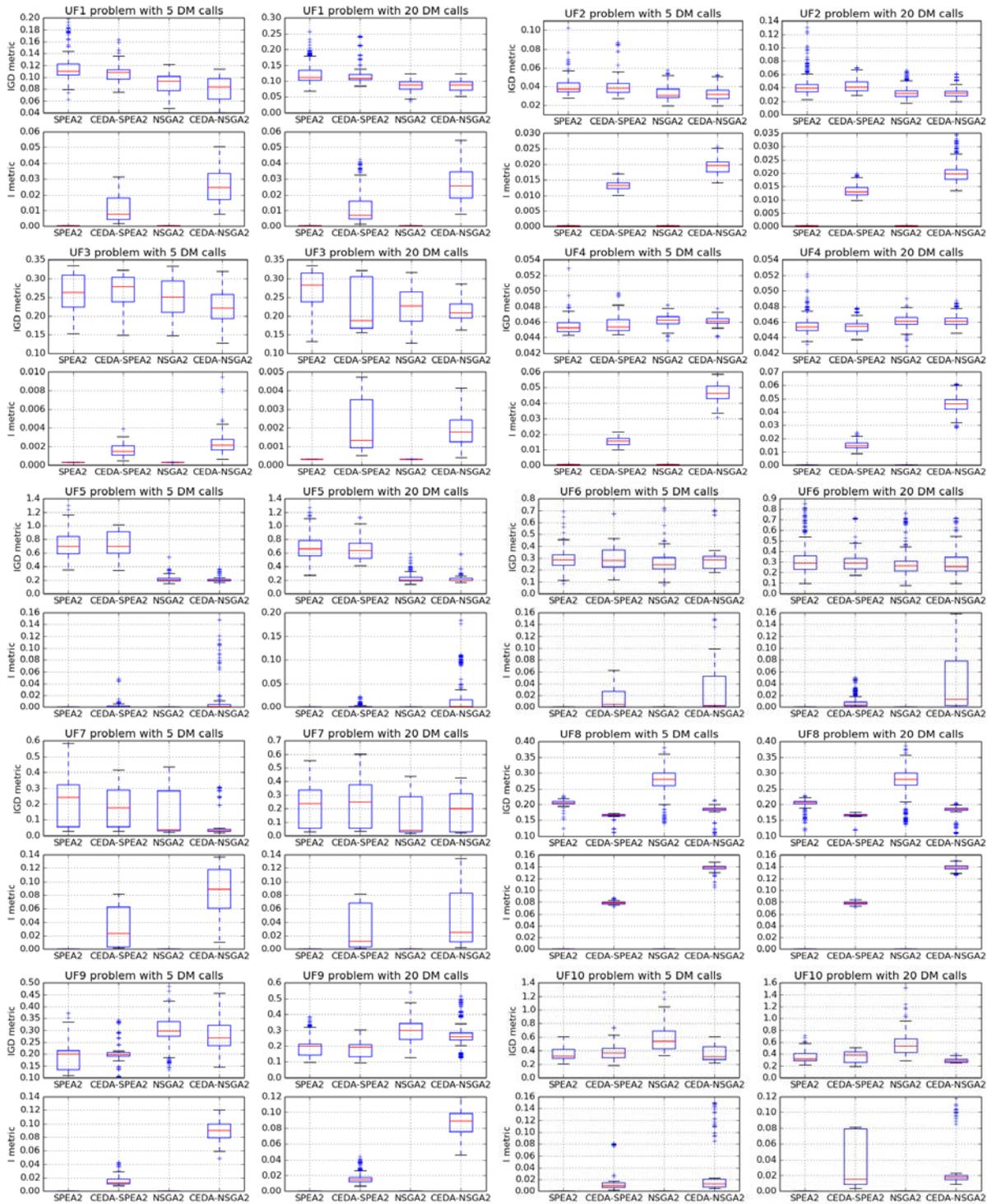


Fig. 5. Mean/Deviation of the *IGD* and *I_{new}* metrics for the unconstrained problems benchmarks

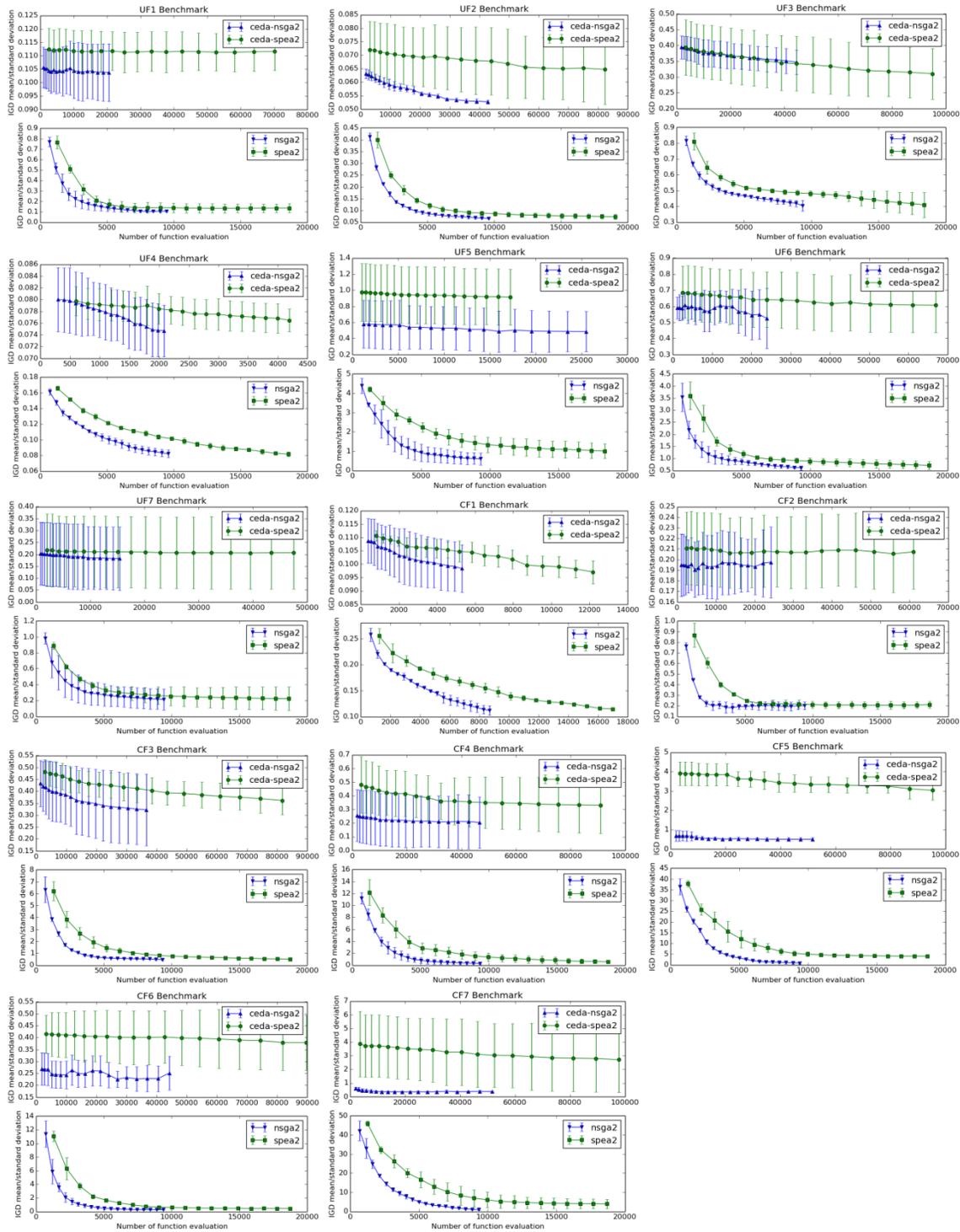


Fig. 6. The evolution of the means/standard deviations of IGD values of the approximate solution sets obtained with the number of function evaluations for the test instances

REFERENCES

- [1] Deb K. Multi-objective optimization using evolutionary algorithms. John Wiley & Sons, 2001.
- [2] Hauschild M, Pelikan M. An introduction and survey of estimation of distribution algorithms. *Swarm Evol Comput* 2011; 1: 111–128.
- [3] Auger A, Doerr B. Theory of Randomized Search Heuristics: Foundations and Recent Developments. vol. 1. World Scientific; 2011.
- [4] Nelsen RB. An introduction to Copulas. Springer; 2006.
- [5] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans Evol Comput* 2002; 6: 182–197.
- [6] Zitzler E, Laumanns M, Thiele L. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. In: *Evolutionary Methods for Design, Optimisation, and Control*; 2002; Barcelona, Spain. pp. 95–100.
- [7] Zitzler E, Deb K, Thiele L. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evol Comput* 2000; 8: 173–195.
- [8] Fonseca CM, Fleming PJ. Multiobjective optimization and multiple constraint handling with evolutionary algorithms II Application example. *IEEE Trans Syst Man Cybern Part A Syst Humans* 1998; 28: 38–47.
- [9] Zhou A, Qu BY, Li H, Zhao SZ, Suganthan PN, Zhang Q. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm Evol Comput* 2011; 1: 32–49.
- [10] Zhang Q, Li H. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Trans Evol Comput* 2007; 11: 712–731.
- [11] Zitzler E, Simon K. Indicator-Based Selection in Multiobjective Search. In: Yao X, Burke EK, Lozano J, Smith J, Merelo-Guervos JJ, Bullinaria JA, Rowe JE, Tino P, Kabán A, Schwefel HP, editors. 8th International Conference on Parallel Problem Solving from Nature. Berlin, Heidelberg: Springer, 2004. pp. 832–842.
- [12] Brockhoff D, Zitzler E. Improving hypervolume-based multiobjective evolutionary algorithms by using objective reduction methods. In: *IEEE Congress on Evolutionary Computation*; Sept 2007; Singapore. IEEE. pp. 2086–2093.
- [13] Bader J, Zitzler E. HypE: an algorithm for fast hypervolume-based many-objective optimization. *Lect Notes Comput Sc* 2011; 19: 45–76.
- [14] Branke J, Deb K. Integrating User Preferences into Evolutionary Multi-Objective Optimization. In: Jin Y, editor. *Knowledge Incorporation in Evolutionary Computation*. vol. 167. Berlin, Heidelberg: Springer, 2005. pp. 461–477.
- [15] Deb K, Jain H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Trans Evol Comput* 2014; 18: 577–601.
- [16] Sinha A, Korhonen P, Wallenius J, Deb K. An interactive evolutionary multi-objective optimization algorithm with a limited number of decision maker calls. *Eur J Oper Res* 2014; 233: 674–688.
- [17] Fleetwood K. An Introduction to Differential Evolution, New ideas in optimization. UK, Maidenhead, UK: McGraw-Hill Ltd, 1999.
- [18] Coello CAC, Cortés NC. Solving multiobjective optimization problems using an artificial immune system. *Genet Program Evolvable Mach* 2005; 6: 163–190.
- [19] Reyes-Sierra M, Coello CC. Multi-objective particle swarm optimizers: A survey of the state-of-the-art. *International journal of computational intelligence research* 2006; 2: 287–308.
- [20] Ünveren A, Acan A, editors. Multi-objective optimization with cross entropy method: Stochastic learning with clustered pareto fronts. In: *IEEE Congress on Evolutionary Computation*; 25–28 Sept. 2007; Singapore. IEEE. pp. 3065 – 3071.
- [21] [21] Han KH, Kim JH. Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. *IEEE Trans Evol Comput* 2002; 6: 580–593.
- [22] Muhlenbein H, Paaß G. From Recombination of Genes to the Estimation of Distributions I. Binary Parameters. In: Voigt HM, Ebeling W, Rechenberg I, Schwefel HP, editors. *Parallel Problem Solving from Nature PPSN IV*. Springer Berlin Heidelberg, 1996. pp. 178–187.
- [23] Salinas-Gutiérrez R, Hernandez-Aguirre A, Villa-Diharce ER. Estimation of Distribution Algorithms Based on Copula Functions. In: *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*; 12–16 July 2011; New York, NY, USA: ACM. pp. 795–798.
- [24] Wang L, Wang Y, Zeng J, Hong Y. An estimation of distribution algorithm based on Clayton copula and empirical margins. In: Li K, Li X, Ma S, Irwin GW, editors. *Life System Modeling and Intelligent Computing*. Berlin, Heidelberg: Springer, 2010. pp. 82–88.
- [25] Salinas-Gutiérrez R, Hernández-Aguirre A, Villa-Diharce ER. Using copulas in estimation of distribution algorithms. In: Aguirre AH, Borja RM, García CAR, editors. *MICAI 2009: Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer, 2009. pp. 658–668.
- [26] Gao Y, Peng L, Li F, Liu M, Hu X. EDA-Based Multi-objective Optimization Using Preference Order Ranking and Multivariate Gaussian Copula. In: Guo C, Hou ZG, Zeng Z, editors. *Advances in Neural Networks*. Berlin, Heidelberg: Springer, 2013. pp. 341–350.
- [27] Gao Y, Peng L, Li F, Liu M, Hu X. Multiobjective Estimation of Distribution Algorithms Using Multivariate Archimedean Copulas and Average Ranking. In: Wen Z, Li T, editors. *Foundations of Intelligent Systems*. Berlin, Heidelberg Springer, 2014. pp. 591–601.
- [28] Wang LF, Zeng JC, Hong Y. Estimation of distribution algorithm based on archimedean copulas. In: *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*. ACM; June 12–14 2009; Shanghai, China. New York, NY, USA :ACM. pp. 993–996.
- [29] Salinas-Gutiérrez R, Hernández-Aguirre A, Villa-Diharce ER. D-vine EDA: a new estimation of distribution algorithm based on regular vines. In: *Proceedings of the 12th annual conference on Genetic and evolutionary computation*; 7–11 July 2010; Portland, Oregon. New York, NY, USA :ACM. pp. 359–366.
- [30] Gao Y. Multivariate estimation of distribution algorithm with laplace transform archimedean copula. In: *Information Engineering and Computer Science*, 2009. ICIECS 2009. International Conference on. IEEE; 2009. pp. 1–5.
- [31] Wang L, Wang Y, Zeng J, Hong Y. An estimation of distribution algorithm based on clayton copula and empirical margins. In: *Life System Modeling and Intelligent Computing*. Springer; 2010. pp. 82–88.
- [32] Wang X, Gao H, Zeng J. Estimation of Distribution Algorithms Based on Two Copula Selection Methods. *Int J Comput Sci Math*. 2012 Jan; 3: 317–331.
- [33] Chang C, Wang L. A multi-population parallel estimation of distribution algorithms based on Clayton and Gumbel copulas. In: Deng H, Miao D, Lei J, Wang FL, editors. *Artificial Intelligence and Computational Intelligence*. Berlin, Heidelberg: Springer, 2011. pp. 634–643.
- [34] Wang L, Guo X, Zeng J, Hong Y. Using gumbel copula and empirical marginal distribution in estimation of distribution algorithm. In: *Advanced Computational Intelligence (IWACI)*, 2010 Third International Workshop on; 25–27 Aug 2010; Suzhou, Jiangsu IEEE. 2010. pp. 583–587.
- [35] Zhang, Q., Zhou, A., Zhao, S., Suganthan, P.N., Liu, W., and Tiwari, S.: ‘Multiobjective optimization test instances for the CEC 2009 special session and competition’, University of Essex, Colchester, UK and Nanyang technological University, Singapore, special session on performance assessment of multi-objective optimization algorithms, technical report, 2008, pp. 1–30
- [36] Mashwani, W.K., and Salhi, A.: A decomposition-based hybrid multiobjective evolutionary algorithm with dynamic resource allocation, *Applied Soft Computing*, 2012, 12, (9), pp. 2765–2780
- [37] Mashwani, W.K.: ‘Hybrid Multiobjective Evolutionary Algorithms: A Survey of the State-of-the-art’, *International Journal of Computer Science Issues (IJCSI)*, 2011, 8, (6), pp. 374–392
- [38] Mashwani, W.K.: ‘Comprehensive Survey of the Hybrid Evolutionary Algorithms’, *Int. J. Appl. Evol. Comput.*, 2013, 4, (2), pp. 1–19
- [39] Mashwani, W.K., and Salhi, A.: ‘Multiobjective memetic algorithm based on decomposition’, *Applied Soft Computing*, 2014, 21, pp. 221–243

Vitality Aware Cluster Head Election to Alleviate the Wireless Sensor Network for Long Time

P. Thiruvannamalai Sivasankar

Research Scholar

Department of Computer Science and Engineering
Sathyabama University, India

Dr. M. RamaKrishnan

Chairperson and Professor
School of Information Technology
Madurai Kamaraj University
India

Abstract—The Wireless Sensor Networks (WSN) motivated by its unique characters such as it is capable of enduring callous ecological circumstances, and grant better scalability. The wireless sensor network is composed of insignificant sensors and a base station. The battery supplies the energy for the sensors. Hence, the lifetime of the network gets tainted while overworking for transmission. Since the WSN is being utilized for the dangerous purpose, we have to swell the lifespan of the network. The clustering is one of the foremost mechanisms to maximize the network's lifespan. The cluster head assortment plays an imperative role given the fact that clusters head was answerable for the transformation of data between cluster member and the base station. This present article deals with the novel scheme for the cluster head selection entitled as vitality aware cluster head election. In this scheme, the sensor nodes are being clustered into an optimal number. Subsequently, the cluster head is selected by a ballot for each and every group based on its remaining energy. To weigh up the performance of the proposed method, a Network Simulator (NS-2) has been employed.

Keywords—Wireless Sensor Networks(WSNs); Residual energy; Clustering; Life span; Sensor

I. INTRODUCTION

Tiny sensor nodes and a base station are the principal components of a wireless sensor network. The resource is being bounded by small sensor nodes, capable of sensing the environmental circumstances like pressure, heat, and dampness. The action course of WSN consists of: sensing the environment; sending out the gathered information to the base station and hence processing the information gathered. If the sensor nodes are being disseminated in the sensing meadow, only some of the nodes are in the vicinity of the base station while others may be distant. The energy utilization of the sensor node occurs only during sensing and transmitting it to the base station. Sensor nodes' lifespan is of great significance due to their applications in critical areas.

Numerous schemes were proposed to maximize the lifespan of the network. The clustering is one such scheme exploited to increase the lifespan of the network. The spatially dispersed nodes collectively cluster in a manner that every cluster has a head node (known as the cluster head) which is closer to the base station and its members. Because the energy used for transmission is directly proportional to the distance, lifetime of the network gets increased [1–8].

The paper is structured as follows where Section 1 provides an introduction to the wireless sensor network. Section 2 describes previous studies related to clustering, sensor, and WSN. Section 3 deliberates about the proposed method. Section 4 elaborates the results that obtained through proposed method, and finally the paper concludes with the conclusion and future direction.

II. RELATED WORK

Extensive literature survey reveals that numerous schemes were employed for this purpose. For instance, Bai et al. [9] proposed the energy efficient clustering mechanism of LEACH (Low Energy Adaptive Clustering Hierarchy). It is used to shun the contention and diminish the traffic load in the channel. In LEACH, the nodes form the clusters locally, and each and every cluster has one cluster head. The LEACH has two phases namely, (i) a setup stage and (ii) a steady state stage. During the process of a setup stage, nodes of a cluster decide the presence of a cluster head in the cycle. The cycle in WSN points towards the fact that every cluster head gathers information from all their members and relays it to the base station [10]. The choice of cluster head selection is based on the random likelihood of a node to become a cluster head. Every cluster head receives the gathered information sent by sensor nodes of a cluster followed by each cluster head's relay of the collected data to a base station in steady state stages. Election of

$$T(n) = \begin{cases} \frac{p}{1 - p(r \bmod (1/p))}, & \text{if } (n \in G) \\ \\ \text{else} & T(n) = 0 \end{cases}$$

Liang [11] offered the fuzzy logic based cluster head election method in which the base station picks the cluster head. There are three descriptors used in that method- energy, attentiveness, and centrality. These descriptors help calculate the likelihood for each sensor node to be a cluster head. In the base station, a central control algorithm provides the total information about the network using a selected cluster head.

Maraiya et al. [12] proposed an efficient cluster head selection method where the primary focus was on avoiding re-clustering, that minimizes the utilization of energy for

transmission purpose by attempting to shun overload in the cluster head. In this well-organized cluster head selection scheme, the cluster head is chosen with respect to the remaining energy in the node. Novelty of the present work is that, an associate cluster head is selected, and it becomes a cluster head while the leftover cluster head's energy drops below the energy of the other non-CHs in the cluster. Therefore, this scheme avoids re-clustering mechanism.

Lakshmi and Neelima [13] suggested that the importance of cluster head. The authors emphasized that selection based on the hit sets where the lifespan of the network is augmented by dipping the number of active nodes contributing to the transmission. The following are two chief ways to enhance the extent of the network's lifetime: optimizing the communication and reducing the energy usage. An efficient way to optimize the communication is to elect the cluster head efficiently. The hit set identifies the active nodes, and one of the same becomes the cluster head by its node degree. Similarly, Kumar et al. [14] and Kumar and Prabha [15] proposed that the location based clustering mechanism. The sensor that is near the base station is being chosen as cluster head.

Findings of these studies indicated that the clustering is the efficient way to enhance the network's lifespan, and the cluster head selection plays a crucial role. In this paper, the vitality aware cluster head selection mechanism was used to pick out an appropriate cluster head to reduce selection overhead of the cluster head. The choice of the cluster head is based on its residual energy and the probability of being a cluster head. The performance of the anticipated method was being proved by the analysis of the simulation results obtained from NS2 [16].

III. PROPOSED WORK

The wireless sensor network finds several usages which include disaster recovery, military application, and fire detection system and temperature monitoring system. Tiny sensors and a base station couple together to constitute a wireless network. The sensor runs on the battery power. Therefore, the lifetime of the network gets degraded while overloading the system.

As the WSN has been used widely in the critical application, there is a need to increase the network's lifetime. The clustering is the primary technique for enhancing the longevity of the network. The choice of a cluster head plays a vital part as every cluster head is accountable on the statement of data sensed by its members to its base station. Figure.1 illustrates clearly the method adopted in the present study. In our proposed method, all sensor nodes present in the network group together as a cluster into a best possible number, followed by the cluster head election for each cluster.

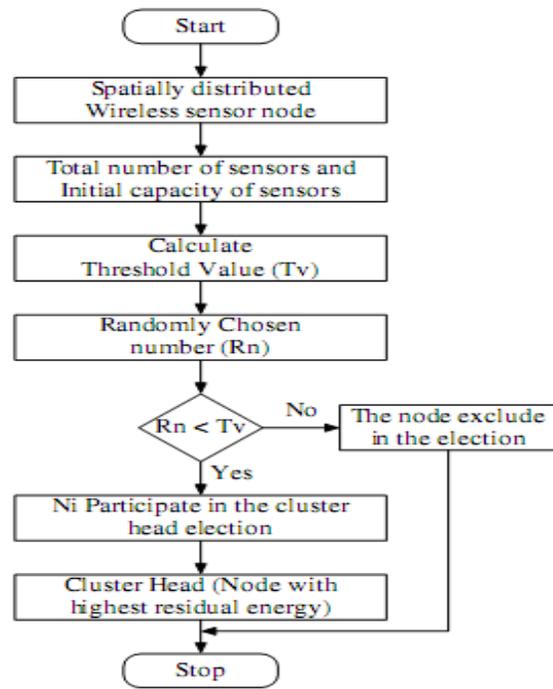


Fig. 1. Block diagram of Vitality aware cluster head election method

To frame the network, the optimal number of clusters required is guesstimated by using the formula given below

$$N_{cluster} = \frac{\text{Total number of sensors}}{\text{Initial capability of a sensor}}$$

The present research work considers the distribution of sensors in a circular region with the base station at the center.

A. Cluster formation

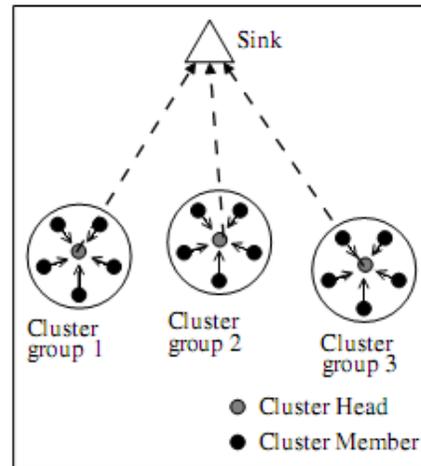


Fig. 2. Cluster Head to Base Station

The nodes are being grouped into an appropriate number of clusters with regard to its base station. The circular region is divided equally by using the following formula:

$$Degree_{partition} = 360 / N_{cluster}$$

The circular region is being partitioned into equal spaces with the angle between each region $Degree_{partition}$. The optimal number of clusters is formed based on the geographical area.

1) Cluster Head Selection Algorithm

Step1: Cluster head election

for ($j \leftarrow 0$ to k) do $j \leftarrow j + 1$

Where k - means number of cluster groups

{
for ($i \leftarrow 0$ to n) do $i \leftarrow i + 1$

Where n - means the number of cluster members per cluster group

{
Finding the cluster head in the clusters was based on the high initial energy through the comparisons between the clusters of the initial energy

if ($C_M[0] < C_M[i]$)

{
 $C_M[0] = C_M[i];$

}
}
 $C_H[j] = C_M[0];$ // Selected highest initial energy as a Cluster Head

Step 2: Cluster head formation

for ($j \leftarrow 0$ to m) do $j \leftarrow j + 1$

Where m - Number of Cluster heads

{
for ($i \leftarrow 0$ to n) do $i \leftarrow i + 1$

Where n - Number of cluster members

{
 $C_H = ClusterHead$

$C_M = ClusterMember$

$C_H[j] \leftarrow C_M[i]$
Information sent from Cluster member to Cluster Head

}
}
Step 3: Energy calculation at the time of Information sent from $C_M \rightarrow C_H$

$$C_M[AE] = C_M[IE] - C_M[EU]$$

Where AE-Available energy, IE-Initial energy, and EU - Energy used

Step3: Energy calculation at the time of Received information sent from C_H to Base Station

for ($j \leftarrow 0$ to m) do $j \leftarrow j + 1$

{
 $BS \leftarrow C_H[j]$

$$C_H[AE] = C_H[IE] - C_H[EU]$$

Where BS -Base Station AE-Available energy, IE-Initial energy and EU -Energy used

}
Step 4: Total Energy used

for ($i \leftarrow 0$ to n) do $i \leftarrow i + 1$

begin
 $C_M[T_{EU}] = T_{CM}[EU] - T_{CH}[EU]$

Where T_{EU} -Total Energy Used

// T_{CH} -Total cluster member

// T_{CH} -Total cluster heads

end.

2) Distance Estimation from the cluster member to cluster head

Distance matrix (DM) form is given as follows

$$DM = \begin{pmatrix} d_{CH_1, a_1} & \dots & d_{CH_n, a_n} \\ \vdots & & \vdots \\ d_{CH_m, a_m} & \dots & d_{CH_m, a_n} \end{pmatrix}$$

Distance Estimation Algorithm

Step1: Initialize the cluster member n in each cluster within a circular region.

cluster $k = 1$

Step2: Select the cluster head based on the highest residual energy within cluster members in each cluster group. Assume the number of cluster head is m in a circular region

Step 3: The distance of a cluster member to its base station is determined through its cluster heads

for ($i = 0; i < m; i ++$)

Where m -means the number of cluster heads

{
K: = 1;

{
for ($j = 0; j \leq n; j ++$)

// Where n is the number of members in each cluster

```
{
d1(0,0) = 0;
di(j,BS) = dij(CH[i],j) + dij(BS,CH[j]);
}
K=K+1;
}
```

B. Energy Consumption Model

Heinzelman et al. [17, 18] applied the Energy Consumption Model as shown in the Figure 3.

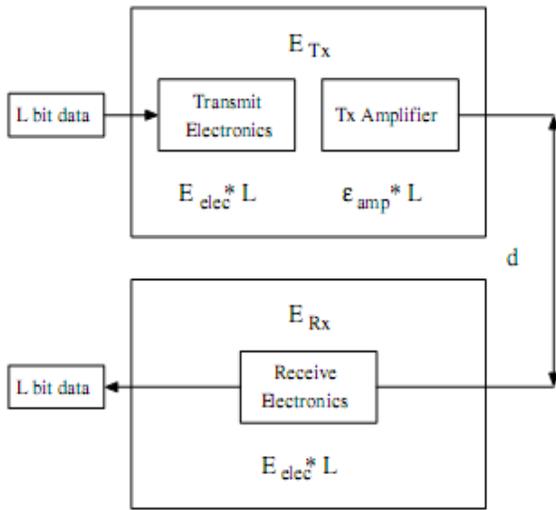


Fig. 3. Energy Consumption Model

The Energy consumption Model in Wireless Sensor Network consists of three parts. The parts are as follows:

1) The sending l -bit message at distance d requires energy.

$$E_{Tx}(l, d) = \begin{cases} l * E_{ele} + l * \epsilon_{fs} * d^2, & d \leq d_0 \\ l * E_{ele} + l * \epsilon_{mp} * d^4, & d \geq d_0 \end{cases}$$

Where E_{ele} is circuit fatigue of sender and receiver.

d_0 is the critical distance of multipath fading model and free space model

ϵ_{fs} is the amplifier coefficient of free space model

ϵ_{mp} is the multipath fading model

2) Receiving l -bit message at a distance d , requires energy

$$E_{RX}(l) = l * E_{ele}$$

3) Data aggregation of energy is

$$E_{DA} = l * E_{da}$$

C. Vitality aware cluster head election

After grouping the sensors into clusters, each and every node in the cluster was selected at a random number between zero and one. In case, the random number is lower than the threshold value (n), each node participates in the process of election otherwise it excludes itself from the election process. The threshold value is ascertained by using the formula given below

$$Threshold(n) = \begin{cases} p(E_{Current}/E_{Average}) & \text{if } n \in C_i \\ 0 & \text{Otherwise} \end{cases}$$

Where p is the node as n is the probability of being cluster head. $E_{Current}$ Denotes the current energy of a node n . $E_{Average}$ Indicates that the average energy of nodes in the cluster C_i .

If a cluster consists of several nodes, a particular node is chosen in a manner that, that node becomes a cluster head. Otherwise the node that is very close to the base station with high residual energy is elected as a cluster head.

For each and every round, the cluster head is reelected. The elected cluster head, gathers the sensed information from all of its members in the cluster. Then the cluster head transmits the collected information to the base station. As the cluster head is selected based on its energy level distance to the base station and its probability of being the cluster head, the proposed method increases the existence of the network and also reduces the clustering overhead which takes place in the wireless sensor network.

IV. SIMULATION RESULTS

The efficiency of the adopted method was analyzed by using a discrete event time driven Network Simulator (NS2). Each and every function of the network is called the event. The programmer explicitly gives the time at which the event should occur. This simulator displays the result in an animated format using the Network animator. The network simulator traces all the network events dynamically to a trace file by which the graph plotted. The following graphs obtained from the simulation of the vitality aware cluster head election scheme (VACS) are used to grade the performance of the network. The parameters like packet delivery rate, packet loss rate, and residual energy has been plotted.

The packet delivery rate was calculated by using the following formula,

$$= \frac{\text{Packet delivery rate}}{\text{Time}} = \frac{\text{No. of packets sent} - \text{No. of packets dropped}}{\text{Time}}$$

Figure 4 shows the graph of the packet delivery rate.

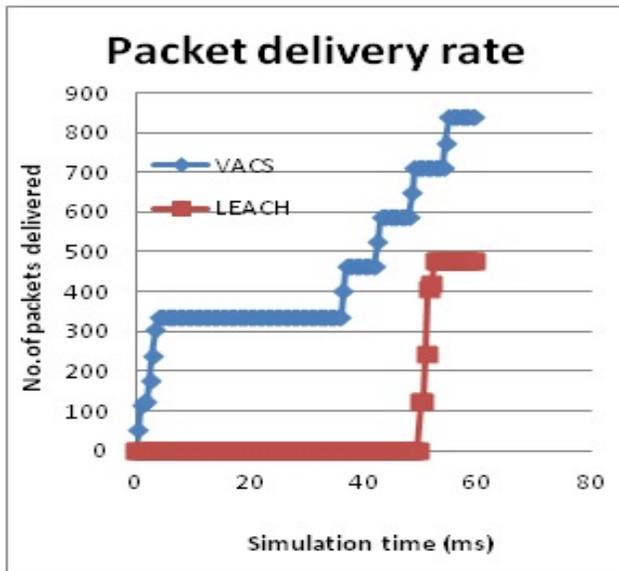


Fig. 4. Packet delivery rate analysis

The proposed system is evaluated by comparing its performance with the standard clustering scheme LEACH. The graph shows that the packet delivery rate of the proposed method is 35% higher than LEACH clustering scheme. Therefore, the throughput of the system is increased.

The packet loss rate of the system was calculated using the formula

$$\text{Packet loss rate} = \frac{\text{No. of packets sent} - \text{No. of packets received}}{\text{Time}}$$

The packet loss occurs due to the destination being out of range and the packet arrival rate being higher than the queue size.

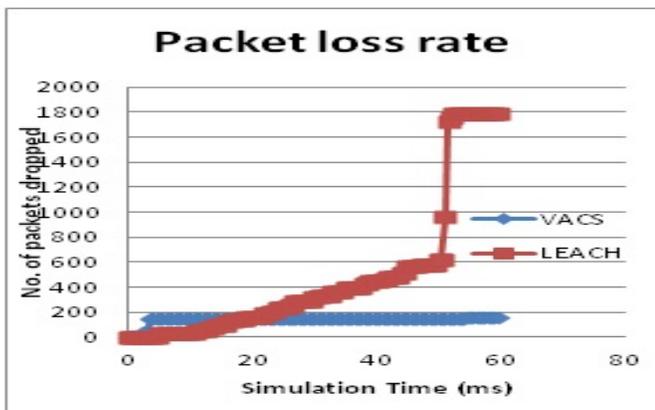


Fig. 5. Packet loss ratio analysis

The graph of the packet loss rate in Figure 5 shows that the proposed system supersedes the standard clustering mechanism LEACH. The residual energy was used to calculate the lifetime of the network. The residual energy was the remaining energy in a node after some events occur. The following formula was used to calculate the Residual energy:

$$\text{Residual energy} = E_{\text{initial}} - E_{\text{Transmission}}$$

Where,

E_{initial} denotes the initial energy

$E_{\text{Transmission}}$ denotes the energy used for transmission

The mean value of residual energy in every node of a network was calculated for each 0.5 ms until the simulation ends.

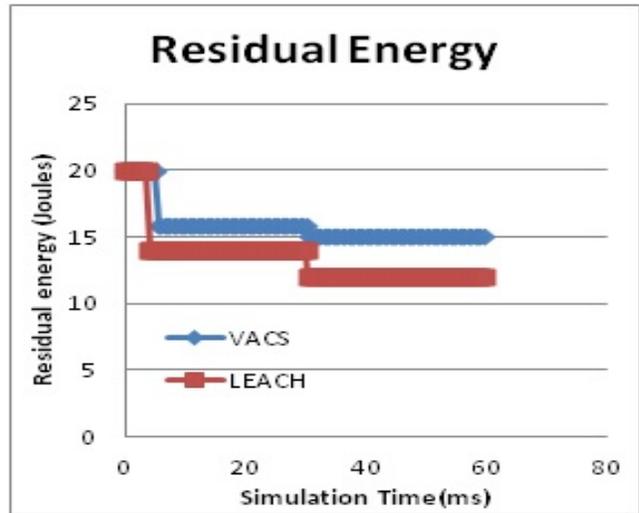


Fig. 6. Residual energy of the sensors after one round

The calculated energy has plotted as shown in Figure 6 and compared to the LEACH clustering mechanism. The findings showed that the proposed scheme outperforms than the LEACH.

V. CONCLUSION

This paper illustrated the expectation of a novel cluster based head election method. The method considered the likelihood value of a node being the cluster head to formulate the sensor node and to contribute to the election process. Remaining energy and the remoteness of the base station are used during the process of election head, which thereby prevents the participation of sensor nodes. Therefore, this method shuns the disqualified node to participate in the election process and subsequently, increases the lifespan of the network and reduces the cluster head selection overhead.

REFERENCES

- [1] I.F. Akyildiz, Weilian Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," IEEE Commun. Mag., vol. 40, pp. 102–114 2002.
- [2] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey," Computer Networks, 38(4): 393-422, 2002.
- [3] .K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," Ad Hoc Networks, vol. 3, no. 3, pp. 325–349, 2005..
- [4] C.Y. Chong, S.P. Kumar, B.A. Hamilton "Sensor Networks: Evolution, Opportunities, and Challenges," Proceedings of IEEE, 91(8):1247-1256, 2003.
- [5] B.K. Debroy, M.S. Sadi and M. Al Imran, "An Efficient Approach to Select Cluster Head in Wireless Sensor Networks," J. Commun., vol. 6, pp. 529–539, 2011.

- [6] A.A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Comput. Commun.*, 30, pp. 2826–2841, 2007.
- [7] J.-Y. Chang and P.-H. Ju, "An efficient cluster-based power saving scheme for wireless sensor networks," *EURASIP J. Wirel. Commun. Netw.*, vo. 2012, pp. 172, 2012.
- [8] B. Singh and D. Lobiyal, "A novel energy-aware cluster head selection based on particle swarm optimization for wireless sensor networks," *Human-centric Comput. Inf. Sci.*, vol. 2, pp. 13, 2012.
- [9] F.e. Bai, H.h. Mou and J. Sun, "Power-efficient zoning clustering algorithm for wireless sensor networks," In: *International Conference on Information Engineering and Computer Science (ICIECS 2009)* pp. 1–4, 2009.
- [10] H. Zhang, S. Zhang and W. Bu, "A Clustering Routing Protocol for Energy Balance of Wireless Sensor Network based on Simulated Annealing and Genetic Algorithm," *Int. J. Hybrid Inf. Technol.*, vol. 7, pp. 71–82, 2014.
- [11] Q. Liang, "Cluster head election for mobile ad hoc wireless network," In: *Proc. 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC)*, pp. 1623 –16, 2003.
- [12] K. Maraiya, K. Kant and N. Gupta, "Efficient Cluster Head Selection Scheme for Data Aggregation in Wireless Sensor Network," *Int. J. Comput. Appl.* (0975 – 8887). vol. 23, pp. 10–18, 2011.
- [13] B.J. Lakshmi and M. Neelima, "Maximising Wireless Sensor Network lifetime through cluster head selection using Hit sets," *IJCSI Int. J. Comput. Sci.*, issue 9, pp. 328–331, 2012.
- [14] A. Kumar, N. Chand and V. Kumar, "Location Based Clustering in Wireless Sensor Networks," *World Acad. Sci. Eng. Technol.*, vol. 5, pp. 1313–1320, 2011.
- [15] N. Kumar and V.S. Prabha, "Comparative analysis of energy-efficient cluster-based routing protocols for wireless sensor networks," *Sensors and Transducers*, vol. 142, pp. 23–32, 2012.
- [16] J. Ferdous, M.J. Ferdous and T. Dey, "A Comprehensive Analysis of CBCDACP in Wireless Sensor Networks," *J. Commun.*, vol. 5, pp. 627–636, 2010.
- [17] W.B. Heinzelman, A.P. Chandrakasan and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wirel. Commun.*, vol. 1, pp. 660–670, 2002.
- [18] W.R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, vol 2, pp. 10, IEEE Comput. Soc., 2000.

Designing an IMS-LD Model for Collaborative Learning

Fauzi El Moudden/ Ph.D student
in computer sciences
Abdelmalek Essaâdi University
Faculty of Sciences, LIROSA
Tetouan, MOROCCO

Prof. Mohamed Khaldi
Abdelmalek Essaâdi University
Faculty of Sciences, LIROSA
Tetouan, MOROCCO

Prof. Aammou Souhaib
Abdelmalek Essaâdi University
Faculty of Sciences, LIROSA
Tetouan, MOROCCO

Abstract—The context of this work is that of designing an IMS-LD model for collaborative learning. Our work is specifically in the field or seeking to promote, by means of information technology from a distance, a collective knowledge construction. Our approach is to first think about the conditions for creating a real collective activities between learners, and designing the IT environment that supports these activities. We chose to use the pedagogy project as a basis for teaching these collective activities. This pedagogy has already proven itself, mostly in traditional learning situations in the classroom.

Keywords—*Collaborative Learning; Pedagogy Project; Socio-constructivist; IMS-LD*

I. INTRODUCTION

In this work, we try to present teachers with a tool for easy generation and management of collaborative educational content online. This tool allows the generation and editing of websites structures through a base educational models rich enough with a variety of choices ensures a better adaptation of the course to pedagogy and learning style. Otherwise, the social constructivism approach is to focus the activity on the learner to support synchronous and asynchronous collaboration, it is therefore necessary to find a method to model all types of activities. To model the activities we have based on IMS LD specification based on collaborative learning online.

First, we'll start the first section by defining some basic concepts such as collaborative learning, socio-constructivist approach, the project-based learning, collaborative learning, IMS-LD then we will address the modeling section in which we present our computer model design.

II. CONCEPTS AND RELATED WORK

A. Collaborative Learning

According to Cuseo [1], cooperative learning is an educational method in which small groups of 3 to 5 learners, made intentionally, working inter-depending on a well-defined and structured task. Learners are responsible for their performance and the teacher is a facilitator, a consultant in the learning process of the group. The group is formed according to educational criteria (such as the heterogeneity of learners' levels); the roles of learners should be assigned so as to be interdependent. The intention to develop social skills is clearly explained in this approach.

The term "collaborative learning" seems to have an English origin, based on the work of teachers who explored how learners could take a more active role in their own learning. [2]. Learners are assumed to be responsible and have social skills. Panitz [2] see collaborative learning as a personal philosophy and not just as a class technique. Learners are responsible for their learning and that of others [3].

Overall, collaborative learning is an approach that gives a lot of freedom to the learner. The activities are not very directed and learners manage their workgroup largely. For example, learners roles are not assigned by the teacher in the case of collaborative learning, but learners negotiate these roles together.

B. Socio-constructivist approach

Although Piaget's theory [4] focuses primarily on the individual aspects in cognitive development, she strongly inspired a group of psychologists - named "Geneva School" - which began in the 1970s a research to know how social interactions affect individual cognitive development [6]. This new approach, highlighting the role of human interaction in learning, is described as socio-constructivist. The role of the interaction in the mental development is explained by the researchers by structuring interaction and processes generated by these interactions called "socio-cognitive conflict". This conflict leads the learner to reorganize its previous designs and incorporate new elements of the situation. The socio-cognitive conflict resulting from the confrontation of representations about a subject from different individuals interact. This reorganization of representations from two types of imbalance: the inter-individual when there is opposition between two subjects; intra-individual, when a subject questions his own representations. An opposition between two subjects during situations of social interaction, allows to generate a socio-cognitive conflict whose resolution will generate a cognitive progress. Learning is, therefore, stimulated by socio-cognitive conflicts, knowledge is developed when learners reconsider their own views through negotiation and argument phenomena. This work helped to highlight the link between the cognitive and the social, stressing the importance of dialogue and shared experiences in the construction of knowledge.

C. Pedagogy Project

1) For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

2) Approach and specifications

The pedagogy project is part of what is conventionally called active pedagogies; it refers to a learning model that we will characterize, and according to LAFORTUNE [13] is related to cognitive models, constructivist, and socio-cognitivist.

For William [6], a project is an activity that has a specific purpose, engages in full those who perform it and takes place in a social environment. This method advocates finding solutions to real problems that occur in everyday life.

Today, pedagogy project is a method commonly used by teachers. However, it has many variants and it is difficult to provide a single definition. We retain this method in that learners work together in small teams from a specification to actual production. This work requires authentic project management (task management, time management). We retain that unlike problem solving, there is no single solution to the project and that it takes place over a longer period.

3) Characterization of the pedagogy project

The literature highlights a number of characteristics of the project pedagogies, and their positive effects on learning. A project must put the learner in a situation that is a challenge [7]; it is initiated from a concrete theme of life [8]; it should move towards a concrete and evaluable output ([9]; [10]; [7]).

The assessment consists of several stages that take place at different stages of project process and not just at the final stage ([10]; [11]).

Project-based learning:

- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".
- Develops knowledge, management skills, interpersonal skills and knowledge to be taking place in action, know-becoming ([12]).
- Develops also transversal competences ([13]; [10]).
- Strengthens the autonomy and responsibility.
- Develops self-confidence in learners ([12] ;[10])
- Promotes teamwork ([9]).
- Allows implementation of interdisciplinarity ([7]).

4) The phases of a project

As has been said, a project takes place in the following time in several phases, which can be described as follows.

a) Preparation phase: The teacher must be able to offer a wide range of project topics, for which clarifies and explains his pedagogical intentions. It should also allow a breakdown of these subjects among learners in a democratic process [16]. Finally he gives each project a synoptic blank. Learners perform needs analysis, feasibility study and establish the specifications [11].

b) A project structuring phase: A project must first be broken down into stages, and then cut each step into tasks [7]. These should then be planned ([10]; [11]).

c) Project implementation phase: To implement a project, is to launch it in reality. It is a long period for which it is necessary to confront the tasks to the outlook reference schedule. About that, we refers to deviations of management.

d) A project evaluation phase : The assessment consists of several moments that take place at different phases of project process and not just the terminal phase. The assessment consists of several moments that take place at different phases of project process and not just the terminally ill. According to FORREST [14], formative assessment is frequent and immediate during the project. Summative assessment is carried out through the evaluation of final products.

D. Work objective

The pedagogy project is to make learners work in small teams on common projects to achieve collective production. The first goal of our research is to cover the modeling of a system of learning in a collaborative online context.

To create a collaborative learning environment online, you must prepare the environment to give learners the ability to view projects, tasks, create and participate in discussions with his group to share ideas and improve the way of thinking.

E. Online collaborative learning

The online collaborative learning was experienced at the onset of online education in the late 1980s under the name "computer conferencing" email first, then by forums. As online learning, collaborative learning benefitted learner's great flexibility of time and place (stimulating autonomy and reflection) and an excellent asynchronous interaction (source of motivation, support, critical thinking, synthesis..) Therefore [16] reported in 1989 that "The collective nature of computer conferencing may be the single most fundamental and critical underpinning the development of theories as well as the design and implementation of educational activities online."

In this perspective, the online collaborative learning is the most important educational contribution of online education. And irrefutable logic of [17], provide online education without benefit learners who follow the advantages of its "most fundamental" is absurd and devalues the remarkable educational tool that is telematics. This does not mean that online education should be limited to collaborative learning online! But it is important that any online program includes a minimum of collaborative learning and operates an appropriate extent and in a manner appropriate to the program and its students.

F. Related works

Much work has been done in the field of collaborative learning and IMS LD specification, we will be limited to 3 examples such as "**Implications of a cooperation model for the design of collaborative tools**" [18] that explains how models socio-cognitive interaction are related to the properties of collaborative tools, "**Modeling of collaborative learning scenarios**" [19] that expresses collaborative learning scenarios by teachers animating virtual classrooms to promote the re-use and share teaching practices. It proposes an approach led by the models in accordance with the recommendations of the Model

Driven Architecture OMG. It presents a meta-model based on IMS-LD enhanced by the concepts of participation model to capture the richness of the interactions inherent in collaborative activities, and "A system to advise the teachers of collaborative learning situations" [20] that targets the development of a system of assistance to the teachers in collaborative distance learning. This system is based on an ontology modeling the different components of tutoring (actors, their characteristics, activities, their parameters and resources.) A rule-based inference engine reasons about the ontology to infer advice to the tutor to help adapt learning situations to learners and learning groups by taking into account their behavior and interactions.

For us, this is not the same case and the same vision as our model is more general, it aims on one hand to create a system from which teachers can animate virtual groups for the re-use and sharing of teaching practices and on the other hand the re-use of the content created in other frameworks.

III. IMS-LD

IMS-LD was published in 2003 by the IMS / GLC. (Instructional Management Systems Global Learning Consortium: This specification allows representing and encoding learning structures for learners both alone and in groups, gathered by roles, such as "learners" and "Team"[21]. We can model a lesson plan in IMS-LD, defining roles, learning activities, services and many other elements and building learning units. The course outline is modeled and built with resources assembled in a compressed Zip folder and initiated by an executable ("player"). The latter coordinates the

teachers, students and activities as long as the respective learning process progresses. A user takes a "role" to play and perform activities related with it to achieve satisfactory learning unit. In all, the unit's structure, roles and activities build the learning scenario to be executed in a system compatible with IMS LD.

IMS-LD does not impose a particular pedagogical model but can be used with a large number of scenarios and pedagogic models, demonstrating its flexibility. Therefore IMS-LD is often called a meta-pedagogic model. Previous e-learning initiatives claim to be pedagogically neutral, IMS-LS is not intended to pedagogical neutrality but seeks to raise awareness of e-learning on the need for a flexible approach.

IMS-LD has been developed for e-learning and virtual classes, but a course face to face can be done and integrated into a structure created with this specification, as an activity of learning or support activity. If the ultimate goal to create rich learning units, with support to achieve the learning objectives by providing the best possible experience, face-to-face meetings, and any other learning resource are permitted such as video conferencing, collaborative table or any field action research.

IMS LD uses the theatrical metaphor, which implies the existence of roles, resources and learning scenario itself: one room is divided into one or more acts and conducted by several actors who can take on different roles at different times. Each role is to carry out a number of activities to complete the learning process. In addition, all roles must be synchronized at the end of each act before processing the next act.

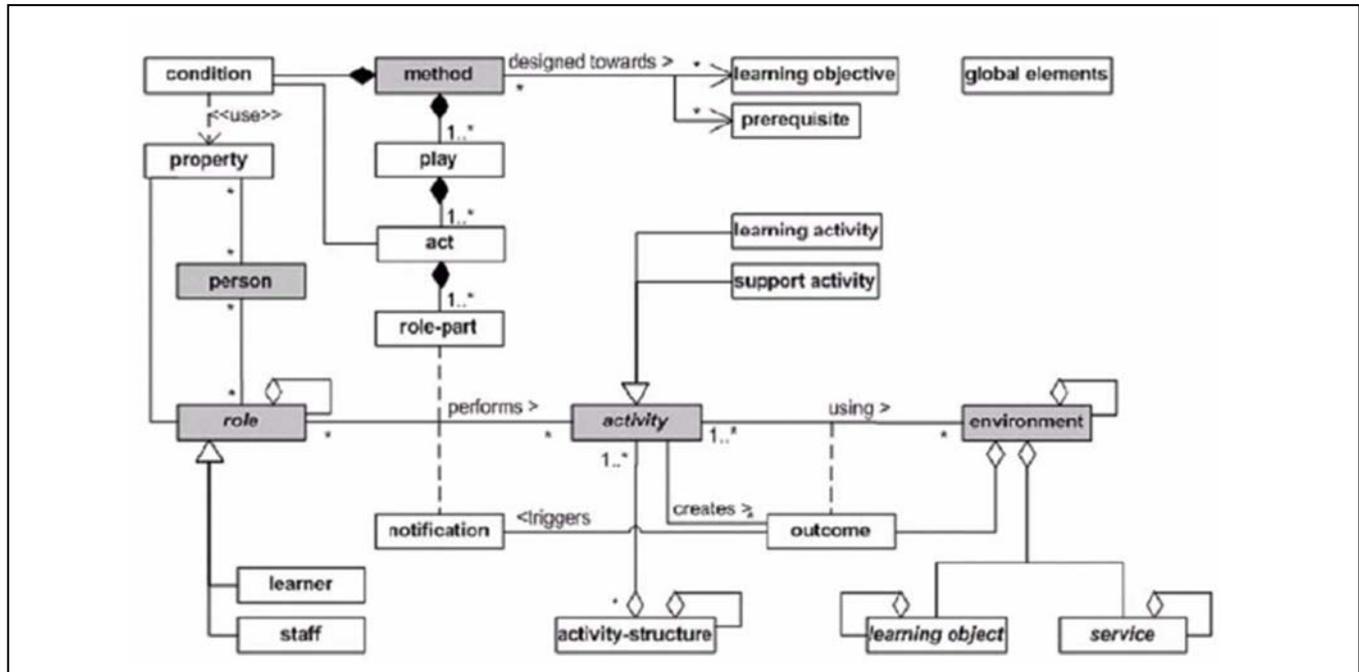


Fig. 1. The conceptual model of IMS LD [21]

IV. MODEL DESIGN

In our previous work [22] we tried to design a collaborative model of learning online beginning with the study of the IDM

approach (Model based Engineering) based on four stages of implementation:

TABLE I. TABLE SERVICE DESCRIPTION

Actor	Service function	Description
<i>Learners</i>	Consult the project	The learner can view the project and its objectives at any time
	Create discussions	The learner may at any time create discussions
	Browse calendar	The learner can browse the defined task calendar.
	Check notifications	The learner can always check for notifications.
	View documents	Learners read the downloaded documents.
	Download documents	The learner can download the documents
	Contact the teacher	The learner may contact the teacher
<i>Teacher</i>	Supervise the learners	The teacher adds, modifies or deletes his learners
	Manage groups	The teacher adds, modifies or deletes groups
	Assign students to groups	The teacher can assign learners to groups
	Create projects	The teacher can create projects
	Set objectives	The teacher can set objectives for projects
	Set phases	The teacher can set project phases
	Set tasks	The teacher can set tasks
	Assign tasks	Teachers can assign tasks to students
	Set calendar	The teacher can set schedules for the tasks and phases
	Upload documents	The teacher can upload documents
	Download documents	Teachers can download the documents
	Initiate discussions	The teacher can start discussions
	Create notifications	The teacher can create notifications
Assess productions	The teacher can assess the productions undertaken by learners.	
<i>Admin</i>	Manage teachers	The admin adds, modifies, or deletes a teacher
	Manage access rights	The admin can manage the access rights of teachers and learners.

- The development of a model without IT preoccupation (CIM: Computer Independent Model).
- Its manual transformation into a model in a particular technological environment (PIM: Platform Independent Model);
- The automatic transformation into a model associated with the target implementation platform (PSM: Platform Specific Model) model must be refined;
- Its implementation in the target platform.

In this section we will talk about the IT design of our collaborative model without using the same approach that we have adopted in previous works, because we have detected the real problems of semantic loss during the transformation of the model.

This led us to develop our model through the outline of the diagram in which we will eventually identify the features of the constituent entities of our model and also the classes diagram in which we will specify the different classes constituents our collaborative model.

A. Use case diagram

The use case diagrams identify the functionality provided by the model (use case); users interact with the system (actors), and the interactions between them.

The main objectives of the use case diagrams are:

- Provide high-level view of the model.
- Identify users ("actors") Model.
- Define the roles of the actors in the model.

Table 1 describes the service function for each actor:

Here are the use case diagrams of the model representing the external actors who will interact with the system and how they will use it:

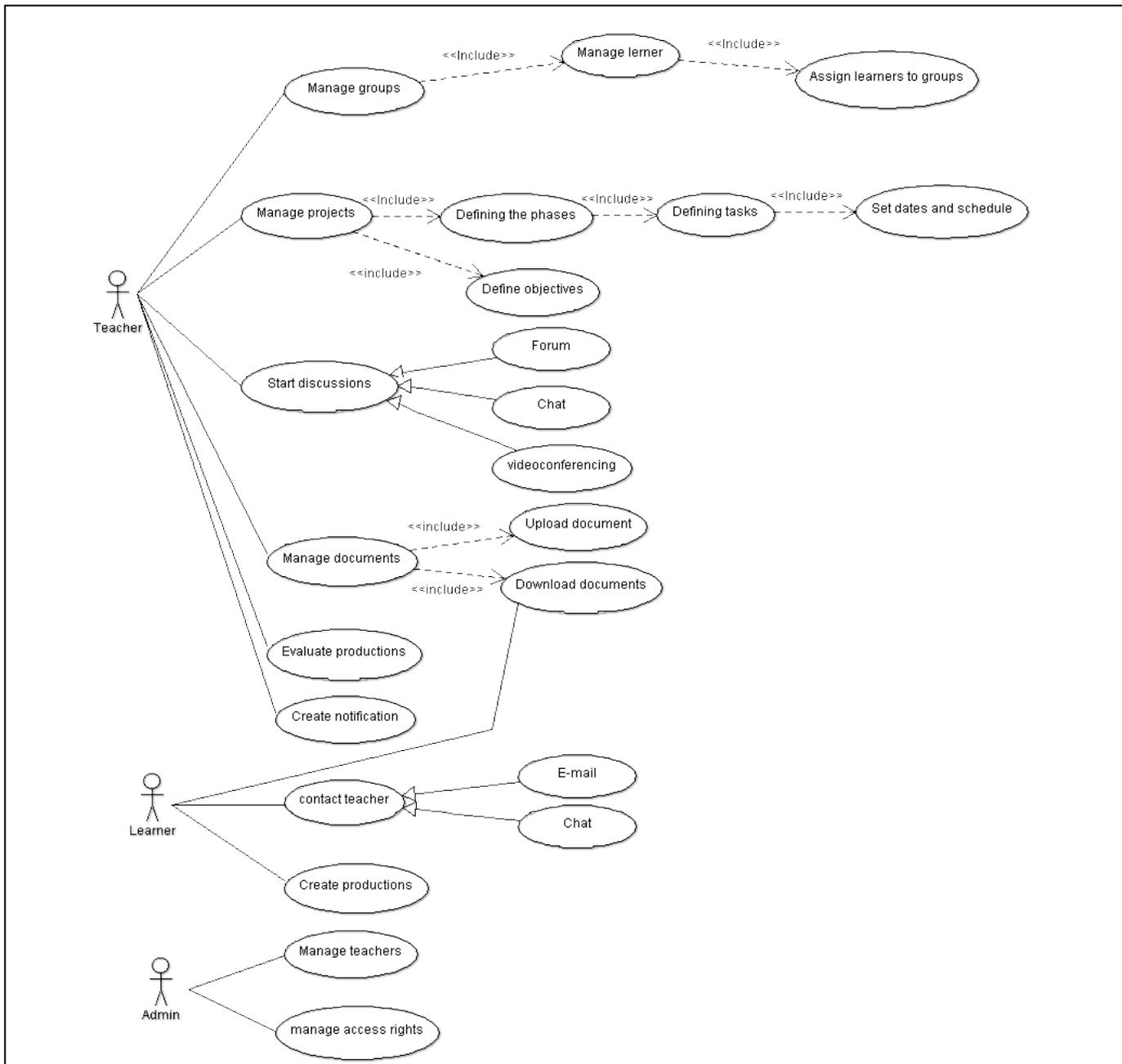


Fig. 2. Use Case Diagram

B. Class Diagram IMS-LD

In this part we will try to create a model based on the theoretical study of our current work and allows simultaneously to ensure its projection to the model, on top of that we will try to recognize our model with the IMS model - LD, this compatibility will not be a direct way c to d, one will use the same terminology but IMS-LD for all classes of our proposed collaborative model, there is an equivalent class in the IMS-LD, which will greatly help us in the implementation level, we present in the following table the different classes of our model and the IMS-LD model:

TABLE II. CORRESPONDENCE BETWEEN THE TERMINOLOGY OF IMS-LD AND THAT OF THE COLLABORATIVE MODEL

Collaborative Meta-Model	IMS-LD
Project	Activity
Task	Role
Subtask	Activity structure
Phase	Play
Members, Team	Person
Teacher	Staff
Learner	Learner
Production	Outcome
Notification	Notification
Objective	Learning Objective
Tools	Services

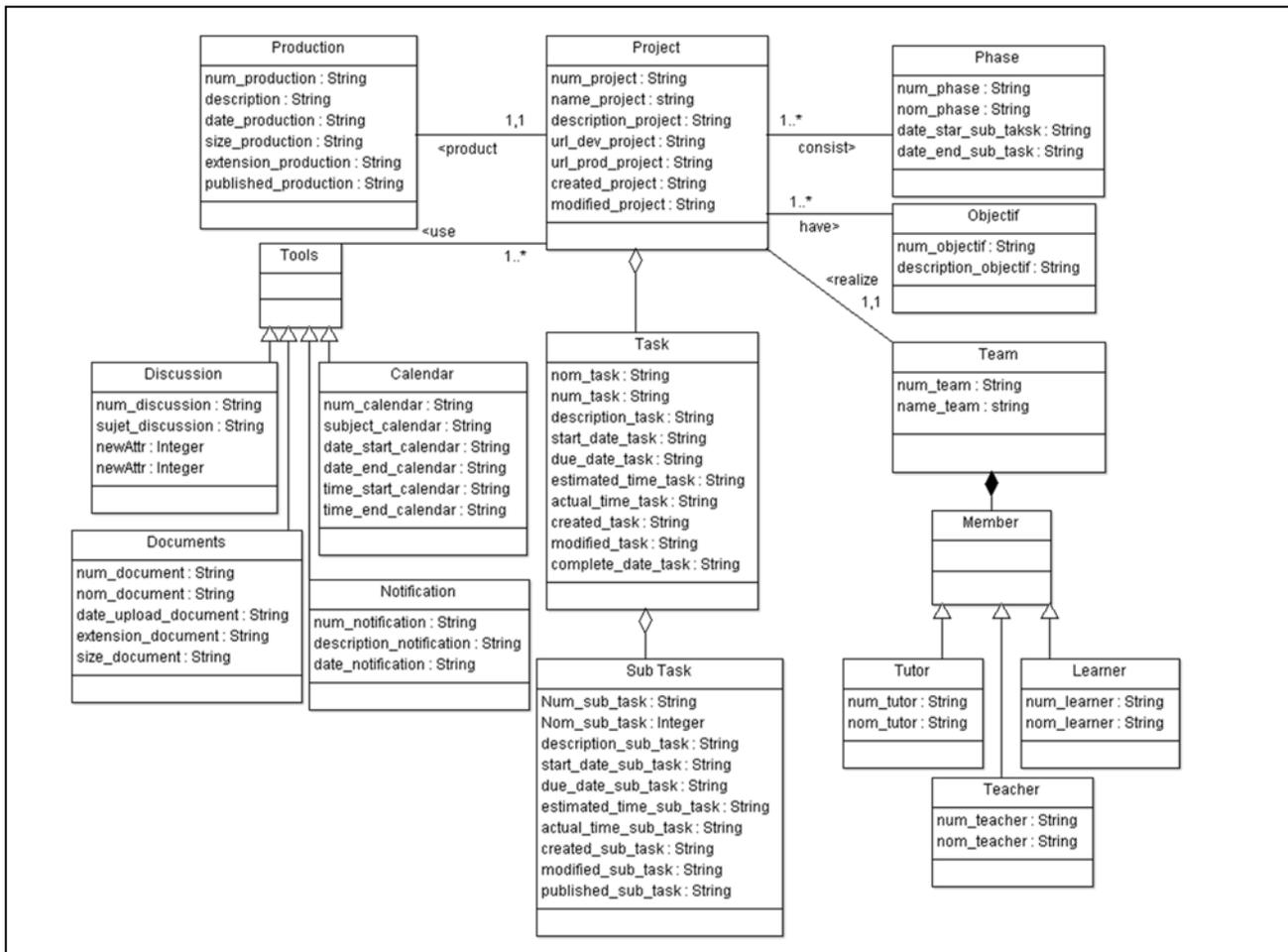


Fig. 3. Class Diagram

V. CONCLUSION AND PERSPECTIVES

In our work, we are on the way to the design and modeling of a collaborative online learning model compatible with IMS-LD. This design is based on active teaching learner-centered, and as an example of the pedagogy we opted for pedagogy project that allows us to reach a teaching object through the implementation of projects that are divided into tasks performed by students in collaboration.

To achieve this goal we need to reach the model validation step, which is one of the tasks to be performed in our future work, also we seek a relevant tool among the models validation tools that will guide us better to start the development part.

REFERENCES

[1] Joseph Cuseo, Marymount College, Cooperative Learning and College Teaching ,2.3 (1992): 5-10.
 [2] Panitz, T. (1997). Collaborative versus cooperative learning. A comparison of the two concepts which will help us understand the underlying nature of interactive learning. Cooperative Learning and College Teaching, 8(2). Recuperado de http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S0254-92472011000100001&lng=es&nrm=iso&tlng=es#
 [3] Gokhale 1995, Journal of Technology Education Volume 7, Number 1Fall 1995

[4] Jean Piaget, Psychologie et pédagogie (Folio essais, Paris, ed. 1991, 1935).
 [5] Doise et Mugny 1981 Willem Doise, G. Mugny, Le développement social de l'intelligence (InterEditions,Paris, 1981).
 [6] William Heart Kilpatrick, The Project Method, Teachers College Record, vol. 19 (2), p.319-335. (1918).
 [7] ABDALLAH, F. . Meta-modélisation pour décrire et instrumenter une situation d'apprentissage de pédagogie de projet collectif. Thèse de doctorat en informatique. Le Mans: Université du Maine, 2009.
 [8] LEBRUN, M. Théories et méthodes pédagogiques pour enseigner et apprendre: Quelle place pour les TIC dans l'éducation ? (2e éd.). Bruxelles: De Boeck Supérieur, 2007.
 [9] PERRENOUD, P. Apprendre à l'école à travers des projets : pourquoi ? Educateur, 14, 6-11, 2002.
 [10] William, J. L'apprentissage de projet. Sainte-Foy: PUQ, 2004.
 [11] TILMAN, F. Penser le projet: concepts et outils d'une pédagogie émancipatrice. Lyon: Chronique sociale, 2004.
 [12] HUBER, M. Apprendre en projets: la pédagogie du projet-apprenants (2e éd.). Lyon: Chronique Sociale, 2005.
 [13] LAFORTUNE, L. La pédagogie du projet et développement des compétences transversales : changement de posture pédagogique. Education Canada, 49(5), 16-20, 2009.
 [14] FORREST, C. (dir.) La pratique de la pédagogie de projet (2e éd.). Alma: Axone, 2011.

AUTHOR PROFILE

- [15] DUMAS, B. Les rôles de l'enseignant en pédagogie de projet. Québec français, 126, 64-66, 2002.
- [16] Harasim L., « Online education: a new domain », in Mason R. & al. (ed.), Mindweave, Oxford, Pergamon, 1989, p. 50-52
- [17] Salmon G., E-Moderating - The key to teaching and learning online, Londres, Kogan Page, 2000.
- [18] Gregory Dyke, Kristine Lund. Implications d'un modèle de coopération pour la conception d'outils collaboratifs. M.N. Lamy, F. Mangelot, E. Nissen. Echanger pour apprendre en ligne (EPAL), Jun 2007, Grenoble, France. Université de Grenoble, Consulté le 20/04/2009 à l'adresse: <http://w3.u-grenoble3.fr/epal/actes.html>, 2007.
- [19] Christine Ferraris, Anne Lejeune, Laurence Vignollet, Jean-Pierre David. Modélisation de scénarios pédagogiques collaboratifs. 2005.
- [20] Rencontres Jeunes Chercheurs en EIAH (RJC-EIAH 2006), May 2006, Evry, France. p. 99-106, 2006
- [21] IMS Learning Design www.epi.asso.fr/revue/articles/a0512c.htm
- [22] Fauzi El Moudden, Souhaib Aammou et Mohamed Khaldi. A TOOL TO GENERATE A COLLABORATIVE CONTENT COMPATIBLE WITH IMS-LD. International Journal of Software and Web Sciences, 11(1), December 2014-February 2015, pp. 01-08.

Mr. EL MOUDDEN Fauzi is a PhD candidate in Computer sciences, at the Laboratory of Informatics, Research Operational and Statistic Applied (LIROSA) at Faculty of Sciences, Abdelmalek Essaadi University. He has a Master degree in Instructional design Multimedia engineering at the The École normale supérieure of Martil, Morocco in 2013. His current research focuses on: E-learning, Collaborative Learning and Pedagogy.

Prof. KHALDI Mohamed is a professor at the The École normale supérieure at Abdelmalek Essaadi University, and he is with the Laboratory of Informatics, Research Operational and Statistic Applied (LIROSA) at Faculty of Sciences, Abdelmalek Essaadi University. Tétouan, Morocco.

Prof. AAMMOU Souhaib is a professor at the The École normale supérieure at Abdelmalek Essaadi University. Tétouan, Morocco. He received his PhD in computer science in 2013 within the Laboratory of Informatics, Research Operational and Statistic Applied (LIROSA) at Faculty of Sciences, Abdelmalek Essaadi University. He has a Graduate Diploma (DESA) in Applied Engineering and Technology Education and Training in 2007 at the University of Hassan II, Mohamadía. In research, his current interests include: Cognitive Science and Artificial Thinking, Ontology Engineering, Human-Computer Interaction and Technology Enhanced Learning.

An Enhanced Steganographic Model Based on DWT Combined with Encryption and Error Correction Techniques

Dr. Adwan Yasin¹

Computer Science Department
Arab American University
Jenin, Palestine

Mr. Nizar Shehab²

Computer Science Department
Arab American University
Jenin, Palestine

Dr. Muath Sabha³

Multi Media Department
Arab American University
Jenin, Palestine

Mariam Yasin⁴

Computer Science Department
Arab American University
Jenin, Palestine

Abstract—The problem of protecting information, modification, privacy and origin validation are very important issues and became the concern of many researchers. Handling these problems definitely is a big challenge and this is probably why so much attention directed to the development of information protection schemes. In this paper, we propose a robust model that combines and integrates steganographic techniques with encryption, and error detection and correction techniques in order to achieve secrecy, authentication and integrity. The idea of applying these techniques is based on decomposing the image into three separate color planes Red, Green and Blue and then depending on the encryption key we divide the image into N blocks. By applying DWT on each block independently, this model enables hiding the information in the image in an unpredictable manner. The part of the image where information embedded is a key dependent and unknown to the intruder and by this we achieve blinded DWT effect. To enhance reliability the proposed model that uses hamming code which helps to recover lost or modified information. The proposed Model implemented and tested successfully.

Keywords—Steganography; DWT; LSB; hamming code; encryption and decryption

I. INTRODUCTION

Mark kahn in [1] has defined steganography as the art and main science of communicating such that the existence of communications is unknown. The goal of steganography is hiding messages into another carrier, in a way that does not allow outsiders to detect or recognize that there is a hidden message. Encryption is another technique that can be used to protect the information and provide secure communication, but in this case the outsiders know and can see the cipher text but they could not understand and use it [2].

Cryptography and steganography are both used to protect information from disclosure by unwanted parties. While cryptography is about protecting the contents of messages, the main purpose of steganography is to hide the data in a covered

media, so that others would not be able to get or to notice it, which is preferable it does not attract the attentions or suspicions of hidden message existence. Many experts prefer using both techniques in order to achieve and provide more security and protection. Steganography has different types according to the way that used to hide the data in covered media. The first one is called audio steganography this type embeds secret messages in audio. This technique is the most challenging technique because it is extremely hard to add or remove data from audio file structure. The second type is the text steganography, and it means hiding the secret messages into other texts. It is a very challenging task. This is because of the small amount of redundant information to replace with a secret message in the text files. The most used type is the Image steganography in which it embeds the secret message into digital images. Image steganography has two major techniques according to its domain, spatial domain and the transform domain. Least Significant Bit (LSB) technique is the simplest and most common one used in the spatial domain. The Discrete Cosine Transformation (DCT) and the Discrete Wavelet Transformation (DWT) are the two most common steganographic domain transformation methods used in the transform domain and they are the most complex and efficient techniques. DCT and DWT hides the data in the areas of the image that are less exposed to Cropping, Compression, and Image processing. Steganography faces and has to overcome three main challenges, the Invisibility "Security of Hidden Communication", Robustness and the size of embedded data. There is no steganographic technique, capable of resolving all the three challenges at a high level of accuracy. It is not possible to attain high robustness to signal modifications and high insertion capacity at the same time [3].

This paper addresses the main steganographic challenges in order to achieve a compromise between the capacity of the embedded data and the robustness to certain attacks, while keeping the perceptual quality of the stego-medium at an acceptable level. The rest of the paper organized as follows:

- 1) Section II Literature review.
- 2) Section III discusses the proposed Model.
- 3) Section IV discusses the implementation results.
- 4) Section V conclusion.
- 5) Section VI future work.

II. LITERATURE REVIEW

A. LSB Technique

It's the simplest technique used in image steganography, in LSB technique the data that we want to hide inserted into the least significant bits of the pixel information [4]. This technique is widely used because it has less chance of distortion of the original image, more capacity to hide information and it is less complex. Despite all these advantages LSB techniques have some serious disadvantages like the hidden information can be lost with image manipulation, hidden data can be destroyed by simple attacks and it requires a high transmission rate due to the large size of stego image.

Mamta Juneja et al., in [5] presents two components based on LSB technique for embedding secret information in the LSB's of the blue plan and partial green plan of random locations of pixels at the edges of the cover images, it's integrated with an Advanced Encryption Standard to be more robust.

Shamim Ahmed Laskar et al., in [6] used a method to embed data in the red plan of the image and it select pixel by generating random numbers, changes in image can't be noticed. Stego key used to generate random number in order to select pixel locations. It focuses on increasing the security of the hidden information and reducing distortion rate.

Y. K. Jain et al., in [7] used a method to divide the image pixel range and generates a stego key, this private key has five different ranges of the gray level in the image and each range present the replaced bits number to be embedded in the least significant bits of the image. It has a drawback that it hides extra bits of signature with hidden message

S. Channalli et al., in [8] used a stego key to hide data, it modifies the LSB of the pixel and the secret data bits. A combination of pattern bits of $M \times N$ size and random key value. The first step of embedding is by matching each pattern bit with a secret message bit, if suit it modifies the 2nd LSB bits of cover image "original" otherwise remains the same. This technique has low hidden capacity because each secret bit requires a block of $(M \times N)$ pixels.

H. Motameni et al., in [9] introduced data hiding technique that finds dark areas of the stego image to hide the secret information using the LSB technique. This method required high computation to find dark regions and has not tested on high texture images and it is not useful for gray or color images just for binary images.

V. Madhu Viswanatham et al., in [10] introduced an image steganography technique, based on LSB replaces and selection of random pixel in the cover image area. It generates random numbers and selects the area of interest where the secret data supposed to be hidden. The biggest advantage of this technique

is the security of hidden data and the drawback is data embedding, does not take care of the Visual Quality when pixels selected.

M. Tanvir Parvez et al., in [11] introduced a pixel indicator method with variable bits; it chooses one plan among red, green and blue planes and embeds data into variable LSB of the chosen plan. The plan selection is sequential and the size depends on the cover image "original" bits.

B. DWT

This technique used to convert the spatial domain into the frequency domain; it separates the high frequency and low frequency information.

DWT divides component into four frequency bands called sub bands known as

- LL- Horizontally and vertically low pass
- LH - Horizontally low pass and vertically high pass
- HL - Horizontally high pass and vertically low pass
- HH - Horizontally and vertically high pass



Fig. 1. One phase decomposition using DWT

Human eyes are more sensitive to the low frequency (LL sub band) the other three sub-bands are high frequency they contain unimportant information like the edge and texture details and they are not sensitive to small changes. Accordingly hiding secret data in these sub-bands does not reduce the image quality, at variance of LL sub-band, which is very sensitive to small changes and not used for information hiding.

K. S. Babu et al., in [12] for authentication purposes the proposed a method hides the data into a cover image, which then used to prove the integrity of the embedded secret data. The secret message transformed from the spatial domain to the discrete wavelet transform and then the coefficients of DWT transposed with the verification code before embedding in the spatial domain of the cover image. This method is computationally complex.

Dr.H.Rohil et al., in [13] applied DWT on colored images and its use Arnold transformation to improve the security. The cover image splits into " Red, Green and Blue plan" then DWT is applied to all plans, after that secret image is changed using Arnold transform and every color plan of the changed secrete images is separated. Then secret images plans embedded into HL, HH, and LH sub bands.

Aayushi Verma et al., in [14] proposed an algorithm for embedding and extraction of secret image embedded behind cover gray scale image. 2-level DWT is applied to the original

image and then targeted band is selected to be modified. Then the size of secret image is calculated, after that the five most significant bits of secret image embedded into high frequency bands.

Barnali Gupta Banik et al., in [15], used a method that applies Haar-DWT for decomposing the cover image, it generates random number and the detailed horizontal & vertical coefficients are modified by adding these random numbers when data bit is 0. Then apply Inverse DWT.

L. Tong and Q. Zheng-ding, in [16] proposed a DWT based color image method. In this approach, the secret information embedded in a publicly accessed color image by a quantization-based strategy. However, the latter case method processes grayscale images as cover object to create a subliminal channel and it utilizes transform coefficients of 2-Dimensional Discrete transform for embedding process.

T. Narasimmalou[17] Proposed an optimal discrete wavelet transform (DWT) based steganography. This method shows enhancement in the generated peak signal noise ratio (PSNR).

Nag et al., in [18] proposed a steganographic technique based on DWT and applied Huffman coding on Secret message before embedding it in high frequency components of the 2-D cover image, the low frequency component is kept untouched, which retains the visual quality of the image. The algorithm has high capacity and satisfactory security.

All the above mentioned approaches are very susceptible to steganographic analysis and attacks as the information are embedded in a predefined and expected locations. The goal of the proposed technique is to achieve high resistance against steganographic analysis and sustains well known attacks.

III. PROPOSED MODEL

A. Secret Data Hiding

In order to enhance security, and reliability, we integrate encryption and error detection and correction techniques along with blinded DWT. The blinded DWT achieved by dividing

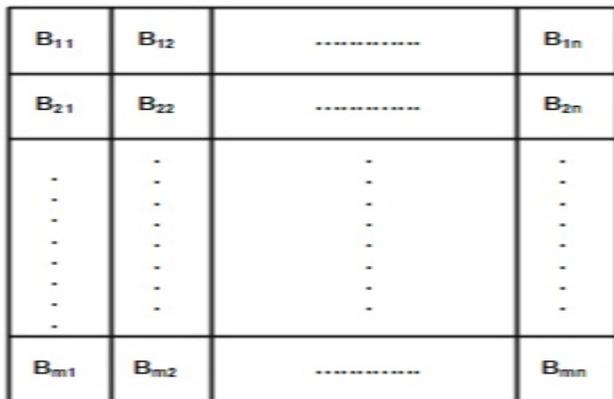


Fig. 2. Image Division

the image into NxN Blocks as shown in figure 2 where N, key dependent or determined by the User and unknown to the outsiders.

The Proposed technique applies DWT on each candidate block that will be used to hide information randomly or in a sequential manner and this is the second level of DWT blinding. The proposed model (see figure 7). Consists of the following main steps:

- 1) Encrypt the secret data.
- 2) Compute the hamming code of the encrypted data.
- 3) Divide the image into N blocks.
- 4) Determine the candidate Blocks
- 5) Decompose the candidate Block into three color planes (R, G and B).
- 6) Apply DWT on each candidate block independently.
- 7) Secret Message and the hamming code embedded into LSB of the HH and LH bands of each candidate block.
- 8) Apply inverse transformations.
- 9) Combine the three color planes that generate the final stego image.

B. Description of the proposed Model

The proposed model enables the user to select any symmetric encryption algorithm, but we recommend to use Advance Encryption Standard (AES) as it is a very secure algorithm and supports larger key sizes and it is faster in both hardware and software.

The cipher text is very susceptible to alteration so we used hamming code error detection and correction technique. We suggest the usage of five check bits for every group of three bytes, which is a compromise between the redundancy and reliability. The image resized or cropped to be squared one as this is necessary for the next processing step. The Block number and the candidate blocks which are should be kept secret can be determined by the user or can be generated from encryption key which is better as it is very difficult for the user to maintain bulky secret information. In the proposed model we suggest to use the following modified linear congruential generator as it is very fast pseudo-random sequence generator and convenient for our model as high-quality randomness is not critical and the duplicated numbers should be excluded in case of sequence duplication:

$$B_0 = K_0 \text{ Mod } N \tag{1}$$

$$B_m = (K_m + 1) * B_{m-1} + g \text{ mod } N \tag{2}$$

Where : N - Blocks Number ; m=1,2,...N-1;

$B_m = m^{\text{th}}$ Block ; $K_m = m^{\text{th}}$ Key Code;

g should be selected carefully to be relatively prime to N

The candidate block decomposed into three colors plans (R, G, B) If the covered image is a color one. (See figure 3).



Fig. 3. Cover Image (R,G, B) color plans

In order to enhance accuracy and to be able to use DWT, which needs to deal with fractions and negative numbers which is not applicable for Pixels values, we use the following scale transformation:

$$\text{New Value} = (D_{max} - D_{min}).(V - S_{min}) / (S_{max} - S_{min}) + D_{min} \quad (3)$$

Where:

$[D_{min}, D_{max}] = \text{new range}$

$[S_{min}, S_{max}] = \text{old range}$

$V = \text{Pixel color value}$

We apply a simplified form of DWT on the scaled data by using the following Transformation algorithm:

For each Row and Column in the Candidate Block(B_c)

```

{
Row length=Column Length
h=Row length/2;
For(i=0; i<h; i++)
  For(j=0; j<h; j++)
    {
k=j*2;
 $B_c[j,i] = (B_c[k,i] + B_c[k+1,i])/2;$ 
 $B_c[j+h,i] = (B_c[k,i] - B_c[k+1,i])/2;$ 
    }
}

```

The output of the described transformation at different levels shown in figure 4 and 5.

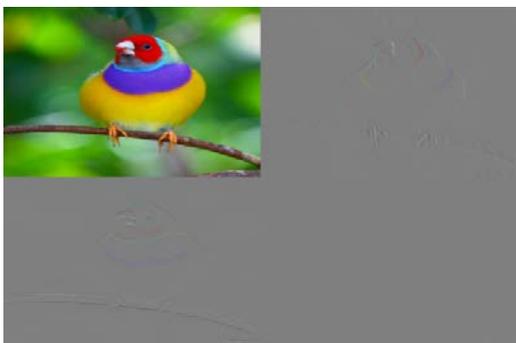


Fig. 4. Cover Image after 1 level of DWT

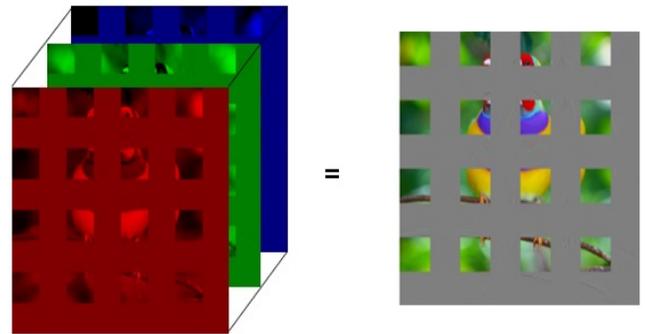


Fig. 5. Cover Image after 4 levels of DWT

At this moment candidates Blocks and the data are ready to start embedding process in which each bit of the data byte inserted in predefined least significant bits of HH color planes (see figure 6) and the hamming code in the LH color plans

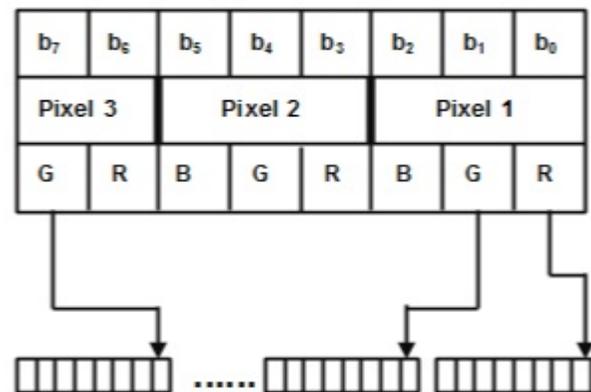


Fig. 6. Data Embedding Diagram

After embedding the data we apply the Inverse DWT and generate the final stego image by combining the three color plans (R, G, B).

C. Secret Data Extraction

The steps of data extraction can be summarized as the following:

- 1) Divide the image into N blocks.
- 2) Determine the candidate Blocks
- 3) Decompose the candidate Block into three color planes (R, G and B).
- 4) Apply DWT on each candidate block independently.
- 5) Extract the Secret Data and the hamming code embedded into LSB of the HH and LH bands of each candidate block.
- 6) Compute the hamming code of the encrypted data
- 7) If the computed hamming=embedded hamming code there are no errors

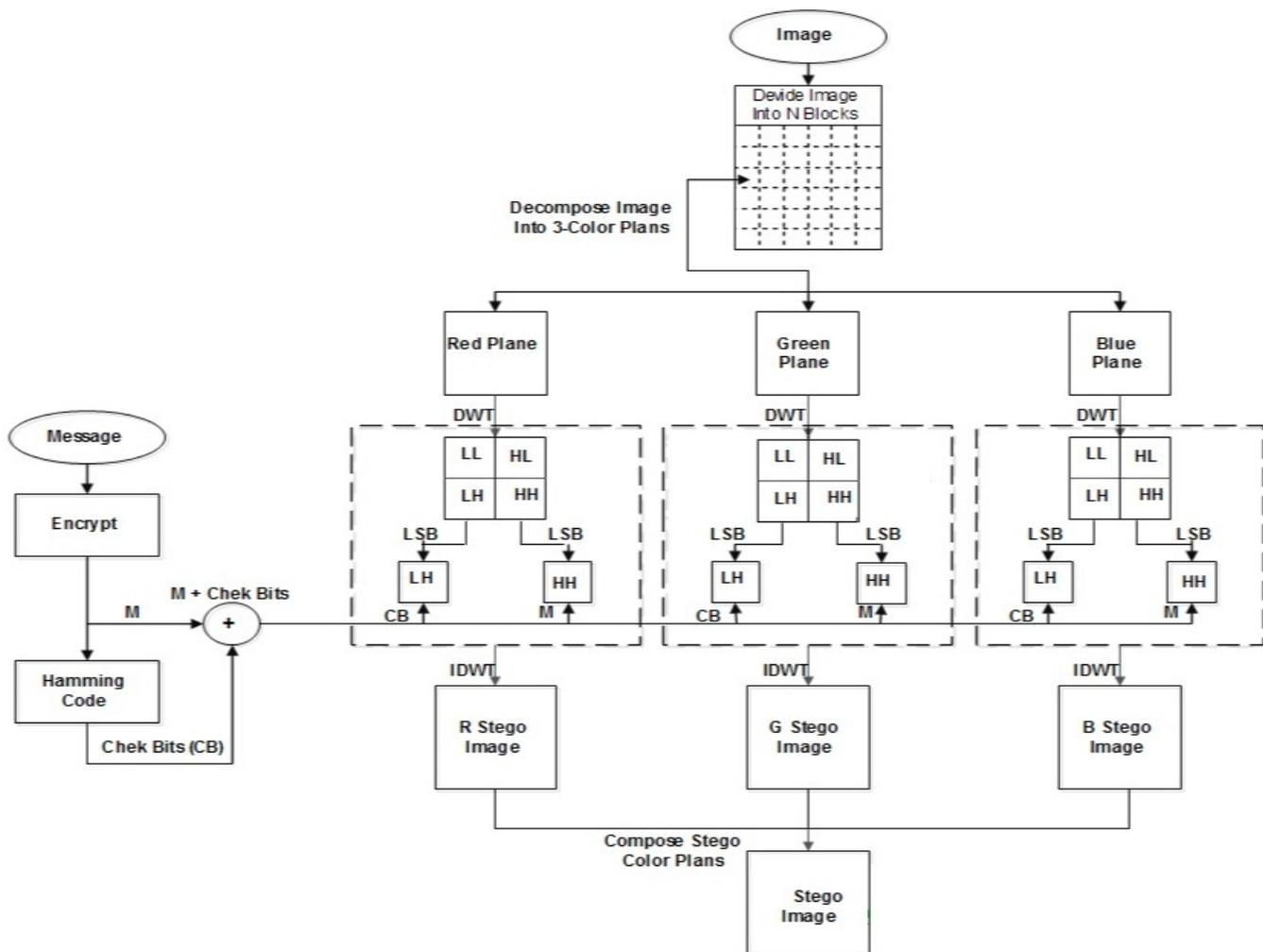


Fig. 7. Proposed Model Structure

Otherwise correct the error.

- 1) Decrypt the cipher
- 2) Apply inverse transformations
- 3) Combine the three color planes that generate the final stego image.

IV. IMPLEMENTATION OF THE PROPOSED MODEL

The proposed model implemented by using C# and tested successfully under various images types: size, color, gray scaled and various levels of DWT transformation, the result shows that it is has a good stego image quality, high level of reliability and security. In Figure 8,9 and 10 illustrated the cover image before and after embedding the data (Stego-Image).



Fig. 8. Cover Image



Fig. 9. 8 Level Blinded DWT



Fig. 10. Stego Image

The proposed technique evaluated and compared by LSB and DWT based-techniques. The evaluation done based on the following parameters:

- Capacity - measures the amount of embedded data with minimum distortion effect on the cover image.
- Invisibility - describes how much the quality of cover image remains intact, which makes the human vision unable to distinguish between the stego-image and the cover one.
- Resistance- measures the hidden data tolerance to image updates and stego-attaches.
- Security- describes how it is difficult for outsider to detect and disclose the hidden data.
- Performance- describe how fast the process of data embedding and extraction.

The evaluation results shown in table 1.

The results show that the proposed technique has better resistance and security characteristics than LSB and DWT, whilst it has the worst capacity characteristic. All techniques have the same degree of invisibility.

TABLE I. COMPARISON RESULTS

Techniques Parameters	LSB Based	DWT Based	Proposed Technique
Security	Low	Medium	Very High
Capacity	High	Medium	Low
Resistance	Low	Medium	High
Performance	Very High	Medium	Medium
Invisibility	High	High	High

V. CONCLUSION

In this Paper, we introduced a new steganographic technique that increases the secrecy and reliability of the hidden data without losing the image quality or losing any data in the image, the secrecy achieved by using encryption algorithm in addition to hiding the part of image where the information embedded. The division of image into key dependent number of blocks, hiding the data into unknown blocks and applying multi levels of DWT increase the confusion of the outsider. The reliability achieved by using error detection and correction technique that enables the recovery of altered data. This technique demonstrates that it is very effective and can resist many steganographic attacks, as it has a high degree of blinding characteristics.

Compared to other steganographic systems, the size of embedded data reduced as result of the embedding redundant hamming code. The user can reduce the redundancy by selecting larger blocks without affecting the security and still maintains a high degree of data integrity.

VI. FUTURE WORK

The proposed technique will be enhanced to increase the size of embedded date and decrease the redundancy in addition to adapting it to be effectively used in image watermarking .

REFERENCES

- [1] Johnson,NeilF.,“Steganography”,2000URL:http://www.ijtc.com/stegdoc/sec201.html.
- [2] Stallings, W. (1995). Network and internetwork security: principles and practice(Vol. 1). Englewood Cliffs: Prentice Hall.
- [3] Hong-Juan Zhang, Hong-Jun Tang,"A Novel Image Steganography Algorithm Against Statistical Analysis",Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.
- [4] Kaur, R., & Singh, B. (2012). Survey And Analysis of Various Steganographic Techniques. International Journal of Engineering Science and Advanced Technology, 2, 561-566
- [5] Juneja, M., & Sandhu, P. S. (2013). A new approach for information security using an improved steganography technique. Journal of Information Processing Systems, 9(3), 405-424.
- [6] Laskar, S. A., & Hemachandran, K. (2013). Steganography based on Random Pixel Selection for Efficient Data Hiding. International Journal of Computer Engineering and Technology, 4(2), 31-44.
- [7] Jain, Y. K., & Ahirwal, R. R. (2010). A novel image steganography method with adaptive number of least significant bits modification based on private stego keys. International Journal of Computer Science and Security, 4(1), 40-49.
- [8] Channalli, S., & Jadhav, A. (2009). Steganography an art of hiding data. arXiv preprint arXiv:0912.2319.
- [9] Viswanatham, V. M., & Manikonda, J. (2010). A novel technique for embedding data in spatial domain. International Journal on Computer Science and Engineering, IJCSSE, 2(2010).

- [10] Motameni, H., Norouzi, M., Jahandar, M., & Hatami, A. (2007, October). Labeling method in Steganography. In Proceedings of world academy of science, engineering and technology (Vol. 24, pp. 349-354).
- [11] Parvez, M. T., & Gutub, A. A. (2008, December). RGB intensity based variable-bits image steganography. In Asia-Pacific Services Computing Conference, 2008. APSCC'08. IEEE (pp. 1322-1327). IEEE.
- [12] Babu, S. K., Raja, K. B., Kiran, K. K., Manjula Devi, T. H., Venugopal, K. R., & Patnaik, L. M. (2008, November). Authentication of secret information in image steganography. In TENCON 2008-2008 IEEE Region 10 Conference (pp. 1-6). IEEE.
- [13] Dr. Harish Rohil, Parul1, Manju2, "Optimized Image Steganography using Discrete Wavelet Transform (DWT)", International Journal of Recent Development of Engineering and Technology, ISSN 2347 - 6435 (Online) Volume 2, Issue 2, February 2014.
- [14] Aayushi Verma, Rajshree Nolkha, Aishwarya Singh and Garima Jaiswal, "Implementation of Image Steganography Using 2-Level DWT Technique", International Journal of Computer Science and Business Informatics
- [15] Banik, B. (2013). Prof. Samir K. Bandyopadhyay, A DWT Method for Image Steganography. International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), pp. 983-989.
- [16] T. Liu and Z. Qiu, "A DWT-Based Color Image Steganography Scheme," in Proc. IEEE, 6th International Conference on Signal Processing, 2002, vol. 2, pp. 1568-1571.
- [17] Narasimmalou, T., & Joseph, R. A. (2012, March). Discrete Wavelet Transform based steganography for transmitting images. In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on (pp. 370-375). IEEE.
- [18] A. Nag, S. Biswas, D. Sarkar and P. P. Sarkar, A novel technique for image steganography based on DWT and Huffman coding, IJCSS, vol. 4, no. 6, pp. 561-570

A Multimedia System for Breath Regulation and Relaxation

Wen-Ching Liao

Department of Computer Science and Information
Engineering
National Taiwan University
Taipei, Taiwan R.O.C.

Han-Hong Lin

Department of Computer Science and Information
Engineering
National Taiwan University
Taipei, Taiwan R.O.C.

He-Lin Ruo

Department of Computer Science and Information
Engineering
National Taiwan University
Taipei, Taiwan R.O.C.

Po-Hsiang Hsu

Department of Computer Science and Information
Engineering
National Taiwan University
Taipei, Taiwan R.O.C.

Abstract—In the hectic life today, detrimental stress has caused numerous illness. To adjust mental states, breath regulation plays a core role in multiple relaxation techniques. In this paper, we introduce a multimedia system supporting breath regulation and relaxation. Features of this system include non-contact respiration detection, bio-signal monitoring, and breath interaction. In addition to illustrating this system, we also propose a novel form of breath interaction. Through this form of breath interaction, the system effectively influenced user breath such that their breathing features turned into patterns that appeared when people were relaxed. An experiment was conducted to compare the effects of three forms of regulation, the free breathing mode, the pure guiding mode, and the local-mapping mode. Experiment results show that multimedia-assisted breath interaction successfully deepened and slowed down user breath, compared with free breathing mode. Besides objective breathing feature changes, subjective feedback also showed that participants were satisfied and became relaxed after using this system.

Keywords—breathing; relaxation; biofeedback; interaction; multimedia

I. INTRODUCTION

Recently, mental stress is becoming one of the major factors causing illness. Health problems caused by stress include hypertension, cardiovascular diseases, increased likelihood of infections and depression [1-3]. To release stress multiple relaxation techniques have been developed over hundreds of years, including yoga, meditation, qi-kung, tai-chi, etc. in eastern culture society. Also in modern psychophysiology, relaxation techniques like autogenic training, diaphragmatic breathing, mindfulness, etc. are also developed [4].

Among those relaxation techniques breath regulation usually plays an essential role. Researchers demonstrated that breath regulation is beneficial for reducing blood pressure [5, 6] and focusing the mind for optimal performance [6].

Over the past decades, computing technologies have been utilized to support breath regulation in multiple applications. For instance, Moraveji proposed peripheral paced respiration, which integrated breath pacing application in the operating system, to slow down user breath during information work [7]. Yu proposed a multimedia biofeedback system for abdominal breath learning [8]. Park et al. designed multiple modes of breath induction to facilitate relaxation and other mental function promotion [9].

Most of these studies tend to influence user breath through providing respiratory guidance for users to follow, and turned their breath slow and deep eventually. However, human breath actually involves randomness between each cycle of inhalation and exhalation. Even in calm status, there exist correlation plus random variation between each breath. Breath features like breath period, breathing depth, and the inhalation/exhalation ratio varies spontaneously. In fact, a range of studies has shown that healthy breathing is characterized by complex variability consisting of considerable structured variability and some random variability [10-13]. Forced breath pacing though mechanically changes user breath, the subjective feeling of relaxation may not be promoted [14]. Therefore, pure guiding style of breath regulation may not be suitable when applied for relaxation.



Fig. 1. The relaxation system appearance

In the breath interaction module of our relaxation system, we propose a target-based breath regulation. Unlike traditional pure breath guiding, in which the guiding multimedia is unrelated to user breath and the user is required to breathe following the guide during the entire process of the breath regulation, the visual or aural feedback reflects user breath in the target-based breath regulation. The user is asked to achieve some feedback target through his or her breathing. For instance, if the visual feedback of user breath is a flower, which opens and closes to reflect inhaling and exhaling. The instruction given to the user may be "Let the flower open to the maximum and close to the minimum through your breathing". After giving the instruction to the user, we can influence users' breathing pattern differently through manage the underlying signal mapping mechanism. The mechanism mapping the breath signal to the visual/aural multimedia feedback can be designed to slow down and deepen user breaths as we expected. The difference between this target-based breath regulation and common pure breath pacing is that user breath is less constrained. Target achievement timing is loosely stipulated. Tension and discomfort caused by forced breath pacing reduce while the influence on breath can still be achieved because of the attempt to accomplish the given target.

Except proposing novel form of breath regulation, we integrate newly developed noncontact breath detection technique, special reflection mechanism, and bio-signal monitoring to implement a relaxation system. Following, we first illustrate the construction of the system in Section II. Then we detailed the proposed breath interaction mechanism and a preliminary experiment conducted in Section III. In Section IV, a formal experiment is introduced. The experiment procedure designed to investigate the effects of three forms of breath regulation - pure guiding mode, target-based mode, and free breathing mode, which means no multimedia feedback is involved, is demonstrated. The experiment results are also reported and illustrated. In Section V, the analysis and comparisons of physiology changes are demonstrated. Subjective reports when experiencing these three modes are also discussed. Finally, conclusions and acknowledgements are represented in Section VI and Section VII.

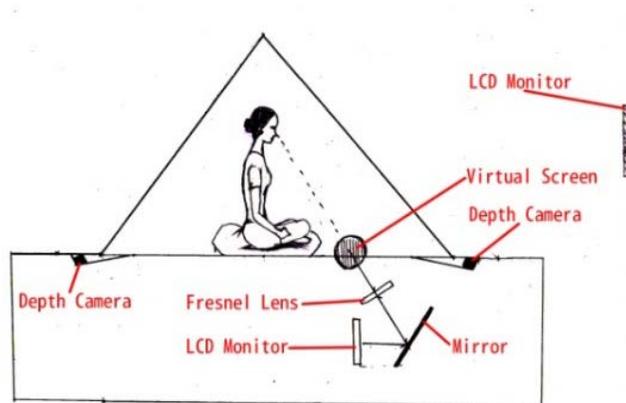


Fig. 2. The platform design diagram

II. SYSTEM DESIGN

A. System Structure

The reason we have the opportunity to implement this multimedia system is because of a collaborative project with a professional music company. The system will be settled at a realistic music store for long term commercial exhibition after this study. Therefore, sometimes the reason of the appearance selection is partially due to the artistic sensation, not entirely for the technical reasons. To deliver the isolated and stable visual sensation, the appearance of this system is specially designed in the shape of a pyramid. Copper tubes are used to construct the pyramid shape frame (Fig. 1). Construction of the system is illustrated in Fig. 2. A user is required to sit at a specified position under the pyramid so that his or her respiration can be detected correctly. Two depth cameras are placed in the front and in the back of the user to detect respiration [15], in the meantime, a blood volume pulse (BVP) sensor and a skin conductance response (SCR) sensor are integrated in a sphere (Fig. 3). The sphere is put aside the seat and meant to be hold by the user when using this system.

There are two monitors in this system. One is embedded in the wall in front of the user, showing immediate detected bio-signals including breaths, heart rate, and SCR (Fig. 4). The other one is under the user's sitting ground displaying a metaphor (i.e. a lotus in our design). Displayed metaphor image is reflected by a mirror and projected through a Fresnel lens so that only people sitting at the specified position under the pyramid can have a clear view of it [16] (Fig. 2). Because the background of the metaphor image is removed and the special optical effect of the Fresnel lens, this projected metaphor looks floating in the air in front of the user.

B. Sensors

1) Respiration Sensor

In the past, the respiratory inductive plethysmography (RIP) sensor is usually the choice of the respiration sensor, and it is widely used in respiration related researches. Although the bound feeling while using RIP makes the user uncomfortable, traditional noncontact respiration detection techniques are too expensive and related settings are complicated. For instance, optoelectronic plethysmography requires the user being attached with multiple light-reflecting markers.

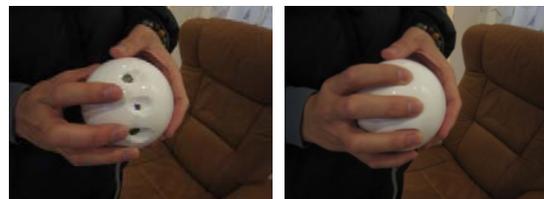


Fig. 3. The remade sphere shape sensor of heart rate and skin conductance

Until recently, a novel respiration detection method using depth camera is proposed [15]. The emitter mounted in the depth camera emits infrared (IR) light beams and the depth sensor reads the IR beams reflected back to the sensor. The

reflected beams are converted into depth information measuring the distance between an object and the sensor. Since we put two depth cameras in front and in back of the user. We can obtain the distance from the front side of the user to the front depth camera and the distance at the back side. Then, given the distance between two depth cameras, we can obtain the thickness of the user torso. Therefore, the respiratory signal (respiratory rate and depth) can be measured through the thickness changes as a result of breathing. This depth information based detecting method is not only cheap but also simple to operate. The only device needed are depth cameras and no other miscellaneous gadget is required. Therefore, we choose this technique as our respiration detection method.

Another advantage of using depth cameras as the respiration sensor is that we can observe the fluctuation at chest and at belly separately. With this information, we can provide user more concrete visual feedback when practicing diaphragmatic breathing (Fig. 4). Diaphragmatic breathing, also known as abdominal breathing, is a well-known breathing technique which activates parasympathetic nervous system (PNS) and induces relaxation [18]. It will be recommended to the users when using this relaxation system in our experiment.

2) Heart Rate and Skin Conductance Sensors

The bio-sensor adopted is the peripheral device of WildDivine, which includes a photoplethysmograph (PPG) sensor and two electrodes so that user's BVP and SCR can be detected [17]. In order to provide comfortable and naturalistic sensor interface, we specially embedded the bio-sensor into a sphere ball with pits for easy contact (Fig. 3).

3) Biofeedback Medium

There are two forms of visual feedback in this system. The first one is the bio-signal monitoring shown on the screen in front of the user (Fig. 4). Signals shown in the order: left up, left down, right up, and right down, are respectively chest breath, belly breath, heartbeat, and SCR. There is a rectangle mask composed of blue and green regions in the middle window. This mask uses blue region to cover the chest part and green region to cover the belly part. A user can adjust the position of this mask and the covering range of the blue and green regions by mouse clicks. This enables identifying the proportion of the abdominal breath. Besides, the SCR curve at the right down window also indicates the relaxing extent of the user.

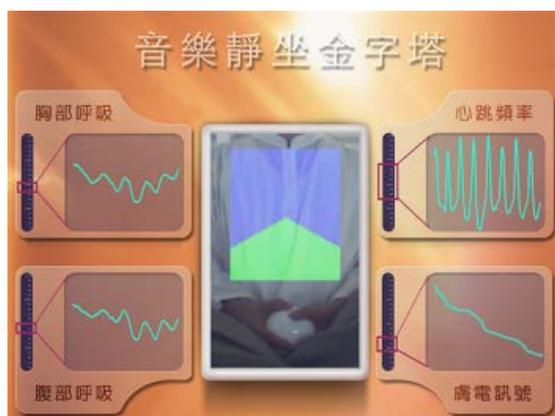


Fig. 4. Bio-signal monitoring

The second form of visual feedback is a metaphor (it is a lotus in our application) projected by the reflection mechanism beneath the sitting ground. The user sitting at the specific position so that he or she can have a special visual effect that the projected metaphor seems floating in the air. The projected metaphor is replaceable, and we choose a lotus as the metaphor of breath in our application because that a lotus is widely used in eastern culture to symbolize a pure, peaceful and clear mental state. Later in the experiment, the lotus is designed to open and close to imply inhalation and exhalation of the user respectively.

III. BREATH INTERACTION

A. The Mapping Mechanism

Traditionally, to influence user breath, users are required to breathe following visual or aural guidance, which is unrelated to their breath. However, pure guiding style of breath regulation though mechanically slow down and deepen user breath, it also induces tension due to unnatural paced breathing.

Therefore, we propose the concept of target-based breath regulation. The target-based breath regulation reflects user breath with the visual or aural multimedia. Users are required to achieve some feedback target through their breathing. Hence, the underlying mapping mechanism which maps breath signal to the feedback multimedia content can affect the result of breathing pattern.

To verify that managing the underlying mapping mechanism can effectively affect the breathing pattern of the user a preliminary experiment is conducted. In this preliminary experiment we use a variable circle to symbolize user breath. The circle expands when the user breathes in and shrinks when breathing out. Two mapping mechanism are designed. The first one, named local-mapping mode, maps the extremes of the circle variation to the local extremes of the breath signal. The second one, named global-mapping mode, maps the extremes of the circle variation to the global extremes of the breath signal. Details are described below.

We first generate n circle images with the radiuses ranging from 1 to n . Then we index these circle images from 1 to n orderly so that the bigger the index is the bigger the circle is.

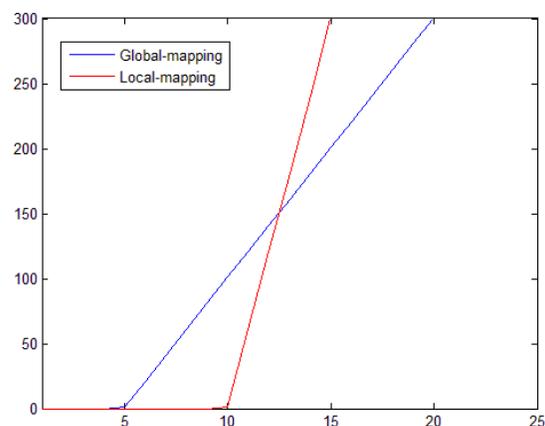


Fig. 5. An example of mapping of the breathing depth and the corresponding image index

Assume the breath signal is $f(t)$ and the corresponding index of the circle image displayed is $g(t)$. Then we let

$$g(t) = \frac{f(t) - \min(f(t))}{\max(f(t)) - \min(f(t))} \times n$$

where $\max(f(t))$ is the local maximum and $\min(f(t))$ is the local minimum in local-mapping mode, and they are global extremes in the global-mapping mode. An example of mapping illustration using $n=300$ and breathing depth signal ranging from 5mm to 20 mm is shown in Fig. 5.

In the local-mapping mode, the extremes are obtained in the period from immediate sampling time t to 15 seconds backward $t-15$. The reason 15 seconds are specified is that we hope to find the extremes in at least one complete breath cycle, and normally people breathe with the frequency higher than 4 times per minute, which means the breath period should be lower than 15 seconds. In the global-mapping mode, the extremes are found in the period from $t=0$ to immediate time t .

B. Preliminary Experiment

We recruit 12 labmates (5 females) as experiment participants. Each participant was required to proceed 3 trials of abdominal breathing practices. Each trial lasts 3 minutes and total 9 minutes for a participant. For all participants, the first trial is free breathing without multimedia. For 6 of the participants the second trial is the local-mapping mode and the third trial is the global-mapping mode. For the other 6 participants the order of the second and the third trial is reversed. Experiment results show that breathing patterns are significantly different in these three modes (Fig. 6, Fig. 7).

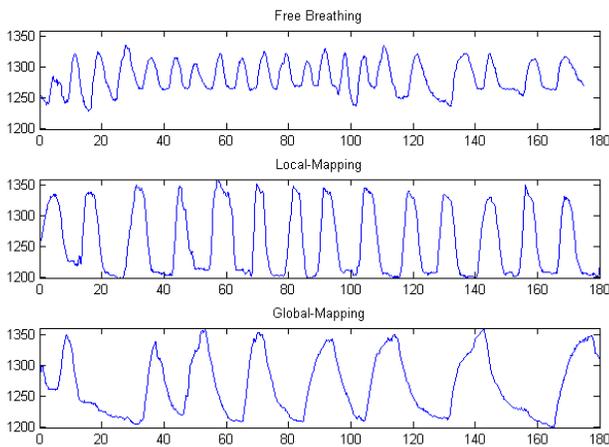


Fig. 6. Three breathing patterns in the preliminary experiment: Free breathing mode, Local-Mapping mode, and Global-Mapping mode

A one-way repeated measure analysis of variance (ANOVA) was conducted to evaluate the average breathing depth and frequency among three modes. The results indicated a significant difference in breathing depth ($N=12$, $F(2, 10) = 10.29$, $p=0.004$). Follow up pairwise paired t-test comparisons also showed significant difference in breathing depth ($p<0.05$), but not in breathing frequency.

There was significant increase in breathing depth when multimedia feedback was used, and the increase in global-mapping mode is even higher than that in the local-mapping

mode. However, the result of ANOVA conducted to compare breathing frequency showed no significant difference. Follow up pairwise paired t-test comparisons showed that the breathing frequency in global-mapping mode was significant lower than those in the other two modes. But there was no significant difference between the free breathing mode and the local-mapping mode (Fig. 7).

According to the research of McCaul et al. in 1979, slowing respiration rate reduced physiological arousal and self-reported anxiety [19]. Although the global-mapping mode performed better in deepening and slowing down user breath, subjects reported that achieving the given target, which was to maximize and minimize the lotus through their breathing, was harder in the global-mapping mode. Some of the participants even feel uncomfortable tension instead of relaxation. To compromise on the effect of slow and deepen user breath and the subjective feeling of tension due to over required effort. We choose the local-mapping mode as our design of the underlying mapping mechanism of the breath regulation. It can effectively deep and slow user breath while the multimedia feedback target provided for the user can be achieved without causing uncomfortable tension instead. In our formal experiment, we compare it with the traditional pure guiding mode and the free breathing mode.

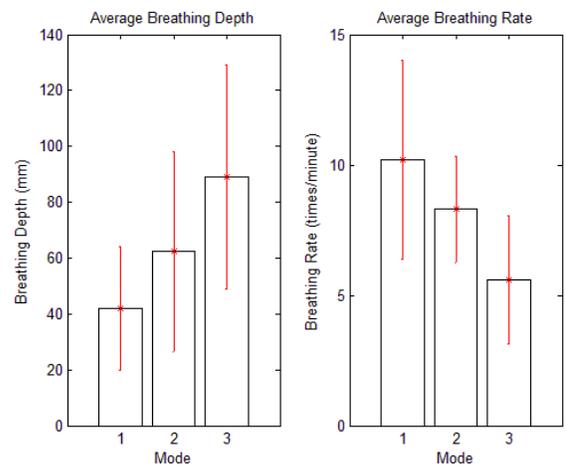


Fig. 7. Comparison of average breathing depth and frequency in three modes. The number 1 to 3 in x-axis represent free breathing mode, local-mapping mode, and global-mapping mode respectively

IV. EXPERIMENT

Among several relaxation techniques, deep breathing or diaphragmatic breathing (DB) is an easy and intuitive evidence-based method for stress management [18]. DB addresses the autonomic nervous system (ANS) imbalance that arises after exposure to a stressor and activation of the sympathetic 'fight-or-flight' response. As DB activates the parasympathetic ANS branch, action of the sympathetic branch becomes inhibited which leads to a calmer, more relaxed state.

Generally, the chest part fluctuates more obviously than the belly part when breathing normally. When practicing DB, the primary motion should be changed to the abdominal part. The aim of the experiment is to investigate the effects of slowing and deepening user's abdominal breath in three different forms

of breath regulation. We also compare the induced tension of the user when using the proposed target-based breath regulation and the common pure guiding style of breath regulation. Following, we describe three modes of breath regulation in the experiment, the experiment procedure, and the analysis of the experiment results.

A. Three Modes of Breath Regulation

In Mode-I, the free breathing mode, the user was required to close eyes and practice DB comfortably. No multimedia assistance was adopted.

In Mode-II, the pure guiding mode, a lotus used as the metaphor for breathing guided user's respiration. At first, the system detected the user's initial respiration frequency, then the lotus started to open and close at the same frequency (e.g. 14 times per minute). After the frequency synchronization, the user was told to follow the visual guide and practice DB. If the user paced their breathing to the guidance well, the guiding frequency decreased eventually (e.g. 13 times per minute).

In Mode-III, the local-mapping mode, the lotus reflected user breathing. It opened and closed corresponding to the inhalation and exhalation. The underlying mapping mechanism

has been detailed in the last section. The users were told that the lotus variation reflected their breathing, and they were instructed to breathe to make the lotus open to the maximum and close to the minimum.

B. Participants

We recruited subjects in our laboratory without considering rigorous demography. Total 17 subjects participated in this experiment, including 12 males and 5 females with ages ranging from 21 to 38. Three of them had ever learned abdominal breathing before the experiment while others were not familiar with it.

C. Experiment Procedure

Every subject was first taught how to correctly proceed DB. Then they ran through these three modes of trials in a random order, five minutes for each trial. Between two trials, there was a two minute rest for users to answer some feedback questions. Total time cost was about 20 minutes. At the beginning of each trial the user was instructed to sit properly and hold the bio-signal sensor fitly. When bio-signals including breathing, heart rate, and skin conductance were correctly shown on the screen (Fig.4), the trial began.

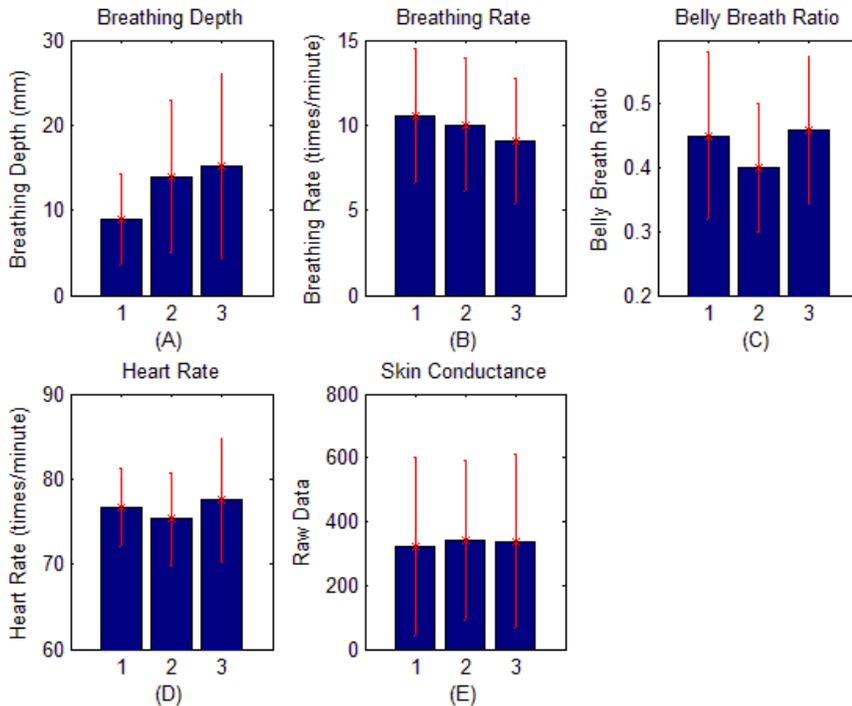


Fig. 8. Bio-signal features of three modes. (A) Breathing Depth (B) Breathing Rate (C) Belly Breath Proportion (D) Heart Rate (E) Skin Conductance Response

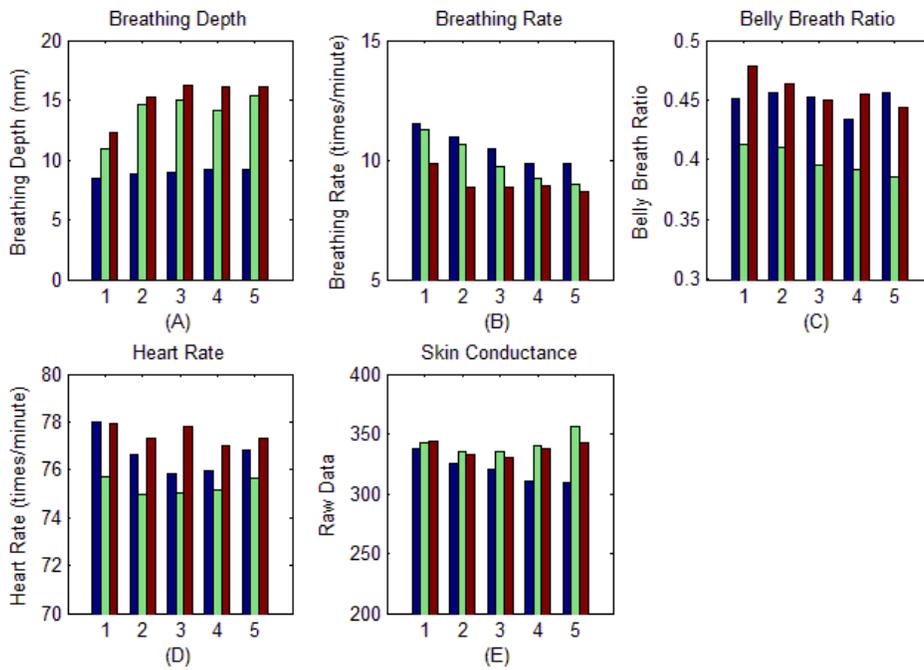


Fig. 9. Bio-signal features during five minutes experience process. (A) Breathing Depth (B) Breathing Rate (C) Belly Breath Proportion (D) Heart Rate (E) Skin Conductance Response. The unit of x-axis is minute and three bars at each minute indicate the results of Mode-I, Mode-II and Mode-III from left to right. (Blue: Mode-I, Green: Mode-II, and Red: Mode-III)

After entire three modes of trials, a five minutes interview was conducted to further realize the user experience of using this system.

D. Experiment Results

Recorded bio-signals include breath, BVP, and SCR. They were analyzed and transformed into five bio-signal features including breathing depth, breathing frequency, belly breath proportion, heart rate, and SCR. The averaged bio-features in three modes are shown in Fig. 8. We ran a one way ANOVA to investigate the difference of these three modes in each bio-signal feature, but there was no overall significant difference found (Breathing Depth: $F=2.48$, $p=0.095$, Breathing Frequency: $F=1.78$, $p=0.18$, Belly Breath Proportion: $F=1.35$, $p=0.27$, Heart Rate: $F=0.59$, $p=0.56$, SCR: $F=0.05$, $p=0.95$).

To further investigate the effects of the two types of multimedia-assisted breath regulation, we conducted paired t-tests to compare the bio-signal features in Mode-II and Mode-III to those in Mode-I respectively.

The results show that the breathing depth in Mode-II and Mode-III is higher than that in Mode-I (Mode-I vs. Mode-II: $p=0.03$, Mode-I vs. Mode-III: $p=0.009$), however, there is no significant difference between Mode-II and Mode-III ($p=0.353$).

In respect of breathing frequency, Mode-III is significantly slower than Mode-I ($p=0.001$), and other comparisons of Mode-II vs. Mode-I and Mode-II vs. Mode-III show no significant difference (Mode-II vs. Mode-I: $p=0.195$, Mode-II vs. Mode-III: $p=0.107$).

These results indicate that multimedia-assisted breath regulation effectively deepen user breath, and that Mode-III, the local-mapping mode, can also slow down user breath significantly.

In respect of the belly breath proportion, which means the proportion of belly fluctuation among entire breath fluctuation involving chest and belly fluctuation, Mode-III and Mode-I are not significant different ($p=0.47$), but Mode-II is significantly lower than Mode-I ($p=0.035$). It means users breathe with fewer belly part and more chest part in the pure guiding mode. The reason for this result will be discussed later. Other comparisons in heart rate and SCR show few differences.

To realize the variation of each bio-signal feature in time, we also calculate averages per minute for each feature (Fig. 9). This figure shows that the average breathing depth remains relatively stable for Mode-I during entire five minutes process, and a raising trend is shown in Mode-II and Mode-III (see Figure 6(A)). This shows multimedia-assisted breath regulation deepen user breath eventually.

In respect of breathing frequency, the decreasing trends are shown in three modes, and the breathing frequency in Mode-III is relatively lower than those in the other two modes. This indicates that there is a trend users slow down their breath rate naturally during practicing DB, and that the local-mapping mode may slow down user breath more rapidly and significantly.

When discussing about belly breath proportion, compared with other features, this feature shows fewer variation during the entire process. Fig. 9(C) shows that proportions of Mode-I

and Mode-III are similar, and Mode-II is with the especially lower proportion from beginning to the end. Finally, in respects of other bio-features - heart rate and SCR, there is few difference in three modes is observed.

V. DISCUSSION

A. Bio-signal Feature Observation

The experiment result indicated that multimedia guidance or feedback indeed facilitated deepening abdominal breathing when practicing DB. The visual guidance (Mode-II) or visual feedback (Mode-III) of respiration reminded users of staying in the process of breath regulation so that their breathing depth continuously increased during the entire process of the trials (Fig.9 (A)(B)). The overall effect increased the amount of oxygen exchange and slowed down the breathing rate, which was beneficial for relaxation [19].

Among three modes of trials, the effect of slowing and deepening breathing of Mode-III is better than that of Mode-II. Some users reported that following the breath guidance in Mode-II caused mental tension because of worrying about falling behind the guidance. In contrast, the lotus reflected user respiration in Mode-III. Although users were given a target that to maximize and minimize the lotus through their breathing, they can accomplish the target at their own paces without rushing. The tension caused in Mode-II may also be the reason that the belly breath ratio of Mode-II is specially lower. According to user feedback, people reflected that they tended to breathe with more chest part forgetting to proceed DB when they focused on following the guidance. To sum up, the local-mapping mode of breath regulation, proposed based on the concept of target-based method, effectively slow and deepen user breath when practicing DB, in the meantime, eliminate the drawback of causing mental tension as in traditional pure guiding style method.

The experiment results in perspective of breathing depth and frequency fulfilled our expectation, but there was some unexpected features observed. For instance, the averaged SCR decreased at first in three modes but increased in the end of the trials in Mode-II and Mode-III (Fig.9 (C)). In Mode-II the phenomenon may be easier to explain. The guiding frequency of the pure guiding mode decreased as long as the user can still catch up with the guidance. Therefore, in the end of the trial the frequency had become so slow that to breathe following the guidance become uncomfortable. This even induced the raise in SCR.

However, the reason for the phenomenon in Mode-III is not clear. The explanation may be similar. In the local-mapping mode, the mechanism mapping the respiratory signal to the opening state of the lotus maps the local extremes of the respiratory signal to the most and the least opening state of the lotus. Once the user increased their breathing depth, they had to remain or rise their breathing depth so that they can achieve the next target. Therefore, in the end of the trials, they may also suffer from uncomfortableness due to the attempt of achieving too deep breathing. This should be noticed and be further improved in the next design of mapping mechanism.

B. User Feedback

After experiencing three modes of trials, some participants reported that they can concentrate on breath regulation better when eyes were closed in the free breathing mode (Mode-I), and others thought it more interesting to have a visual feedback. They reported that the visual feedback was advantageous for reducing wandering minds and calming down.

In the pure guiding mode (Mode-II), some participants thought it difficult to follow the guidance. Others can follow the guiding rhythm and become comparably more concentrative. The reason why the guiding rhythm is hard to follow may be that we simply activate the opening and closing of the lotus by a sinusoidal wave with different frequencies. So the length of the inhaling and exhaling periods are equal, which is usually not the case our voluntary respiration is. This reminds us that a customized ratio of inhalation and exhalation should be considered in a pure guiding style of breath regulation in the future.

In the local-mapping mode (Mode-III), most of people tested the accuracy of the lotus response at first. Sometimes participants thought the reflection inaccurate. We found the reason for that was the inappropriate locating of the respiration detection mask. Looking at a loosely controlled lotus certainly bothered users. However, for most cases of correct breath detection, participants favor this mode because of the connection between their breathing and the motion of the lotus. It made them feel more relaxed and immersive in the process of breath regulation. Those feedback suggest that the interaction form through real time visualization of respiration may be beneficial and suitable when applied in the relaxation application.

There are also user feedbacks about other components in this relaxation system. Most of users thought that the bio-signal feedback on the wall (Fig. 4), was assistive for identifying whether they breathe primarily with the abdominal part. Besides, the reflected image of the lotus also interested multiple users. They like the design of visual feedback and the immersive visual effect created by that.

VI. CONCLUSION

In this paper, we integrate the non-contact breath detection technique, the bio-signal sensors and the specially designed reflection mechanism to construct a multimedia system for breath regulation and relaxation. A novel form of breath regulation named target-based breath regulation is proposed. This concept is between the pure guiding style and the pure reflecting style of breath regulation. Through managing the underlying mapping mechanism we can variously affect the breathing pattern of the user. A preliminary experiment was conducted to prove the concept.

Finally, two forms of multimedia-assisted breath regulation were adopted in the formal experiment to investigate the effects of the pure guiding mode and the local-mapping mode on breathing depth and frequency and the subjective experiences. Quantified analysis of bio-signal features was conducted and the results were discussed. Experiment results

show that the local-mapping mode is actually beneficial for slowing and deepening user breathing, in the meantime, reduces the drawback of inducing mental tension, which is usually the case in the common pure guiding style of breath regulation. At the end, some user feedbacks from interviews were discussed and essential factors worth noticing for constructing a breath-based application for relaxation in the future were suggested.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support and comments from the editor and the reviewers, respectively. This work is in part supported by grants from Ministry of Science and Technology of Taiwan under NSC 104-2627-E-002-001.

REFERENCES

- [1] Steptoe, A., "Invited review: The links between stress and illness." *Journal of Psychosomatic Research*, 35(6), pp. 633-644, 1991.
- [2] Schubert, C., Lambert, M., Nelesen, R., Bardwell, W., Choi, J.-B., and Dimsdale, J., "Effects of stress on heart rate complexity—a comparison between short-term and chronic stress." *Biological Psychology*, 80(3), pp. 325-332, 2009.
- [3] Tsigos, C., and Chrousos, G.-P., "Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress." *Journal of Psychosomatic Research*, 53, pp. 865-871, 2002.
- [4] Varvogli, Liza, and Christina Darviri. "Stress Management Techniques: evidence-based procedures that reduce stress and promote health." *Health Science Journal*, 5(2), pp.74-89, 2011.
- [5] Grossman, E., Grossman, A., Schein, M., Zimlichman, R., Gavish, B., "Breathing-control lowers blood pressure." *Journal of Human Hypertension*, 15(4), pp. 263-269, 2001.
- [6] Ley, R., "The Modification of Breathing Behavior: Pavlovian and Operant Control in Emotion and Cognition." *Behavior Modification*. 23(3), pp.441-479, 1999.
- [7] Moravejji, N., Olson, B., Nguyen, T., Saadat, M., Khalighi, Y., Pea, R., Heer, J., "Peripheral Paced Respiration: Influencing User Physiology during Information Work." *The 24th ACM User Interface Software and Technology (UIST) Symposium*. pp.423-427, 2011.
- [8] Yu, M.-C., Chen, J.-S., Chang, K.-J., Hsu, S.-C., Lee, M.-S., Hung, Y.-P., "i-m-Breath: The Effect of Multimedia Biofeedback on Learning Abdominal Breath." *Multimedia Modaling (MMM)*, pp.548-558, 2011.
- [9] Park, S.-H., Jang, D.-G., Son, D.-H., Zhu, W., Hahn, M.-S., "A biofeedback-based breathing induction system." *The 3rd International Conference on Bioinformatics and Biomedical Engineering (ICBBE)*, pp.1-4, 2009.
- [10] Donaldson, G. C., "The chaotic behaviour of resting human respiration." *Respiration Physiology*, vol. 88, pp. 313–321, 1992.
- [11] Hughson, R. L., Yamamoto, Y., & Fortrat, J. O., "Is the pattern of breathing at rest chaotic? A test of Lyapunov exponent." *Advances in Experimental Medicine and Biology*, vol.393, pp.15-19, 1995.
- [12] Small, M., Judd, K., Lowe, M., & Stick, S., "Is breathing in infants chaotic? Dimension estimates for respiratory patterns during quiet sleep." *Journal of Applied Physiology*, vol. 86, pp. 359–376, 1999.
- [13] Wysocki, M., Fiamma, M.-N., Straus, C., Poon, C.-S., & Similowski, T., "Chaotic dynamics of resting ventilatory flow in humans assessed through noise titration." *Respiratory Physiology & Neurobiology*, vol. 153, pp. 54–65, 2006.
- [14] Vlemincx, Elke, Ilse Van Diest, and Omer Van den Bergh. "Imposing respiratory variability patterns." *Applied psychophysiology and biofeedback*, 37(3), pp.153-160, 2012.
- [15] Yu, M.-C., Liou, J.-L., Kuo, S.-W., Lee, M.-S., Hung, Y.-P.: Noncontact Respiratory Measurement of Volume Change Using Depth Camera. In: *Proc. of IEEE EMBC*, 2371-2374 (2012)
- [16] Chan, L.-W., Chuang, Y.-F., Yu, M.-C., Chao, Y.-L., Lee, M.-S., Hung, Y.-P., Hsu, J.: Gesture-based Interaction for a Magic Crystal Ball. In: *Proc. of the ACM VRST*. 157-164 (2007)
- [17] WildDivine. Finger sensor "The Iom" [online]. Available: <http://www.wilddivine.com/>
- [18] Varvogli, L., Darviri, C., "Stress Management Techniques: evidence-based procedures that reduce stress and promote health," *Health Sci J*, vol. 5, pp. 74-89, 2011.
- [19] McCaul, K., Solomon, S., Holmes, D.: Effects of Paced Respiration and Expectations on Physiological and Psychological Responses to Threat. In: *Journal of Personality and Social Psychology*, vol. 37(4), pp. 564-571, 1979.

A Secure Network Communication Protocol Based on Text to Barcode Encryption Algorithm

Abusukhon Ahmad
Department of Computer Networks,
Al-Zaytoonah University of Jordan,
Amman, Jordan

Bilal Hawashin
Department of CIS, Al-Zaytoonah University of Jordan,
Amman, Jordan

Abstract—Nowadays, after the significant development in the Internet, communication and information exchange around the world has become easier and faster than before. One may send an e-mail or perform money transaction (using a credit card) while being at home. The Internet users can also share resources (storage, memory, etc.) or invoke a method on a remote machine. All these activities require securing data while the data are sent through the global network.

There are various methods for securing data on the internet and ensuring its privacy; one of these methods is data encryption. This technique is used to protect the data from hackers by scrambling these data into a non-readable form. In this paper, we propose a novel method for data encryption based on the transformation of a text message into a barcode image. In this paper, the proposed Bar Code Encryption Algorithm (BCEA) is tested and analyzed.

Keywords—Encryption, Decryption; Algorithm; Secured Communication; Private Key; Barcode Image

I. INTRODUCTION

Nowadays, many applications on the web allow users from the whole world to interact with them. These applications rely on securing the channels between the client and the server while sending data through the global network.

Securing a channel between a server and a client is handled using authentication (i.e. a username and a password) and one of the encryption algorithms.

There are different methods for data encryption, which are used to protect data over a network and thus build a secure channel. These techniques can be classified based on the data type (e.g. text, image, sound) of the encrypted data into three categories; namely, text encryption, image encryption, and sound encryption. Fig. 1 describes the encryption process for private-key encryption. As shown in Fig.1, the data encryption system consists of a plain text (also could be an image or a sound), which is the data before running the encryption algorithm. The encryption algorithm is the algorithm used to transfer the original data (e.g. text message) into an unreadable or a hidden form [1]. The core of the encryption algorithm is a private key used by both encryption and decryption algorithms. The encryption key is used to encrypt and decrypt data.

The decryption algorithm is an algorithm used for transforming the encrypted data into the original data [2], or simply, it is the encryption algorithm working in reverse.

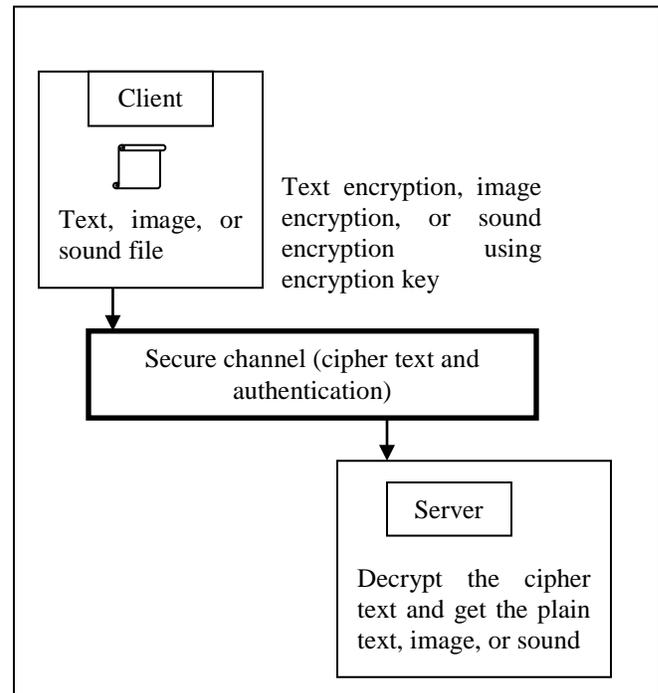


Fig. 1. Encryption process

The Internet is the richest area for hackers to perform their attacks. Hackers are unauthorized users who may attack sensitive data sent through the Internet and use false IP addresses to achieve various goals. Thus, in most of the Internet applications, verification and validation techniques are required to check the user's identity. These techniques include digital signature, and digital certificate [3]. Digital signature and digital certificate are not the focus of this research.

In general, the standard methods used for data encryption are private-key encryption (called symmetric encryption), public-key encryption (called asymmetric encryption), digital signature, and hash functions [4].

In private-key cryptography, both the sender and the receiver agree on a single key to be used for both encryption and decryption. This key is kept secret by sending it through a secure channel to the other side of the network [5].

This paper proposes a new encryption algorithm based on private-key techniques.

As it is mentioned earlier in this paper, an encryption key is used to encrypt images, text, and sounds. Image encryption techniques focus on encrypting a digital image in a specific format (e.g. png, bmp, etc.) into an unreadable image with the same image format using a specific encryption key.

One of the techniques used for image encryption is based on dividing the image into blocks and then encrypting those blocks using an encryption algorithm, the following are examples of this technique.

Nithin, Anupkumar, and Hegde [6] proposed and evaluated an image encryption algorithm (called FEAL) that is based on the DES encryption algorithm. The algorithm divides the original image into a number of blocks (16×16 blocks). Later encryption and decryption algorithms are performed using 12 keys of size 16-bit. Images used in this algorithm are gray scale images of size 256×256 resolution.

AliBaniYounes, and Janta [7] proposed an encryption algorithm based on dividing an image into blocks. These blocks are then rearranged into a transformed image (using their proposed transformation algorithm) and then the transformed image is encrypted using the Blowfish algorithm. Their work showed that increasing the number of blocks by decreasing the block's size resulted in a lower correlation and higher entropy.

Divya, Sudha, and Resmy [8] proposed a simple encryption algorithm based on dividing the image into 8×8 blocks. In their method, they proposed to encrypt a portion of a given image instead of encrypting the whole image to make the encryption process faster. In their algorithm, the resulting blocks are transformed from the spatial domain to frequency domain using the Discrete Cosine Transform (DCT). A selected DCT coefficients are then encrypted and XORed with random bits to make it difficult for hackers to guess the original message.

M.Mishra, P. Mishra, Adhikary, and Kumar [9] proposed a new method for image encryption based on Fibonacci and Lucas series.

Different techniques are used for encrypting text messages into an unreadable form. Examples of this technique are presented next.

Singh and Gilhotra [10] proposed an encryption algorithm based on the concept of arithmetic coding. In this algorithm, a given word in a text is transformed into a floating point between 0 and 1. The resulting floating number is then transformed into a binary number that is in turn encrypted to another binary number, and then the resulting binary number is converted to a decimal number.

Huang, Chi Lee, and Hwang [11] proposed a novel encryption algorithm. This algorithm generates n^2+n common secret keys in one session. It is based on the difficulty of calculating discrete logarithms problem.

Torkaman, Kazazi, and Rouddini [12] proposed a hybrid cryptosystem which is a combination of public and private cryptography. Their technique is based on a combination of cryptographic and steganography techniques. This algorithm provides a secure communication while defeating the up to

date attacks. In their work, steganography algorithm is based on DNA algorithm and is used to hide a secret key. This secret key is distributed among two parties once a network communication is established.

Krishna [13] proposed a new mathematical model in which the output of the Elliptic Curve Cryptography (EEC) algorithm, a variable value, and a dynamic time stamp are used to generate the cipher text. They compared the results from their proposed model with the results from RSA and ECC algorithms. The results from their work showed that the security strength of their proposed model is more than RSA and ECC's security strength.

Other techniques for text encryption are proposed. These techniques are used to encrypt text into musical notes. Examples of other techniques are presented next.

Dutta, Chakraborty, and Mahanti [14] proposed a novel method for encrypting a text into musical notes. In their work, they used MATLAB in which 26 alphabets and 0 to 9 numbers are considered as -12 to 23 as musical notes. A sender encrypts the text message into musical notes and sends it to a receiver. The receiver, when receiving the encrypted message, decrypts the musical notes into the original text message (i.e. the plain text).

Yamuna, Sankar, Ravichandran, and Harish [15] proposed an encryption algorithm based on the transformation of a text message into musical notes. The encryption algorithm consists of two phases; in the first phase, the text message is encrypted into a traditional Indian music. In the second phase of encryption, the Indian music notes are encrypted again into western music notes.

Dutta, Kumar, and Chakraporty [16] proposed an encryption algorithm that encrypts a text message into musical notes. The text characters of a message are replaced by mathematically generated musical notes. These musical notes and the seed value for encryption/decryption key are sent to the receiver using the RSA algorithm.

The reset of this paper is organized as follows. Section II presents the related work. Section III presents our work, including research methodology, experiments, and analysis of the proposed algorithm. Finally, section IV presents the conclusions and future work.

II. RELATED WORK

Bh, Chandravathi, and PROja [17] presented Koblitz's method and used it to map a message to a point in the implementation of Elliptic Curve Cryptography [18, 19]. A given character in a text is mapped into its ASCII code, and then this ASCII code is encrypted into a point on a curve.

Singh and Gilhorta [5] proposed an encryption algorithm which is based on the transformation of a word of text into a floating point number (n) where, $1 \geq n \geq 0$. The resulting floating point number (n) is then encrypted into a binary number (b), and then (b) is encrypted using an encryption key.

Kumar, Azam, and Rasool [20] proposed a new technique of data encryption. In this technique, three random numbers are generated, say (D1), (D2), and D3. The random number D1 is

used for rows transformation in a matrix (V). D2 is used for columns transformation, and D3 is converted into a binary number. Rows and columns transformation is based on the value of the individual bits of that binary number. Three operations are defined in order to perform the matrix transformation namely, circular left shift, circular right shift, and reverse operation.

Abusukhon and Talib [21], and Abusukhon, Talib, and Issa [22] proposed the Text-to-Image Encryption algorithm (TTIE). In their algorithm, a given text file is encrypted into an image. Each individual character in the text file is transformed into an individual pixel (a pixel with a specific color). Each pixel in the resulting image consists of three integers; namely, Red, Green, and Blue, and each integer represent a specific color density. Having a matrix of integers, they were able to perform columns and rows shuffling making it difficult for hackers to guess the plain text (i.e. the original text message).

Abusukhon [23] investigated using block cipher encryption with TTIE encryption algorithm. In their work, the plain text is divided into number of blocks say $\{b_1, b_2 \dots b_n\}$, and then each block is encrypted into an image. All images from all blocks are combined into one image. This image represents the plain text.

Abusukhon, Talib, and Nabulsi [24] analyzed the encryption time for the TTIE encryption algorithm. They divided the total time of their experiment into six parts. The results from their work showed that the most significant time is the time required to store the encrypted data into the hard disk.

Abusukhon, Talib, and Almimi [25] proposed the Distributed Text-to-Image Encryption Algorithm (DTTIE) in order to improve the speed of the TTIE algorithm when a large scale data collection is used. They proposed to distribute the Text-to-Image Encryption Algorithm (TTIE) proposed in [21, 22] among seven nodes, where each node encrypts a partition of the data collection. They evaluated the speed up of their system when a large data collection (5.77 Giga Bytes) is used.

Our work differs from the work presented in [21, 22, 23, 24, 25]. In their work, each letter in the plain text is encrypted and mapped into one colored pixel (for example, letter "a" is represented as red pixel, letter "b" is represented as green pixel and so on). In this paper, each letter is encrypted into a black bar. Each black bar consists of a specific number of black pixels. In this paper we propose the Bar Code Encryption Algorithm (BCEA).

III. OUR WORK

In this paper, Java NetBeans is used as a vehicle to carry out our experiments. All algorithms are implemented in Java, and build from scratch including encryption and decryption algorithms, client code, and the server code.

A. Machine Specifications

Our experiments are carried out using a single machine with the following specifications; processor Intel (R) core (TM)2, Duo CPU T5870 @ 2.00GHz, installed memory (RAM) 2.00GB operating system Windows 7 Ultimate and hard disk 24.5 GB (free space).

B. Data Sample

The data sample is created and stored in a notepad file. The data sample is shown Fig. 2.

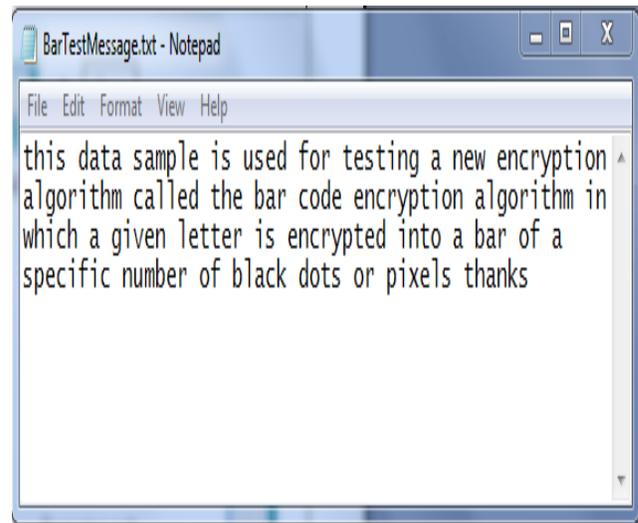


Fig. 2. Tested data

C. Research Methodology and Evaluation

The plain text shown in Fig. 2 is allocated at the client node. The client node encrypts the plain text using the proposed algorithm (BCEA), produces a bar code image, and then the resulting image is sent to the server. The server decrypts the received image and then displays the plain text message. To evaluate our system; the plain text message is checked and compared with the original one (i.e. the message sent by the client).

D. Our Experiment

In this experiment, encryption and decryption algorithms, a client code, and a server code are built from scratch using java. The system architecture is shown in Fig. 3.

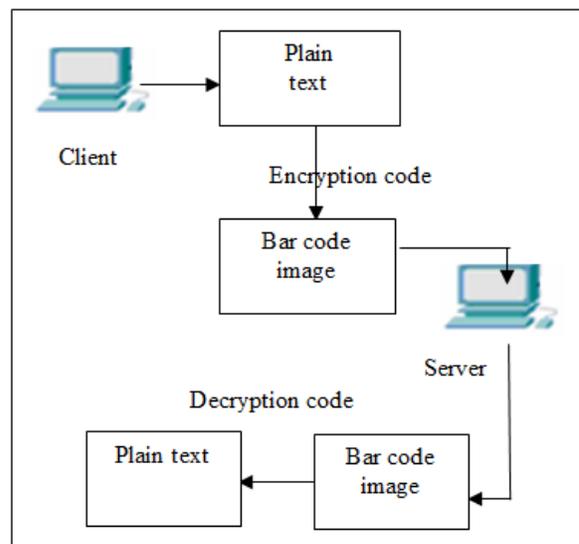


Fig. 3. The system architecture

In this experiment, the plain text shown in Fig. 2 is placed on the client side. The client uses the proposed encryption algorithm (BCEA) for encrypting the plain text. The output of the BCEA algorithm is an image of type ".png" as shown in Fig.4.

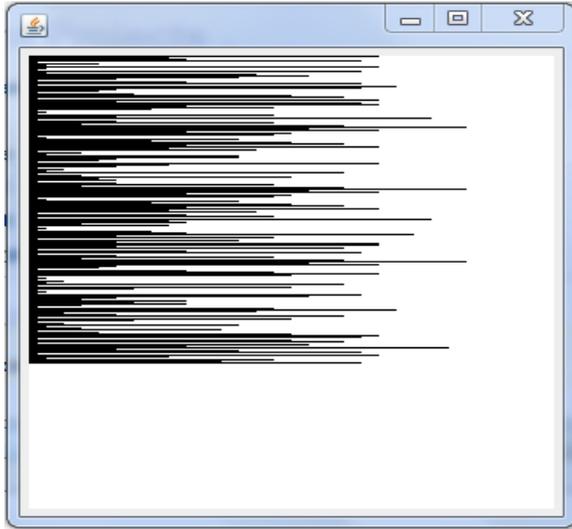


Fig. 4. The Barcode image results from running the BCEA algorithm

Using the proposed algorithm (BCEA), each letter from the plain text is encrypted into one bar. Each bar consists of a number of black pixels and has a specific length (the bar's length is measured in pixels). For example, in our experiment the letter "a" is encrypted as one bar of length = 10 black pixels. Letter "b" is encrypted as another bar of length = 20 pixels, and so on. We leave two white bars between each two black bars in order to clarify the bar code shape.

To verify our algorithm, the client encrypts the sample shown in Fig. 2, and then the encrypted text (.png file) is sent to the server. The server decrypts the .png file, and gets the original message shown in Fig. 5.

Encrypting the plain text into a bar code image makes it difficult for hackers to guess that each black bar in the image represents a specific letter from the plain text.

The main steps of encryption and decryption for BCEA algorithms are described in Fig. 6 (a) and (b).

In addition, we test the efficiency of our algorithm (BCEA) with respect to encryption time when different data collection sizes are used as shown in Fig. 7.

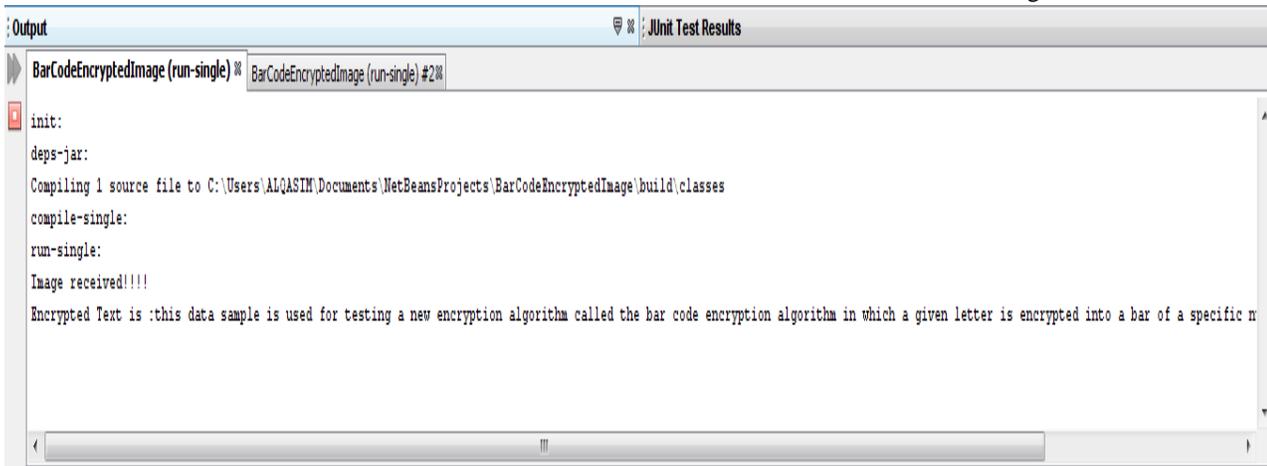


Fig. 5. Encryption algorithm running in reverse

(a) Encryption Algorithm

```
// System set up
1. Determine the minimum size (X) and the maximum size (Y) of the bar.
2. Select an integer number between (X) and (Y) for each letter in the alphabet set [A to Z].
   //This number represents the bar length corresponding to a specific letter.
   // letter A → 10 black Pixels, letter B → 20 black pixels and so on.

// do the encryption
3. Read the plain text and store it in an array of characters (chr)
4. For (int i=1; i<= chr.length; i++)
{
  Read chr [i]           // read a letter (L) from chr
  Search for the bar length (L) correspond to the current letter
  Create a black bar whose length is (L) // (see step 2)
  Draw the bar on the result image (. png)
  Draw two white bars on the image // in order to separate the black bars from each other
}
```

(b) Decryption Algorithm

```
1. Read the image (the cipher text)
2. Let the String "OriginalMessage" = null
3. While not the end of image // determined by the image size
{
  Extract a bar from the image

  If the extracted bar is a white bar then ignore // discard white bars since they do not
                                                    //represent any letter from the plain text

  Else
  Calculate the bar length // count the number of black pixels
  Search for the bar length and retrieve the corresponding letter
  OriginalMessage = OriginalMessage + the current letter // + means concatenation
}
```

Fig. 6. The main steps of encryption and decryption for the BCEA algorithm

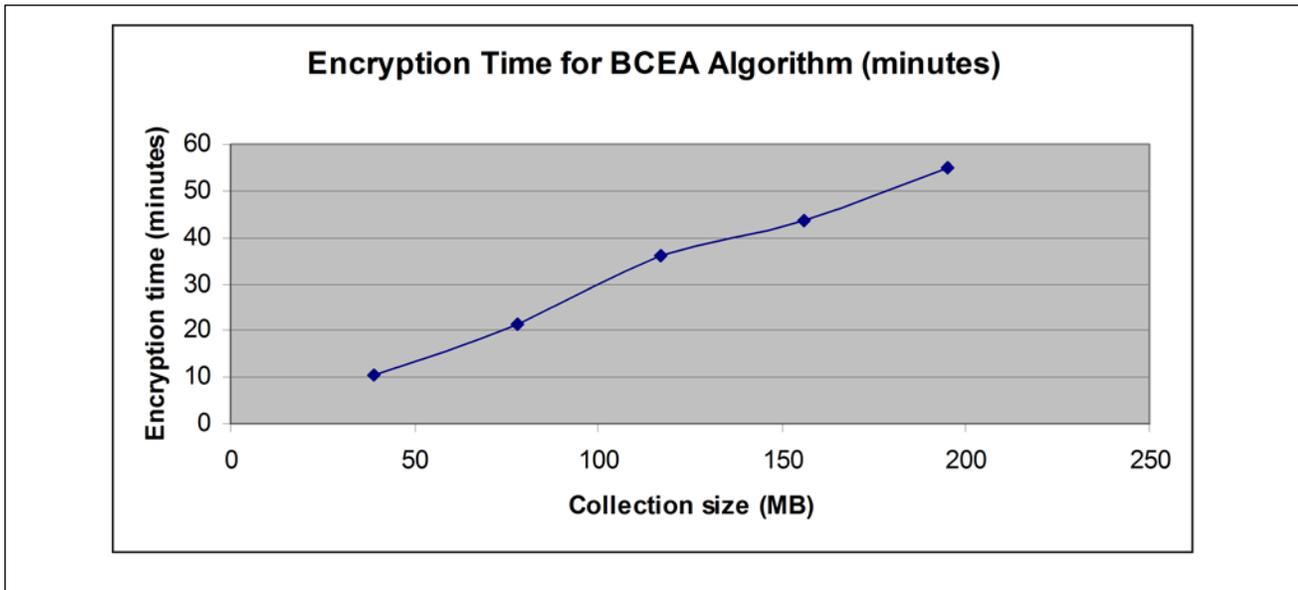


Fig. 7. Measuring the efficiency of (BCEA) encryption algorithm

To achieve our goal, five data collection are generated randomly. The first data collection consists of 100 groups and each group contains 100 text files. Each text file consists of 10 words and each word consists of seven characters generated randomly from the alphabet list. The second data collection consists of 200 groups, and the third data collection consists of 300 groups and so on. As shown in Fig. 7, the encryption time is proportional to collection size.

E. Analysis of the proposed algorithm (BCEA)

In this section, the maximum number of permutations (How many times a hacker may try before guessing the encryption key and getting the original text) is calculated.

The image resulting from the BCEA algorithm is a matrix of integers (M) having (R) rows and (C) columns. Suppose that the maximum row length (i.e. the number of columns in a row) is (mR), and the maximum number of rows is (nR); then the length of each bar in the image is limited by (mR). Each letter (L) from the plain text is encrypted as one bar and thus it allocates one row of the matrix (M) regardless the bar length.

The maximum number of key permutations (the range or the domain from which a key is picked out) is limited to (mR).

The number of letters (nL) in the plain text is limited to (nR). In other words, the number of letters in the plain text is limited by the maximum image size supported by Netbeans (in this paper). Thus, the maximum number of key permutations (P) provided by the BCEA is calculated as shown in (1).

$$P = \frac{mR!}{(mR - 26)!} \tag{1}$$

Thus, suppose that mR = 1000 pixels, then P is calculated as follows.

$$P(1000, 26) = (1000 \times 999 \times 998 \dots \times 974!) / 974! \\ = 7.2 e +77.$$

In our experiment, we use the key permutation where the letter "a" is represented as a bar of length equals 10 black pixels, the letter "b" is represented as a bar of length equals 20 black pixels and so on, Table 1 shows one of the key permutations.

TABLE I. POSSIBLE KEY PERMUTATION

Letter	a	b	c	d	...	z
Bar length (in pixels)	10	20	30	40	...	260

IV. CONCLUSIONS AND FUTURE WORK

In this paper, a novel encryption algorithm, the Bar Code Encryption Algorithm (BCEA), is proposed and tested. The BCEA is based on encrypting the plain text into a bar code image, where each letter in the plain text is encrypted into black bar consists of a specific number of black pixels.

The decryption algorithm is also tested where the plain text (the original message) is produced from the bar code image. Also, in section III-D, we measured the efficiency of the (BCEA) on encryption time, where different sizes of data collections are used.

Section III-E showed that the maximum number of key permutations is limited by the maximum row length (mR) of the resulting image.

The (BCEA) algorithm could be used for e-mail encryption, off-line data encryption, as well as online data encryption. For example, it can be used as a logistics barcode system (in packaging system), or as online Quick Response (QR) barcode for E-commerce.

In future, we propose to investigate the efficiency of the (BCEA) algorithm when a huge data size (multi Gigabytes) is used as well as to compare the efficiency of our proposed algorithm with the efficiency of other algorithms such as the TTIE algorithm with respect to the encryption time.

ACKNOWLEDGMENT

We would like to acknowledge and extend our heartfelt gratitude to Al-Zaytoonah University of Jordan.

REFERENCES

- [1] K.Lakhtaria "Protecting computer network with encryption technique: a study", International Journal of u- and e-service, Science and Technology, Vol. 4, No. 2, pp 43-52, 2011.
- [2] A.Chan, "A security framework for privacy-preserving data aggregation in wireless sensor networks", ACM transactions on sensor networks, Vol. 7, No. 4, 2011. [Available online at: <http://individual.utoronto.ca/aldar/paper/2011/cda-journal-tosn.pdf>]. Accessed on 25-03-2015.
- [3] S. Goldwasser, S.Micali, R. L.Rivest, "A digital signature scheme secure against adaptive chosen-message attacks", SIAM Journal of Computing Vol. 17, No.2, pp 281-308,1998.
- [4] B. Zaidan, A.Zaidan, A. Al-Frajat, and H. Jalab, "On the differences between hiding information and cryptography techniques: an overview", Journal of Applied Sciences Vol. 10, No. 15, pp 1650-1655,2010.
- [5] A. Singh, R. Gilhorta, "Data security using private key encryption system based on arithmetic coding", International Journal of Network Security and its Applications (IJNSA) Vol. 3, No. 3, pp. 58-67,2011.
- [6] N. Nithin, M.B. Anupkumar, G. P. Hegde, "Image encryption based on FEAL algorithm". International Journal of Advances in Computer Science and Technology, Vol.2, No.3, pp 14-20,2013.
- [7] M. Ali BaniYounes, A. Jantan, "Image encryption using block-based transformation algorithm". International Journal of computer science (IJCS). Vol.35 No. 1. pp 407-415, 2008.
- [8] V.V Divya, S.K. Sudha, and V.R. Resmy, "Simple and secure image encryption". International Journal of Computer Science Issues (IJCSI). Vol. 9, No. 3, pp 286-289, 2012.
- [9] M. Mishra, P. Mishra, M.C. Adhikary, S. Kumar, "Image encryption using Fibonacci-Lucas Transformation". International Journal on Cryptography and Information Security (IJCIS). Vol.2, No.3, pp 131-141, 2012.
- [10] A. Singh, and R. Gilhorta, "Data security using private key encryption system based on arithmetic coding". International Journal of Network Security and its Applications (IJNSA). Vol. 3, No. 3, pp 58-67,2011.
- [11] L. Huang, C. Chi Lee, and M. Hwang, "A n^2+n MQV key agreement protocol". The International Arab Journal of Information Technology. Vol. 10, No. 2, pp 137-142,2013.
- [12] M.R.N. Torkaman, N.S.Kazazi, and A. Rouddini, "Innovative approach to improve Hybrid Cryptography by using DNA steganography". International Journal on New Computer Architectures and Their Applications (IJNCAA). Vol.2 No. 1, pp 224, 235,2012.
- [13] A.V. Krishna, "Time stamp based ECC encryption and decryption". The International Arab Journal of Information Technology. Vol. 11, No. 3. pp 276-281, 2014.
- [14] S. Dutta, S. Chakraborty, and N.C. Mahanti, "A novel method of hiding message using musical notes". The International Journal of Computer Applications . Vol. 1, No. 16. pp 76-79, 2010.
- [15] M. Yamuna, A. Sankar, S.Ravichandran, and V. Harish, "Encryption of a Binary String using music notes and graph theory". International Journal of Engineering and Technology (IJET). Vol. 5, No. 3. pp 2920-2925, 2013.
- [16] S. Dutta, C. Kumar, and S. Chakraborty, "A Symmetric Key algorithm for cryptography using music". International Journal of Engineering and Technology (IJET). Vol. 5, No. 3. pp 3109- 3115,2013.
- [17] P. Bh, D. Chandravathi, P. PROja, "Encoding and decoding of a message in the implementation of Elliptic Curve cryptography using Koblitz's method", International Journal of Computer Science and Engineering, Vol. 2, No. 5, pp 1904-1907, 2010.
- [18] N. Koblitz, "Elliptic Curve cryptosystems", Mathematics of computation Vol. 48, No. 177, pp 203-209, 1987.
- [19] N. Koblitz, "A course in number theory and cryptography". 2nd. ed. Springer-Verlag, 1994.
- [20] K.M. Kumar, M.S.Azam, S.Rasool, "Efficient digital encryption algorithm based on matrix scrambling technique", International Journal of Network Security and its Applications (IJNSA) Vol. 2, No. 4, pp 30-41,2010.
- [21] A. Abusukhon, M.Talib, "A novel network security algorithm based on Private Key encryption", International Conference on Cyber Security, Cyber Warfare and Digital Forensic. Kuala Lumpur, Malaysia, 2012.
- [22] A. Abusukhon, M. Talib, and O. Issa, "Secure network communication based on text to image encryption", International Journal of Cyber-Security and Digital Forensics (IJCSDF), The Society of Digital Information and Wireless Communications (SDIWC) Vol. 1, No. 4, pp 263-271, 2012.
- [23] A. Abusukhon, "Block cipher encryption for Text-to-Image Encryption algorithm", International Journal of Computer Engineering and Technology (IJCET) Vol. 4, pp 50-58, 2013.
- [24] A. Abusukhon, M. Talib, and M. Nabulsi, "Analyzing the efficiency of Text-to-Image Encryption algorithm", International Journal of Advanced Computer Science and Applications (IJACSA) Vol. 3, No. 11, pp 35 – 38,2012.
- [25] A. Abusukhon, M. Talib, and H. Almimi, "Distributed Text-to-Image Encryption algorithm", International Journal of Computer Applications Vol. 106, No. 1. [Available online at : <http://research.ijcaonline.org/volume106/number1/pxc3899518.pdf>]. Accessed on 25-03-2015,2014.

Comparison Contour Extraction Based on Layered Structure and Fourier Descriptor on Image Retrieval

Cahya Rahmad

Departemen of Information Technology
State Polytechnics of Malang
Malang East Java, Indonesia

Kohei Arai

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract—In this paper, a new content-based image retrieval technique using shape feature is proposed. A shape features extracted by layered structure representation has been implemented. The approach is extract feature shape by measuring the distance between centroid (center) and boundaries of the object that can capture multiple boundaries in the same angle, an object shape that has some points with the same angle. Once an input taking into account, the method will search most related image to the input. The correlation between input and output has been defined by specific role. Firstly the input image has to be converted from RGB image to Grayscale image and then follow by edge detection process. After edge detection process the boundary object will be obtained and then calculate the distance between the center of an object and the boundary of an object and put it in the feature vector and if there is another boundary on the same angle then put it in the different feature vector with a different layer. The experiment result on the plankton dataset shows that the proposed method better than other conventional Fourier descriptor method.

Keywords—Cbir; Mlccd; extract features; rgb; Fourier descriptor; shape; retrieval

I. INTRODUCTION

The Content-Based Image Retrieval (CBIR) technique uses image content to search and retrieve digital images. Basically, CBIR systems try to retrieve images similar to a user-defined specification or pattern (e.g., shape sketch, image example). Their goal is to support image retrieval based on content properties (e.g., shape, color, texture) [1], CBIR is also based on the idea of extracting visual features from the image and using them to index images in a database. Content-based image retrieval systems were introduced to address the problems associated with text-based image retrieval. CBIR is a set of techniques for retrieving semantically-relevant images from an image database based on automatically-derived image features[2]. Content-based image retrieval also known as query by image content is a technique which uses visual content that well known as features for extracting similar images from an image in a database[3][4][5]. Image database every time become bigger and it makes a problem dealing with database organization so the necessity of efficient algorithm is obvious needed [6].

On The Content-based Image Retrieval local feature of an image is computed at some point of interest location. In order to recognize the object firstly the image has to be represented by a feature vector. These feature vectors are converted to a different domain to make simple and efficient image

characteristic, classification and indexing. Many techniques to extract the image feature is proposed [7][8][9][10].

The shape is one of the primary features in Content-Based Image Retrieval (CBIR). The shape is also one of an important visual feature of an image and used to describe image content. Among them is methods based Fourier descriptors (FDs), Fourier descriptors are obtained by applying Fourier transform on shape boundary, The concept of Fourier descriptor (FD) has been widely used in the field of computational shape analysis Fourier descriptor [11] [12]. The idea of the FD (Fourier Descriptor) is to use the Fourier transformed boundary as a shape Feature. Suppose a shape signature $Z(u)$ is a 1-D function that represents 2-D areas or boundaries. The discrete Fourier transform of a signature $z(u)$ is defined as follows:

$$a_n = \frac{1}{N} \sum_{u=0}^{N-1} Z(u) e^{-j2\pi nu/N} \quad (1)$$

where $n = 0, 1, 2, \dots, n-1$. The coefficients a_n ($n=0, 1, \dots, N-1$) are called the Fourier descriptors (FDs) of the shape.

II. PROPOSED METHOD

The algorithm of proposed method is described as follow:

- 1) Input image from database image / Query Image
- 2) Convert RGB image To Gray Image
- 3) Edge detection
- 4) Morphology Filter
- 5) Construct feature vector using multi-layer centroid contour distance (MLCCD)
- 6) Comparison for similarity retrieval
- 7) Display Result based on distance measure

Image from database image or from query image convert from RGB image into Grayscale image then implement the canny filter to detect edge position then use morphology filter to ensure the shape of object clear. Then local feature of an image at some point at interest location is computed. The feature vector is computed by measuring a distance between a center of object and point in the boundary object Then the result is placed to the feature vector layer by layer (see fig.1). and then the feature vector that obtained from database image and Query Image be compared each other (similarity process) then display the result. The retrieval result is not a single image but a list of image ranked by their similarity. in these

case if the distance between feature representation of an image in database image and feature representation of image query small enough then it to be considered as similar.

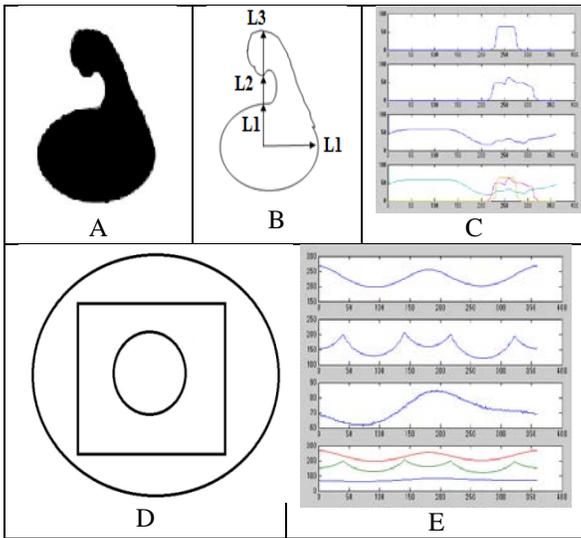


Fig. 1. An object and its feature vector layer by layer

Figure 1A is example object before edge detection process. Figure 1B is a boundary of an object after edge detection process, when the angle 0 there is one point have to be captured, the distance between centroid and boundary object is placed in the layer one. However, when the angle is 270 degree there are three points have to be captured (clockwise direction), the distance between centroid and boundary object is placed in the layer one, layer two and layer three (see figure 1C). Figure 1D is example shape of an object with tree point for every angle. Similarly, the distance between centroid and boundary object in the Figure 1D is placed in the layer one, layer two and layer three (see Figure 1E).

In order to obtain the MLCCD firstly position of the centroid have to be computed (see equation 1) then calculate the distance between the centroid and the boundary of an object by using Euclidean distance, see equation 2) repeat this method for another boundary in the same angle and different angle.

The position of the centroid is:

$$X_c = \frac{X_1+X_2+X_3+\dots+X_n}{n}, \quad Y_c = \frac{Y_1+Y_2+Y_3+\dots+Y_n}{n} \quad (2)$$

where:

X_c = position of the centroid in the x axis

Y_c = position of the centroid in the y axis

n = Total point in the object

(every point have x position and y position)

After the centroid was obtained then calculate a distance between centroid to every point in the boundary, Suppose there is t point in the boundary of an object the distance every point in the boundary with centroid is:

$$Dis(n) = \sqrt{(x(n) - x_c)^2 + (y(n) - y_c)^2} \quad (3)$$

where:

n = number point in the boundary of object (1,2,..t)

t = total point in the boundary

x_c = position center in the x axis

y_c = position center in the y axis

$x(n)$ = position point number n in the x axis

$y(n)$ = position point number n in the y axis

The computed distances are saved in a vector layer by layer. In order to achieve rotation invariance, scale invariance and translation invariance implementation shifting and normalization to this vector is needed.

III. SIMILARITY AND PERFORMANCE

To test the effectiveness of the approach, the similarity and performance measure is conducted. The comparisons that determine a similarity between images depend on the representations of the features and the definition of an appropriate distance function. The similarity metric is very important on the retrieval result.

The similarity measure is computed by using Euclidean distance (See Equation 3) between feature representation of an image in database image and feature representation of image query. This feature representation is image feature that refer to the characteristics which describe the contents of an image.

The retrieval result is a list of image ranked by their similarity. Suppose $S1$ and $S2$ are shape of object represented layer by layer of feature vectors each ($Q1, Q2, \dots, Qn$) and ($D1, D2, \dots, Dn$) then the Distance between $S1$ and $S2$ is:

$$dis(FQ, FD) = \sqrt{\sum_{j=1}^n (Q_j - D_j)^2} \quad (4)$$

where:

FQ = Feature vector of a query image.

FD = Feature vector of image in dataset

n = Number element of the feature vector

If the distance between feature representation of image query and feature representation of an image in dataset small enough then it to be considered as similar, For example, a distance of 0 have meant an exact match with the query and 1 mean totally different. Base the ranked of that similarity then the retrieval results be displayed.

The performance of Content-Based Image Retrieval: CBIR system is calculated by showing an image with X top ranking from the dataset. Precision is The common way method to evaluate the performance of the CBIR system, Formula Precision is:

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \quad (5)$$

The precision measures the retrieval accuracy, it is the ratio between the number of relevant images retrieved and the total number of images retrieved (See Equation 4).

IV. EXPERIMENT AND ANALYSIS

Image database of phytoplankton [13] for experiment to real data In order to show the feasibility of the shape recognition scheme is used. Red tide occurs in a nutrition rich ocean. Nutrition rich water makes chlorophyll-a then

phytoplankton is increase thus red tide occurs. Algal blooms (red tides) are a phenomenon of clear ecological importance in many regions of the world.

Caused by a nutrient influx (e.g. agricultural pollution) into the ocean, by either natural or anthropogenic causes, Red tide can be toxic to marine life and humans under certain conditions. They are a significant problem not only for fisherman but also ocean biologist. Red tide is one of measure for representation of ocean healthy[14] [15].

Figure 2 shows Example of phytoplankton image. In order to detect red tide, many researchers check phytoplankton in water sampled from the ocean with a microscope. Immediately after they check phytoplankton, they have to identify the species of phytoplankton. Image retrieval is needed for identification. The proposed method is to be used for image retrieval and identification.

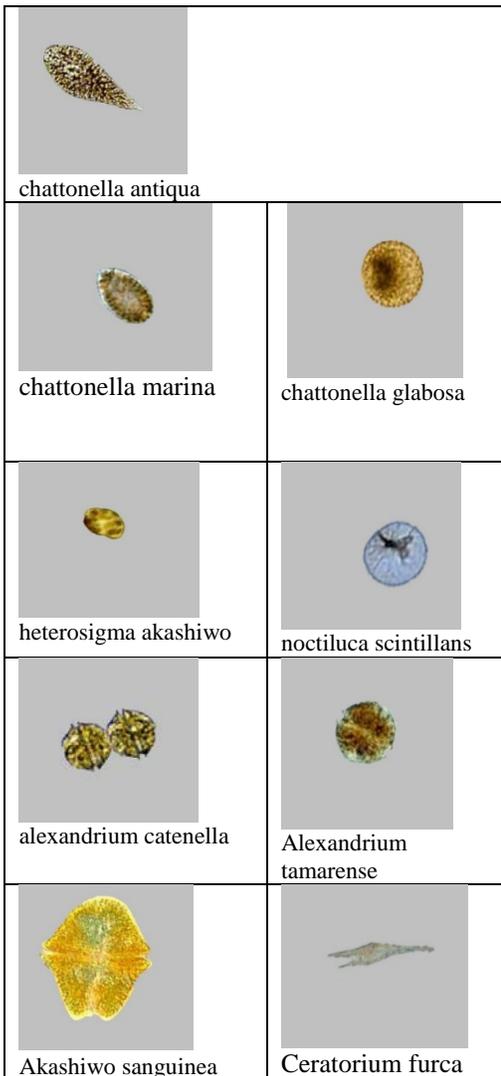


Fig. 2. Example of phytoplankton image

TABLE I. AVERAGE PRECISION ON PHYTOPLANKTON DATASET

Number Group	Total Image	Phytoplankton name	Fourier descriptor	Proposed Method
1	18	chattonella antiqua	85	91
2	17	chattonella marina	86	90
3	17	chattonella glabosa	83	87
4	17	heterosigma akashiwo	84	88
5	17	noctiluca scintillans	84	87
6	20	alexandrium catenella	85	93
7	22	Alexandrium tamarense	84	94
8	23	Akashiwo sanguinea	85	93
9	24	Ceratorium furca	83	94
Average			84.333	90.777

The experiment on phytoplankton image dataset in Table 1 is precision measure base on equation 5, Average precision result by using the proposed method is higher 3 percent (see in group 5) up to 11 percent (see in group 9) rather than another method. From this table and figure, it may say that the proposed method is superior to the conventional method for all cases by approximately 6.444 %.

V. CONCLUSION

In The comparison between the proposed contour extraction based on layered structure representation and the Fourier descriptor based on image retrieval have propose a new approach to extract features of an object shape that have some points with the same angle.

a novel approach feature shape by measuring a distance between centroid (center) and a boundary of an object that can capture multiple boundaries in the same angle is developed. The experiment results on phytoplankton dataset demonstrate a new approach better than another method. the proposed method is superior to the conventional method for all cases by approximately 6.444 %.

REFERENCES

- [1] João Augusto da Silva Júnior, Rodney Elias Marçal, Marcos Aurélio Batista, "Image Retrieval: Importance and Applications ", Workshop de Vis~ao Computacional - WVC 2014.
- [2] F. Long, H. Zhang, and D. Feng, "Fundamentals of content based image retrieval," "Multimedia Information Retrieval and Management. Technological Fundamentals and Applications,"Multimedia Signal Processing Book, Chapter 1, Springer-Verlag, Berlin Heidelberg New York, pp. 1–26, 2003.
- [3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," ACM Computing Surveys, vol. 40, no. 2, pp. 1–60, Apr. 2008.
- [4] C. Carson and S. Belongie, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," Pattern Analysis and ..., 2002.

- [5] P. B. Thawari and N. J. Janwe, "CBIR BASED ON COLOR AND TEXTURE," vol. 4, no. 1, pp. 129–132, 2011.
- [6] Fotopoulou F and Economou G, "Multivariate angle scale descriptor of shape retrieval," Proc. Signal. Process Appl. Math. Electron ..., pp. 105–108, 2011.
- [7] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," Pattern Recognition, vol. 40, no. 1, pp. 262–282, Jan. 2007.
- [8] N. Singhai and S. K. Shandilya, "A Survey On: Content Based Image Retrieval Systems," International Journal of Computer Applications, vol. 4, no. 2, 2010.
- [9] D. Zhang and G. Lu, "Review of shape representation and description techniques," Pattern Recognition, vol. 37, no. 1, pp. 1–19, Jan. 2004.
- [10] D. Zhang and G. Lu, "A Comparative Study of Three Region Shape Descriptors," no. January, pp. 1–6, 2002.
- [11] FUNTOURA Costa and C. R. M., "Shape Analysis and Classification: Theory and Practice," Boca Raton, FL.: CRC Press, 2001.
- [12] W. Jun, G. E. Q. Jeffrey, G. Feng, and S. Hai-Jun, "Fourier Descriptors with Different Shape Signatures: a Comparative Study for å," vol. 47, no. 50728503, 2011.
- [13] "Kohei Arai and Yasunori Terayama, Polarized radiance from red tide, Proceedings of the SPIE Asia Pacific Remote Sensing, AE10-AE101-14, Invited Paper, 2010."
- [14] "Kohei Arai, Red tides: combining satellite- and ground-based detection. 29 January 2011, SPIE Newsroom. doi: 10.1117/2.1201012.003267, <http://spie.org/x44134.xml?ArticleID=x44134>."
- [15] C. J. Walsh, S. R. Leggett, B. J. Carter, and C. Colle, "Effects of brevetoxin exposure on the immune system of loggerhead sea turtles," Aquat. Toxicol. 97, no. 4pp. 293-303, 2010. doi:10.1016/j.aquatox.2009.12.014.

AUTHOR PROFILE

Cahaya Rahmad: He received BS degrees from Brawijaya University Indonesia in 1998 and MS degrees from Informatics engineering at Tenth of November Institute of Technology Surabaya Indonesia in 2005. He is a lecturer in The State Polytechnic of Malang Since 2005. He received Doctoral degrees at Saga University japan in 2013, His interest researches are image processing, data mining and patterns recognition.

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 and 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.

Arabic Sentiment Analysis: A Survey

Adel Assiri^{1,2}, Ahmed Emam Ph.D^{1,3}, Hmood Aldossari¹

¹Information Systems Department, King Saud University, Riyadh, Saudi Arabia

²Abha Technology College, TVTC

³Menoufia University, Menoufia, Egypt

Abstract—Most social media commentary in the Arabic language space is made using unstructured non-grammatical slang Arabic language, presenting complex challenges for sentiment analysis and opinion extraction of online commentary and micro blogging data in this important domain. This paper provides a comprehensive analysis of the important research works in the field of Arabic sentiment analysis. An in-depth qualitative analysis of the various features of the research works is carried out and a summary of objective findings is presented. We used smoothness analysis to evaluate the percentage error in the performance scores reported in the studies from their linearly-projected values (smoothness) which is an estimate of the influence of the different approaches used by the authors on the performance scores obtained. To solve a bounding issue with the data as it was reported, we modified existing logarithmic smoothing technique and applied it to pre-process the performance scores before the analysis. Our results from the analysis have been reported and interpreted for the various performance parameters: accuracy, precision, recall and F-score.

Keywords—Arabic Sentiment Analysis; Qualitative Analysis; Quantitative Analysis; Smoothness Analysis

I. INTRODUCTION

Sentiment analysis is a type of natural language processing (NLP), where NLP or computational linguistics, is the scientific study of human languages from a computational perspective [1]. Natural language processing is an extensive field covering such applications and investigations as human language translation/generation/comprehension, speech & named entity recognition, question answering and information retrieval, word/topic segmentation, and relationship extraction. Sentiment Analysis (SA) is using natural language processing, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit [2]. Sentiment analysis has also been referred to as opinion mining (OM) and is concerned with the analysis of human opinion, sentiment, and emotion about specific entities (such as food, products, organizations, etc.) and issues (politics, news, etc.) [3][4][5].

Sentiment analysis, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features [38][4][5]. This paper reviews efforts to build SA systems for Arabic. The rest of this paper has arranged as following: After a brief discussion of the

properties of Arabic language in Section 2, we review sentiment analysis process in Section 3. Related work and qualitative analysis for Arabic presented in Section 4, we presented quantitative analysis in Section 5, conclusion and future work in Section 6.

II. ARABIC LANGUAGE CHALLENGES

As an important player in international politics and the global economy, the Arab world is the focus of many multi-national interest groups and analysts who endeavour daily to decipher sentiments on issues like oil and gas prices, stock market movements, politics and foreign policy, emanating from this part of the world. The resulting chatter being in the Arabic language, there is a great need for natural language analysis of large amounts of Arabic language text and documents to support the required sentiment extraction. As described in the foregoing, the relative importance of the Arabic language in global communications demands a proportional amount of interest and research for natural-language processing of large amounts of Arabic language text and documents to facilitate sentiment extraction for industrial use [6][7][8].

The reality, however, is that there is relatively little available support for Arabic-language sentiment analysis, majorly for the following reasons: (1) relatively limited scholarly work and research funding in this area, when compared to other-language studies, especially English. (2) Morphological complexities and dialectal varieties of the Arabic language which require advanced pre-processing and lexicon-building steps beyond what is applicable for the English language domain [7][8]. This limits the potential applications of current tools and custom tools for Arabic SA may not be easy to come by, may be limited in current functionality, or may not be freely available. Farra et al [9] illustrated the challenges of Arabic-language sentiment analysis: the existence of many inflectional and derivation forms - where words have transitional meanings depending on position within a sentence, and the type of sentence (verbal or nominal). Multiple word prefixing, suffixing, affixing, and diacritical forms add high-order dimensionality for words, where the same three-letter root can generate different words in each case [9]. The nature of the Arabic language identifies the need for custom tools for Arabic SA that will be capable of identifying these diacritics and performing efficient automated POS tagging for Arabic text. As explained, morphological analyzers should be used in tandem with POS taggers to carry out root extraction as well as prefix, suffix and affix extraction. Currently, tools like MADA (Morphological Analysis and Disambiguation for Arabic) and BAMA

(Buckwalter Arabic Morphological Analyzer) are being used by Arabic language sentiment analysis researchers but these tools are far from being advanced, and there is still a need for complex and more capable POS taggers to be developed for this domain, among other issues.

III. SENTIMENT ANALYSIS PROCESS

Sentiment Analysis generally consists of three main steps: pre-processing, feature selection and sentiment classification.

A. Preprocessing

The text documents contain rich textual information such as words and phrases, punctuation, abbreviation, emoticons etc. They also tend to have misspelling, duplicate-characters (such as "coool"), especially for social media text. Direct application of SA methods on such text usually leads to poor performance. Therefore, pre-processing is typically conducted to convert the text into textual features that could be fit into the SA methods. Once the pre-processed text features are extracted, they are ready to be fit in the next phase of SA – Feature Selection [10][11]. Pre-processing is usually based on NLP techniques such as tokenization (splitting the sentences into words), de-noising (remove special characters, capture symbols for emotions), normalization (remove duplicate characters, identify root words etc.), stop-words removal (remove the stop words and the words which are of no use to sentiment analysis), stemming (return the word to its stem or root), lemmatization (convert inflected words to their root form) etc.

Haddi et al. [10] studied the role of text pre-processing in sentiment analysis, including online text cleaning, white space removal, expanding abbreviation, stemming, negation and stop words removal. For stop words, they constructed list of domain specific stop words which are not standard stop words but carry no information for the specific domain. Bao et al. [11] evaluated the effects of text pre-processing in twitter sentiment analysis. They first considered username, hashtags, emotions, digital symbols, single letters, punctuations and other non-alphabetic symbols for de-noising. Then they conducted five steps for pre-processing: URLs features reservation, negation transformation, repeated letters normalization, stemming and lemmatization. They showed that sentiment classification accuracy rises when URLs features reservation, negation transformation and repeated letters normalization are employed while descends when stemming and lemmatization are applied.

B. Feature Selection

The outputs of pre-processing are the extracted text features. Many text features are considered for SA: unigram (individual words), bigram (two consecutive words), or n-grams (n consecutive words) and either their presence for binary weighting or their frequency to indicate their relative importance; words and phrases commonly used to express opinions words and phrases commonly used to express intensification of opinions negative words that change the opinion orientation; part-of-speech (POS) to find adjectives that contains opinion information, emoticon (special characters to represent emotions). Many words in the text do not have an impact on the general orientation of it. Therefore,

keeping those words makes the dimensionality of the classification problem high and hence the classification more difficult. These words may also contain noise for the classification problem [12][13][14]. The goal of feature selection is to select important text features out of the pool of all extracted ones. Generally speaking, feature selection methods can be categorized into filter methods and wrapper methods. Filter methods rank the features according to certain metric and select the top-ranked features. Wrapper methods, on the contrary, select the best subset of features by generation and evaluation of different subsets with a classifier. Therefore, the selected features tend to be classifier specific, namely they might perform well using the specific classifier that is used for the selection, but not necessarily well with other classifiers.

The work by Yu and Wu [12] presented a 'contextual entropy model' based on basic point-wise mutual information (PMI) to perform seed word expansion originating from a small corpus of stock market news articles. The model estimates the similarity between words and seed words by comparing their relative contextual distributions using an entropy system and selecting high-match entries. Elawady et al. [13] evaluated the performance of mRMR (minimum redundancy maximum relevance), IG (information gain) and hybrid method based on Rough set theory and IG. They showed that mRMR has better performance compared with IG and the hybrid method has the best performance for sentiment analysis tasks. Agarwal and Mittal [14] considered using text features such as unigram, bigrams, the concatenation of them and POS (parts of speech). They also compared the performance of mRMR and IG and showed that mRMR is superior to IG for sentiment analysis tasks.

C. Sentiment Classification

Sentiment classification techniques are usually divided into supervised, unsupervised and semi-supervised approaches. Supervised learning uses training data to process extracted text features by adopting machine learning techniques. Unsupervised learning in the sentiment analysis context relies on robust sentiment lexicons with a sizeable number of terms with known polarity and the application of statistical-semantic weighing and distribution schemes to apply polarities to unknown words and determine the polarity of blocks of text. We can further divide unsupervised methods into dictionary-based and corpus-based relative to how the lexicon is built. [15][16][17]. Dictionary-based approach carries out a forked distributed search (two forks: antonym and synonym) for each opinion word in the dictionary. The corpus-based approach guarantees context specificity of word orientations by searching a large corpus. Lexicon-based approaches require manual collection of the opinion words and has been criticized for requiring too much human effort [15][16][17]. As a solution, the semi-supervised approach uses an initial list of seed words with annotated polarities and uses synonym-based label propagation to map polarities to unknown words [15][16][18][9].

IV. RELATED WORK

Many studies have presented several different approaches for sentiment analysis. In general, many of these studies focus on sentiment analysis for the English language and other

languages (Chinese, Italy, Ordo). There are comparatively few studies for sentiment analysis for the Arabic language. In this section we first present some important sentiment analysis studies in different languages before going on to survey the Arabic sentiment analysis studies.

A. Sentiment analysis In General

Moraes et al. [19] compared the performance of SVM (support vector machines) and NN (neural networks) for a document-level SA analysis. They showed that NN achieves better performance than SVM on balanced datasets. Rui and Liu [20] investigated pre-consumer (prior to purchase) and post-consumer (after purchase) opinion differences using NB and SVM classifiers on twitter data from both classes of users. Li and Li [21] addressed subjectivity and expresser credibility in opinion studies using SVM as the classifier. Wang et al [22] studied the performance of three popular ensemble methods (bagging, boosting, random subspace) based on five basic learners (Naive Bayes, Maximum Entropy, Decision Tree, K-Nearest Neighbour, and Support Vector Machines) on sentiment classification tasks. They showed that random subspace achieves the best results.

New developments in supervised learning show a heavy dependence on conceptual analysis. Formal Concept Analysis and Fuzzy Formal Concept Analysis (FCA/FFCA) specifically were employed in works by Li and Tsai [23] showing an abstract conceptual classification system of documents and use of training (FFCA-based conceptual classifier training as opposed to document-based training) examples to boost accuracy. Kontopoulos et al. [24] have used FCA also to build an ontology domain model. In their work, they proposed the use of ontology-based techniques toward a more efficient sentiment analysis of twitter posts by breaking down each tweet into a set of aspects relevant to the subject. Poria et al. [25] proposed a novel paradigm to concept-level sentiment analysis that merges linguistics, common-sense computing, and machine learning for improving the accuracy of tasks such as polarity detection. Yang and Cardie [26] proposed an approach that allows structured modelling of sentiment by considering both local and global contextual information. They encode intuitive lexical and discourse knowledge as expressive constraints and integrate them into the learning of conditional random field models via posterior regularization. The paper by Tang et al. [27] shows a joint sentence-level segmentation and classification system. Latent Dirichlet Allocation (LDA) was used by Xiang and Zhou [28] in the creation of topic-specific information, before going on to divide the data into several subsets based on topic distribution. In the last wave, they presented a semi-supervised training system to further increase classification accuracy. They showed that the framework can better handle the inconsistent sentiment polarity between a phrase and the words it contains. Tang et al. [29] applied neural network to learn sentiment-specific word embedding (SSWE), which encodes sentiment information in the continuous representation of words. Unsupervised approaches also have a long history for SA. Xianghua and Guo [30] presented work in the Chinese-language domain. Their work used an unsupervised approach to automatically segment Chinese social reviews into aspects - and compute the sentiment expressed in each aspect. They

used Latent Dirichlet Allocation (LDA) for aspect discovery and employed a sliding-window context over the review text to generate local topics and the linked sentiment. In [31] by Cruz and Troyano presented a taxonomy-based approach where knowledge about how people express opinions in a given domain is catalogued. They showed that this domain-specific knowledge improves opinion mining accuracy. Huang et al. [32] considered words, symbols or phrases with emotional tendencies as input features. They studied the phenomenon of polysemy in single-character emotional word in Chinese and discussed single-character and multi-character emotional word separately. Kiritchenko et al. [33] conducted SA for short informal texts on both message-level and term-level. They generated novel high-coverage tweet-specific sentiment lexicons from tweets with sentiment word hashtags and from tweets with emoticons. Pablos et al. [34] used a set of raw texts from a specific domain (the corpus) to build a list of opinion terms for that domain using seed-list propagation based on rules that featured dependency relations and POS restrictions. In unsupervised approach a significant methods are introduced in [35][36].

Semi-supervised approaches for SA have recently attracted lots of attention. A semi-supervised approach was proposed by Tang et al [37] to evaluate different types of emotional signals in Twitter data using a correlated model. The model presents dual learning based on controlled alternating propagating and fitting processes operating on labelled and unlabeled data. Zhou et al. [38] applied a semi-supervised approach Fuzzy Deep Belief Network (FDBN) on SA. The deep architecture of FDBN consists of a set of unsupervised hidden layers and a final layer of supervised training. They did a comprehensive evaluation on the state-of-the-art semi-supervised methods for SA, including semi-supervised spectral learning(Spectral), transductive SVM(TSVM), deep belief networks(DBN), personal/impersonal views(PIV), active learning(Active), mine the easy classify the hard(MECH), active deep networks(ADN), fuzzy deep belief networks(FDBN), active FDBN(AFD). A hybrid study was performed by Ortigosa et al. [39] that combined machine learning and lexicon-based approaches with a selective logic that uses machine learning when a sufficient level of labelled data is available, and a lexicon-based system when not available. They believed their approach will not only extract sentiment but also identify significant changes in emotional signatures. As we have seen in the foregoing section, there has been a lot of advancement in sentiment analysis for the English-language domain. Many highly conceptual and experimental methods have been developed to improve the performance of basic classifiers, also more work has been done to advance the scope and applicability of supervised, unsupervised, semi-supervised, and hybrid techniques. This could be the result of an abundant level of research focus in this area, as well as favorable linkages between the research and profitable industrial applications.

B. Arabic Sentiment Analysis

There are many studies have been done in opinion mining field. Most of these studies have been done in English language context, and a little in Arabic language context. In this paper we will present some studies of Arabic language

context. We present a comprehensive review of recent Arabic sentiment analysis research using a component-by-component approach.

We study the following components: approach used, methods (classifiers) used, data sources used, Arabic dialects processed, and sentiment analysis level. We also provide a merit-based assessment of the advantages and disadvantages of the sentiment analysis systems used in each research work surveyed. As we have seen in the introduction and related work, the approaches in sentiment analysis are usually divided into four classes: supervised, unsupervised, semi-supervised, and hybrid. Table 1 below categorizes the surveyed Arabic SA studies into these classes, and fig.1 shows the result.

TABLE I. ARABIC SA STUDIES BY APPROACH

S/N	Approach	Studies
1	Supervised	[1], [7], [9], [10], [11], [13], [19], [20], [27], [28], [29], [33], [41], [42], [48], [49], [52], [65], [69], [70]
2	Semi-supervised	[23], [39], [46]
3	Unsupervised	[4], [6], [21]
4	Hybrid	[5], [22], [24], [32], [51]

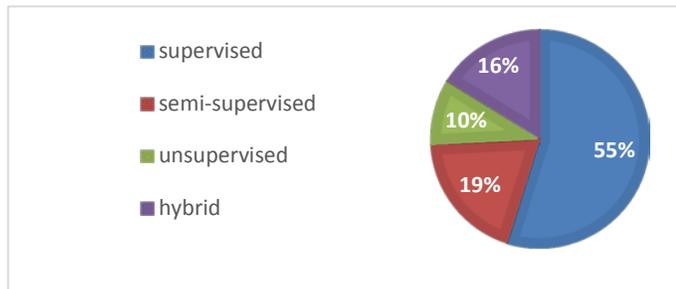


Fig. 1. Arabic SA studies by approach

From table 1, it is clear that there is a dominance of supervised learning over other techniques (semi-supervised, unsupervised, and hybrid techniques).

Arabic sentiment analysis studies used different method based on the used approach, some of these methods considered as a dominance. We have collected the various methods used in the different Arabic SA studies for supervised, semi-supervised, unsupervised and hybrid experiments and presented the result in Table 2.

From the above table, the most widely used methods (by far) appear to be based on Support Vector Machines (SVM), Naive Bayes (NB), and K-Nearest Neighbors (KNN). Lexicon-based approaches are generally prevalent across the majority of works sampled. Ensemble methods (comprising a variety of techniques) are also gaining significance.

In Arabic sentiment analysis studies several different text sources have been used, based on the objective of study, as outlined in Table 3 where the researchers in this domain appear to use Tweets, reviews/opinions & comments almost exclusively as datasets for their work on sentiment analysis. This may indicate a focus on social media.

We also investigated the size and diversity of the datasets used for the various Arabic sentiment analysis studies. We found that, there is significant variety in the quantity (size) of the datasets used in the various studies. It is more common to find studies where a single type of data was used, but there are a number of cases where multiple data types were combined.

TABLE II. ARABIC SA STUDIES BY METHOD

S/N	Method	Studies
1	SVM	[13], [24], [29], [70]
2	SVM + NB	[7], [9], [10], [11], [27], [39], [41], [42], [48], [49], [52]
3	SVM + NB + KNN	[5], [11], [19], [20], [28], [33], [69]
4	Binary Classifier	[1]
5	Maximum Entropy	[9], [22]
6	Decision Tree	[5], [9], [11], [27], [33]
7	Ensemble	[9], [11],[42]
8	Bayes Net	[9]
9	Bayes Point Machine	[29]
10	Lexicon-based	[4], [5], [6], [21], [22], [23], [24], [32], [46], [51]
11	Corpus-based	[5]
12	Grammar-based	[24]
13	Rocchio Classifiers	[42]
14	KNN (without SVM, NB)	[22]
15	NB (without SVM, KNN, D-tree)	[32], [65]

TABLE III. ARABIC SA STUDIES BY DATA

S/N	Source	Studies
1	Tweets/Twitter	[1], [5], [6], [9], [20], [21], [46], [51], [52], [65], [69], [70]
2	Wiki Pages	[1], [3]
3	Web Forums	[1], [3]
4	Reviews/Opinions	[7], [11], [19], [23], [24], [33], [39], [42], [48], [49]
5	Social Comments/Social News	[2], [4], [6], [10], [27], [65], [69], [70]
6	Lemmas	[13]
7	Website Comments	[28],[33],[41]
8	Biographic Information	[29]
9	Posts	[22], [32]
10	Documents	[24]

The Arabic language has many dialects, and no study of Arabic sentiment analysis is complete without a review of the different dialects covered in the studies. Table 4 below presents an overview of the Arabic dialect distribution in the studies surveyed. As we can see from the above table, modern Standard Arabic (MSA) sources are widely used throughout the studies sampled in this survey. Where dialects are used, Egyptian (MSA/Egyptian) was more favorable. There are also works with Levantine, Khaliji, Arabizi, Mesopotamian, Syro-Palestinian, Middle East Region, and Informal (Lebanese, Syrian, Iraqi, Libyan, Algerian, Tunisian, and Sudanese) dialects.

Also we can see that the Saudi dialect was not given attention by the researchers.

TABLE IV. ARABIC SA STUDIES BY LANGUAGE

S/N	Language	Studies
1	MSA	[1], [2], [3], [4], [5], [6], [9], [13], [19], [20], [22], [23], [24], [33], [39], [41], [42], [46], [48], [49], [65], [70]
2	MSA (Egyptian)	[7], [10], [11], [21], [27], [28], [51], [70]
3	MSA (Levantine)	[7]
4	MSA (Khaliji)	[7], [11], [65]
5	MSA (Arabizi)	[7]
6	MSA (Mesopotamian)	[11]
7	MSA (Syro-Palestinian)	[11]
8	MSA (Islamic)	[29]
9	MSA (Middle East Region)	[52]
10	MSA (Informal)	[32]

We were also interested in investigating the scope of the sentiment analysis carried out in the various Arabic language studies, so as to classify them as sentence-level, document-level, or sentence-level & document-level. Our findings are presented in Table 5 below.

TABLE V. ARABIC SA STUDIES BY PROCESSING LEVEL

S/N	Processing Level	Studies
1	Sentence-level	[2], [3], [23], [52], [69], [70]
2	Document-level	[1], [4], [5], [6], [7], [9], [10], [11], [13], [19], [20], [21], [22], [27], [28], [29], [32], [33], [39], [41], [42], [46], [48], [49], [51], [65],
3	Document-level + Sentence-level	[24]

From the previous table, it is rare to find projects in this domain that feature a combination of document-level and sentence-level sentiment analysis. On the contrary, nearly all the works sampled are focused on document-level sentiment analysis. There are also a few cases of sentence-level sentiment analysis.

One of the most crucial aspects of this work is the critical review of the various Arabic language sentiment analysis

studies surveyed, with the goal of identifying positive highlights, shortcomings, and areas of improvement, after a comprehensive review of each of the studies. Our comments are provided in Tables 6 and 7. As summarized from the above table, we have made some conclusive observations about studies in the field of Arabic Sentiment Analysis through our review of current Arabic SA research works. We found that most Arabic sentiment analysis works focus on the use of supervised methods as opposed to other classes of sentiment analysis including unsupervised, semi-supervised and hybrid or experimental systems. This method requires a huge amount of corpus and manually labeling for training and testing purpose this can be expensive, time-consuming, and difficult due to sarcasm especially in Arabic text [40][41]. The main disadvantage of this approach, it is a domain-biased which mean it give low accuracy when it is applied in different domain that was trained. This approach usually use machine learning methods such as Support Vector Machines, Naïve Bayes Classifiers and Maximum Entropy approaches [40][41]. In the other hand some studies employed the lexicon-based approach using different techniques to generate sentiment lexicons that would contribute to the task of sentiment analysis. This approach is based on a list of sentiment words with their polarities to determine the sentiment of review. This approach is considered practical since it is not domain-biased, recently some researchers intended to use the ontology in this approach, and such ontology may be used for different tasks: Arabic NLP tools, information retrieval [42]. Dialects are not supported in many of the Arabic SA studies surveyed in this paper. This presents a major disadvantage because the Arab language is dialectically rich and its diverse structural properties in the various dialects need to be fully captured in order to derive maximum benefit from Arabic SA, especially for less-formal channels like Social Media, whose corpora are principally not in Modern Standard Arabic (MSA).

It was also noticed that a limited set of classifiers (techniques) were repeatedly used for sentiment analysis in many of the papers surveyed. While researchers probably choose these same set of classifiers because they are proven to be effective, value is not being added to the field of Arabic SA if more experimental or conceptually novel techniques are not implemented or investigated. There is very little focus on sentence-level sentiment analysis for many of the studies. Most of the observations recorded during this survey generally lead to the conclusion that Arabic sentiment analysis is in its growing phases.

TABLE VI. ADVANTAGES ANA DISADVANTAGES SUMMARY

Paper ID	Advantages	Disadvantages
[1]	showed extensive list of features, studied the importance of different features	disregarded neutral and mixed classes
[2]	The annotations are extensive	No Sentiment Analysis evaluations on the corpus
[3]	Multi-genre corpus	No Sentiment Analysis evaluations on the corpus
[4]	Multiple lexicons constructed, integrated lexicon achieves best performance	Dialects are not considered
[5]	Negation and Intensification are considered	Neutral class is not included. Sarcasm is not considered
[6]	Advanced lexicon construction	Using individual words polarities technique
[7]	showed extensive list of features, studied the importance of different features	could try more classification methods, no details given on how sampling is conducted to obtain a balanced subset of data
[9]	Pre-processing leads to improvement	Tags need to be added manually
[10]	Introduced Social Network specific features	Dialects are not considered
[11]	Besides classification on subjectivity and polarity, also considered intensity classification	Does not deal with Emoticons, chat language and Arabizi
[13]	Large-scale lexicon	Could try more classifiers, dialects not considered
[19]	Studied the effects of pre-processing and the characteristics of the dataset	Could try more classifiers, dialects not considered
[20]	Developed three lexicons as well as a negation library, the dataset was large	Intensifications were not considered
[21]	Evaluated methods to learn the weights of the words and combine such weights	Dialects are not considered
[22]	Combination of multiple methods improves the performance	Considered posts from only three domains
[23]	label propagation is effective for lexicon construction	Only considered sentence level
[24]	Considered both grammar and lexicon	Dialects, suffix and prefix extraction not extracted, small dataset
[27]	Developed three lexicons as well as a negation library, the dataset was large	Only sentence level
[28]	Particularly addressed slang language	No benefits for non-slang cases
[29]	NLP is used, word presence feature leads to better performance	Could use more classifiers

TABLE VII. ADVANTAGES AND DISADVANTAGES SUMMARY

Paper ID	Advantages	Disadvantages
[32]	Considered both supervised and unsupervised approaches	Evaluated limited supervised methods
[33]	SentiStrength has better performance than SocialMention	Dialects are not considered
[39]	Semi-supervised lexicon construction	Dialects, Franco Arabic and compound phrases are not considered (single word match only)
[41]	Addressed unbalanced classification	Proposed methods didn't show advantage
[42]	Ensemble classifier achieves better classification	More classifiers can be added to the ensemble
[46]	Extensive feature categories, addressed topic shift	semi-supervised approach improves subjectivity analysis but not sentiment analysis
[48]	The Corpus has good quality	Could include more features
[49]	Determines the polarity of an Arabic corpus using English translation	SA depends on the quality of the translation
[51]	Very detailed investigation on the processing techniques	pre-processing techniques could be improved by cross-validation, lexicon might not be extensive
[52]	n -gram features are used	The corpus is small and low frequency terms are ignored
[65]	Sizeable dataset used	Could have used more classifiers
[69]	Used an ensemble of classifiers with a relatively comprehensive dataset	The size of the dataset is small
[70]	Supported dialectal Arabic in addition to MSA	Could have used more classifiers, relatively limited dataset

V. QUANTITATIVE ANALYSIS OF RECENT ARABIC SA RESEARCH

Our primary concern for performing a quantitative analysis on the performance data provided by the different Arabic sentiment analysis studies is to determine if, and the degree to

which, there is any significant difference in the performance outputs (evaluated across accuracy, precision, recall and F-score) for each of the methods used in the research works being surveyed, as this knowledge will put us in a position to potentially identify areas for improvement in current approaches. Table 8, catalogues reported statistics collated

from the various research publications being surveyed. Note: where multiple results were provided in these works, we selected only the best results. Every attempt was made to state the results as they were originally published by their various authors.

TABLE VIII. REPORTED STATISTICS FROM SURVEYED STUDIES

Paper ID	Accuracy	Precision	Recall	F-score
[1]	95.83%			
[4]	74.60%			
[6]	61.20%	60.60%	63.90%	
[9]	87.43%			
[10]	61.40%			
[11]	96.90%	95.00%	97.00%	
[13]				71.10%
[19]	97.20%	99.60%	94.80%	
[20]	76.78%			
[21]	83.80%	44.40%	57.10%	49.00%
[22]	84.34%	87.20%	89.62%	85.57%
[24]	87.00%			
[27]		87.40%	33.80%	
[28]		88.60%	78.00%	88.54%
[29]	95.91%			
[32]	91.20%			
[33]	99.20%			
[39]	97.81%	98.00%	98.00%	
[41]	96.00%	98.00%	98.00%	
[42]		98.60%	98.60%	97.60%
[49]		90.00%	95.00%	90.73%
[51]	75.90%	76.90%	75.90%	76.20%
[65]	80.60%	86.10%	99.90%	83.20%
[69]			83.00%	72.00%
[70]	99.90%	99.90%	99.90%	99.90%

A. Analysis Technique – Smoothness Analysis

Smoothness analysis is based on arithmetic series in discrete mathematics [42]. For any arithmetic series, we have a first term S_1 , last term S_n and common difference d such that any member of the series can be represented as:

$$S_i = S_1 + (i - 1)d$$

Because real-life data may not always behave as an arithmetic series, the smoothness of a distribution is simply an estimation of the error in the real distribution relative to the projected arithmetic series distribution [42]:

$$\text{smoothness} = \left(1 - \frac{S_1 + S_n}{2\bar{S}}\right) \quad (1)$$

Where: S_1 = first term in the real data series when arranged in increasing order, S_n = last term in the real data series when arranged in increasing order, and \bar{S} = average of the real data series.

Benefit: the smoothness of a distribution as calculated by equation (1) gives us the % error (percentage error) in the straight-line form of the data, and tells us how the data has

changed with respect to the different input values (that is, we can evaluate the impact or significance of the different methods used by the research works on the performance scores reported).

B. Local Optimum Problem in Smoothness Analysis

When evaluating the impact of studies using equation (1) and the data from Table 7, we run into the problem of local optimum: an approximately constant score for all performance categories. This is because all the performance scores are less than 1 and are therefore, similar.

Lemma: For all pairs of similar values, the smoothness function will return a zero (no impact) result.

$$\forall (S_1, S_2) : S_1 \approx S_2 \\ \approx S, \lim \left(1 - \frac{S_1 + S_2}{2\bar{S}}\right) \rightarrow \left(1 - \frac{2S}{2\bar{S}}\right) \rightarrow 0$$

This invalidates the analysis unless a solution can be obtained to proportionally amplify the input values (performance scores), so that the validity condition (shown below) can be met:

$$\lim \left(1 - \frac{S_1 + S_2}{2\bar{S}}\right) \neq 0 \\ \text{(Validity condition for smoothness function)}$$

C. Solution to Local Optimum: Logarithmic Smoothing

To solve the local optimum problem described above which will invalidate our analysis according to Lemma due to the closely-bounded performance scores, we explore the use of the logarithmic smoothing technique described in [43], a procedure for proportionally expanding individual elements within the space of a closely-bounded range.

$$\Gamma(\ln r, \ln \theta, \ln \phi) \\ = r' \theta' \phi' e^{\frac{r}{r_{\max}} + \frac{\theta}{\theta_{\max}} + \frac{\phi}{\phi_{\max}} - 3} \quad (2)$$

The logarithmic smoothing process is shown above equation (2) for a three-dimensional smoothing problem in spherical coordinates (r, θ, ϕ).

Where: r = component of a point in r -coordinate, θ = component of a point in θ -coordinate, ϕ = component of a point in ϕ -coordinate, r' = target projection of r , θ' = target projection of θ , ϕ' = target projection of ϕ , r_{\max} = maximum value of r , θ_{\max} = maximum value of θ , ϕ_{\max} = maximum value of ϕ

Benefit: At any point within the sphere (3D space), the function $\Gamma(\ln r, \ln \theta, \ln \phi)$ gives a smooth projection that is continuous in r, θ , and ϕ directions [43].

This means that with this transformation, the problem of local optimum can be reasonably avoided because input values are transformed to their smooth projections $r \rightarrow r_{\text{smooth}}, \theta \rightarrow \theta_{\text{smooth}}, \phi \rightarrow \phi_{\text{smooth}}$ and these values will pass the validity condition because $r_{\text{smooth}} > r, \theta_{\text{smooth}} > \theta, \phi_{\text{smooth}} > \phi$ and $\frac{r_{\text{smooth}}}{r} \neq \frac{\theta_{\text{smooth}}}{\theta} \neq \frac{\phi_{\text{smooth}}}{\phi}$ such that we have a valid analysis (by the validity condition for smoothness function).

For our purpose in this analysis, we present a simplification of this idea as follows:

As we only have 1-dimensional data (each performance parameter is evaluated on a case-by-case basis – accuracy only, precision only, recall only and F-score only), for which only the r-coordinate is sufficient, we need to remove unnecessary coordinates (θ, ϕ) by setting these values to 1: $\theta = \theta_{\max} = 1, \phi = \phi_{\max} = 1$

This reduces equation (2) to a form that is applicable for our analysis, which is:

$$\Gamma(\ln r, \ln 1, \ln 1) = \Gamma(\ln r, 0, 0) = r'(1)(1)e^{\frac{r}{r_{\max}}+1+1-3} = r'e^{\frac{r}{r_{\max}}-1}$$

Which we can write as:

$$\Gamma(\ln r) = r'e^{\frac{r}{r_{\max}}-1} \quad (3)$$

Conclusion: equation (3) above is the logarithmic smoothing function that we will use in our analysis to solve the problem of local optimum.

Fig.2 below shows the effect of applying the logarithmic smoothing function equation (3) in transforming data from closely-bounded spaces (x-space) to loose-bounded spaces (L-space):

Outcome: there is proportional amplification in the data, such that the behavior of the data remains unchanged, while small differences are much easily visualized and evaluated.

D. Results of The Analysis

Comparative Results – Accuracy: Table 9 shows the raw accuracy scores and the converted logical scores for use in the smoothness analysis.

To arrive at the logical scores shown in the table above (used for the analysis), we used the function of logarithmic smoothing, by setting $r' = 1000$, calculating converted scores and arranging in increasing order. For this table, max: r (largest element in r) is 0.999.

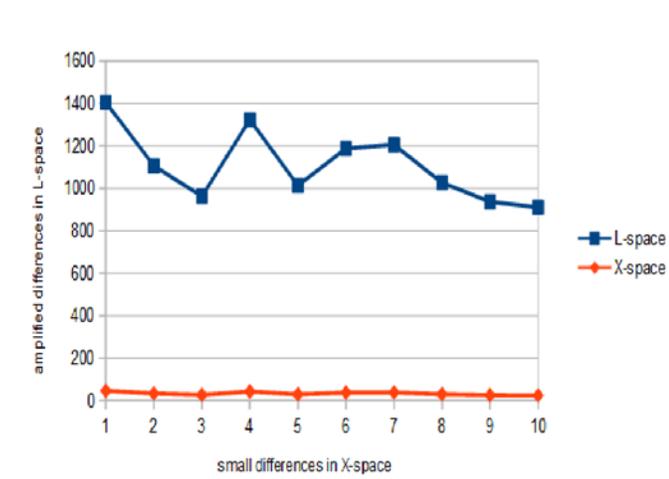


Fig. 2. Transformation from closely-bound x-space to widely-bound l-space

The smoothness = 0.0463, for this dataset (*slightly rough*), as calculated by using the smoothness function, indicating that: there is no significant impact of the different methods used on the accuracy. See Fig. 3 for a visualization (plot correlates well with trend-line). By using the same process, we obtained: 0.10973023, 0.12700562, and 0.045964546, as smoothness result for precision, recall, and F-score, respectively. The results lead to the following conclusions: there is slight impact of the different methods used on the precision and recall (see *slightly rough* curves in Fig.3, Fig.4, and Fig.5 – the plots do not correlate very well with their trend-lines) but there is no significant impact on accuracy and F-score (see *smooth* curves in Fig. 3, and Fig.4 – the plots correlate well with their trend-lines).

Table 10 presents a summary of smoothness results obtained from the experiments. As can be seen from this analysis, *Accuracy & F-score* are not impacted by the different methods adopted by the various researchers in the studies surveyed. *Precision & Recall*, however, show slight response to the different methods used by the researchers in the studies surveyed. **Table 8:** Accuracy distribution by logical score and raw values

TABLE IX. ACCURACY DISTRIBUTION BY LOGICAL SCORE AND RAW VALUES

Study	raw accuracy score	Converted logical score
[6]	0.6120	678.828
[10]	0.6140	680.188
[4]	0.7460	776.271
[51]	0.7590	786.438
[20]	0.7678	793.397
[65]	0.8060	824.322
[21]	0.8380	851.154
[22]	0.8434	855.768
[24]	0.8700	878.860
[9]	0.8743	882.651
[32]	0.9120	916.597
[1]	0.9583	960.078
[29]	0.9591	960.847
[41]	0.9600	961.713
[11]	0.9690	970.416
[19]	0.9720	973.334
[39]	0.9781	979.296
[33]	0.9920	993.017
[70]	0.9990	1000

TABLE X. SUMMARY OF SMOOTHNESS RESULTS

Slight significance	No significance
Precision (0.1097)	Accuracy (0.0463)
Recall (0.1270)	F-score (0.0459)

VI. CONCLUSION AND FUTURE WORK

In this paper we have surveyed the important Arabic sentiment analysis studies qualitatively and quantitatively. We have presented detailed analyses of methods used and results obtained in the current Arabic sentiment analysis studies, as well as a rich discourse on the direction of current research, present limitations. In our qualitative evaluation, we found that, the majority of Arabic SA uses established supervised methods as opposed to more progressive or experimental unsupervised and semi-supervised approaches. The dialects are not processed in many of the Arabic SA studies surveyed, which is a major drawback on the effectiveness of current Arabic SA because most of the available Arabic language text in the social media and other spaces represent a wide range of distinct, autonomous, and morphologically complex Arabic language dialects. It was also observed that many of the studies surveyed used the same limited set of classifiers - raising questions about reasonable value added to the field if every study essentially repeats the same experiment on a Different dataset. There is a definite need for more inventiveness and creativity in the design of experiments as well as the development of novel classification and analysis techniques beyond the established algorithms.

For our quantitative evaluation, we applied rigorous data modelling and statistical procedures to investigate the effectiveness of methods adopted by the various researchers in the Arabic SA works surveyed. We collected performance data (accuracy, precision, recall and f-score) for the various studies and applied advanced techniques including logarithmic smoothing field analysis and a relative smoothness function, to uncover deep patterns in the performance data. Our approach was based on the reasoning that similar processes will produce similar results. The various studies conducted for Arabic SA will not be differentiable if they all produce similar results across the various performance classes - accuracy, precision, recall and f-score. But where we have significant variance of results, then there is opportunity for improvement. The analysis performed yielded the following conclusion: there is only a slight impact of the different methods used on the Precision & Recall of results obtained while there was no significant impact on the Accuracy & F-score. This ultimately leads us to the conclusion that Arabic SA researchers should employ a more diverse set of techniques and approaches that do more to improve scoring across the full range of performance parameters.

In the future work, we believe that there is a promising trend to obtain optimal Arabic SA system. We intend to propose and develop a new hybrid method using deep learning technique and big data technique such as Hadoop and MapReduce to solve some of the existing problems in Arabic sentiment analysis as highlighted in this survey as well as to obtain optimal system for Arabic SA. As we have seen, most of the work in the field of Arabic sentiment analysis has focused on the use of supervised learning techniques, and are largely lexicon-based approaches with the characteristic limitations. We believe that the opportunity space for growth in this field will be driven by the exploration of unsupervised learning techniques, principally through hybrid method.

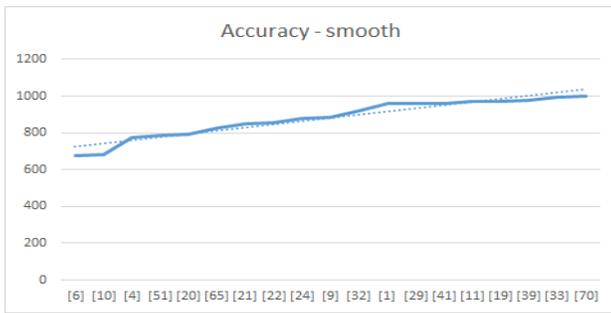


Fig. 3. Smoothness accuracy result

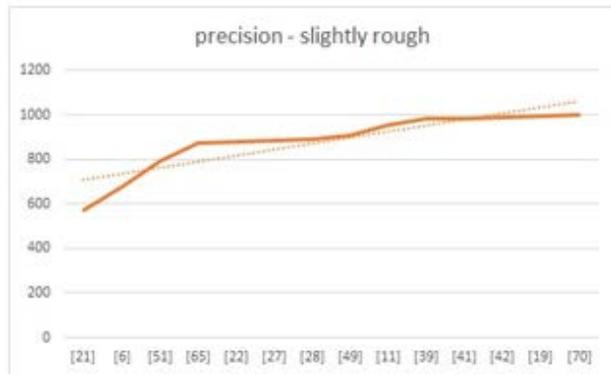


Fig. 4. Smoothness precision result

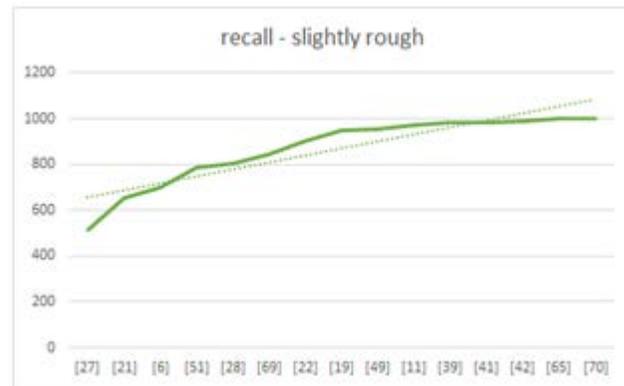


Fig. 5. Smoothness recall result



Fig. 6. Smoothness f-score result

REFERENCES

- [1] Association for computational linguistics. What is Computational Linguistics? <http://www.aclweb.org/archive/misc/what.html> (accessed 18 September 2015).
- [2] Web. Introduction to Sentiment Analysis, <http://ict-master.org/files/mullensentimentcourseslides.pdf> (accessed 18 September)
- [3] Liu B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*; 5(1): pp. 1-167, 2012.
- [4] Tsytzarou M and Palpanas T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*; 24(3): pp. 478-514, 2012.
- [5] Spanish society for natural language processing. Natural Language Processing and Opinion Mining, http://www.aepia.org/aepia/files/evia/transparencias/urea_evia2014.pdf (accessed 18 September 2015).
- [6] Al-kabi MN, Abdulla NA and Al-ayyoub M. An analytical study of arabic sentiments: maktoob case study. In: 8th international conference for internet technology and secured transactions, IEEE, London, UK, , pp. 89-94, 2013.
- [7] Sarah O. Alhumoud, Mawaheb I. Altuwaijri, Tarfa M. Albuhairei, Wejdan M. Alohaideb. Survey on Arabic Sentiment Analysis in Twitter. *International Science Index*, , 9 (1), pp. 364-368, 2015.
- [8] Korayem M, Crandall D and Abdul-mageed M. Subjectivity and sentiment analysis of arabic: a survey. In *Advanced Machine Learning Technologies and Applications*, , 128-139, 2012.
- [9] Farra N, Challita E, Abou-assi R and Hajj H. Sentence-level and document-level sentiment mining for arabic texts. In: International conference on data mining workshops, IEEE, , pp. 1114-1119, 2010.
- [10] Haddi E, Liu X and Shi Y. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, pp. 26-32, 2013.
- [11] Bao Y, Quan C, Wang L and Ren F. The role of pre-processing in twitter sentiment analysis. *The 10th International Conference on Intelligent Computing*, Taiyuan, China.; pp. 615-624, 2014.
- [12] Yu LC, Wu JL, Chang PC and Chu HS. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-based Systems*, 41, pp. 89-97, 2013.
- [13] Arafat H, Elawady RM, Barakat S and Elrashidy NM. Different feature selection for sentiment classification. *International Journal of Information Science and Intelligent Systems*; 3(1): pp. 137-150, 2014.
- [14] Agarwal B and Mittal N. Optimal feature selection for sentiment analysis. In: *Computational linguistics and intelligent text processing*, Springer, , pp. 13-24, 2013.
- [15] Abdulla N, Majdalawi R, Mohammed S, Al-ayyoub M and Al-kabi M. Automatic lexicon construction for Arabic sentiment analysis. *International Conference on Future Internet of Things and Cloud (FiCloud-2014)*, Barcelona, Spain, , pp. 547-552, 2014.
- [16] Abdulla NA, Ahmed NA, Shehab MA and Al-ayyoub M. Arabic sentiment analysis: lexicon-based and corpus-based. In: *Applied electrical engineering and computing technologies*, IEEE Jordan Conference, , pp. 1-6, 2013.
- [17] Abdulla NA, Ahmed NA, Shehab MA, Al-ayyoub M, Al-kabi MN and Al-rifai S. Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and web Engineering*; 9(3): pp. 55-71, 2014.
- [18] El-beltagy SR and Ali A. Open issues in the sentiment analysis of arabic social media: a case study. In: 9th international conference on innovations in information technology, IEEE, , pp. 215-220, 2013.
- [19] Moraes R, Valiati JF and Neto WPG. Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Systems with Applications*; 40(2): pp. 621-633, 2013.
- [20] Rui H, Liu Y and Whinston A. Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems*; 55(4): pp. 863-870, 2013
- [21] Li YM and Li TY. Deriving market intelligence from microblogs. *Decision Support Systems*; 55(1): pp. 206-217, 2013.
- [22] Wang G, Sun J, Ma J, Xu K and Gu J. Sentiment classification: the contribution of ensemble learning. *Decision Support Systems*; 57, pp. 77-93, 2014.
- [23] Li ST and Tsai FC. A fuzzy conceptualization model for text mining with application in opinion polarity classification. *Knowledge-based Systems*; 39, pp. 23-33, 2013.
- [24] Kontopoulos E, Berberidis C, Dergiades T and Bassiliades N. Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*; 40(10): pp. 4065-4074, 2013.
- [25] Poria S, Cambria E, Winterstein G and Huang GB. Sentic patterns: dependency-based rules for concept-level sentiment analysis. *Knowledge-based Systems*; 69, pp. 45-63, 2014.
- [26] Yang B and Cardie C. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In: *Proceedings of the ACL*, Baltimore Maryland, , pp. 325-335, 2014.
- [27] Tang D, Wei F, Qin B, Dong L, Liu T and Zhou M. A joint segmentation and classification framework for sentiment analysis. In: *Proceeding of EMNLP conference*, 2014.
- [28] Xiang B and Zhou L. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics*, ACL, , pp. 434-439, 2014.
- [29] Tang D, Wei F, Yang N, Zhou M, Liu T and Qin B. Learning sentiment-specific word embeddings for twitter sentiment classification. In: *proceedings of the 52nd annual meeting of the association for computational linguistics*, ACL, pp. 1555-1565, 2014.
- [30] Xianghua F, Guo L, Yanyan G and Zhiqiang W. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-based Systems*; 37, pp. 186-195, 2013.
- [31] Cruz FL, Troyano JA, Enriquez F, Ortega FJ and Vallejo CG. Long autonomy or long delay? the importance of domain in opinion mining. *Expert Systems and Applications*; 40(8): pp. 3174-3184, 2013.
- [32] Huang Z, Zhao Z, Liu Q and Wang Z. An unsupervised method for short-text sentiment analysis based on analysis of massive data. In: *Intelligent computation in big data era*, Springer, , pp. 169-176, 2015.
- [33] Kiritchenko S, Zhu X and Mohammad SM. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*; 50, pp. 723-762, 2014.
- [34] García-Pablos A., Cuadros M., Gaines, S. Rigau G. V3: Unsupervised Acquisition of Domain Aspect Terms for Aspect Based Sentiment Analysis. 8th International Workshop on Semantic Evaluation (SemEval 2014) at Third Joint Conference on Lexical and Computational Semantics and the 25th International Conference on Computational Linguistic (COLING'14). Dublin, Ireland. 2014.
- [35] Hassan S, Miriam F, Yulan H and Harith A. Senticircles for contextual and conceptual semantic sentiment analysis of twitter. In: *proceedings of the 11th ESWC conference*, Crete, Greece, 2014.
- [36] Hassan S, Yulan H and Harith A. Semantic sentiment analysis of twitter. In: *proceedings of the 11th international semantic web conference*, Boston, USA, 2012.
- [37] Tang J, Nobata C, Dong A, Chang Y and Liu H. Propagation-based sentiment analysis for micro-blogging data. In: *SIAM international conference on data mining*, 2015, Vancouver, British Columbia, Canada. pdf [185].
- [38] Zhou S, Chen Q and Wang X. Fuzzy deep belief networks for semi-supervised sentiment classification. *Neurocomputing*; 131, pp. 312-322, 2014.
- [39] Ortigosa A, Mart JM and Carro RM. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*; 31, pp. 527-541, 2014.
- [40] Abdul-mageed M, Diab M and Kubler S. Samar: subjectivity and sentiment analysis for Arabic social media. *Journal of Computer Speech & Language*; 28(1): pp. 20-37, 2014.
- [41] Al-kabi MN, Alsmadi IM, Gigieh AH, Wahsheh HA and Haidar MM. Opinion mining and analysis for arabic language. *International Journal of Advanced Computer Science and Applications*; 5(5), pp. 181-195, 2014.

- [42] Maha A, Mona A, Nehal A, Wafa A, Asma A and Mesheal A. Semtree ontology for enriching arabic text with lexical semantic annotations. In: Proceedings of the 9th conference on semantic computing, IEEE, 2015.
- [43] Schmidt P and Ayres F. Schaum's outline of college mathematics. 4th edition. McGraw-Hill, 2010.
- [44] Spiegel MR. Schaum's Theory and Problems of Theoretical Mechanics. McGraw-Hill, 1967.
- [45] Abdul-mageed M and Diab MT. Subjectivity and sentiment annotation of modern standard Arabic newswire. In: Proceedings of the 5th linguistic annotation workshop, Association for Computational Linguistics, , pp. 110-118, 2011.
- [46] Abdul-mageed M and Diab MT. Awatif: a multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In: proceeding of the 8th International conference on language resources and evaluation (LREC), pp. 3907-3914, 2012.
- [47] Abdulla NA, Al-ayyoub M and Al-kabi MN. An extended analytical study of arabic sentiments. International Journal of big Data Intelligence; 1(1): pp. 103-113, 2014.
- [48] Ahmed S, Pasquier M and Qadah G. Key issues in conducting sentiment analysis on arabic social media text. In: 9th international conference on innovations in information technology, IEEE, , pp. 72-77, 2013.
- [49] Badaro G, Baly R, Hajj H, Habash N and El-hajj W. A large-scale arabic sentiment lexicon for arabic opinion mining. In: Proceedings of ANLP 2014, EMNLP 2014 Doha, Qatar, , p. 165, 2014.
- [50] Bai X. Predicting consumer sentiments from online text. Decision Support Systems; 50(4): pp. 732-742, 2011.
- [51] Deng L and Dong Y. Deep learning: methods and applications. Foundations and Trends in Signal Processing; 7(3-4): pp. 197-387, 2014.
- [52] Ding C and Peng H. Minimum redundancy feature selection from micro-array gene expression data. Journal of Bioinformatics and Computational Biology; 3(2): pp. 185-205, 2005.
- [53] Duwairi R and El-orfali M. A study of the effects of preprocessing strategies on sentiment analysis for arabic text. Journal of Information Science; 40(4): pp. 501-513, 2014.
- [54] Duwairi RM, Marji R, Sha'ban N and Rushaidat S. Sentiment analysis in arabic tweets. In: 5th international conference on information and communication systems, IEEE, , pp. 1-6, 2014.
- [55] El-halees A. Arabic opinion mining using combined classification approach. In: proceeding of the international Arab Conference on Information Technology. ACIT (2011).
- [56] Elhawary M and Elfeky M. Mining arabic business reviews. In: International conference on data mining workshops, IEEE, pp. 1108-1113, 2010.
- [57] A. E.-D. A. Hamouda and F. E. El-taher. Sentiment Analyzer for Arabic Comments System. International journal Adv. Comput. Sci. Appl., , 4 (3), pp. 100–103, 2013.
- [58] Taysir Hassan A. Soliman, M. Ali M. Abdel Rahman Hedar, M. M. Doss, "MINING SOCIAL NETWORKS' ARABIC SLANG COMMENTS". In Proceedings of IADIS European Conference on Data Mining (ECDM'13), Prague, Czech Republic, 2013.
- [59] Helmy T and Daud A. Intelligent Agent for Information Extraction from Arabic Text without Machine Translation. CEUR, Vol. 687, 2010.
- [60] Itani MM, Hamandi L, Zantout RN and Elkabani I. Classifying sentiment in arabic social networks: naive search versus naive bayes. In: 2nd international conference on advances in computational tools for engineering applications, IEEE, , pp. 192-197, 2012.
- [61] Khasawneh RT, Wahsheh HA, Al-kabi MN and Alsmadi IM. Sentiment analysis of arabic social media content: a comparative study. In: International conference on information science and technology, IEEE, 2013, pp. 101-106.
- [62] Mahyoub FHH, Siddiqui MA and Dahab MY. Building an arabic sentiment lexicon using semi-supervised learning. Journal of King Saud University-computer and Information Sciences; 26(4): pp. 417-424, 2014.
- [63] Mountassir A, Benbrahim H and Berrada I. Some methods to address the problem of unbalanced sentiment classification in an arabic context. In: Colloquium in information science and technology, IEEE, , pp. 43-48, 2012.
- [64] Omar NO, Albared M, Al-shabi A and Al-moslimi T. Ensemble of classification algorithms for subjectivity and sentiment analysis of arabic customer reviews. International Journal of Advancements in Computing Technology; 14(5): pp. 77-85, 2013.
- [65] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©1988
- [66] Refaee, E. and Rieser, V. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. The 9th edition of the Language Resources and Evaluation Conference LREC'2014. The European Language Resources Association. Reykjavik, Iceland, May, pp.26-31, 2014.
- [67] Rushdi-saleh M, Martin-valdivia MT, Urena-lopez LA and Perea-ortega JM. Oca: opinion corpus for arabic. Journal of the American Society for Information Science and Technology; 62(10): pp. 2045-2054, 2011.
- [68] Rushdi-saleh M, Martin-valdivia MT, Urena-lopez LA and perea-ortega JM. Bilingual experiments with an Arabic-english corpus for opinion mining. In Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, pp.740-745, 2011.
- [69] Safavian AR and Landgrebe D. A survey of decision-tree classifier methodology. IEEE Transactions on Systems, man and Cybernetics; 21(3): pp. 660-674, 1991.
- [70] Shoukry A and Rafea A. Preprocessing egyptian dialect tweets for sentiment mining. In: Proceedings of the 4th workshop on computational approaches to arabic script-based languages, p. 47, 2012.
- [71] Shoukry A and Rafea A. Sentence-level arabic sentiment analysis. In: International conference on collaboration technologies and systems, IEEE, , pp. 546-550, 2012.
- [72] Specht DF. Probabilistic neural networks. Neural Networks; 3(1): pp. 109-118, 1990.
- [73] Mourad A and Darwish K. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, , pp. 55-64, 2013.
- [74] Duwairi R and Qarqaz I. Arabic sentiment analysis using supervised classification. In: 1st international workshop on social networks analysis management and security, 2014.
- [75] Hossam I, Sherif A and Mervat G. Sentiment analysis for modern standard arabic & colloquial. International Journal on Natural Language Computing; 4(2): pp. 95-109, 2015.
- [76] Sallab A, Hajj H, Badaro G, Baly R, El-hajj W and Shaban K. Deep learning models for sentiment analysis in arabic. In: Proceedings of the 2nd workshop on arabic natural language processing, Beijing, China. Association for Computational Linguistics, pp. 9-17, 2015.

A Novel Ball on Beam Stabilizing Platform with Inertial Sensors

Part II: Hybrid Controller Design: Partial Pole Assignment & Rapid Control Prototyping

Ali Shahbaz Haider

Control Systems Laboratory
Electrical Engineering Dept
COMSATS Institute of Information Technology,
Beta Facility for Advance Automatic Control Tech.
Wah, Pakistan

Samter Ahmed

Electrical Engineering Dept
COMSATS Institute of Information Technology
Wah, Pakistan

Muhammad Bilal

Electrical Engineering Dept
COMSATS Institute of Information Technology
Wah, Pakistan

Saqib Raza

Electrical Engineering Dept
COMSATS Institute of Information Technology
Wah, Pakistan

Imran Ahmed

Electrical Engineering Dept
COMSATS Institute of Information Technology
Wah, Pakistan

Abstract—This research paper presents a novel controller design for one degree of freedom (1-DoF) stabilizing platform using inertial sensors. The plant is a ball on a pivoted beam. Multi-loop controller design technique has been used. System dynamics is observable but uncontrollable. The uncontrollable polynomial of the system is not Hurwitz hence system is not stabilizable. Hybrid compensator design strategy is implemented by partitioning the system dynamics into two parts: controllable subsystem and uncontrollable subsystem. Controllable part is compensated by partial pole assignment in the inner loop. Prediction observer is designed for unmeasured states in the inner loop. Rapid control prototyping technique is used for compensator design for the outer loop containing the controlled inner loop and uncountable part of the system. Real-time system responses are monitored using MATLAB/Simulink that show promising performance of the hybrid compensation technique for reference tracking and robustness against model inaccuracies.

Keywords—stabilizing platform; ball on beam; multi-loop controller; inertial sensors; rapid control prototyping; partial pole assignment

I. INTRODUCTION

Stabilizing platforms are among challenging control systems. One of such systems is the single degree of freedom (1-DoF) ball on beam mechanism. Plant of this control problem consists of a ball capable of rolling on a beam under the action of gravity due to the inclination of the beam. The control objective is to stabilize the positions of the ball on the beam in the presence of external disturbances and to achieve

ball position reference tracking. The system is open loop unstable so feedback is inevitable [1], [13].

Owing to the significance of ball on beam system a lot of research work has been dedicated to it. Classical PID controller has been implemented in [13] treating system a single input single output plant without taking in to account the internal states of the system. The observer-based model reference adaptive iterative learning controller has been demonstrated in [2]. A new technique based on geometric control has been implemented in [3], which involves designing immersion and invariance based speed and rotation angle observer for the ball and beam system. Decoupled neural fuzzy sliding mode control of the nonlinear ball on beam system has been considered in [4]. Nonlinear model predictive control for a ball and beam has been implemented in [5]. MATLAB based modeling and modulation of nonlinear ball-beam system controller has been demonstrated in [6]. A new adaptive state feedback controller for the ball and beam system is presented in [7]. Augmented state estimation and LQR control for a ball and beam system are implemented in [8]. Adaptive Neural Network for stabilization of ball on beam system has been studied in [9]. Human simulated intelligent control for ball and beam system is implemented in [10]. The Lyapunov direct method for the stabilization of the ball is presented in [11] and Energy-based balance control approach to the ball and beam system is presented in [12].

The majority of research work in the literature takes into account a reduced order model of the system by neglecting certain states in the system. In this research paper full order

model of the system is stabilized using a novel method that is a hybrid of partial pole assignment and rapid control prototyping using feedback from inertial sensors. Rapid Control Prototyping is a controller testing and tuning strategy on the actual plant in the feedback loop. With the availability of low-cost high processing capability digital processors and software suits, responses of real plants can directly be obtained and evaluated for a given control law. Nowadays rapid control prototyping is industry-wide adopted because the behavior of control algorithm can directly be tested on real world plants.

This research paper is the second part of two parts research. Part-I described geometrically accurate and detailed nonlinear model of the ball on beam system followed by linearization and state space conversion. In this part-II of the research work, controller is designed for the model developed in part-I.

Organization of the paper is as follows, section-II gives a brief overview of system dynamics. Section-III comprises of multi-loop hybrid compensation design involving partial pole assignment for inner loop and rapid control prototyping for the outer loop. Section-IV presents simulation and experimental results followed by section-V describing conclusions and future work.

II. OVERVIEW OF SYSTEM DYNAMICS

Hardware platform is shown in Figure 1. Functional description for this plant is given in [1]. Position of a metallic ball capable of rolling on a beam is to be controlled. Beam consists of two parallel rods. Both rods are hollow thin cylindrical. One rod is wound by a chromium wire and the other rod has metallic conducting surface. Position of the ball is monitored by a linear potentiometer mechanism which consisting of aforementioned two rods shorted by metallic ball hence producing a voltage proportional to position of ball on the beam. An accelerometer and a rate gyro on an inertial measurement unit (IMU) board measure beam inclination angle and angular velocity respectively as shown in Figure 2.

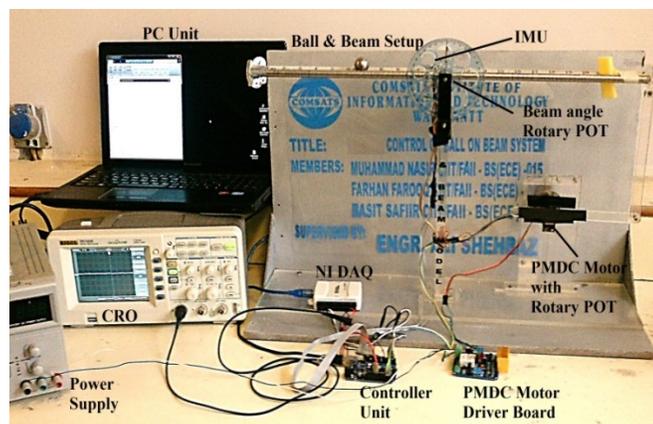


Fig. 1. Hardware platform



Fig. 2. Sensing mechanism

Inclination of the beam is actuated by a permanent magnet DC motor (PMDC) with its shaft coupled to a rotary potentiometer. Motor is driven by driver board. Control strategy is implemented by a digital micro controller and data acquisition card (DAQ) interfaced with MATLAB/Simulink for real time data monitoring and processing. The continuous time state space of the plant is given by (1), which has been derived in [1].

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.82 & 9.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -341.3e3 & 0 & 0 & 54.6e3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -13.65 & 1.36 \\ 0 & 0 & 0 & 0 & 0 & -61.2e-3 & -3.26 \end{bmatrix}, \quad (1)$$

$$B = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1]^T,$$

$$C = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.5 \end{bmatrix},$$

$$D = [0 \ 0 \ 0 \ 0 \ 0 \ 0]^T.$$

The model (1) is discretized in the MATLAB using c2d command with zero-order-hold and 0.01sec sampling interval. The discretized state space model is given by (2).

III. CONTROLLER DESIGN

System dynamics in (2) are observable but uncontrollable. In order to stabilize the system and to achieve control objects, system in (2) is partitioned in block upper triangular configuration given by (3). The partitioning has created two subsystems as shown in Figure 3. One of these subsystems is completely controllable and observable. This subsystem is

named subsystem 2 given by (4). The other subsystem is termed subsystem 1 given by (5). This partitioning into subsystems is shown in Figure 4. Our controller design strategy involves hybrid compensation in multi-loop control topology. Subsystem 2 is controlled in inner loop by unmeasured state observation followed by partial pole assignment. Controlled subsystem 2 along with subsystem 1 is compensated in outer loop using rapid control prototyping.

$$\begin{aligned} \underline{x}_b(k+1) &= G_{bb}\underline{x}_b(k) + H_b e(k), \\ y_b &= C_{bb}\underline{x}_b(k) + D_{bb}e(k), \end{aligned} \quad (4)$$

$$C_{bb} = \begin{bmatrix} 3.5 & 0 & 0 \\ 0 & 0 & 4.5 \end{bmatrix}, \quad D_{bb} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$\underline{x}(k+1) = G\underline{x}(k) + He(k),$$

$$y(k+1) = C\underline{x}(k) + De(k),$$

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \\ x_5(k+1) \\ x_6(k+1) \\ x_7(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 9.9591e-03 & 4.8866e-04 & 1.4309e-09 & 0 & -3.8549e-11 & 2.5842e-07 \\ 0 & 9.9183e-01 & 9.7599e-02 & 2.8588e-07 & 0 & -1.5289e-08 & 7.7286e-05 \\ 0 & 0 & 1 & 2.9300e-06 & 0 & -4.6263e-07 & 1.5735e-03 \\ 0 & 0 & 0 & 0 & 0 & -8.9984e-05 & 1.5485e-01 \\ 0 & 0 & 0 & 0 & 1 & 9.3475e-03 & 6.4300e-05 \\ 0 & 0 & 0 & 0 & 0 & 8.7240e-01 & 6.4300e-05 \\ 0 & 0 & 0 & 0 & 0 & -5.6263e-04 & 9.6792e-01 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ x_5(k) \\ x_6(k) \\ x_7(k) \end{bmatrix} + \begin{bmatrix} 6.4718e-10 \\ 2.5842e-07 \\ 7.9080e-06 \\ 1.5735e-03 \\ 2.1735e-07 \\ 6.4300e-05 \\ 9.8387e-03 \end{bmatrix} e(k), \quad (2)$$

$$\begin{bmatrix} y_1(k) \\ y_2(k) \\ y_3(k) \\ y_4(k) \\ y_5(k) \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.5 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ x_5(k) \\ x_6(k) \\ x_7(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} e(k).$$

$$\begin{aligned} \underline{x}_a(k+1) &= G_{aa}\underline{x}_a(k) + G_{ab}\underline{x}_b(k) + H_a e(k), \\ G_{ba} &= \underline{0}, \end{aligned} \quad (5)$$

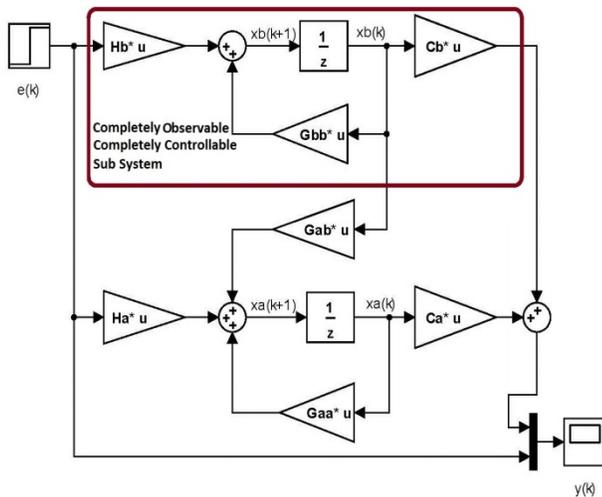
$$y_b(k) = \begin{bmatrix} y_4(k)/3.5 \\ y_5(k)/4.5 \end{bmatrix}.$$

$$\begin{bmatrix} \underline{x}_a(k+1) \\ \underline{x}_b(k+1) \end{bmatrix} = \begin{bmatrix} G_{aa} & G_{ab} \\ G_{ba} & G_{bb} \end{bmatrix} \begin{bmatrix} \underline{x}_a(k) \\ \underline{x}_b(k) \end{bmatrix} + \begin{bmatrix} H_a \\ H_b \end{bmatrix} e(k),$$

$$y(k) = [C_a \mid C_b] \begin{bmatrix} \underline{x}_a(k) \\ \underline{x}_b(k) \end{bmatrix} + D e(k),$$

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \\ x_5(k+1) \\ x_6(k+1) \\ x_7(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 9.9591e-03 & 4.8866e-04 & 1.4309e-09 & 0 & -3.8549e-11 & 2.5842e-07 \\ 0 & 9.9183e-01 & 9.7599e-02 & 2.8588e-07 & 0 & -1.5289e-08 & 7.7286e-05 \\ 0 & 0 & 1 & 2.9300e-06 & 0 & -4.6263e-07 & 1.5735e-03 \\ 0 & 0 & 0 & 0 & 0 & -8.9984e-05 & 1.5485e-01 \\ 0 & 0 & 0 & 0 & 1 & 9.3475e-03 & 6.4300e-05 \\ 0 & 0 & 0 & 0 & 0 & 8.7240e-01 & 1.2503e-02 \\ 0 & 0 & 0 & 0 & 0 & -5.6263e-04 & 9.6792e-01 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ x_5(k) \\ x_6(k) \\ x_7(k) \end{bmatrix} + \begin{bmatrix} 6.4718e-10 \\ 2.5842e-07 \\ 7.9080e-06 \\ 1.5735e-03 \\ 2.1735e-07 \\ 6.4300e-05 \\ 9.8387e-03 \end{bmatrix} e(k), \quad (3)$$

$$\begin{bmatrix} y_1(k) \\ y_2(k) \\ y_3(k) \\ y_4(k) \\ y_5(k) \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.5 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ x_5(k) \\ x_6(k) \\ x_7(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} e(k).$$



$$T = \begin{bmatrix} 1/3.5 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1/4.5 & 0 \end{bmatrix}. \quad (6)$$

The new system $G_h = T^{-1}G_{bb}T, H_h = T^{-1}H_b, C_h = C_{bb}T$ is given by (7).

$$G_h = \begin{bmatrix} 1 & 5e-5 & 2.27e-2 \\ 0 & 9.67e-1 & -2.53e-3 \\ 0 & 2.77e-3 & 8.72e-1 \end{bmatrix} = \begin{bmatrix} G_{haa} & G_{hab} \\ G_{hba} & G_{hbb} \end{bmatrix},$$

$$H_h = \begin{bmatrix} 7.61e-7 \\ 4.43e-2 \\ 6.43e-5 \end{bmatrix} = \begin{bmatrix} H_{ha} \\ H_{hb} \end{bmatrix}, \quad (7)$$

$$C_h = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = [C_{ha} \mid C_{hb}].$$

Let $K_e = [\alpha \ \beta]$ be the observer state gain matrix. Placing the pole of observer at origin puts condition (8) on observer closed loop characteristic polynomial.

Fig. 3. System partitioning into two subsystems

A. Prediction observer for subsystem 2

In order to accomplish pole assignment for subsystem 2, we have to design observer for unmeasured states. State x_6 is unmeasured [1] in vector \underline{x}_b . Following the standard procedure for minimum order prediction observer design in [13], we define a similarity transformation matrix T for system in (4) such that $C_h = C_{bb}T = [I \ 0]$ and $\underline{x}_b(k) = Tq(k)$.

$$|zI - G_{hbb} + K_e G_{hab}| = z \quad (8)$$

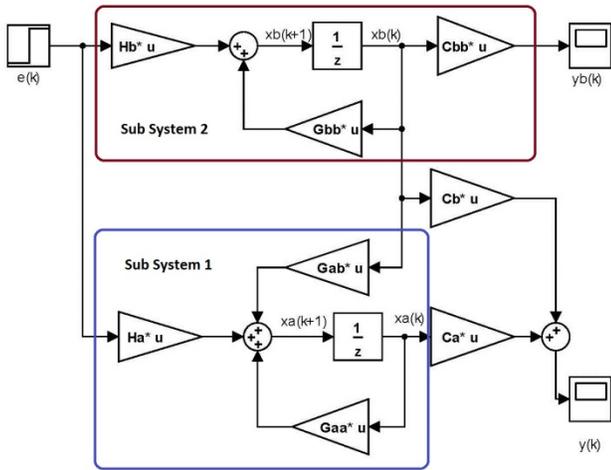


Fig. 4. Block diagram representation of subsystem1 & subsystem 2

Solution of (8) is non-unique. Assigning $\alpha = 1$ we get $\beta = -335.88$. Value $K_e = [1 \quad -335.88]$ is used in observer design algorithm (9).

$$q(k+1) = \{H_{hb} - K_e H_{ha}\} e(k) + \{G_{hba} - K_e G_{haa}\} y_b(k) + \{G_{hbb} - K_e G_{hab}\} \{q(k) + K_e y_b(k)\} \quad (9)$$

Observed state vector is given by (10).

$$\tilde{x}_b(k) = T \left\{ \begin{bmatrix} C_{hb} \\ 1 \end{bmatrix} q(k) + \begin{bmatrix} C_{ha} \\ K_e \end{bmatrix} y_b(k) \right\} \quad (10)$$

The procedure for minimum order prediction observer design is presented diagrammatically in Figure 5.

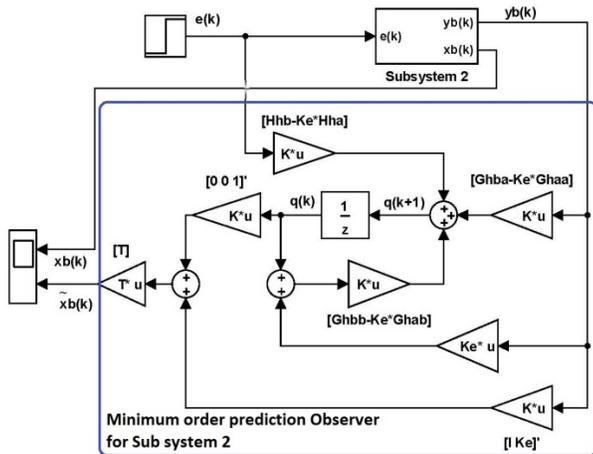


Fig. 5. Minimum order prediction observer for subsystem 2

B. Pole assignment to subsystem 2

We assign one pole at 0 and two poles at 0.8, an experimental optimal for fast response within actuator capacity. The characteristic polynomial becomes $z(z - 0.8)(z - 0.8)$.

Let K_b be the state gain for pole assignment then from [13] we have $K_b = \varphi(G_{bb})M^{-1}[0 \quad 0 \quad 1]^T$ where M is the controllability matrix of subsystem 2. This expression results in state gain given by (11).

$$K_b = [3.198e4 \quad 2.036e3 \quad 1.12e2] \quad (11)$$

Stabilized closed loop subsystem 2 is shown in Figure 6 with new reference input $v(k)$ and signal $e(k)$ given by (12).

$$e(k) = v(k) + K_b \tilde{x}_b(k) \quad (12)$$

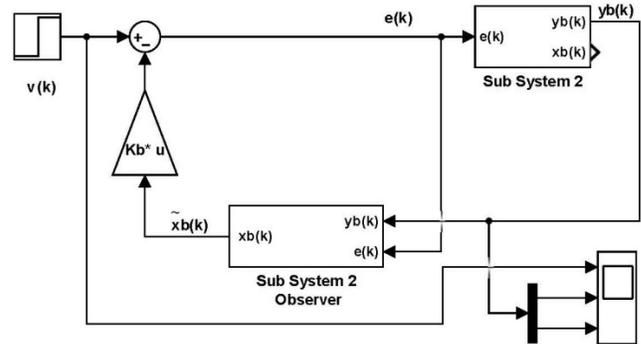


Fig. 6. Subsystem 2 stabilized by pole assignment and prediction observer

Step response of inner loop system containing observer and pole assignment is shown in Figure 7.

C. Inner loop system dynamics with stabilized subsystem 2

Using (12), (4) and (5) we get the dynamics of the overall system given by equation (13).

$$x(k+1) = G_{s2s} x(k) + H v(k) \quad (13)$$

$$y = C x(k)$$

System matrix in (13) with subsystem 2 stabilized is given by (14).

$$G_{s2s} = \begin{bmatrix} G_{aa} & G_{ab} \\ \underline{0} & G_{bb} - H_b K_b \end{bmatrix} \quad (14)$$

To implement rapid control prototyping we treat system in (13) as single input single output system with input $v(k)$ and distance covered by ball on beam $y_1(k)$ as an output and we

consider it as inner loop system given by transfer function in (15).

$$G_{il}(z) = K_{il} \frac{a_1 z^5 + a_2 z^4 + a_3 z^3 + a_4 z^2 + a_5 z + a_6}{z(z^5 + b_1 z^4 + b_2 z^3 + b_3 z^2 + b_4 z + b_5)} \quad (15)$$

Values of various parameters of transfer function (15) are tabulated in Table 1.

TABLE I. INNER LOOP SYSTEM PARAMETERS

Parameter	Value	Parameter	Value
a_1	0.0324	b_1	-4.5918
a_2	0.2928	b_2	8.4106
a_3	-0.2827	b_3	-7.6805
a_4	-0.3168	b_4	3.4965
a_5	0.2467	b_5	-0.6348
a_6	00276	K_{il}	1e-7

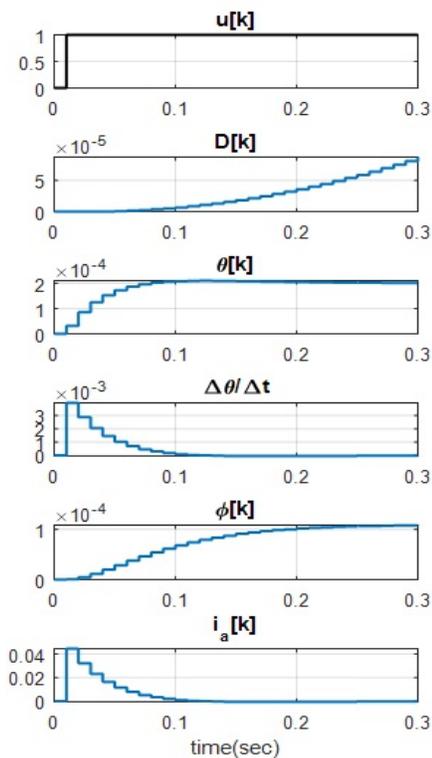


Fig. 7. Step responses for inner loop system

D. Rapid Control prototyping for inner loop

RCP implementation strategy is elucidated in Figure 8. Using hardware/software interface module i.e. NI DAQ, real plant is put into the software control loop with model compensator to be tuned. Responses of the system against various test commands are evaluated and controller parameters are adjusted accordingly until satisfactory performance is achieved.

Compensator model that has been used is given by (16).

$$C_{ol} = K_{ol} \frac{z - \zeta}{z - \rho} \quad (16)$$

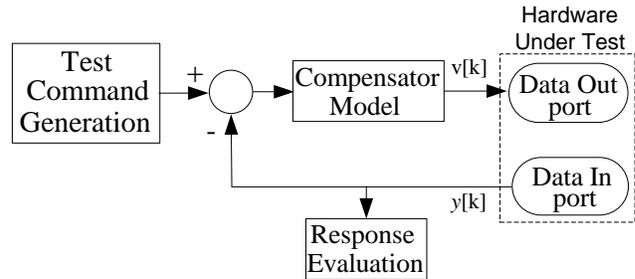


Fig. 8. Rapid Control Prototyping implementation strategy

Tuned parameter values for C_{ol} are given by (17).

$$\begin{aligned} K_{ol} &= 2262.4 \\ \zeta &= 0.9912 \\ \rho &= 0.9673 \end{aligned} \quad (17)$$

Overall implementation of multi loop control law that is hybrid of pole assignment and rapid control prototyping has been explained diagrammatically in Figure 9. Partial pole assignment is implemented on digital controller in inner loop followed by rapid control prototyping strategy implemented in outer loop using real time data acquisition, processing and monitoring in MATLAB. Figure 10 shows simulation of the hybrid multi-loop control algorithm. This simulation is used to obtain simulated responses in section IV. Figure 11 shows actual implementation of RCP strategy in Simulink.

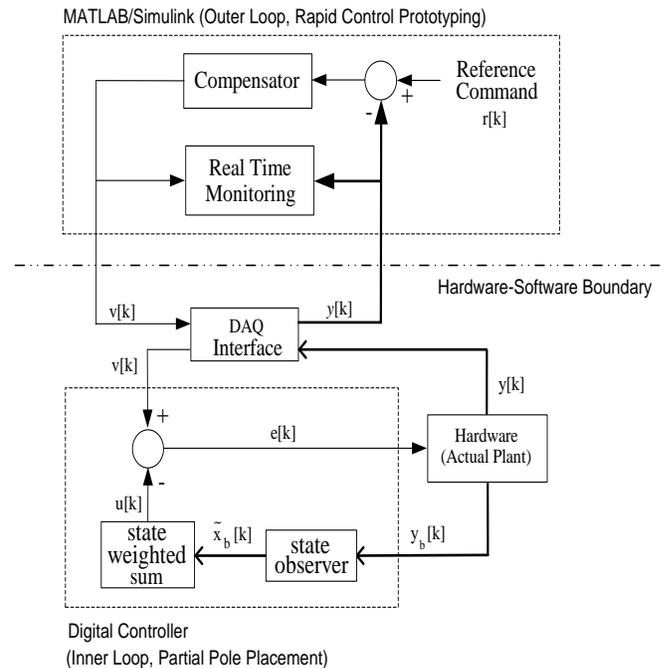


Fig. 9. Block diagram of multi loop hybrid control law implementation

IV. SIMULATION AND EXPERIMENTAL RESULTS

The proposed hybrid multi loop control law is simulated and experimentally tested. Figure 12 shows the step response

of position of the ball on beam. Actual response nearly follows simulation result. Response settles down in 1.5sec.

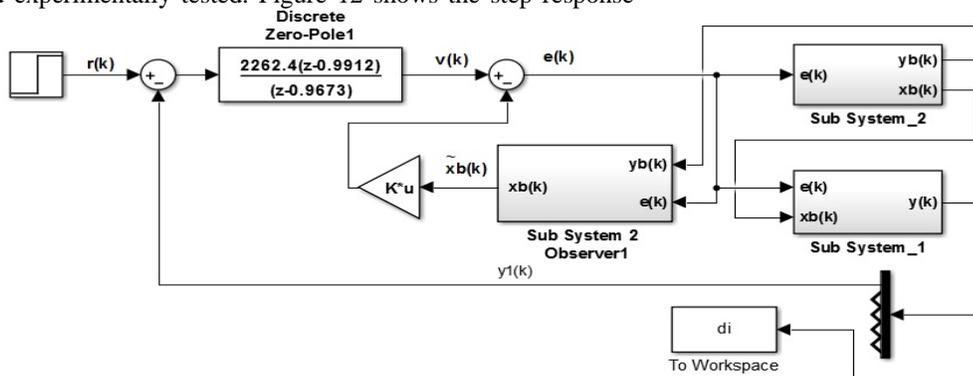


Fig. 10. Simulation of hybrid multi-loop control algorithm in Simulink

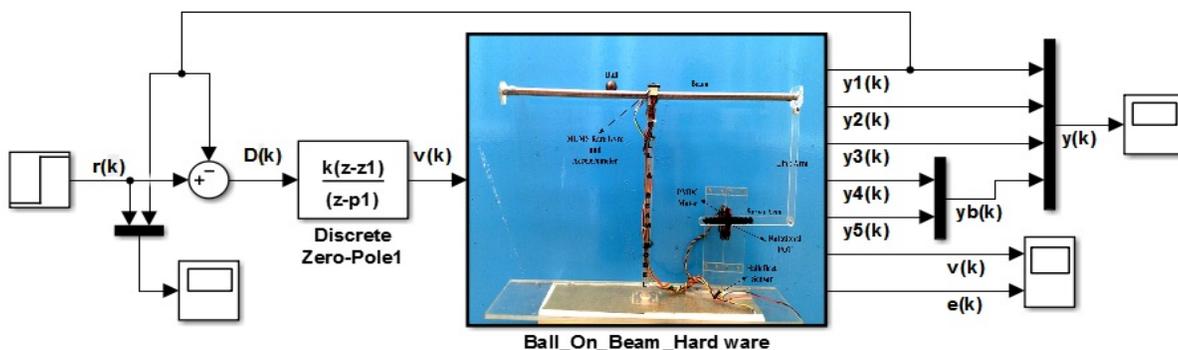


Fig. 11. Actual Rapid Control Prototyping implementation in MATLAB/Simulink

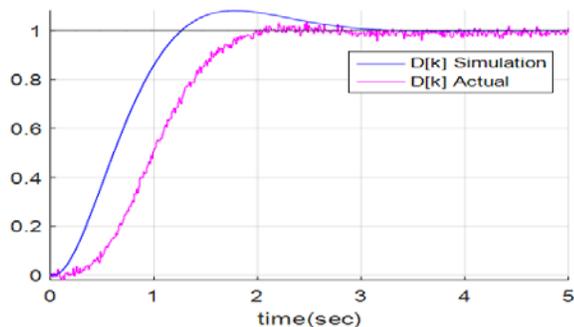


Fig. 12. Unit step response of position of ball on the beam

Unit step response in Figure 12 has zero steady state error. Figure 13 shows unit step response of the beam angle. The supply limitations result in the lag in the actual response during fast transients, however the steady state response well follows the simulation response.

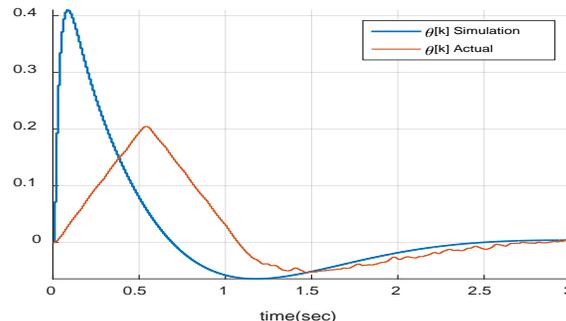


Fig. 13. Unit step response of the angle of the beam

Figure 14 shows unit step response of servo arm angle. The lag in the actual response during fast transients is due to the supply limitations. The steady state response follows simulation response.

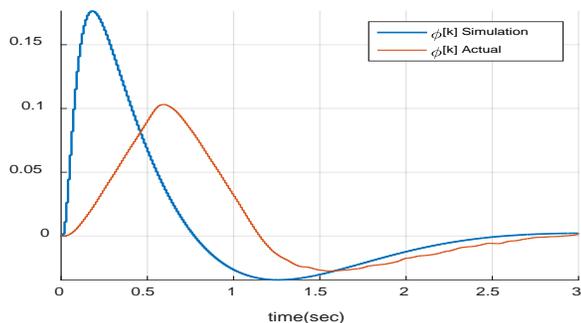


Fig. 14. Unit step response of angle of the servo arm

Figure 15 shows unit step response of PMDC motor current. Actual current waveform is limited within $\pm 3A$ power supply current bounds. Figure 16 shows unit step response of beam angular velocity. Actual angular velocity of the beam is bounded by $\pm 3A$ current limits of supply as shown in Figure 15. Figure 17 shows unit step response of motor input voltage. Actual input voltage waveform is bounded by $\pm 24V$ power supply limits for PMDC motor driver board. Figure 18 shows the unit step response of the control algorithm signal $v(k)$ from Figure 9. The supply limitations are not included in the simulations so that we may compare actual response with ideal conditions of the simulation and monitor ideal compensator robustness against practical limitations.

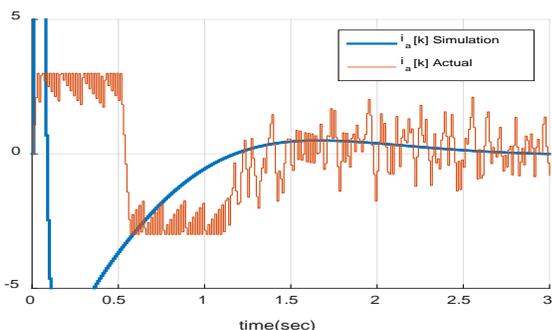


Fig. 15. Unit step response of the current of PMDC motor

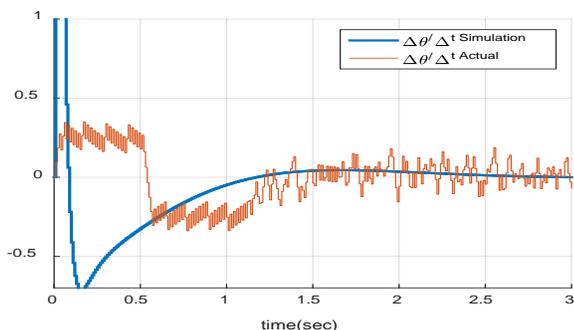


Fig. 16. Unit step response of the angular velocity of the beam

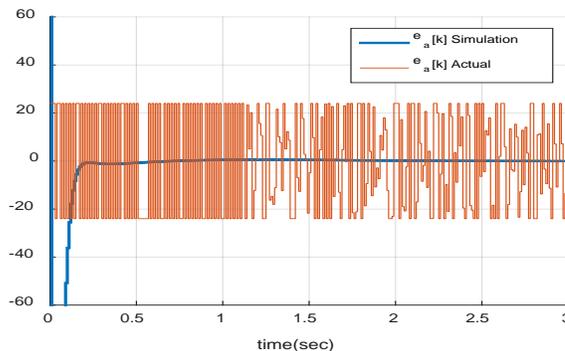


Fig. 17. Unit step response of the voltage applied to the PMDC motor

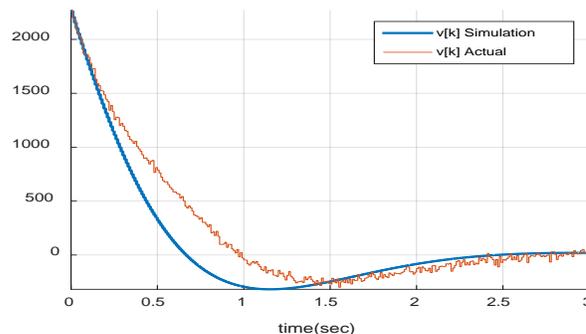


Fig. 18. Unit step response of signal $v[k]$

The Sinusoidal and Sawtooth reference tracking responses are shown in Figure 19 and Figure 20. Trapezoidal reference tracking response is shown in Figure 20. Actual response well follows the simulation responses with a constant steady state error for the ramp part of the reference input signal. Despite actual model has saturation limits for current and voltage yet responses well follow the simulation results. This tantamount to robustness of proposed technique against model inaccuracies.

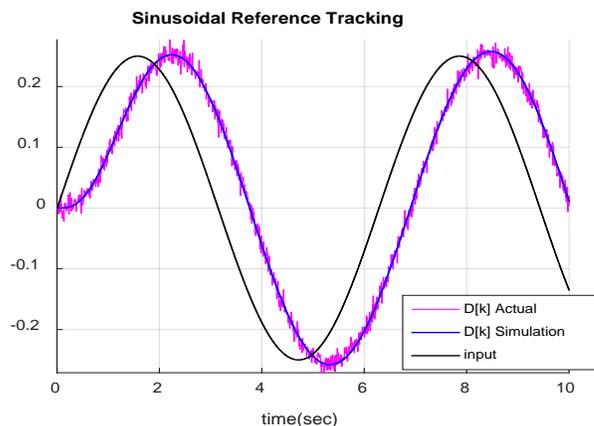


Fig. 19. Sinusoidal reference tracking response of the position of the ball

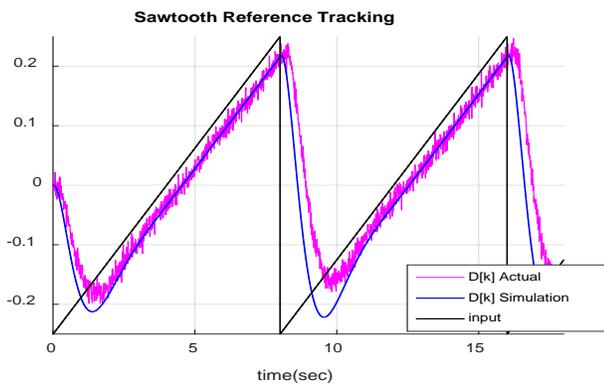


Fig. 20. Sawtooth reference tracking response for the position of the ball

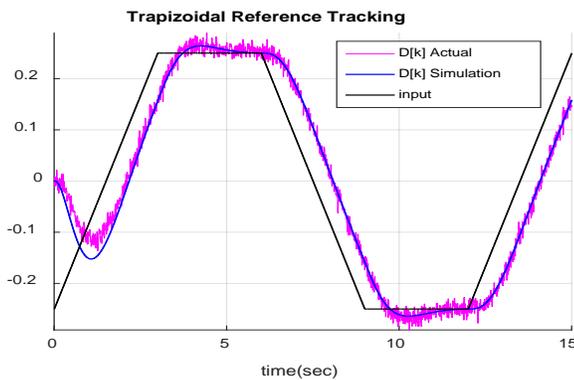


Fig. 21. Trapizoidal reference tracking response for the position of the ball

V. CONCLUSIONS

A novel compensator for the ball on beam platform is presented. Full order dynamic model of system is broken down into two parts. One part is controlled by partial pole assignment. The resulting system is compensated by rapid control prototyping. Experimental results validate that this hybrid compensator design strategy has given full control on all system outputs with system order reduction and it has given excellent results, especially regarding reference tracking and robustness against model inaccuracies.

REFERENCES

- [1] A. S. Haider, M. Nasir, B. Safir, and F. Farooq, "A Novel Ball on Beam Stabilizing Platform with Inertial Sensors," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 8, pp. 54–61, July 2015.
- [2] W. Y. Chung, and C. C. Ju, "An observer-based model reference adaptive iterative learning controller for nonlinear systems without state measurement," *Proc. of the 2013 International Conference on Fuzzy Theory and Its Applications*, Taipei, Taiwan, pp. 171–176, Dec 2013.
- [3] P. Rapp, O. Sawodny, and C. Tarin, "An immersion and invariance based speed and rotation angle observer for the ball and beam system," *Proc. of the 2013 American Control Conference*, Washington DC, pp. 1069–1075, June 2013.
- [4] R. M. Nagarale, and B. M. Patre, "Decoupled neural fuzzy sliding mode control of nonlinear systems," *Proc. of the 2013 IEEE International Conference on Fuzzy Systems*, Hyderabad, India, pp.1–8, July 2013.
- [5] D. Martinez, and F. Ruiz, "Nonlinear model predictive control for a Ball&Beam," *Proc. of the 2012 IEEE 4th Colombian Workshop on Circuits and Systems*, Barranquilla, Colombia, pp. 1–5, Nov 2012.
- [6] C. Wenzhuo, S. Xiaomei, and X. Yonghui, "Modeling and modulation of nonlinear ball-beam system controller based on matlab," *Proc. of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, Shenyang, China, pp. 2388–2391, May 2012.
- [7] D. N. Kouya, and F. A. Okou, "A new adaptive state feedback controller for the ball and beam system," *Proc. of the 2011 24th Canadian Conference on Electrical and Computer Engineering*, Niagara Falls, Canada, pp. 247–252, May 2011.
- [8] P. Z. Hua, Z. Geng, and L. C. Xiang, "Augmented state estimation and LQR control for a ball and beam system," *Proc. of the 2011 6th IEEE Conference on Industrial Electronics and Applications*, Beijing, China, pp. 1328–1332, June 2011.
- [9] W. Wei, and X. Peng, "A Research on Control Methods of Ball and Beam System Based on Adaptive Neural Network," *Proc. of the 2010 International Conference on Computational and Information Sciences*, Chengdu, China, pp. 1072–1075, Dec 2010.
- [10] P. Xu, L. Lei, and Z. Tan, "Human simulated intelligent control for ball and beam system," *Proc. of the 2010 29th Chinese Control Conference*, Beijing, China, pp. 96–100, July 2010.
- [11] A. I. Carlos, "The Lyapunov direct method for the stabilisation of the ball on the actuated beam," *International Journal of Control*, vol 12, no. 82, pp. 2169–2178, Oct 2009.
- [12] E. Lia, Z. Z. Lianga, Z. G. Hou, and M. Tana, "Energy-based balance control approach to the ball and beam system," *International Journal of Control*, vol. 6, no. 82, pp. 981–992, May 2009.
- [13] K. Ogata. *Modern Control Engineering*. 3rd ed., New Jersey: Prentice Hall, 1997.

A Feature Analysis of Risk Factors for Stroke in the Middle-Aged Adults

Focused on Perception of Sudden Speech and Language Problem

Haewon Byeon

Department of Speech Language Pathology & Audiology
Nambu University, Gwangju, Republic of Korea

Hyeung Woo Koh*

Jeju Seogwipo Medical Center
Jeju, Republic of Korea

Abstract—In order to maintain health during middle age and achieve successful aging, it is important to elucidate and prevent risk factors of middle-age stroke. This study investigated high risk groups of stroke in middle age population of Korea and provides basic material for establishment of stroke prevention policy by analyzing sudden perception of speech/language problems and clusters of multiple risk factors. This study analyzed 2,751 persons (1,191 males and 1,560 females) aged 40–59 who participated in the 2009 Korea National Health and Nutrition Examination Survey. Outcome was defined as prevalence of stroke. Set as explanatory variables were age, gender, final education, income, marital status, at-risk drinking, smoking, occupation, subjective health status, moderate physical activity, hypertension, and sudden perception of speech and language problems. A prediction model was developed by the use of a C4.5 algorithm of data-mining approach. Sudden perception of speech and language problems, hypertension, and marital status were significantly associated with stroke in Korean middle aged people. The most preferentially involved predictor was sudden perception of speech and language problems. In order to prevent middle-age stroke, it is required to systematically manage and develop tailored programs for high-risk groups based on this prediction model.

Keywords—C4.5; stroke; decision tree; risk factor; speech problem

I. INTRODUCTION

Stroke is a generic term for both cerebral infarction caused by the blockage of blood vessel in the brain and cerebral hemorrhage caused by the rupture of blood vessel (in the brain). As of 2013, death rate from cerebrovascular diseases was 50.2 persons per 100,000, which is the second highest right after cancer [1]. This order of death rate has not changed over the last 10 years and especially, stroke is serious in that it takes the second place in the cause of death regardless of gender.

Incidence of stroke is high in old age. According to 2013 Annual Report on the Cause of Death Statistics, death rate of cerebrovascular disease was 10.1 persons per 100,000 for people in their 40s compared to 277.4 for 70s, which is approximately 27 times higher [1]. In terms of life cycle, however, death rate of stroke skyrockets from 40s and over the recent 20 years, increase rate of stroke is the highest in 40s and 50s [2]. In addition, it has been reported that health risk behaviors causing stroke is most frequent in middle age [3]. Therefore, in order to maintain health during middle age and

achieve successful aging, it is important to elucidate and prevent risk factors of middle-age stroke.

In particular, in the case of stroke, even though operation is performed successfully, not only is the disease highly likely to accompany disabilities such as speech impediment during rehabilitation process but the patients also are likely to experience loss of labor. Middle age is the period when one accomplishes his/her goal of life. Acute diseases such as stroke not just are the direct cause of loss of job but cause enormous economic loss as well [4]. As of 2011, socio-economic loss from stroke (e.g. medical cost, transportation, nursing care, loss of production, etc.) in Korea surpassed US\$ 3.5 billion and among them, social cost for middle-aged people from age 40 to 50 (45% of total cost) was reported to be the greatest [5].

Although it is important to comprehend and systematically manage high-risk groups of middle-age stroke, risk factors of middle-age stroke are less known than old-age stroke and there is also lack of studies on its risk groups. So far, chronic diseases such as diabetes, hyperlipidemia and high blood pressure and life style factors such as smoking, drinking, eating habits and exercise and social and economic status are known to be risk factors of middle-age stroke [6][7][8][3].

However, since preceding studies which investigated risk factors of stroke did not adjust socio-economic factors such as occupation and level of income, it is difficult to find out social factors of middle-age stroke [9][10]. Moreover, as health risk behaviors tend to cluster together rather than individually exist (separate from other factors) [11], investigation on individual risk factor has a limitation in identifying high-risk groups of cardiocerebrovascular diseases with various characteristics.

Especially, recent studies reported that perception of sudden speech/language problems are major warning signs of stroke and in a survey on Korean adults, 80% of stroke patients perceived speech/language problems as a warning sign of stroke and 98% of stroke patients visited medical institutions due to speech/language problems as a warning sign, which is translated that perception of speech/language problem is a major factor of warning sign for stroke [12]. If high-risk groups are comprehended and managed by considering risk factors and warning signs of stroke, significant portion of strokes can be prevented and the time required to respond to emergency situation can also be reduced.

Recently, as a method of exploring multiple risk factors of diseases, data-mining analysis such as decision tree is being

used [13]. Use of data-mining can facilitate comprehension of attributes of diseases as well as multiple risk factors.

Since tendency of occurrence and risk factors of stroke differ depending on ethnicity and culture, in order to prevent stroke in Korea, it is necessary to develop a stroke prediction model reflecting demographic characteristics of middle age population of Korea and, based on it, manage them systematically.

This study investigated high risk groups of stroke in middle age population of Korea and provides basic material for establishment of stroke prevention policy by analyzing sudden perception of speech/language problems and clusters of multiple risk factors. Organization of this study is as follows; chapter 2 explains data resources and definition of variables and chapter 3 explains procedure for development of prediction model; chapter 4 suggested results of developed prediction model and chapter 5 presents results and suggests direction for future studies.

II. METHODS

A. Sources of data

Study subjects were adults aged 40–59 who participated in the 2009 Korea National Health and Nutrition Examination Survey (KNHANES), a nationwide representative survey of the non-institutionalized population in the Republic of Korea, and who then participated in a health survey [14].

The KNHANES is a nationwide cross-sectional survey conducted annually by The Korea Centers for Disease Control and Prevention. It employs a rolling sampling design that uses a complex, stratified multistage probability cluster survey of representative non-institutionalized civilians. The KNHANES sampling process is described in detail elsewhere [14]. Briefly, the creators of the survey redesign the KNHANES from once every years to once every year in order to provide timely health statistics for monitoring changes in health risk factors and diseases and developing associated public health policies and health programs. The 2009 KNHANES, conducted in January to December, was composed of three component surveys: a health interview, health examination, and nutrition survey. Trained medical staff and interviewers performed the health interview and health examination at a mobile examination center and at participants' households. The 2009 KNHANES was conducted on 12,722 persons out of 4,000 households with a participation rate of 82.8% (n=10,533).

This study targeted 2,885 persons who completed both the health survey and examination. Of these, 134 persons whose nonrespondents were excluded from the research, and data from 2,751 persons (1,191 males and 1,560 females) were analyzed.

B. Measurements

Outcome was defined as prevalence of stroke. Explanatory variables were included as age (40~49, 50~59), sex, final education (high school and lower, over college), Occupation

(economically inactive, manual workers, non-manual workers), income (quartiles), marital status (living with spouse, living without spouse, unmarried person), at-risk drinking (yes, no), smoking (non-smoker, past smoker, current smoker), subjective health status (good, fair, poor), moderate physical activity (yes, no), Diabetes (yes, no), hypertension (yes, no), sudden perception of speech and language problems (yes, no).

High-risk drinking was classified into normal (less than 12 points) and high-risk drinking (over 12 points) by using alcohol use disorder identification test (AUDIT) [15]. Regular moderate physical activity was defined as practicing moderately breathless exercise for more than 30 minutes per session over 5 days a week. Occupations classified based on the Korean Standard Classification of Occupations (KSCO-06)[16] were reclassified into economically inactive (unemployed person, homemaker), non-manual (managers & professionals, clerical support workers, service & sales workers), and manual (skilled agricultural & forestry & fishery workers, craft & plant and machine operators and assemblers, and unskilled laborers) occupations.

III. STATISTICAL ANALYSIS

A. Exploration on factors related to the stroke

For general characteristics, mean and percentage were presented and difference between groups based on stroke was analyzed by Chi-square test.

B. C4.5 algorithm

C4.5 is a decision tree algorithm developed by Quinlan [17], purpose of which is to create a tree which can exactly classify outcomes even with small number of tests. This algorithm constructs the simplest decision tree by using the concept of entropy based on information theory [18] (Figure 1).

In general, entropy means numbers representing disorder. As data sources are mixtures of proper cases and improper ones, they are very high in the degree of disorder. However, degree of disorder becomes 0 since terminal nodes are decided with one grade after decision tree is learned. Thus, it calculates information gain of each factor while it classifies data, keeping entropy close to 0.

Then, if the attribute with highest discerning power is selected as standard of classification, it makes as many branches as the number of kinds of given attribute values. Cases are divided according to the value of each branch and same processes are repeated in each branch. If there is no more decrease in information, the division stops [19].

Method of dividing tree by C4.5 algorithm is as follows; First, information gain of root node is acquired at input variables where target variables are composed of p and n.

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \quad (1)$$

Second, the gain is acquired which decreases degree of disorder in the case it is divided by attribute A or variable A at root node.

$$gain(A) = I(p, n) - E(A) \tag{2}$$

Third, among various attributes, node is divided by the attribute with greatest gain. If the divided node is composed only of either p or n, the node stops multiplying.

In case incidence rate is low as the outcome of this study, (which is) prevalence rate, there may be problems due to unbalanced data distribution [20]. In order to complement this unbalanced distribution, this study adjusted data balance by asymmetrically setting weight of misclassification costs considering prevalence rate of middle-age stroke in Korea [21]. Validity of the developed model was assessed with 10-fold cross-validation method.

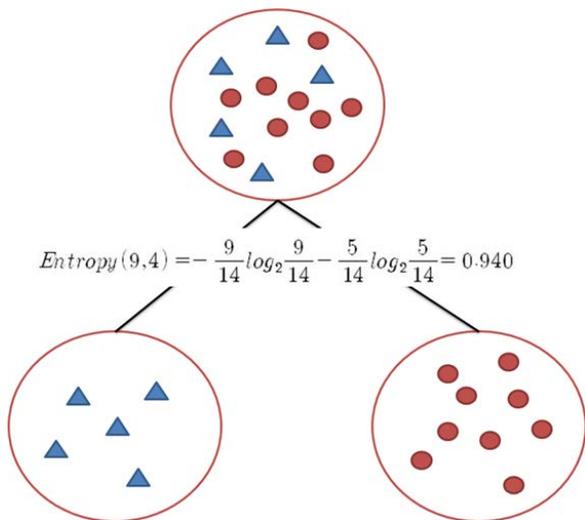


Fig. 1. Calculation of entropy

IV. RESULTS

A. Characteristics characteristics of subjects and potential factors related to stroke

General characteristics of subjects and factors related to stroke are presented in Table 1. Among the total of 2,751 subjects, number of those who have stroke was 33 (1.2%).

As the result of chi-square test, prevalence of stroke has statistically significant difference in age, gender, final education, income, marital status, diabetes, hypertension, and sudden perception of speech and language problems (p<0.05).

The prevalence of stroke was higher in aged 50~59 (1.8%), man (1.8%), high school and lower (1.5%), Groups the lowest income (3.4%), unmarried person (3.7%), those with bad subjective health (2.4%), diabetes (4.2%), hypertension (4.5%), and sudden perception of speech and language problems (83.4%).

TABLE I. GENERAL CHARACTERISTICS OF THE SUBJECTS BASED ON STROKE (UNIVARIATE ANALYSIS), N (%)

Characteristics	Stroke		P
	No (n=2,718)	Yes (n=33)	
Age			0.014
40~49	1,488 (99.3)	11 (0.7)	
50~59	11,230 (98.2)	22 (1.8)	
Sex			0.018
Male	1,170 (98.2)	21 (1.8)	
Female	1,548 (99.2)	12 (0.8)	
Education			0.036
High school and lower	2,032 (98.5)	30 (1.5)	
Over college	676 (9.6)	3 (0.4)	
Occupation			0.070
Economically inactive	740 (98.0)	15 (2.0)	
Non-manual workers	1,084 (99.1)	10 (0.9)	
Manual workers	877 (99.1)	8 (0.9)	
Income (quartiles)			<0.001
Q1	315 (96.6)	11 (3.4)	
Q2	583 (98.3)	10 (1.7)	
Q3	819 (99.0)	8 (1.0)	
Q4	977 (99.6)	4 (0.4)	
Marital status			0.002
Living with spouse	2,366 (99.1)	22 (0.9)	
Living without spouse	300 (97.1)	9 (2.9)	
Unmarried person	52 (96.3)	2 (3.7)	
At-risk drinking			0.769
No	1,812 (99.0)	19 (1.0)	
Yes	586 (98.8)	7 (1.2)	
Smoking			0.284
Non-smoker	1,626 (99.0)	33 (1.2)	
Past smoker	475 (98.1)	9 (1.9)	
Current smoker	617 (98.7)	8 (1.3)	
Moderate physical activity			0.250
Yes	449 (99.3)	3 (0.7)	
No	2,261 (98.7)	30 (1.3)	
Subjective health status			0.009
Good	1,175 (99.0)	12 (1.0)	
Fair	966 (99.3)	7 (0.7)	
Poor	568 (97.6)	14 (2.4)	
Diabetes			<0.001
Yes	160 (95.8)	7 (4.2)	
No	2,558 (99.0)	26 (1.0)	
Hypertension			<0.001
Yes	425 (95.5)	20 (4.5)	
No	2,293 (99.4)	13 (0.6)	
Sudden perception of speech and language problems			<0.001
Yes	2 (16.6)	10 (83.4)	
No	2,718 (99.2)	21 (0.8)	

B. Prediction model for stroke using C4.5 algorithm

Prediction model for stroke using C4.5 algorithm is presented in Figure 2. As the result of constructing statistical classification model using C4.5 algorithm after including variables set as factors related to stroke through chi-squared test, factors having significant effect were sudden perception of speech and language problems, hypertension, and marital

status. The most preferentially involved predictor was sudden perception of speech and language problems.

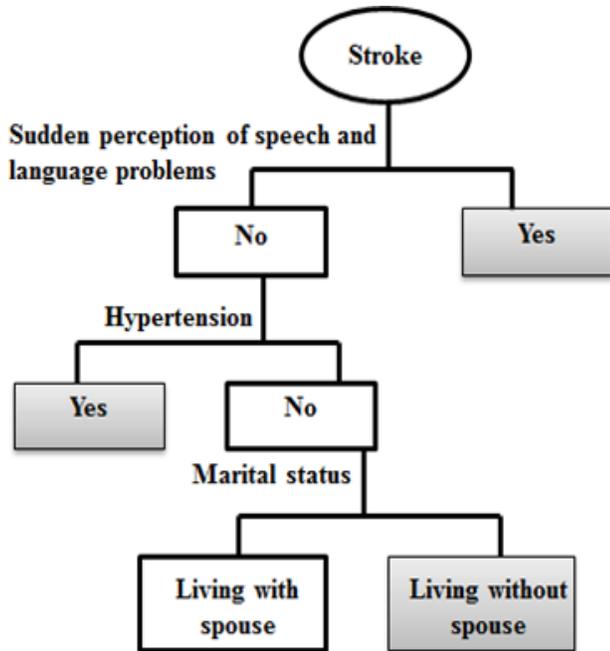


Fig. 2. Prediction model for stroke among Korean middle aged people

Table 2 is a profit chart of prediction model for stroke by C4.5 algorithm suggested in the higher order of path for subjects' improved gain. When this study drew out profit indicator for each node to seek out prediction paths for stroke, 3 nodes were confirmed as significant paths which effectively predict the stroke.

The first path with the biggest profit indicator for the prediction of the stroke was “middle-aged persons from the age of 30 to 58 who currently perceive sudden language problems” and its profit indicator was 8336.4%.

The second path was “middle-aged persons from the age of 30 to 58 who currently do not have sudden language problems or high blood pressure and do not live with spouse due to divorce or bereavement” and its profit indicator was 735.6%.

The third path was “middle-aged persons from the age of 30 to 58 who currently do not have sudden speech/language problems but have high blood pressure” and its profit indicator was 210.3%.

When the analysis on the prediction model by CART algorithm was completed, this study conducted 10-fold cross-validation test to assess developed prediction model. As the result of the 10-fold cross-validation test to compare stability of drawn-out model, drawn-out risk index was 0.360 and misclassification rate was 36% for cross classification model, showing the same risk index 0.352 and misclassification rate 35% of prediction model.

TABLE II. GAINS CHART OF PREDICTOR VARIABLE BY C4.5 ALGORITHM

Node no	Node n (%) ¹	Gain n (%) ²	Response % ³	Gain Index % ⁴	Group
2	12 (0.4)	12 (36.4)	83.4	1336.4	Middle-aged persons from the age of 30 to 58 who currently perceive sudden language problems
6	34 (1.2)	3 (9.1)	8.8	735.6	Middle-aged persons from the age of 30 to 58 who currently do not have sudden language problems or high blood pressure and do not live with spouse due to divorce or bereavement
3	436 (15.8)	11 (33.3)	2.5	210.3	Middle-aged persons from the age of 30 to 58 who currently do not have sudden speech/language problems but have high blood pressure

¹ Node n(%); node number, % to 2,751

² Gain n(%); gain number, % to 33

³ Response (%): The fraction of the stroke

⁴ Gain index (%):= 1336.4 in total 4 node

V. CONCLUSION

Early detection and management of high-risk groups of stroke enables healthy and happy aging. This study developed prediction model for middle-age stroke by using C4.5 algorithm. As the result of constructing stroke prediction model considering multiple risk factors, perception of sudden speech/language problems, high blood pressure and marital status were significant prediction factors for middle-age stroke and among them, perception of sudden speech/language problem was the most prioritized prediction factor. Numerous preceding studies have reported that perception of sudden speech/language problems is a major risk factor of stroke and those who perceived speech problem had higher rate of stroke [22][23]. However, these studies were limited to exploring individual risk factors while this study confirmed as the result of exploring multiple risk factors that combination of individual risk factors causes a synergy effect.

Another finding of this study was that marital status is major prediction factor for stroke. This study found out that middle-aged people from the age of 30 to 58 who do not live with spouse due to divorce, bereavement or separation are high-risk group for stroke. It is supposed that middle-aged

people who do not live with spouse have high risk of stroke since the middle-aged living alone not only have frequent health risk behaviors such as smoking but also are more vulnerable in health management. According to studies which researched on the relationship between marital status and health, married men who lived away from family had higher risk of accidents, alcohol and substance addiction, depression, death and cardiocerebrovascular diseases than men with stable marriage life and had 2.3 times more suicide rate, 4.7 times more death rate from alcohol and alcohol addiction and 1.7 times more death rate from cardiocerebrovascular diseases [24].

Especially, it has been reported that unstable marriage states such as divorce, separation and bereavement have negative effect on cardiocerebrovascular system by causing depression, which in turn increases death risks [25]. Hence, in order to prevent middle-age stroke, it is necessary to develop health management programs for the middle-aged without spouse. Furthermore, it is also necessary to prescribe guidelines for the prevention of middle-age stroke so that they will immediately visit medical institutions when they perceive sudden speech/language problems even if they do not have stroke-related diseases such as high blood pressure and diabetes.

Results of this study are expected to be an important ground to be considered in the strategy to prevent and manage stroke. In order to prevent middle-age stroke, it is required to systematically manage and develop tailored programs for high-risk groups based on this prediction model.

ACKNOWLEDGMENT

The author wish to thank the Korea Centers for Disease Control and Prevention that provided the raw data for analysis.

REFERENCES

- [1] Statistics Korea, Cause of death statistics 2013. Daejeon, Statistics Korea, 2013.
- [2] Ministry of Health and Welfare, 2001 National Health and Nutrition Survey. Seoul, Ministry of Health and Welfare, 2002.
- [3] S. Kaffashian, A. Dugravot, E. J. Brunner, S. Sabia, J. Ankri, M. Kivimäki, and A. Singh-Manoux, Midlife stroke risk and cognitive decline: A 10-year follow-up of the Whitehall II cohort study. *Alzheimer's & Dementia*, vol. 9, no. 5, pp. 572–579, 2013.
- [4] H. J. Lee, and M. Yi, Adjustment of middle-aged people with hemiplegia after a stroke. *Journal of Korean Academy of Nursing*, vol. 36, pp. 792–802, 2006.
- [5] National Rehabilitation Center, Report on the socio-economic costs of disorders. Seoul, National Rehabilitation Center, 2015
- [6] R. Behrouz, and C. J. Powers, (2015). Epidemiology of classical risk factors in stroke patients in the Middle East. *European Journal of Neurology*, E-pub, DOI: 10.1111/ene.12742, 2015.
- [7] N. Allen, J. D. Berry, H. Ning, L. Van Horn, A. Dyer, and D. M. Lloyd-Jones, Impact of blood pressure and blood pressure change during middle age on the remaining lifetime risk for cardiovascular disease: the cardiovascular lifetime risk pooling project. *Circulation*, vol. 125, no. 1, pp. 37–44, 2012.
- [8] J. Addo, L. Ayerbe, K. M. Mohan, S. Crichton, A. Sheldenkar, R. Chen, C. D. Wolfe, and C. McKeivitt, Socioeconomic status and stroke an updated review. *Stroke*, vol. 43, no. 4, pp. 1186–1191, 2012.
- [9] J. B. Olesen, G. Y. Lip, M. L. Hansen, P. R. Hansen, J. S. Tolstrup, J. Lindhardsen, C. Selmer, O. Ahlehoff, A. M. Olsen, G. H. Gislason, and C. Torp-Pedersen, Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *British Medical Journal*, vol. 342, doi: <http://dx.doi.org/10.1136/bmj.d124>, 2011.
- [10] L. Friberg, L. Benson, M. Rosenqvist, and G. Y. Lip, Assessment of female sex as a risk factor in atrial fibrillation in Sweden: nationwide retrospective cohort study. *British Medical Journal*, vol. 344, doi: 10.1136/bmj.e3522, 2012.
- [11] H. Byeon, and Y. Lee, Laryngeal pathologies in older Korean adults and their association with smoking and alcohol consumption. *Laryngoscope* vol. 123, no. 2, pp. 429–433, 2013.
- [12] Y. H. Lee, Y. T. Kim, G. J. Oh, N. H. Kim, K. H. Cho, H. Y. Park, H. S. Lee, Y. S. Ha, J. Cheong, J. K. Park, K. S. Lee, and H. S. Kim, Effects of community-based education and advocacy intervention on public awareness about the warning signs of stroke and the golden window of time. *Korean Journal of Health Promotion*, Vol.32, No.1, pp.1–10, 2015.
- [13] H. Wimmer, and L. Powell, A comparison of the effects of K-anonymity on machine learning algorithms. *International Journal of Advanced Computer Science and Application*. Vol. 5, No. 11, pp. 155–160, 2014.
- [14] Korea Centers for Disease Control and Prevention. The Korea national health and nutrition examination survey 2008, Seoul, Korea Centers for Disease Control and Prevention, 2009.
- [15] D. F. Reinert, and J. P. Allen, The alcohol use disorders identification test (AUDIT): a review of recent research. *Alcoholism: Clinical and Experimental Research*, vol. 26, no. 2, pp. 272–279, 2012.
- [16] Korea National Statistical Office. The Korean standard classification of occupations, Daejeon, Korea National Statistical Office, 2007.
- [17] J. R. Quinlan, C4. 5: programs for machine learning. Burlington, Elsevier, 2014.
- [18] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 13–19, 2014.
- [19] H. Byeon, Development of prediction model for endocrine disorders in the Korean elderly using CART algorithm. *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, pp. 125–129, 2015.
- [20] P. Tan, M. Steinbach, and V. Kumar, Introduction to data mining, Boston, Addison Wesley, 2006.
- [21] H. Byeon, The risk factors of laryngeal pathology in Korean adults using a decision tree model. *Journal of Voice*, vol. 29, no. 1, pp. 59–64, 2015.
- [22] D. W. Dietrich, N. J. Okon, D. V. Rodriguez, and A. M. Burnett, J. A. Russell, M. J. Allen, C. C. Fogle, S. D. Helgerson, D. Gohdes, T. S. Harwell. Rural community knowledge of stroke warning signs and risk factors. *Preventing Chronic Disease*, vol. 2, no. 2, pp. 1–8, 2005.
- [23] T. G. Robinson, A. Reid, V. J. Haunton, A. Wilson, and A. R. Naylor, The face arm speech test: does it encourage rapid recognition of important stroke warning symptoms?. *Emergency Medicine Journal*, vol. 30, no. 6, pp. 467–471, 2013.
- [24] W. G. Ringback, B. Burstom, and M. Rosen, Premature mortality among lone fathers and childless men. *Social Science & Medicine*, vol. 59, no. 2, pp. 1449–1459, 2004.
- [25] P. L. Morris, R. G. Robinson, and J. Samuels, Depression, introversion and mortality following stroke. *Australian and New Zealand Journal of Psychiatry*, Vol. 27, No. 3, pp. 443–449, 1993.

Analysis on Existing Basic Slas and Green Slas to Define New Sustainable Green SLA

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—Nowadays, most of the IT (Information Technology) and ICT (Information and Communication Technology) industries are practicing sustainability under green computing hoods. Users/Customers are also moving towards a new sustainable society. Therefore, while getting or providing different services from different ICT vendors, Service Level Agreement (SLA) becomes very important for both the service providers/vendors and users/customers. There are many ways to inform users/customers about various services with its inherent execution functionalities and even non-functional/Quality of Service (QoS) aspects through SLAs. However, these basic SLAs actually do not cover eco-efficient green issues or ethical issues for actual sustainable development. That is why green SLA (GSLA) should come into play. GSLA is a formal agreement incorporating all the traditional/basic commitments as well as respecting the ecological, economical and ethical aspects of sustainability. This research would survey on different basic SLA parameters for various services in ICT industries. At the same time, this survey would focus on finding the gaps and incorporating basic SLA parameters with existing green computing issues and ethical issues for different services in various computing domains. This research defines future GSLA in relationship with ICT product life and three pillars of sustainability. The proposed definition and overall survey could help different service providers/vendors to define their future GSLA as well as business strategies for this new transitional sustainable society.

Keywords—SLA; GSLA; Green ICT; Sustainability; IT ethics; ICT Product Life

I. INTRODUCTION

SLA is defined as a formal document between an IT service provider and one or more customer outlining Service Commitment [1]. The main issue is that most of these traditional/basic SLA actually do not cover eco-efficient green issues. Currently, cloud and grid computing and many data centers acts 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 [2]. At each level, they maintain a SLA with respect to their parties. Therefore, this shows the growth rate of SLA in recent time as well as the need of GSLA for actual sustainability achievement in the industry. Presently, the revolution of ICTs and ITs in daily average life has also resulted in the increase of Green House Gas (GHG), due to continual increase in global “carbon footprint”. In 2007, the ICT sector produced as much GHG as the aero industry and is

projected to grow rapidly [3, 4]. If ICT has a negative impact on environment, it can be also be used for greening the other human activities (logistic, city, industry etc) in this new society. Indeed, the dimensions of Green Informatics contributions are: the reduction of energy consumption, the rise of environmental awareness, the effective communication for environmental issues and the environmental monitoring and surveillance systems, as a means to protect and restore natural ecosystems potential [5]. At the same time, many IT and ICT industries or service providers need to think about their business scope in the light of green perspective. However, the IT and ICT sectors mostly concern about energy or power consumption, carbon, recycling and productivity issues under greening computing lens. On the contrary, most of the recent industries overlooked many green parameters under sustainability lens. Therefore, with the increase attention that green informatics and sustainability practice within our society, it is timely to not only conduct SLAs for traditional/basic computing performance metrics or only on energy or carbon footprint issues, but also to relate the effort of conducting green computing with respect to 3Es of (Ecology, Economy and Ethics) sustainability pillars. Therefore, the journey of GSLA is getting importance in ICT business world. This research did thorough review on existing basic SLA indicators for network, storage, compute and multimedia domain in IT industry. Then, it goes deep down for finding more current green performance indicators in some datacenter’s SLAs. In addition, a new future GSLA definition proposed, which shows the importance and relationships of ICT product life cycle. Moreover, the GSLA should be designed considering three pillars of sustainability. Finally, GSLA research briefly describes the management complexities and some challenges.

The rest of the work is organized according to 4 sections—the next Research Review section discusses and analyses some existing scientific theory and practical works based on basic SLA for four different services in the industry. Empirical Work Review section identifies all basic SLA indicators for network, compute, storage and multimedia services, which do not cover any eco-efficient parameters. Next, the following subsection discovers most of the green indicators for various services, usually used in grid and cloud computing, datacenters etc. Basic SLA and existing GSLA parameters are also derive and organize in details through existing empirical viewpoint. The existing GSLA subsection actually shows currents trends of the industry to practice sustainability under greening lens. The future GSLA definition sections describe the gap between greening and sustainability in the current

industry. In addition, this section gives some hints about future indicators for sustainable future GSLA. Moreover, it also depicts the relationships of future GSLA with ICT product life in the industry. Finally, the conclusion gives brief discussion about few challenges for the ICT engineer to incorporate and trade-off between all existing indicators and new indicators for sustainable achievement in the ICT industry.

II. RESEARCH REVIEW

This GSLA work did rigorous literature review and analysis based on existing work in the field of SLA, GSLA, green computing, energy optimization in IT industry, impact of ICT on environment and natural resource, IT ethics issues, IT for Sustainability etc. In the findings, GSLA research divides its work based on basic SLA and then existing GSLA for various types of services from their providers. The existing theory work on basic SLA and GSLA discusses in the following sub sections.

A. Basic SLAs

S. A. Baset [6] gave an idea for presenting SLA for different cloud service providers. He surveyed on some well known public IaaS providers and found a common anatomy of basic SLA with some common metrics. In [7], H. Lee *et al.* offered a general SLA monitoring system architecture that could be used to monitor service levels provided by some

network, Internet and application service providers. Their work showed much clear idea of finding some QoS parameters, measurement metrics for various services. In contrast, L. Jin *et al.* [8] presented another approach to model and understand the relationship between customers and some web service providers, which is very important for designing basic SLA and Green SLA. A. Paschke *et al.* contributed to a systematic categorization of basic SLA contents with a particular focus on SLA metrics in IT industry [9]. They categorized five basic IT object classes and their performance indicators in SLA. J. Lankinen *et al.* [10] surveyed on security profiles of some existing well known storage service providers like Amazon, Apple iCloud, Dropbox etc. In [11], the paper presented SLA for voice and Internet services covering basic performance indicators. Most of the paper found on basic SLA discussed performance based indicators for various services in recent ICT arena. Some empirical work found on SLA implementation, management, automation, template design and assessment in the context of business requirement. Very few scientific works found on interesting aspects such as security and privacy issues on traditional SLA, which could be important for green SLA research under IT ethics concept. Table I shows the brief idea of basic SLA work through some interesting criteria of SLAs as column subheads. The cell identified with “X” symbol means that, the authors mentioned and worked on that criteria of basic SLAs.

TABLE I. ANALYSIS OF EXISTING BASIC SLA WORKS

Author Lists	Analysis Criteria						
	Services	Information	Methodology	Implementation	Assessment	Monitoring	Reuse to Green SLA
S. A. Baset [6]	X		X	X		X	
H. Lee <i>et al.</i> [7]	X	X		X	X	X	
L. Jin <i>et al.</i> [8]	X	X	X	X			
A. Paschke <i>et al.</i> [9]		X	X		X		
J. Lankinen <i>et al.</i> [10]	X	X			X		X
Anonymous [11]	X	X					X
C. Raibulet <i>et al.</i> [12]				X	X		
V. Stantchev <i>et al.</i> [13]		X	X				X
N.J. Dingle <i>et al.</i> [14]		X				X	X
T. Unger <i>et al.</i> [15]		X	X				
E. Marilly <i>et al.</i> [16]	X	X		X		X	
T. Onali [17]	X	X		X			
H. Ludwig <i>et al.</i> [18]	X			X			X
P. Hasselmeyer <i>et al.</i> [19]	X			X	X	X	
Anonymous [20]		X	X				
E. Wustenhoff [21]						X	
Anonymous [22]		X					X

B. Green SLAs

S. Klingert *et al.* [23] introduced the notion of Green SLAs. However, their work focused on indentifying known hardware and software techniques for reducing energy consumption and integrating green energy. In [4] and [5], the authors showed the impact of ICT in a natural environment and resources in this world. Z. S. Andreopoulou [5] proposed a model ICT for Green and Sustainability whereas SMART 2020 report [4] gave the idea of GHG emission from the ICT sector. G. V. Laszewski *et al.* [24] invented a framework towards the inclusion of Green IT metrics for grids and cloud computing. According to Md. E. Haque *et al.* [25], high performance computing cloud providers offer a new class of

green services in response to practicing explicit sustainability goals in their field. R. R. Harmon *et al.* [26] defined the term Green Computing as the practice of maximizing the efficient use of computing resources to minimize environmental impact. They also discovered that, sustainable IT services require the integration of green computing practice such as power management, virtualization, cooling technology, recycling, electronic waste disposal and optimization of IT infrastructure. Finally, the white paper [22] provided some qualitative parameters in cloud service SLA which was very important for proposing Green SLA. In [27] and [28], the authors discussed one of the most promising concepts in Green SLA- IT Ethics issues. In their research, they showed the concepts of organizing ethics programs in IT industry. The

existing scientific work on green SLA is mainly based on cloud and grid computing environment. Some works have been found on green services, operation and framework for the cloud infrastructure [30]; few work done on green performance indicators for designing SLA. The next Table II demonstrates the analysis of exiting green SLA works with

some criteria, such as green services and operations, greening practice, green metrics, framework development and monitoring. Here some papers also discussed IT ethics issues briefly. Therefore, IT ethics need to include here as an important analyzing criteria in the table.

TABLE II. ANALYSIS OF EXISTING GSLA WORKS

Author Lists	Analysis Criteria					
	Green Services & Operations	Greening Computing Practice	Metrics Information	Framework/ Implementation	Assessment	IT Ethics issue
L. Wu <i>et al.</i> [4]		X				X
Z. S. Andreopoulou [5]		X				X
Klingert <i>et al.</i> [23]	X	X				
G. V. Laszewski <i>et al.</i> [24]	X	X	X	X		
Md. E. Haque <i>et al.</i> [25]				X	X	
R. R. Harmon <i>et al.</i> [26]	X	X	X			
Fritz H. Grupe <i>et al.</i> [27]				X		X
R. Herold [28]						X
N. Agarwal <i>et al.</i> [29]	X	X				
Ahmed <i>et al.</i> [30]	X	X		X	X	
Li <i>et al.</i> [31]		X		X		
Kien Le <i>et al.</i> [32]				X		
M. Nichollas [33]						X
A. P. Bianzino <i>et al.</i> [34]	X	X		X		
A. Atrey <i>et al.</i> [35]		X	X	X		
A. Orgerie [36]	X	X				

III. EMPIRICAL WORK REVIEW

In the findings on existing empirical work, green SLA research splits its work based on basic SLA and then existing green SLA for various types of services from their providers such as Network, Compute, Storage and Multimedia [37].

In the basic SLA section, findings are divided into four main services as network, compute, storage and multimedia [37]. Most of the performance indicators in basic SLA sections were quantitative parameters and they were simple to evaluate, control and monitor.

A. Basic SLAs for Network, Compute, Storage and Multimedia domain:

Usually network services domain include connectivity and switching as well as advanced network systems and management functions for well known network service providers.

The basic SLA for network specifies service level commitments which are applied to measure and evaluate network performance and give proper support for their clients. Usually, from different network service provider, the following performance indicators [7, 9, 11, 24] found in their SLAs are- *Network Availability, Delay, Latency, Packet Delivery Ratio, Jitter, Congestion, Flow Completion time, Response time, Bandwidth, Utilization, MTBF (Mean Time Between Failure), MTRS (Mean Time to Restore Services), Solution time, Resolution time, LAN/WAN period of operation, LAN/WAN Service Time, Internet access across Firewall, RAS (Remote access Services)* (Table III).

TABLE III. BASIC SLAS FOR NETWORK SERVICES

Sl.No.	Performance Indicator Name		Unit
1.	Network Availability	Connectivity (IPPM)	% (Percentage)
		Functionality	
2.	Delay	One way delay	Time in Milliseconds
		RTT delay (Round Trip Time)	
3.	Latency		Time in Milliseconds
4.	Packet Delivery Ratio(PDR) or Packet Loss Ratio(PLR)		% (Percentage)
5.	Jitter		Time in Milliseconds
6.	Congestion		% (Percentage)
7.	Flow Completion Time (FCT)		Time in Milliseconds/ Seconds
8.	Response Time		Time in Milliseconds
9.	Bandwidth		Hertz (Hz)
10.	Utilization		% (Percentage)
11.	LAN/WAN period of Operation		Time in Milliseconds/ Seconds
12.	LAN/WAN Service Time		Time in Milliseconds
13.	MTBF (Mean Time between Failure)		Time in Milliseconds
14.	MTRS (Mean Time to Restore Services)		Time in Milliseconds
Sl.No.	Performance Indicator Name		Unit
15.	Solution Times		Time in Seconds/Minutes/ Hours
16.	Internet access across Firewall		YES/NO
17.	RAS (Remote Access Service)		YES/NO
18.	Resolution Time (TTR)		Time

Among these performance indicators, only *Internet access across Firewall* and *RAS* are subjective indicators- there is no standard procedure to evaluate or calculate these indicators. Some indicators like *Bandwidth, Utilization, and Congestion* are related to link capacity whereas *Availability, Delay, Jitter, Response Time* etc. associated with time related information for different network service providers.

Most the cloud, grid service companies provides computing service to their consumers. In recent time, the Service Oriented Architecture (SOA) also comes into the computing field. The main point is that there is research on building middleware SLA infrastructure for computing services. Some of the current work: the European Union-funded Framework 7 research project, SLA@SOL, which is research on aspects of multi-level, multi-provider SLAs within service-oriented infrastructure and cloud computing [38]. The basic SLA parameter [9, 11, 22, 24] for computing domains are:- *Broad Network Accessibility, Multi-tenancy, Rapid Elasticity, Scalability, Resource Pooling Time, Solution Time, Response Time, Availability (MTBF & MTTR), Capacity, Virtualization, Delay, Resolution Time and Logging & Monitoring*. Here, *Broad Network Accessibility, Multi-tenancy* and *Logging & Monitoring* are informative indicators presented in their SLAs (Table IV).

TABLE IV. BASIC SLAS FOR COMPUTE SERVICES

Sl.No.	Performance Indicator Name	Unit
1.	Broad Network Accessibility	% (Percentage) Or YES/NO
2.	Multi-tenancy	YES/NO
3.	Rapid Elasticity	% (Percentage)
4.	Scalability	% (Percentage)
5.	Resource Pooling Time	Time in Milliseconds Or Seconds
6.	Solution Time	Time in Seconds/Minutes/ Hours
7.	Response Time	Time in Milliseconds Or Microseconds
8.	Availability	MTBF MTTR Time in Milliseconds Or Seconds
9.	Capacity	Number Or Request per Minutes
10.	Virtualization	% (Percentage)
11.	Delay	Time in Milliseconds
12.	Service Time	Time
13.	Logging & Monitoring	YES/NO
14.	Resolution Time (TTR)	Time

The storage domains are typically handled by cloud storage provider. Interestingly, today's cloud storage SLAs just ensure uptime guarantee but not data availability and data protection. In some case, traditional SLAs just mention about data storage security and backup but there is no proper authority or standard to check their commitments. Some common basic SLA performance indicator [7, 9, 11] for storage services are as follows:- *Availability, Response Time, Maximum Down Time, Uptime, Failure Frequency, Period of Operation, Service Time, Accessibility, Backup, Physical Storage Backup, Transportation for Backup, Size, Data Accessibility, Security*. Among all these parameters, some of them are just informative such as *Accessibility, Backup,*

Physical Storage Backup, Transportation for Backup, and Security (Table V). These parameters might vary according to human perspective.

TABLE V. BASIC SLAS FOR STORAGE SERVICES

Sl.No.	Performance Indicator Name	Unit
1.	Availability	% (Percentage) per time
2.	Response time	Time in Milliseconds
3.	Maximum down time	Time Or % (Percentage)
4.	Failure Frequency	% (Percentage)
5.	Periods of Operation	Time in Milliseconds/ Seconds
6.	Service Time	Time in Hours/Day
7.	Accessibility	YES/NO
8.	Back up	YES/NO
9.	Physical Storage Back up	YES/NO
10.	Transportation of Back up	YES/NO
11.	Size	Number in Bytes
12.	Data accessibility	Number per seconds
13.	Security	YES/NO

Multimedia service domain SLAs are classified into three broad application areas- Audio, Video and Data. It is challenging to monitor and evaluate some qualitative indicator such as *Mean Opinion Score (MOS)* and *Lip Synchronization* for one way video, conferencing or in videophone. These could vary among different consumers at the same time. Most of the SLA indicators for multimedia domain for different applications are *Information Loss (PLR), Jitter, One way Delay, MOS, Lip Synchronization, and Security Policy* [17]. Next Table VI shows all performance indicators for multimedia services in their SLAs.

TABLE VI. BASIC SLAS FOR MULTIMEDIA SERVICES

Media	Application Name	Performance Indicator Name	Unit
Audio	Conversational Voice	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
		Delay Variation (Jitter)	Time in Milliseconds
	Voice Messaging	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
		Delay Variation (Jitter)	Time in Milliseconds
Video	One way Video	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
		Mean Opinion Score (MOS)	Number (0 to 5)
Media	Application Name	Performance Indicator Name	Unit
Video	Videophone	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
		Mean Opinion Score (MOS)	Number (0 to 5)
		Lip Synchronization	Time in

			Milliseconds
Data	Still Images	One way Delay	Preferred or Acceptable
		Information Loss (Packet Loss Ratio)	% (Percentage)
	Interactive Game	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
	E-mail	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
	Web-browsing	One way Delay	Preferred or Acceptable
		Information Loss (Packet Loss Ratio)	% (Percentage)
	Transaction Services e.g. e-commerce, ATM	Information Loss (Packet Loss Ratio)	% (Percentage)
		One way Delay	Time in Milliseconds
		Security Policy	YES/NO

B. Existing Green SLA (GSLA)

Most of the GSLA performance indicator corresponds to traditional high performance distributed computing environment such as grid and cloud computing industry. Currently, several IT and ICT industries provide their GSLAs with green computing practice. GSLA survey shows that most of existing GSLAs are mainly focused on energy/ power, carbon footprint, green energy, recycling issues. Additionally, several existing GSLA also demonstrates their productivity issues with necessary monitoring unit. In recent days, various research draws attention only on minimizing energy consumption while improving networking performance on wireless connection under green computing hood [39, 40].

Table VII depicts the performance indicators and their unit for different services considering green computing practices. The table has several headings. *Green Computing Domain* mentions the category of green computing practices in IT industry; *Performance Indicator Name* is the notion which used an evaluating, monitoring metric for defining performance in GSLAs, and then their measurable unit as *Unit* column. All these performance indicators help various service providers and consumers either to design or to choose services mainly with respect to energy consumption, renewable energy usages, carbon emission issues and productivity issues in recent time. However, the IT industry needs to find out new services for achieving sustainability as current trends of the society shows that people are much more concerned about new issues, such as recycling, obsolescence, ICT pollution, ethical aspects etc. It is also important to mention that, monitoring of GSLA is vital to respect the services by concerned parties.

TABLE VII. PERFORMANCE INDICATOR FOR DIFFERENT SERVICES CONSIDERING EXISTING GSLA

Green Computing Domain	Performance Indicator Name	Unit
Energy/ Power	Total Power Consumption [26, 41]	kW-h (Kilowatt-hour)
	PUE (Power Usages Effectiveness) [24, 35, 37, 42]	Number (1.0 to ∞) Or Dimensionless
	DCIE (Data Center Infrastructure Efficiency) [24, 38, 42]	% (Percentage)
	CPE (Compute Power Efficiency) [35]	Watts
	SPECPower [24, 35]	Watt
	JouleSort [26]	kW/J
	WUE (Water Usages Effectiveness) [35]	Liter/kW-h
	TDP (Thermal Design Power) [42]	Watts
	ERF (Energy Reuse Factor) [35]	Number [0 to 1.0]
	ERE (Energy Reuse Effectiveness) [35]	Number [0 to ∞]
	GEC (Green Energy Co-efficient) [35]	Number [0 to 1.0]
	ITEE (IT Equipment Energy Efficiency) [43]	% (Percentage)
	ITEU (IT Equipment Utilization) [43]	Number
	HVAC (Heating, Ventilation, Air-conditioning) Effectiveness [42]	Dimensionless
	Cooling System Efficiency [42]	kW/ton
Carbon footprint	CUE(Carbon Usages Effectiveness) [35]	KgCO2 per kW-h
	DPPE (Data Center Performance Per Energy) [43]	Number [0 to 1]
Recycling	e-Wastage Or IT Wastage [42]	Gm (Gram)
	Recycling [37,44]	% (Percentage)
Productivity	DCP (Data Center Productivity) [35]	Not Available
	DCeP (Data Center Energy Productivity) [24,35]	Not Available
	Analysis Tool [26]	Not Known
	EnergyBench [26]	Numeral Rating
Costing Information	ScE (Server Compute Efficiency) [35]	% (Percentage)
	Energy/Power Cost [41]	Currency [according to country]
Others	SWaP (Space, Wattage and Performance) [24, 35]	Not Available
	User Satisfaction [11, 24]	Number [0 to 5]
	Mean Opinion Score (MOS) [11, 24, 45]	Number [1 to 5]
	Reliability [24]	Number [0.0 to 1.0]
	Air Management Metric [42]	F (Fahrenheit)
	UPS System Efficiency [42]	% (Percentage)
	Risk Assessment [11, 24]	% (Percentage)

IV. FUTURE GREEN SLA (GSLA) DEFINITION

In existing GSLAs, most of the performance indicators mainly concentrate on energy consumption issues and productivity concern in cloud and grid computing industry (Table VII). Most of the existing GSLA do not consider recycling, radio wave, toxic material usage, noise, light pollution for sustainable development. Moreover, people's interaction and IT ethics issues, such as user satisfaction, intellectual property right, user reliability, confidentiality etc are also missing in current GSLA under green computing lens. Next section discusses the proposed new performance indicators of GSLA for achieving sustainability from 3Es perspectives (Ecological, Economical and Ethical). Fig.1 shows the concepts of 3Es relationship, that ICT engineer can use as a guideline to respect all the facets of sustainable development.



Fig. 1. 3Es for Sustainability

The proposed definition according to Fig.1 could be “the GSLA should aggregate and satisfy all three main pillar of sustainability achievement- *Ecology Pillar*, *Economy Pillar*, and *Ethics Pillar*”. There must be trade-offs between 3Es to achieve sustainable development under green computing domains. Under *Ecology Pillar* the following new indicators should take into consideration while developing new services or application in the ICT field, such as, *Recycling*, *ICT Toxic Material Usage limit*, *ICT Radio Wave guideline*, *Pollution level* and *Obsolescence Indication* etc. Moreover, at the same time, *Economic Pillar* needs to aggregate some new indicators in future GSLA;- *Carbon Taxation*, *ICT Product Life Cycle Cost*, *Civil Engineering Cost*, *Cooling cost*, *Energy Cost* etc. Moreover, research shows that, in most industries the green computing practice focuses on the ecological, economical point but usually neglect human's interaction and ethical

aspects [37]. The use of ethics in IT and ICT field covers many new indicators such as *Satisfaction level*, *Intellectual Property Right*, *Reliability*, *Confidentiality*, *Security and Privacy*, *Gender/Salary/Productivity Information*. The ICT companies should also analyze their social responsibilities towards their customer, employee and community through developing IT Ethics program and guideline [37, 46]. All of these indicators are usually subjective and informative, thus making GSLA assessment difficult in future. On the other hand, ecological and economic indicators seem might be easy to evaluate and monitor.

In this section, this research gives some idea most of the important missing performance indicators with respect to three pillars of sustainability and this will definitely help ICT and IT service providers to develop and design their existing GSLA more greener for achieving sustainability as well as making more profit in their businesses. However, ICT engineer would face some challenges to incorporate, manage and finding the relationship between all new indicators for GSLA under three pillars of sustainability in future. To achieve sustainability, the future *GSLA* should aggregate and satisfy all three entities in their existing GSLA model- *Ecology Pillar*, *Economy Pillar*, *Ethics Pillar*. Now, it the matter of urgency that, to achieve sustainability the ICT industry need to indentify more new services from users perspective under this three pillar too. It is important to indicate that, the *ICT Product Life Cycle* must need to include at the first level of GSLA model as this entity have direct relationship to calculate existing ecological, economical and ethical indicators, such as carbon/GHG emission, energy consumption, recycling, energy cost, pollution level, comfort level etc [Table VII]. The ICT product life cycle and its relationships with sustainability pillars coexist while developing future GSLA. In future, *ICT Product Life Cycle* also needs to define as new services for achieving sustainability in the ICT industry. The whole life cycle of an ICT product consists of following four main entities, - *manufacturing*, *transportation*, *usage* and *dismantling* entities. All these entities should directly connect to future *GSLA* design to respect global analysis of sustainable development. The total GHG emission, total energy consumption and total costing of energy could not be estimated without considering all these product life cycle entities. The interaction between ICT product life cycle and GSLA are shown in Fig.2 using UML notation [47].

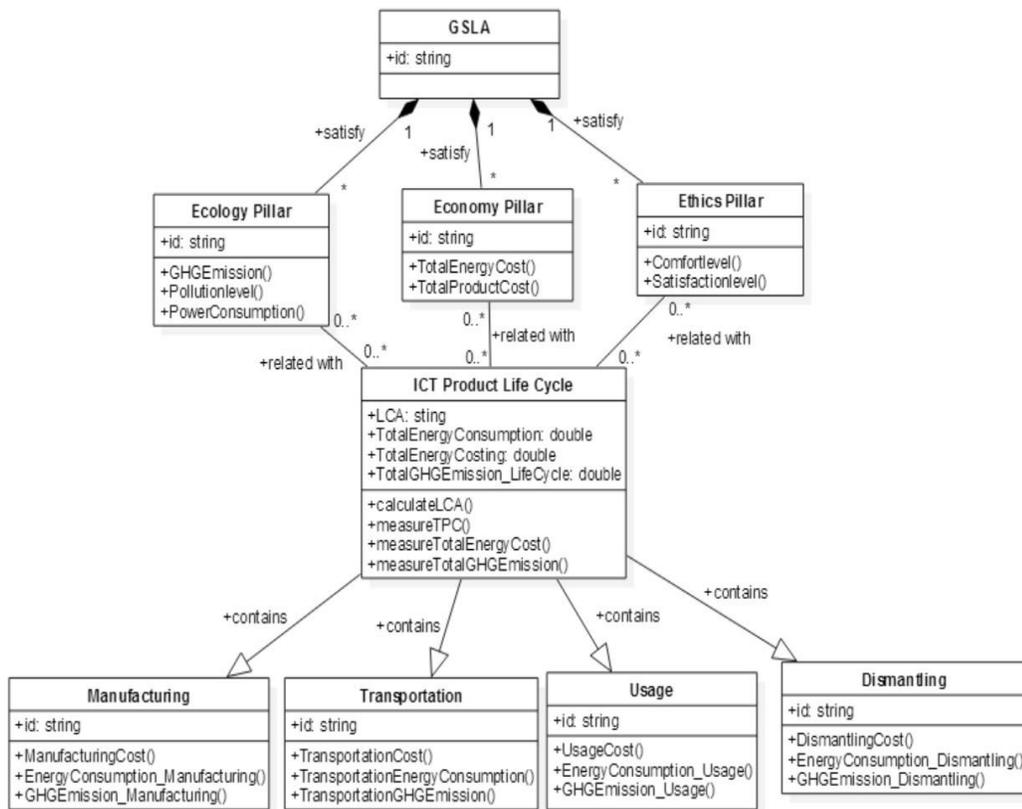


Fig. 2. Relationship between GSLA and ICT Product Life Cycle

At the bottom level of Fig.2 depicts all four main entities under ICT product life cycle. For example, *manufacturing* entities needs to calculate manufacturing cost, then energy/power consumption during manufacturing process and also total GHG emission during that time. The same way, the other entities would calculate their own costing, energy consumption and GHG emission. Therefore, total costing, total energy consumption and total GHG emission could be accumulated during the whole life cycle of each ICT product. Additionally, an environmental closed-loop supply (ECLS)[48] chain would need to be added with the proposed relationships as currently ICT products remanufacturing are getting importance in the industry. The ECLS chain would be helpful to improve economic and environmental performance of every product [46]. This UML notation actually shows the importance of ICT Product Life cycle assessment, while providing any new services. The next step of this research would be identifying and analyzing all new entities under 3E for achieving a sustainable GSLA. In future, this research would work on validation and evaluation of the proposed UML notation (Fig.2) by taking some case study on ICT product. An overall framework and survey could be designed after the case study analysis.

In addition, to define and design new green SLA (GSLA), the ICT engineer should analyses different level of cascading effect of any ICT product- direct, indirect important and indirect small effects. Moreover, finding measurement units and their assessment for all new indicators need to be defined through proper standardization and authority. At the same time, user awareness and knowledge about sustainable

development should need to incorporate at different level of the society. Moreover, to reach the sustainability, industries should invest in design and planning of their products, & also to optimize their logistic network considering the trade-off between cost and environmental effects.

V. CONCLUSION

This GSLA research did survey and review on different basic SLA parameters for network, compute, storage and multimedia domain of IT and ICT industry. The analysis of existing theory work on basic SLAs and GSLAs are mentioned in Table I and II. Empirical Work Review section demonstrates most of the basic SLAs performance indicators and their measurable unit for all mentioned services (Table III to Table VI). Moreover, existing GSLA survey covers most of the recent days green indicators and their measurable unit which are presented using Table VII from different computing industry. In addition, Table VII also discovers today's concerns are mainly on energy issues and productivity through the greening lens in many industries. These industries actually overlooked practicing sustainable development in their scope. This research believe, incorporating all new and existing indicators for future new GSLA might be difficult and cumbersome work for the ICT engineers. The management complexity of some proposed indicators in future GSLA would be the most challenging task. It is worth mentioning here that, ICT product life cycle need to consider at the first level of new GSLA design. The research shows the relation of ICT product life cycle with future GSLA for achieving sustainability. Some challenges exist for designing sustainable

future GSLA such as, new performance indicators need to be defined accurately which has association with other indicators; most of the subjective, qualitative indicators related with ethics issue need standardization or governed and authorized by proper laws and directives. In addition, it is very important to mention here that the definition of GSLA is crucial in development of Green ICT solutions and requires long time to be standardized. The standardization of green indicators is one of the main issues as mentioned by ITU-T report (2012). Also, further research is necessary on monitoring the indicators which depend on human interactions. However, this research illustrates a rigorous survey and analysis to provide a new dimension and strategy for defining future GSLA under sustainability lens in ICT arena.

ACKNOWLEDGMENT

The authors would like to thank *PERCCOM* Consortium for giving the theme of Green SLA. The authors would like to show their gratitude and thanks to European Union for supporting *PERCCOM*. Few Parts of this work submitted in International SEEDS Conference 2015, Leeds Beckett University, UK.

REFERENCES

- [1] L. Wu, and R. Buyya, "Service Level Agreement (SLA) in Utility Computing Systems," Performance and Dependability in Service Computing: Concepts, Techniques and Research Directions, V. Cardellini et. al. (eds), ISBN: 978-1-60-960794-4, IGI Global, Hershey, PA, USA, July 2011, pp.1-25.
- [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] J. Mankoff, R. Kravets, and E. Blevis, "Some Computer Issues in Creating a Sustainable World" Computer, Vol. 41, No. 8, 2008
- [4] SMART 2020 Report, "Enabling the low carbon economy in the information age," The Climate Group, GeSI, 2008.
- [5] Z. S. Andreopoulou, "Green Informatics: ICT for Green and Sustainability," Journal of Agriculture Informatics (EIFTA), Vol. 3, No. 2, 2012.
- [6] S. A. Baset, "Cloud SLAs : Present and Future," IBM Research, 2011, pp. 57-67.
- [7] H. Lee, M. Kim, and J. W. Hong, "Mapping between QoS Parameters and Network Performance Metrics for SLA monitoring," Conference on the Asia-Pacific Network Operations and Management Symposium (APNOMS), South Korea, September 2002
- [8] L. Jin, V. Machiraju, and A. Sahai, "Analysis on Service level Agreement of Web Services," Software Technology Laboratory, HP Laboratories Palo Alto, HPL-2002-180, June 2002.
- [9] A. Paschke, and E. Schnappinger-Gerull, "A Categorization Scheme for SLA Metrics," Multi-Conference Information Systems (MKWI06), Passau, Germany, 2006.
- [10] J. Lankinen, and J. Porras, "Survey on security profiles of existing cloud services," Wireless World Research Forum (WWRF) conference, October, 2012.
- [11] Anonymous, "Green Cloud Technologies: Service Level Agreement-Voice and Internet," SLA 1. 2VO, May 2013.
- [12] C. Raibulet, and M. Massarelli, "Managing Non-functional aspects in SOA through SLA," 19th International IEEE Workshop on Database and Expert Systems Application, DEXA, September 2008.
- [13] V. Stantchev, and C. Schropfer, "Negotiating and Enforcing QoS & SLAs in Grid & Cloud Computing," 4th International Conference on Grid and Pervasive Computing, GPC, LNCS 5529, 2009, pp. 25-35.
- [14] N. J. Dingle, W. J. Knottenbelt, and L. Wang, "Service level Agreement Specification, Compliance prediction and monitoring with Performance Trees," 22nd Annual European Simulation and Modelling Conference (ESM'08), October 2008, pp. 137-144.
- [15] T. Unger, F. Leymann, S. Mauchart, and T. Scheibler, "Aggregation of Service Level Agreements in the Context of Business Processes," 12th International IEEE Enterprise Distributed Object Computing Conference, 2008.
- [16] E. Marilly, O. Martinot, H. Papini, and D. Goderis, "Service Level Agreements: A main challenges for next generation network," 2nd European Conference on Universal Multiservice Networks, ECUMN, April 2002.
- [17] T. Onali, "Quality of Service technologies for Multimedia Applications in Next Generation Networks," Ph. D Thesis, University of Cagliari, Italy, 2007.
- [18] H. Ludwig, A. Keller, A. Dan, and R. King, "A Service Level Agreement for Dynamic Electronic Services," IEEE international workshop on E-Commerce and Web-Based Information Systems (WECWIS), 2002.
- [19] P. Hasselmeyer, H. Mersch, B. Koller, H. N. Quyen, L. Schubert, and P. Wieder, "Implementing an SLA negotiation framework," Proceedings of eChallenges Conference, 2007.
- [20] White Paper, "Comparing Public Cloud: Service Level Agreements," Dimension Data, 2013.
- [21] E. Wustenhoff, "Service level Agreement in the Data Centre," Sun Microsystems, April 2002.
- [22] C-SIG-SLA Subgroup Members, "Cloud Service Level Agreement Standardization Guidelines," White Paper, Brussels, June 2014.
- [23] S. Klingert, T. Schulze, and C. Bunse, "GreenSLAs for the energy-efficient management of data centres," International Conference on Energy-Efficient Computing and Networking, May, 2011.
- [24] G. von Laszewski, and L. Wang, "GreenIT Service Level Agreements," Grids and Service-Oriented Architectures for Service Level Agreements, Springer Science, LLC, 2010, pp. 78-88.
- [25] Md. E. Haque, K. Le, I. Goiri, R. Bianchini, and T. D. Nguyen, "Providing Green SLAs in High Performance Computing Clouds," International Green Computing Conference, June, 2013.
- [26] R. R. Harmon and N. Auseklis, "Sustainable IT Services: Assessing the Impact of Green Computing Practices," IEEE xplore, Proceeding of Portland International Centre for Management of Engineering and Technology, PICMET, August, 2009.
- [27] F. H. Grupe, T. Gracia-Jay, and W. Kuechler, "Is It Time For An IT Ethics Program?," Information Management: Strategy, Systems and technologies, Auerbach Publications, CRC Press LLC, 2002.
- [28] R. Herold, Introduction to Computer Ethics, Source: http://www.infosectoday.com/Articles/Intro_Computer_Ethics.htm, Retrieved on March 2015.
- [29] N. Agrawal, and K. N. Agrawal, "Current trends in Green ICT," Journal of Administration & Governance (JOAAC), Vol. 7, No. 1, 2012.
- [30] A. Amokrane, M. F.Zhani, Qi Zhang, R. Langar, R. Boutaba, and G. Pujolle, "On Satisfying Green SLAs in Distributed Clouds" 10th International Conference on Network and Service Management (CNSM), November 2014, pp. 64-72.
- [31] C. Li, A. Qouneh, and T. Li, "iSwitch: Coordinating and Optimizing Renewable Energy Powered Server Clusters," International Symposium on Computer Architecture, May 2011.
- [32] K. Le, O. Bilgir, R. Bianchini, M. Martonosi, and T. D. Nguyen, "Managing the Cost, Energy Consumption, and Carbon Footprint of Internet Services," Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems, 2010, pp. 57-58.
- [33] M. Nichollas, "Why Users still see Red even when SLAs are Green," Fujitsu Service Limited, United Kingdom, 2010.
- [34] A. P. Bianzino, C. Chaudet, D. Rossi, and J. Rougier, "A Survey of Green Networking Research," IEEE Communication Surveys and Tutorials, Vol. 14, Issue. 1, December 2010, pp. 3-20.
- [35] 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.

- [36] A. C. Orgerie, "A Survey on Techniques for Improving the Energy Efficiency of Large Scale Distributed Systems," ACM Computing Surveys (CSUR), Vol. 46, Issue 4, April 2014.
- [37] 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.
- [38] SLA@SOI, Source: <http://sla-at-soi.eu/>, retrieved on April 2015.
- [39] D. Jiang, X. Zhengzheng, and L.V Zhihan, "A multicast delivery approach with minimum energy consumption for wireless multihop networks" Journal of telecommunication Systems, 2015, pp. 1-12.
- [40] D. Jiang, X. Ying, Y. Han Y, and L.V Zhihan, "Collaborative multi-hop routing in cognitive wireless networks" Journal of Wireless Personal Communications, 2015, pp. 1-23.
- [41] R. L. Sawyer, "Calculating Total Power Requirement for Data Centers," White Paper, Schneider Electric white paper library, Retrieved on March 2015.
- [42] 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.
- [43] T. Shiino, "Green IT by all Parties," PhD Presentation at Nomura Research Institute, Tokyo, Japan, March 2010.
- [44] N. Drouant, E. Rondeau, J. P. Georges, and F. Lepage, "Designing green network architectures using the Ten Commandments for a mature ecosystem," Computer Communications, Vol. 42, April 2014, pp. 38-46.
- [45] Anonymous, "Application Note: Voice Quality Measurement, Series: Voice over IP Performance Management," Telchemy Incorporation, USA, November 2014.
- [46] M. B. Uddin, M. R. Hassan, and Kazi M. Tarique, "Three Dimensional Aspects of Corporate Social Responsibility," Daffodil Intenational University Journal of Business and Economics, Bangladesh, Vol.3, No.1, January 2008.
- [47] IBM Technical Libaray, "UML Basics: The Class Diagram," Source: http://www.ibm.com/developerworks/rational/library/content/RationalE_dge/sep04/bell, Retrieved on April 2015.
- [48] M. A. Ruimin, Y. A. O Lifei, J. I. N Maozhu, R. E. N Peiyu, and L. V Zhihan, "Robust environmental closed-loop supply chain design under uncertainty" Journal of Chaos, Solitons & Fractals, Elsevier, November 2015.

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.

EMCC: Enhancement of Motion Chain Code for Arabic Sign Language Recognition

Mahmoud Zaki Abdo

Electronic, communication, and computer department
Faculty of engineering, Helwan University
Cairo, Egypt

Sameh Abd El-Rahman Salem

Electronic, communication, and computer department
Faculty of engineering, Helwan University
Cairo, Egypt

Alaa Mahmoud Hamdy

Electronic, communication, and computer department
Faculty of engineering, Helwan University
Cairo, Egypt

Elsayed Mostafa Saad

Electronic, communication, and computer department
Faculty of engineering, Helwan University
Cairo, Egypt

Abstract—In this paper, an algorithm for Arabic sign language recognition is proposed. The proposed algorithm facilitates the communication between deaf and non-deaf people. A possible way to achieve this goal is to enable computer systems to visually recognize hand gestures from images. In this context, a proposed criterion which is called Enhancement Motion Chain Code (EMCC) that uses Hidden Markov Model (HMM) on word level for Arabic sign language recognition (ArSLR) is introduced. This paper focuses on recognizing Arabic sign language at word level used by the community of deaf people. Experiments on real-world datasets showed that the reliability and suitability of the proposed algorithm for Arabic sign language recognition. The experiment results introduce the gesture recognition error rate for a different sign is 1.2% compared to that of the competitive method.

Keywords—*image analysis; Sign language recognition; hand gestures; HMM; hand geometry; and MCC*

I. INTRODUCTION

The incident deficiencies in the language for deaf people make there it difficult to translate thoughts and feelings into words and phrases understandable and aware. The normal people translate ideas into words audible, but the deaf people translate ideas into visual signs through the fingers and hands movement.

Normally, there is no problem when deaf persons communicate with each other by using their common sign language. The problem appears when a deaf people want to communicate with a non-deaf people. Usually both will be disgruntled in a very short time [1]

Since the beginning of the use of the deaf people sign language, they have created specific language among themselves. Those signs of these languages were the only form of communication between deaf people. Within the diversity of cultures of deaf people, signing developed to complete languages. It is a form of communication with deaf people. There has been interest in recognizing human hand gestures.

The target of the sign language recognition is to introduce an accurate mechanism to convert sign gestures into speech or

meaningful text so that communication between deaf and non-deaf society. Sign language is not uniform on the world, but different from country to other country. the researchers attempt to unify the sign language in each country separately have been carried out such as Jordan, Egypt and Saudi Arabia to support members of the deaf for each community [2].

Many previous researchers have been working on hand gestures recognition in many sign languages such as the Dutch Sign Language, the American Sign Language (ASL) [3], the Australian Sign Language (Auslan) [4], and the Chinese Sign Language (CSL) [5], the Arabic Sign Language (ASL) has less attention [6].

In this section we focus the discussion of the previous researchers on sign language gesture recognition, and especially on Arabic sign language (ArSL) recognition. The sign language recognition can be classified into signer-independent and signer-dependent according to the signer sensitivity. Also Most of the previous studies on sign languages are based on vision method or glove based method [7]. In the glove based method, the person needs to wear special electronic devices, like gloves or markers. While in vision based method, it uses image processing methods to recognize the gestures without setting any limitation on the user, to supply the system with data related to the motion and hand shape [6].

Cyber gloves are used in most of previous works on Sign Language Recognition. Research [8] developed a system depends on power gloves. It recognizes a set of 95 isolated signs on Australian sign languages with accuracy 80%. Research [9] developed a system to recognize 262 isolated sign with accuracy 91.3% by using HMM. The use of cyber gloves or other input devices conflicts with recognizing and is very difficult to running in real time [10]. The researchers presented several Sign language Recognition systems based on vision methods [2, 10, 11, 12, 13, 14,15,16].

Some of vision research works recognize the Arabic alphabet using vision based as research work [2]. It created an automatic translation system for gestures of manual alphabets in the Arabic sign language recognition. It does not rely on

using any visual markings or gloves. The extracted features phase depends on two stages only, the first stage is edge detection and the second stage is feature-vector-creation. It used multilayer perceptron (MLP) classifier and minimum distance classifier (MDC) to detect 15 characters only of 28 characters.

The research work in [11], a system of the recognition and translation of the numbers were designed. The system is composed of four main phases; Pre-processing phase, Feature Extraction phase, interpolation phase and Classification phase. The extracted features are scale invariant and make the system more flexible. The experimental results revealed that the system was able to recognize a representing numbers from one to nine based on the minimum Euclidean distance between the numbers.

The research work in [12] investigated appearance-based features for the deaf person- vision-based on sign language recognition. It does not depend on a segmentation of the input images and he used the image as a feature. The system used a combination of features including PCA, hand trajectory, hand position, and hand velocity. The rwth-boston-104 database is used for the grey scale image with a reduced frame size 195x165 pixels and downscaling to 32x32 pixels.

The research work in [13], a system of the recognition and translation of the Arabic letters was designed. The system depends on the inner circle position on the hand contour and divides the rectangle surrounding by the hand shape into 16 zones. The extracted features are scale invariant. Experiments revealed that the system was able to recognize Arabic letters based on the hand geometry. The experiment results shown that the different signs gesture recognition rate of Arabic alphabet for were 81.6 %.

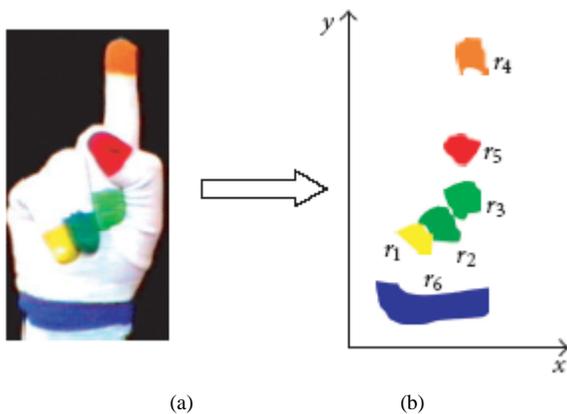


Fig. 1. (a) Colored gloves and (b) output image segmentation

The research work in [14] used Adaptive Neuro-Fuzzy Inference system (ANFIS). The system used 30 Arabic sign language alphabets visually. The recognition rate of the system was 93.55%. The research work in [15] built an ArSL system and measures the performance of ArSL data collected. The system based on Polynomial classifiers. It collected a 30 letter of ArSL. It collected the data by using gloves marked with six different colours at different regions as shown in Fig. 1 [15]. The recognition rate is 93.41 %.

The research work in [16] used new two features are introduced for American Sign Language recognition: those are kurtosis position and principal component analysis PCA. Principal component analysis was used in this research as a descriptor that represents features of image to provide a measure for hand orientation and hand configuration. PCA has been used before in sign language as a dimensionality reduction. Kurtosis position is used as a local feature for measuring edges and reflecting the position of articulation recognition. It used motion chain code that represents the movement of hand as feature. The system input is a sign from RWTH-BOSTON-50 database, and the recognition error rate of the output is 10.90%.

In this paper the motion chain code used in [16] to recognize Arabic sign language is to be enhanced through an EMCC algorithm. Applying the EMCC on forty different Arabic words, as Fig. 2, the conducted results showed the enhancement compared to the MCC algorithm.

The rest of the paper is organized as follows. Section two presents a Motion chain code (MCC). Section three explains HMM classifier. Section four presents the proposed system. Section five shows the experimental data. Section six explains the experimental results. Section seven presents the conclusions.

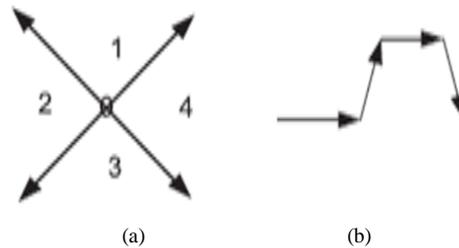


Fig. 2. MCC. (a) CC values. (b) Sample MCC 4143

II. MOTION CHAIN CODE (MCC)

This method provides a representation of hand trajectory. It is a sequence of numbers {0,1,2,3,4}, to represent the motion directions of the hand, zero to no motion, one to up, two to left, three to down, and four to right [17] as Fig. 2. The chain code is extracted from the relative motion of the hand by subtracting a centroid of the hand in two frames.

III. HMM CLASSIFIER

HMM is used as a classifier for speech [18] and used in sign language recognition systems. In HMM-based approaches, the information of each sign is modelled by a different HMM. The model that gives the highest likelihood is selected as the best model and the test sign is classified as the sign of that model [19]. It consists of a set of N states where the transition from each state to another state. It is denoted by Eq. 1:

$$\lambda = (A, B, \pi) \quad (1)$$

- **The state transition probability distribution** $A = \{a_{ij}\}$ where its elements represent the transition probability from each state to another state. State transition coefficients having the properties Eq. 2 and Eq. 3.

$$a_{ij} \geq 0 \quad (2)$$

$$\sum_{j=1}^N a_{ij} = 1 \quad (3) \quad i \geq 1, j \leq N$$

- **The observation symbol probability distribution in state j**, $B = \{b_j(k)\}$ where its elements represent the probability of certain observation to occur at a particular state $\{1 \leq j \leq N, 1 \leq k \leq M\}$, where M is a number of observation sequence O1 O2... OM

The initial state distribution = $\{\pi_i\}, 1 \leq i \leq N$

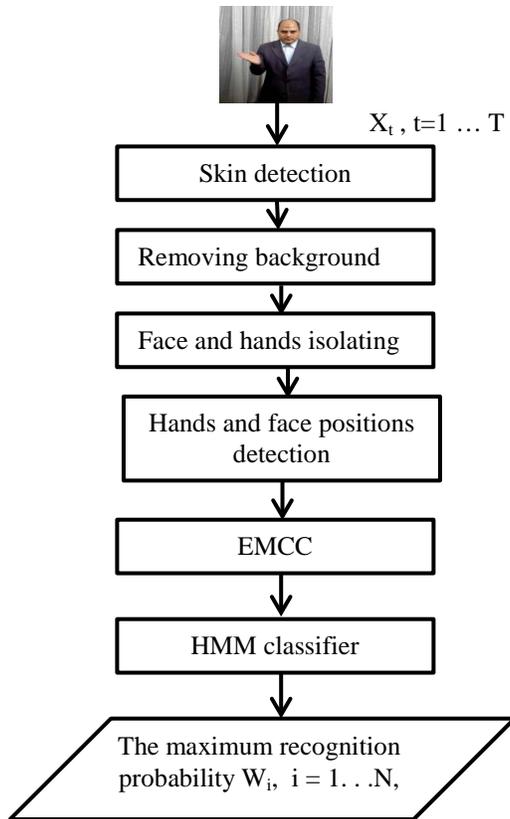


Fig. 3. Proposed system architecture

IV. PROPOSED SYSTEM

The proposed system, as shown in Fig. 3, consists of six phases, skin detection, removing background, face and hands isolating, hand and face positions detection, Enhancement Motion Chain Code EMCC, and Hidden Markov Model HMM classifier.

A sequence of input video frames $X_t, t=1..T$, where T is a number of video frames and the output is Maximum recognition probability $W_i, i = 1..N$, where N is a number of signs, is corresponding to sign detection.

The system components described in the following subsections: Sub section 4.1 presents skin detection and removing background. Sub section 4.2 presents face and

hands isolating. Sub section 4.3 presents hand and face position detection. Sub section 4.4 presents a Proposed Enhancement of Motion Chain Code (EMCC) and HMM.

A. Skin Detection and Background Removal

The algorithm uses skin detection [20]. The algorithm adopts skin colour detection as the first step. Due to YCbCr color space transform, YCbCr is faster than other approaches [21, 22]. The algorithm calculates the average luminance Y_{avg} of the input image as given in Eq.4.

$$Y_{avg} = \sum y_{ij} \quad (4)$$

Where $y_{ij} = 0.3 R + 0.6 G + 0.1 B$ is normalized to the range $\{0 \text{ to } 255\}$, where i, j are the indices of the pixel in the image. According to Y_{avg} , the algorithm can calculate the compensated image C_{ij} by the following equations Eq.5 and Eq.6 [20]:

$$\begin{aligned} R'_{ij} &= (R_{ij})^\tau \\ G'_{ij} &= (G_{ij})^\tau \end{aligned} \quad (5)$$

$$C_{ij} = \{R'_{ij}, G'_{ij}, B_{ij}\}$$

Where

$$\tau = \begin{cases} 1.4, & Y_{avg} < 64 \\ 0.6, & Y_{avg} > 192 \\ 1, & \text{otherwise.} \end{cases} \quad (6)$$

It should be noted that the algorithm compensates the colour of R and G to reduce computation. Due to chrominance (Cr) which can well represent human skin, the algorithm only consider Cr factor for colour space transform to reduce the computation. Cr is defined as follows Eq. 7 [22]:

$$Cr = 0.5R' - 0.419G' - 0.081B \quad (7)$$

Accordingly, the human skin binary matrix can be obtained as follows:

$$S_{ij} = \begin{cases} 0, & 10 < Cr < 45 \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

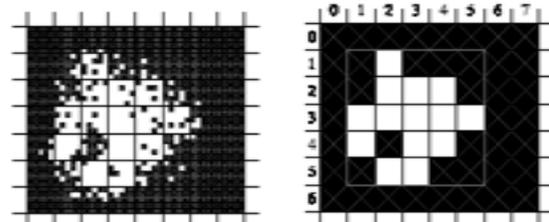


Fig. 4. (a) An example of S_{ij} (b) Noise removal by the 5×5 filter

Where '0' is the white point and '1' is the black point. The algorithm implements a filtration by a 5×5 mask. First, the algorithm segments S_{ij} into 5×5 blocks, and calculate show many white points in a block. Then, every point of a 5×5 block is set to white point when the number of white points is greater than half the number of total points. Otherwise, if the number of black points is more than a half, this 5×5 block is modified to a complete black block, as shown in Fig. 4 [21].

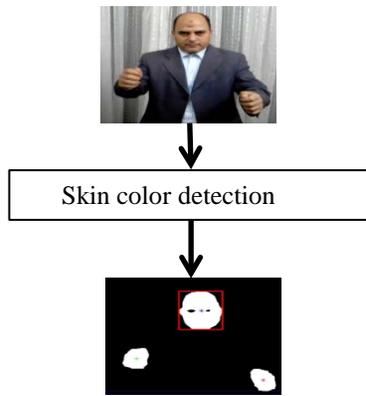


Fig. 5. Skin colour detection and removing background

Fig. 5 shows the resultant image shapes after skin detection and removing the background [23] of image.

B. Face and Hand Isolating

The algorithm tracks the objects in each image. The algorithm neglected the small objects, and then detects the largest objects as hands and the face. The algorithm isolates the hand and face as in Fig. 5. After detecting the skin colour and removing background the position of the face and hands can be isolate and detected as Fig.6.

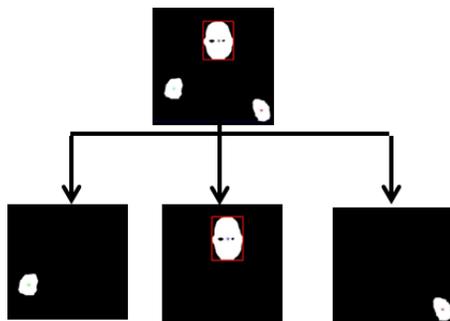


Fig. 6. Isolating the face and hands

Figure 5 shows the detected skin with background removal. The image contains a right hand and a face. The algorithm detects the hand and a face by the position and shape of each. Fig.6 shows isolating the face and hands, then isolate the right hand to detect the letter.

C. Hands and Face position detection

Figure 6 shows the skin detected with background removal. The image contains two hands and a face. The algorithm detects the hands and a face by the position and shape of each. Figure 7 shows three images at times $\{t-1, t, t+1\}$. The algorithm detects the hand position of each $\{U_{t-1}, U_t, U_{t+1}\}$ to recognize the changes of hand position for each frame from video sequence.



Fig. 7. Hand position detection and tracking

D. Proposed Enhancement of Motion Chain Code (EMCC) and HMM

The algorithm of EMCC depends on a two factors as Fig.8:

- Column number: The algorithm detects the column number that has a position hand as Fig 8(a).
- Angle direction: the algorithm detects the angle direction by calculating the angle between the hand positions for sequence frames as Fig 8(b).

The algorithm calculates the observation number by change of column number and angle direction for hand position in each in a video sequence.

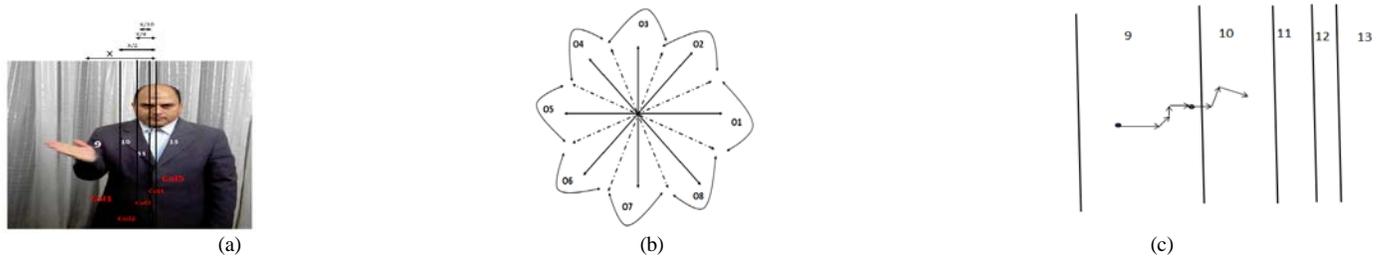


Fig. 8. (a) The column distribution, (b) The eight directions of the right hand motion. The dotted lines represent the decision boundaries between different directions, (c) A sample EMCC 9 1 2 3 1 10 3 8

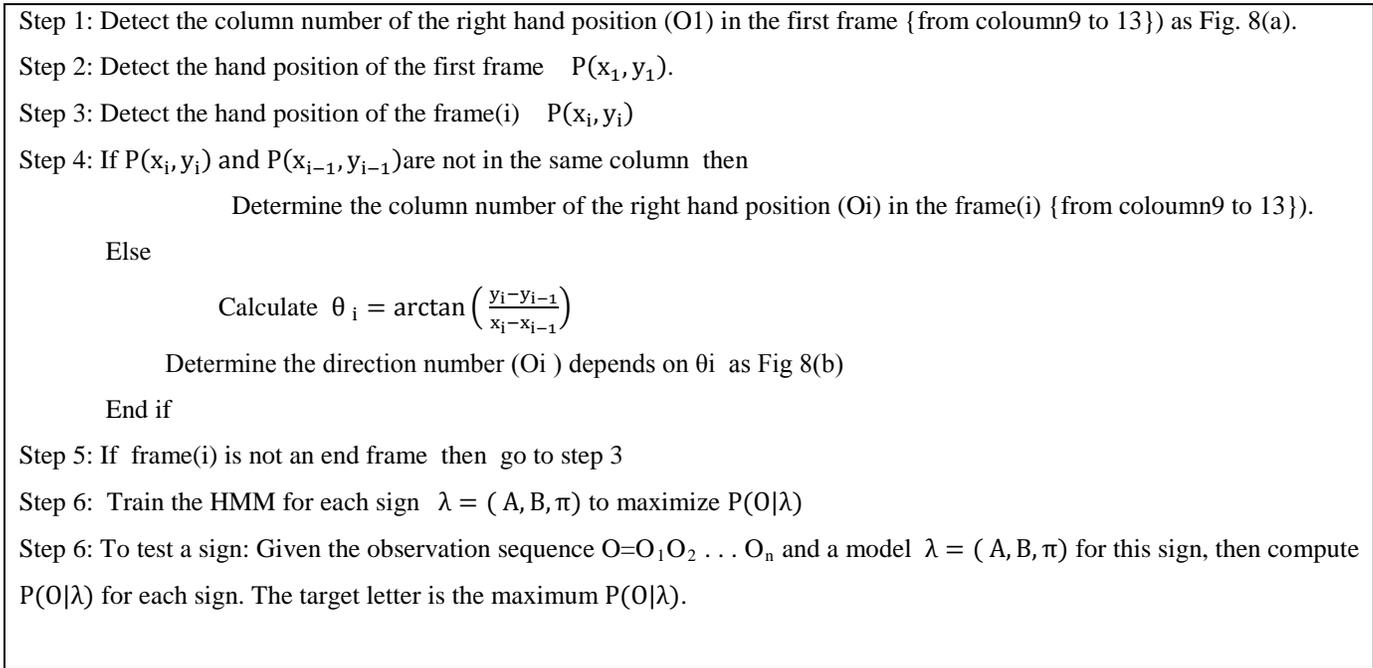


Fig. 9. The proposed algorithm of calculating the observation vector and using HMM to train and test the signs

Figure 9 shows the proposed algorithm to detect an observation (O) of the sign or word detected. The first step detects the column number of the first frame of the sign (O1) and detects the hand position $P(x_1, y_1)$. For each frame of a sign the hand position $P(x_i, y_i)$ is detected. If the hand position in the same column, calculate the angle between the hand position of the previous frame and the recent frame to detect the observation number (Oi) from Fig.7 (b). If the hand position changes to other column, the observation number (Oi) is the same number of a column as Fig. 7(a). After calculating the observation of the sign, apply the HMM algorithm.

V. EXPERIMENTAL DATA

To tune and test the proposed system, Arabic sign database EMCC database (EMCCDB) is generated as follows. The EMCCDB corpus consists of 40 Arabic words as Fig 10. The words were signed by three signers: one female and two male signers. All of the signers are dressed differently and the brightness of their clothes is different.

The video frames of the database are sampled at 30 frames per second and the size of the frames is 640 x 480 pixels. The implementation is carried out using the following as table 1:

- Number of words: 40.
- Number of videos: 1288.
- Number of training videos: 1045.
- Number of testing videos: 243.
- Average videos per word: 32.2.
- Average training videos per word is: 26.125.
- Percentage of training videos per word is: 81.13%.
- Percentage of testing videos is: 18.87%.

The prototype is implemented using a Windows based MATLAB (R2013a).

TABLE I. EMCC DATABASE (EMCCDB) DETAILS

Sign name	English Sign Name	Number of videos	Training videos	Test videos	Accuracy%	
1	يشترى	'Buy'	31	25	6	100
2	يذهب	'go'	35	28	7	100
3	كبير	'big'	40	32	8	100
4	يشوي	'grill'	26	21	5	100
5	ماذا	'what'	30	24	6	66.67
6	يصل	'arrival'	43	35	8	100
7	يوم	'day'	24	20	4	100
8	سيارة	'car'	33	27	6	100
9	ذبابه	'fly'	14	12	2	100
10	يتميز ب	'featuring'	31	25	6	100
11	ذراع	'Arm'	31	25	6	100
12	كبد	'liver'	31	25	6	100
13	يبيع	'sell'	32	26	6	100
14	يختار	'selection'	36	29	7	100
15	ذهول	'stupor'	33	27	6	100
16	خطينه	'sin'	37	30	7	100
17	يتنامى	'growing'	43	35	8	100
18	كتاب	'book'	38	31	7	100
19	يحصد	'reaps'	36	29	7	100
20	يسبح	'swim'	34	28	6	100
21	يكسر	'breaks'	35	28	7	100
22	ينفت	'puffed'	36	29	7	100
23	يمشى	'walking'	39	32	7	100
24	استقلال	'freedom'	27	22	5	100
25	إقتدى ب	'followed'	27	22	5	100
26	الاتحاد العالمى للصم	'wfd'	22	18	4	100
27	الاسوأ	'worst'	31	25	6	100
28	الأضحيه	'sacrifice'	27	22	5	100
29	امتحان	'exam'	26	21	5	100
30	ينزل	'down'	35	28	7	100
31	يمنع	'prevents'	32	26	6	100
32	يفتح	'opens'	33	27	6	100
33	يعيش	'live'	36	29	7	85.71
34	يصعد	'climb'	28	23	5	100
35	يحرث	'plowing'	36	29	7	100
36	وزع	'distribute'	31	25	6	100
37	وصف	'description'	39	32	7	100
38	وصله	'link'	30	24	6	100
39	وطن	'homeland'	27	22	5	100
40	وطني	'national'	33	27	6	100
Overall					98.81	

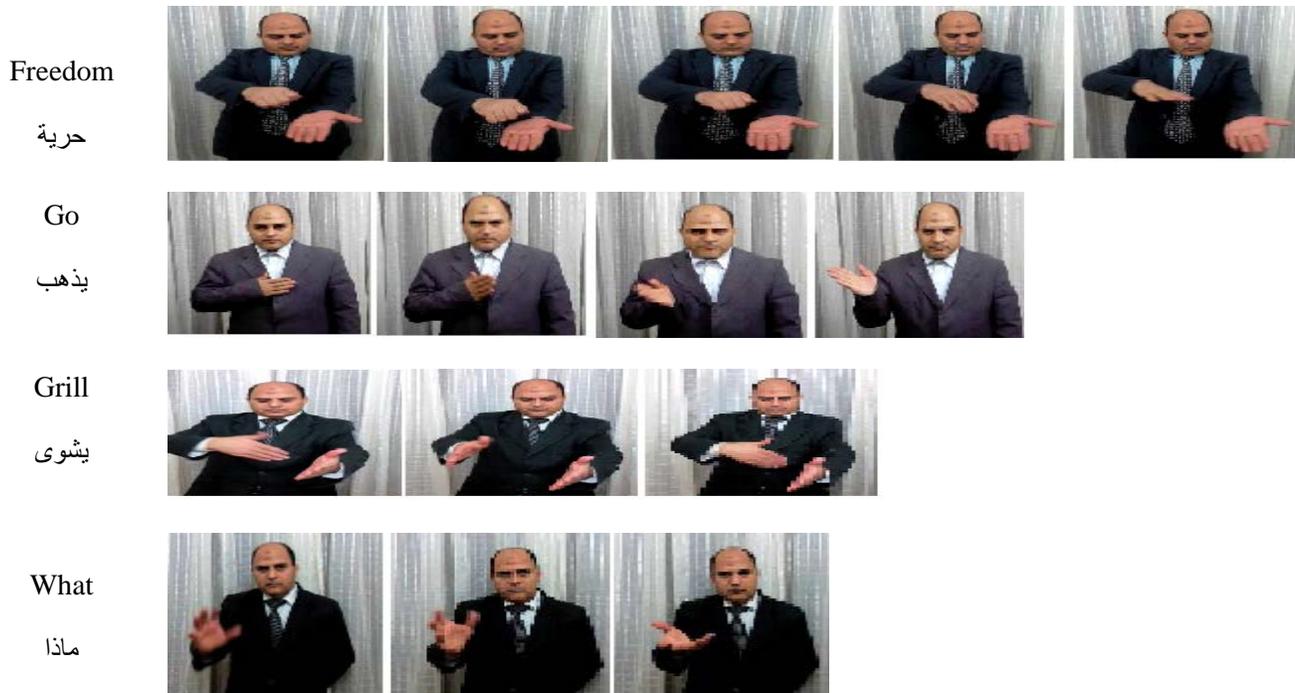


Fig. 10. Arabic Sign Language sample gestures

TABLE II. COMPARISON WITH ARSL RECOGNITION

	Instruments used	Number of signs	Classifier	Recognition Rate
EMCC	None: Free Hands	40 words	HMM	98.8 %
Zaki and shaheen [16]	None:Free Hands	30 words	MCC	65.46%
Mohandes and Deriche [26]	Gloves : pair of colored gloves	50 words	HMM	98 %
Shanableh. [27]	Gloves : pair of colored gloves	23 words	KNN	87 %
EL-Bendary et al. [2]	None: Free Hands	15 Alphabet Letter	MDC MLP	91.3% 83.7 %
Jarrah, et al. [14]	None: Free Hands	30 Alphabet Letters	ANFIS	93.55 %
Assaleh, et al. [15]	Gloves marked with six different colour	30 Alphabet Letters	polynomial classifiers	93.41%
ArSLAT [13]	None: Free Hands	29 Alphabet Letter	Outer of the inner circle zones	83.16

VI. EXPERIMENTAL RESULT

For the purpose of comparisons, MCC [16] is applied on EMCCDB database, and it achieves an error rate with 38.15 %, while EMCC achieves an error rate with 1.2 %. Figure 11 shows the EMCC performance for every sign detected in EMCCDB versus MCC performance. The total recognition rate enhancement is 36.95 %. As shown in table 2, it comparisons between EMCC and previous work on Arabic sign language recognition. In [16], MCC was applied on the American Sign

Language. The error rate is 34.54% over the RWTH-BOSTON-50 database. The implementation in [16] was carried out using the following:

- Number of words: 30.
- Number of videos: 110.
- Number of training videos: 90.
- Number of testing videos: 20

recognition rate of 93.41 %. Reference [12] did not use gloves and used ANFIS to recognize 30 letters by recognition rate of 93.55 %.

VII. CONCLUSIONS

In this paper, a new algorithm, which is called Enhancement Motion Chain Code (EMCC), for Arabic sign language recognition is presented. It has been demonstrated experimentally that the phases of the proposed algorithm includes skin detection, background exclusion, face and hands extraction, hands and face position detection, feature extraction, and also classification using Hidden Markov Model (HMM). Experimental results show that the proposed algorithm achieves 1.2% error rate compared to the other competitive algorithm which achieves 38.15 % error rate.

REFERENCES

- [1] A.Youssif, Amal Elsayed Aboutabl, Heba Hamdy Ali, "Arabic Sign Language (ArSL) Recognition System Using HMM ", IJACSA, Volume. 2, No. 11, 2011.
- [2] Nashwa El-Bendary, Hossam M. Zawbaa, Mahmoud S. Daoud, Aboul Ella Hassanien, and Kazumi Nakamatsu, "ArSLAT: Arabic Sign Language Alphabets Translator", IJCISIM, Volume 3, 2011, pp. 498-506.
- [3] T. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models ", International Workshop on Automatic Face and Gesture Recognition, 189–194, June 1995.
- [4] E.J. Holden, G. Lee and R. Owens, Automatic Recognition of Colloquial Australian Sign Language, IEEE Workshop on Motion and Video Computing 2, December 2005, 183–188.
- [5] C.L. Wang, W. Gao and S.G. Shan, "An approach based on phonemes to large vocabulary Chinese sign language recognition", Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2002, 411–416.
- [6] M. Al-Rousan, O. Al-Jarrah, and M. Al-Hammouri, "Recognition of Dynamic Gestures in Arabic Sign Language using Two Stages Hierarchical Scheme", The International Journal of Intelligent and Knowledge Based Engineering Systems, Volume 14, Number 3, 2010.
- [7] F. Chen, C. Fu and C "Hand gesture recognition using a real time tracking method and hidden Markov models," Image and Vision Computing 21, vol. 21, no. 8, p. 745–758, March 2003.
- [8] M.W. Kadous, "Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language", in: Proceedings of the Workshop on the Integration of Gestures in Language and Speech, 1996, pp. 165–174.
- [9] K. Grobel, M. Assan, Isolated sign language recognition using hidden Markov models, in: Proceedings of the International Conference on System, Man and Cybernetics, 1997, pp. 162–167.
- [10] Al-Rousan, M., Khaled Assaleh, and A. Tala'a. "Video-based signer-independent Arabic sign language recognition using hidden Markov models." Applied Soft Computing 9.3 (2009): 990-999.
- [11] Mahmoud Zaki Abdo, Alaa Mahmoud Hamdy, Sameh Abd El-rahman Salem, El-sayed Mostafa Saad. "C30. An Interpolation Based Technique for Sign Language Recognition." Radio Science Conference (NRSC), 2013 30th National. IEEE, 2013.
- [12] Rybach, D., "Appearance-Based Features for Automatic Continuous Sign Language Recognition", Diploma Thesis, RWTH Aachen University, Aachen, Germany, 2006.
- [13] Mahmoud Zaki Abdo, Alaa Mahmoud Hamdy, Sameh Abd El-rahman Salem, El-sayed Mostafa Saad, "Arabic Sign Language Recognition", International Journal of Computer Applications, 0975 – 8887, Volume 89 – No 20, March 2014
- [14] Al-Jarrah, Omar, and Alaa Halawani. "Recognition of gestures in Arabic sign language using neuro-fuzzy systems." Artificial Intelligence 133.1-2, pp. 117-138, 2001.
- [15] Assaleh, Khaled, and M. Al-Rousan., "Recognition of Arabic sign language alphabet using polynomial classifiers", Journal on Applied Signal Processing , pp. 2136-2145, 2005
- [16] Zaki M. M., Shaheen S. I., "Sign language recognition using a combination of new vision based features, Pattern Recognition Letters", Vol. 32, Issue 4, 1 Mar 2011, pp. 572-577
- [17] Ozer, Omer Faruk, et al. "Vision-based single-stroke character recognition for wearable computing." Intelligent Systems, IEEE 16.3, 2001, 33-37.
- [18] Rabiner, Lawrence R., and Biing-Hwang Juang. "An introduction to hidden Markov models." ASSP Magazine, IEEE 3.1, pp. 4-16, 1986.
- [19] Oya Aran, "Vision Based Sign Language Recognition: Modelling and Recognizing Isolated Signs with Manual and Non-Manual Components", 2008, Phd thesis.
- [20] Pai, Yu-Ting, et al. "A simple and accurate color face detection algorithm in complex background." Multimedia and Expo, IEEE International Conference on. IEEE, 2006.
- [21] S. Gundimada, Li Tao, and v. Asari, "Face detection technique based on intensity and skin color distribution," in 2004 International Conference on Image Processing, vol. 2, pp. 1413–1416. , Oct. 2004
- [22] K. P. Seng, A. Suwandy, and L.-M. Ang, "Improved automatic face detection technique in color images," in IEEE Region 10 Conference TENCON 2004, vol. 1, pp. 459–462. , Nov. 2004
- [23] Khaled, H., Sayed, S. G., Saad, E. S. M., & Ali, H. "Hand gesture recognition using modified 1\$ and background subtraction algorithms". Mathematical Problems in Engineering, 2015.
- [24] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition", In Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium (DAGM), Lecture Notes in Computer Science, volume 3663, pages 401-408, Vienna, Austria, August 2005.
- [25] Mohandes, Mohamed, Junzhao Liu, and Mohamed Deriche. "A survey of image-based arabic sign language recognition." Multi-Conference on Systems, Signals & Devices (SSD), 2014 11th International. IEEE, 2014.
- [26] Mohandes, Mohamed, and Mohamed Deriche. "Image based Arabic sign language recognition." Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on. Vol. 1. IEEE, 2005.
- [27] T. Shanableh and K. Assaleh, "Arabic sign language recognition in user-independent mode", in International Conference on Intelligent and Advanced Systems, ICIAS 2007, pp 597-600, 2007.
- [28] Mahmoud Zaki Abdo, Alaa Mahmoud Hamdy, Sameh Abd El-Rahman Salem and Elsayed Mostafa Saad, "Arabic Alphabet and Numbers Sign Language Recognition" International Journal of Advanced Computer Science and Applications (ijacsa), 6(11), 2015.

A Novel Approach for Ranking Images Using User and Content Tags

Arif Ur Rahman, Muhammad
Muzammal, Humayun Zaheer
Ahmad
Department of Computer Science,
Bahria University, Islamabad,
Pakistan

Awais Majeed
Department of Software
Engineering,
Bahria University, Islamabad,
Pakistan

Zahoor Jan
Department of Computer Science,
Islamia College University,
Peshawar, Pakistan

Abstract—In this study, a tag and content-based ranking algorithm is proposed for image retrieval that uses the metadata of images as well as the visual features of images, also known as “visual words” to retrieve more relevant images. Thus, making the retrieval process more accurate than the keyword-based retrieval approaches. Both tag and content-based image retrieval techniques have their own advantages and disadvantages. By combining the two, their disadvantages have been offset. The proposed system has been developed to bridge the gap between the existing techniques and the desired user requirements. Initially, the system extracts the metadata of images and stores them into a custom designed dictionary dataset. Then, the system creates a visual vocabulary and trains a classifier on a dataset of images belonging to different categories. Next, for any given user-query, the system makes a decision to display a class of images that best matches the query. These class images are processed in a way that we compute the relevance scores for each image and display the result based on the score.

Keywords—Image retrieval; search engine; user-tags; relevance scores; visual words; multi-class classification

I. INTRODUCTION

The process of obtaining information resources relevant to data need, from a collection of information materials is the main idea presented in this paper, commonly known as “Information Retrieval”. The process of information retrieval has evolved over time as we can see that search engines today are much more efficient than they were years back. Among the current mainstream image search engines such as Google Image Search and Bing Image Search, Google is one of the most efficient, but it is tag-based [1] that takes keywords as queries and relies on the tags associated with images to search them. The tags are referred to as metadata of images that includes information such as name of an image, etc. While the tag-based image retrieval (TBIR) is a useful approach to improve the results of image retrieval, it suffers from the inconsistency of metadata that often results in more irrelevant images [2] making it a limited retrieval strategy because the tags may or may not be correct.

In this paper, the idea of “content tags” (C-Tags) is proposed where we tag the image using image content. The proposed technique therefore relies not only on the user-tags associated with the images but also on the visual features of images, also known as “visual words”. So, in the system,

given a user-query, the image dataset is processed in a way that the relevance score of each image is computed on the basis of user-tags and content-tags and a pool of images based on the relevance scores is displayed.

The ranking algorithm produces a result set based on user-tags and content-tags.

This paper is organized as follows. Section II critically analyzes the existing work and builds an argument for the proposed work. In Section III we give an overview of the proposed system and propose a new image ranking algorithm. Section IV and V discusses the implementation of the proposed system and result analysis, respectively, followed by conclusion, future work and references.

II. RELATED WORK

All mainstream image search engines like Google Image Search, and Bing Image Search, generally rely on the textual information linked to an image, such as the user-tag and other image rounding-text in order to rank images. The ranking algorithms rank images based on how relevant the query is to the data associated with images. While tag-based image retrieval (TBIR) is often effective, the resulting images also contain irrelevant images making the results less accurate. Therefore, the current image search engines are constrained by the dissimilarity between the relevance of an image and its significance understood from the related textual information. Due to this reason, the search engines like Google show irrelevant results in text query based image search [3]. This is a significant downside of user-tag based image retrieval systems.

III. PROPOSED WORK

In this work, a novel ranking system is proposed which is based on the user tags associated with images and their visual features in order to make the retrieval results more efficient and relevant to the provided user-query. Many techniques were adopted like extraction of metadata, creation of dictionary dataset, removal of stop words, tag relatedness [4], tag length normalization [4], image representation and classification by high level features, computation of relevance scores, Jaccard similarity [4], etc. These techniques helped improve the ranking of more relevant images in search. We now briefly review the relevant concepts.

A. Tag-Based Retrieval

Tag-based retrieval is a simple keyword based search in which the images in a dataset are indexed according to their metadata, like filename, alternate tag, etc. For the implementation of this approach, we first extract and store the metadata of images in a specially designed dictionary dataset so as to make the retrieval of large information more efficient. We make use of techniques such as removal of stop words, tag relatedness, tag length normalization, Jaccard similarity and tag refinement to make the retrieval process more efficient and the result set more relevant to the provided user query. Although tag-based methods are fast and reliable when images are well tagged, they are incapable of retrieving results that are relevant to the user query, if the associated user-tags are incorrect, and/or missing from the image collection. Therefore, in the proposed system, we combine the proposed tag-based retrieval technique with the content-based technique to offset the disadvantages of the tag-based approach.

B. Content-Based Retrieval

Image representation by effective features is crucial to the performance of the proposed image ranking system. While researching, we discovered that the most popular image representations have been described by low-level visual features such as color, shape, etc. while removing unimportant details [5]. The category of low-level features includes features such as color histograms, color moments, shape contour, shape region, and homogeneous texture, etc.

Bag-of-Words (BoW) approach is an efficient and effective image representation approach for tasks such as image classification and retrieval [5]. The BoW model is illustrated in fig. 1.

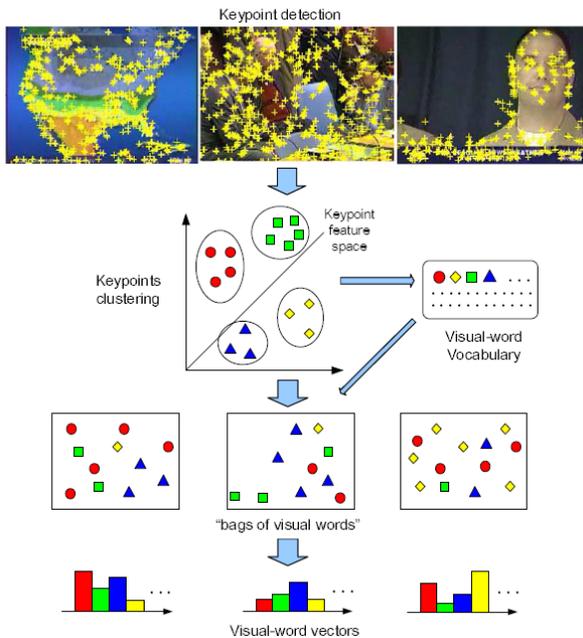


Fig. 1. Visual-word representation based on vector-quantized key-point features (image taken from [6])

The proposed content-based ranking algorithm uses the concept of “visual words” in order to achieve an automatic ranking of images. These visual words are the segments of an image that carry some kind of information related to the features which can be automatically detected and depicted by descriptors such as SIFT (Scale Invariant Feature Transform) [7], its variant PCA-SIFT (Principal Component Analysis-SIFT) [8], SURF (Speeded Up Robust Features) [9], etc.

This approach would make the retrieval of images efficient and thus improve the ranking of images that are more relevant to the query provided by the user. Hence, the system would be capable to compute visual relevance between images that would be more consistent with human expectations. Fig. 2 shows the architecture of the proposed system.

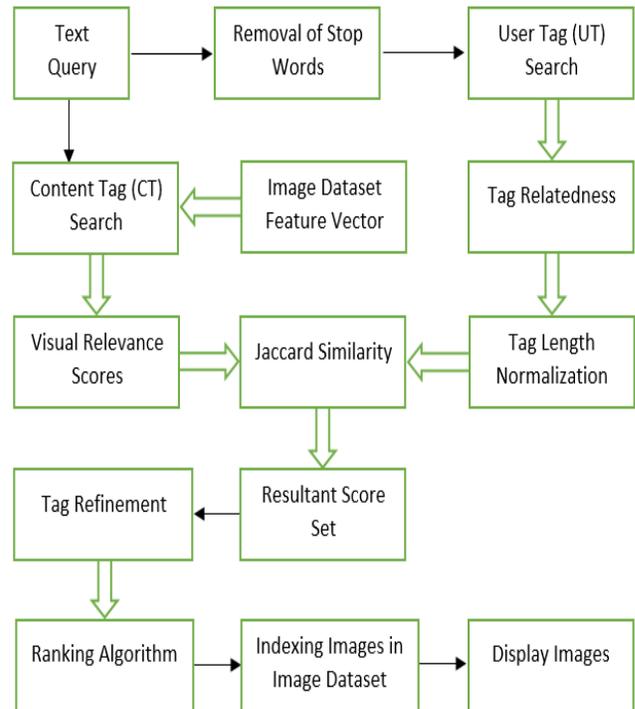


Fig. 2. Architecture proposed

IV. IMPLEMENTATION

A. Tag-Based Approach

1) User Tag Search:

Processing a user query involves the removal of stop words, which are the commonly occurring words of little interest such as “the”, “is”, etc., to reduce the total number of tags and to obtain more meaningful user tags (UTs) associated with the images. Thus we obtain a processed query, which is used to search the image collection. These meaningful UTs are then stored in a specially designed dictionary dataset to make the searching of tags more efficient. For example, when a query is entered, a binary search is applied to search the location of the relevant tag(s) which makes the search process quick and efficient.

TABLE I. PSEUDO CODE OF THE DICTIONARY DATASET

Dictionary Dataset Algorithm	
(1)	Split the list into two equal parts.
(2)	Check the middle node.
(3)	While (Data == Query)
(4)	If (Data > Node.Data)
(5)	Use 2 nd Node
(6)	Else
(7)	Use 1 st Node

2) Tag Relatedness:

To further improve the search process, we measure the degree of effectiveness of a tag describing the tagged image in a collection. This technique is called Tag Relatedness, a technique in which the degree of relatedness between user tags is quantified and the images in the collection. In order to do that, priority is given to the UTs in the initial tag position than the UTs in the middle or last tag positions. For example, initial 40% of the UTs are given 20% more priority than the rest of the 60% UTs found. As the study suggests [4], initial UTs are more important than the UTs at the last positions [10]. Fig. 3 shows the percentage of images that have their most relevant user-tag at their n-th position in the user tag list.

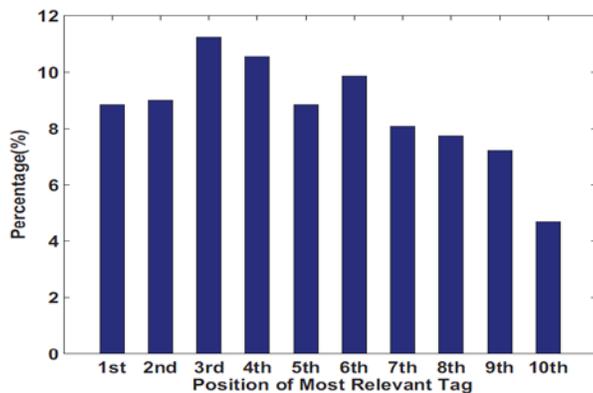


Fig. 3. Tag Relevance (image taken from [4])

3) Tag Length Normalization:

UTs length played an important role during the process of measuring tag relatedness, like if a searched image has just 1 or 2 UTs, then those UTs naturally get higher priority which deteriorates the results in the ranking of images. To overcome this problem, a tag length normalization technique is used to normalize the number of tags in the image. This way we make sure that the number of UTs does not make any difference. This task is achieved by first taking the square root of the total number of UTs and multiplying them with the scores obtained from the tag relatedness process and then by dividing its result by 1.

4) Jaccard Similarity:

Jaccard Similarity is a technique through which we combine both tag-based and content-based approaches, which makes it the most important process of the proposed system. Images in the dataset are ranked on the basis of the resultant scores which are obtained by finding the similarity between the UT scores and the CT scores.

5) Tag Refinement:

In this process, a decision for prioritizing tags between UTs and CTs is made. For example, if both UTs and CTs are matched and the result is true then UTs are given more priority than the CTs and if UTs are not matched with the CTs then CTs are given more priority than UTs. The justification for such a decision comes from the studies in Literature and also from the fact that whenever user and content tags match for an image, it could be the most relevant tags for the image.

B. Content-Based Approach

Content-based Image Retrieval (CBIR) offers a number of approaches and strategies for the retrieval of visual data from huge databases. A careful analysis suggests that for the proposed image ranking system, a high-level visual feature descriptor such as Bag of Visual Words (BoVW) is to be computed. To get this descriptor, we need visual features from the images which can be anything such as SIFT, PCA-SIFT, and SURF, etc. Among all these options we based the proposed system on SURF as it is quick and has performance similar to SIFT [11].

1) Local Feature Vectors:

SURF helps us auto detect key-points from the images and from these points local feature vector (descriptor) is extracted, which is simply a vector comprising of numerical values that describes the visual data of the image region from which it was extracted [6]. The dimensionality of a local feature vector is always the same so the number of descriptors extracted from two different images does not need to be the same.

In traditional image classification, every image is described by a global feature vector that can be seen as a set of numerical attributes in the context of machine learning. It is a requirement of image classification to have a global representation of an image that we didn't achieve when we extracted a set of local feature vectors.

To solve the above problem, a technique was employed that is known as bag of words.

2) Bag of Words:

As the name suggests, the idea of Bag of "Words" is taken from text analysis, which is to represent a document as a "bag" of essential words that carry some kind of information [12]. In computer vision, the idea is very similar. An image is represented as a Bag of "Visual Words" (BoVW) – patches that are described by a certain descriptor:

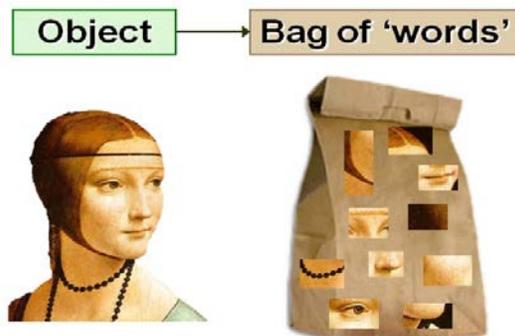


Fig. 4. Illustration of BoW model in images (image from [12])

The above mentioned BoW model for image classification and ranking is used by constructing a large vocabulary of 500 words [13]. We choose the size of the dictionary as 500 based on the processing power of available computing resources.

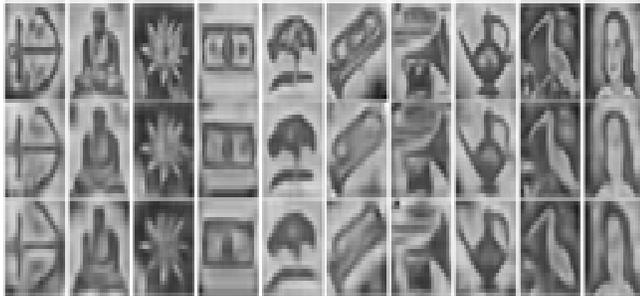


Fig. 5. Images reconstructed from descriptors. Size of the dictionary k is varied

Then we represented every image as a histogram of the frequency of words that are stored in the image. The idea is illustrated in fig. 6.

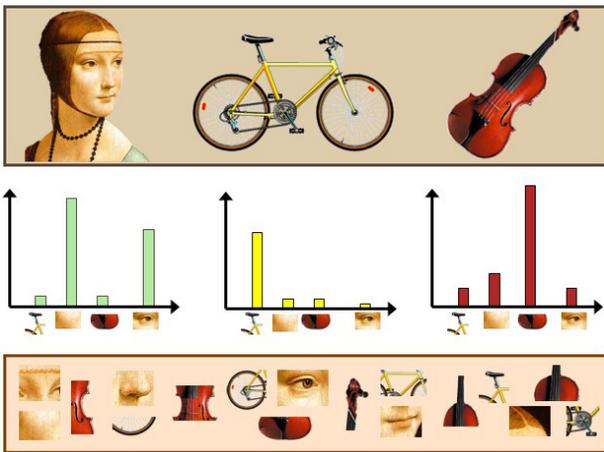


Fig. 6. BoW – representing object as histograms of words occurrences

The visual vocabulary was built by first detecting some key-points in a large image dataset (e.g. 2,400 images) using SURF detector and then descriptors were extracted from the detected key-points using SURF descriptor.

Next, we use k-means clustering algorithm on the computed set of descriptors to find the “centroids” which would be the vocabulary for the BoVW.

For the query image, we again used the same technique which helped us detect the key-points and extracted descriptors from the image around the detected key-points. Next step was to compute nearest neighbor against each extracted descriptor in the dictionary. At the end, we built a histogram of length equal to the number of centroids that represented the frequency of the proposed dictionary words (as illustrated in fig. 7):

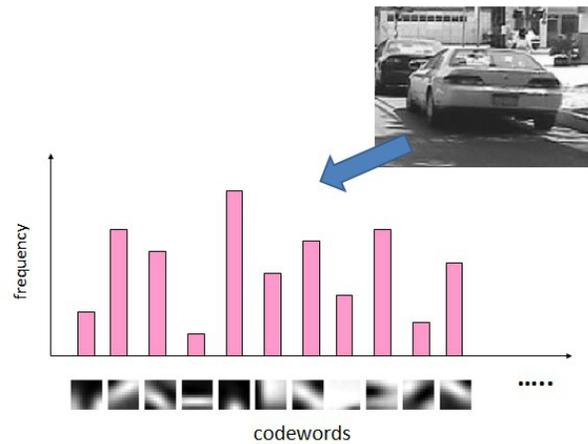


Fig. 7. BoW – representing object as histograms of words occurrences

Table II shows the simplified version of Bag of Words algorithm.

TABLE II. SIMPLIFIED BOW OUTLINE

BoW Outline	
(1)	Extract the SURF descriptors from the image dataset.
(2)	Put all the descriptors of detected keypoints into a single set.
(3)	Apply a k-means clustering algorithm over the set of descriptors to find the centroids. These cluster centers represent the proposed dictionary's visual words. Store these centroids in a YML file (e.g. dictionary.yml)
(4)	The global feature vector will be a histogram that represents the frequency of each centroid occurred in each image. Find the nearest centroid for each local feature vector to compute the histogram.

This model was then used in conjunction with an SVM classifier in order to evaluate query images.

3) Image Classification:

We used a set of images which contained images with class label as input for image classification and classified previously unclassified images on the basis of their visual appearance with the classified images. We used two techniques, with the help of which we classified images.

In the first technique, we used an SVM [14] classifier to classify images. SVM is a binary classifier, which assigns +1 to a positive class and -1 to all other (negative) classes on the basis of some confidence score. SVM predicts the class of a query image using this confidence score. But the OpenCV implementation of SVM does not provide the confidence scores of negative classes. By confidence score, we mean some weight, threshold or distance with the help of which SVM rejects other classes. In order to get the confidence of all classes, we used a second technique.

Table III shows the simplified version of classification algorithm using the first technique.

TABLE III. SIMPLIFIED CLASSIFICATION ALGORITHM (FIRST TECHNIQUE)

Classification Algorithm (First Technique)	
(1)	Set the training data.
(2)	Train the SVM classifier with the training data to build the SVM model.
(3)	Extract the SURF local features from the query image.
(4)	Put all the features of query image into a single set.
(5)	Predict the class of query image using the set on the basis of a confidence score.

In the second technique, we used K-means clustering algorithm to compute the centroids of descriptors of all classes and then used a FLANN-based matching technique to compute the matches (distances) of query image with the centroids of all classes. Next, we calculated the mean value of matches (distances) of each class separately. The result of mean is actually the confidence score, with the help of which we can classify images. The lower the mean value of a class with query image, the more probable it is that the query image belongs to that class. This technique not only classifies image, but also provides confidence of all classes. Thus, with this technique we were able to handle the multi-class problem, which was not possible in the OpenCV implementation of SVM. Table IV shows the simplified version of classification algorithm using the second technique.

TABLE IV. SIMPLIFIED CLASSIFICATION ALGORITHM (SECOND TECHNIQUE)

Classification Algorithm (Second Technique)	
(1)	Set the training data.
(2)	Extract features and descriptors of images of all classes.
(3)	Calculate the centroids of descriptors of each class using K-means clustering algorithm.
(4)	Extract the features and descriptors of query image.
(5)	Match the descriptors of query image with centroids of all classes.
(6)	Calculate mean values of matched descriptors for each class separately.
(7)	Get the minimum 3 mean values.

Using the output from both techniques, we improved the classification of images. Table V shows the confidence scores

TABLE V. RELEVANCE SCORES OF DIFFERENT IMAGES BELONGING TO DIFFERENT CLASSES

Query	Scores			
	Best	2 nd Best	3 rd Best	4 th Best
Plane	0.85118	0.33915	0.33018	0.32292
Bike	0.71795	0.32445	0.32423	0.32395
Face	0.5004	0.29813	0.29686	0.29454
Car	9.48087	0.32434	0.32417	0.32318
Elephant	0.57092	0.36381	0.36137	0.35581
Bread Maker	0.92687	0.32338	0.32263	0.32160
Revolver	0.60420	0.34228	0.34083	0.33813
Average	0.66464	0.33079	0.32861	0.32573

for a few example images which were tested during the implementation. The first best result is the output of the first

technique, and the rest of the results came from the second technique.

4) Ranking of Images:

Once we received the class label from the SVM classifier, we again computed the descriptors for each image belonging to that class. We used FLANN (Fast Approximate Nearest Neighbor Search Library) [15] to match the descriptors of the query image and the class images. We then calculated the mean values of matched descriptors (lower the distance, the better). FLANN based-matcher is quick and efficient in feature matching making it an ideal choice for us [16].

At the end, we sorted the class images in ascending order of their distances so that best matches are at the top. Table VI shows the proposed ranking algorithm.

TABLE VI. SIMPLIFIED RANKING ALGORITHM

Ranking Algorithm	
(1)	Compute descriptors of the query image.
(2)	Predict the class of image using SVM.
(3)	For each image in the class, compute descriptors.
(4)	Use FLANN based-matcher to match the descriptors between query image and all images in the class.
(5)	Calculate mean values of matched descriptors.
(6)	Sort the class images in ascending order of their distances so that best matches (with low distance) come to front.

V. RESULT ANALYSIS

A large number of images belonging to a set of 20 predefined concepts (e.g., plane, car, elephant, etc.) have been exhaustively tested to evaluate the functioning of the system then real-time input was taken.

The datasets were taken from ‘Computational Vision at Caltech’ (2,400 images, 120 per class) [17].

The test with SURF feature detector showed good performance with respect to the system requirements, but it was not very stable to rotation and illumination changes. We came very close to the performance of SIFT (Scale Invariant Feature Transform). According to the maker of SURF detector, it should be quicker than SIFT detector. Our inability to perform at that level is most likely due to the avoidance of pre-processing of images before using SURF algorithm.



Fig. 8. Matching visual words and their histogram representation

Even though the feature detection step is quick, feature extraction is slow and time consuming process as calculation of the SURF descriptors takes a very long time. Hence, the training of image classifier takes quite a while. On the bright side, since training of image classifier is done just once toward the beginning, classification and ranking of images is relatively fast.

We use the standard precision and recall definitions from image retrieval to assess the relatedness of the retrieved results which are given as follows:

$$\text{Precision} = \frac{\text{No.of Relevant Images Retrieved}}{\text{No.of Retrieved Images}} \quad (1)$$

$$\text{Recall} = \frac{\text{No.of Relevant Images Retrieved}}{\text{Total No.of Relevant Images}} \quad (2)$$

Fig. 9 shows the retrieval results obtained for the text-based user query "Cars".

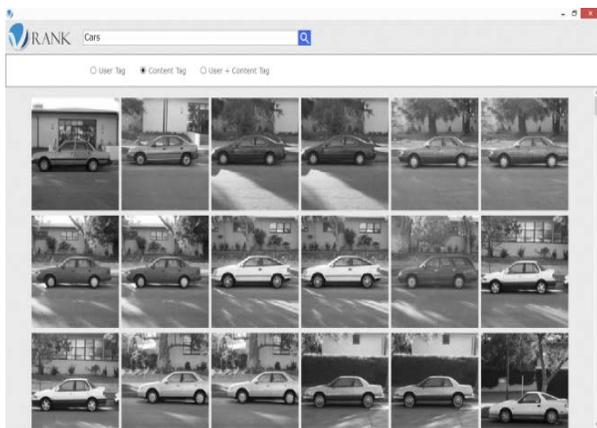


Fig. 9. Search Results

Table VII shows the performance of the proposed system which was measured on the basis of precision and recall of image retrieval.

TABLE VII. PERFORMANCE ANALYSIS

Query	Total Relevant Images	Retrieved Images	Relevant Images Retrieved	Precision	Recall
Text	15	10	9	0.9	0.6
Text and Content	15	13	11	0.85	0.73

Our results show that the proposed system is able to retrieve more than 80% of relevant images from an image collection, given a user-query.

VI. CONCLUSIONS

In this paper, a novel image ranking technique is proposed using the metadata of images and the concept of bag-of-visual-words, which is an effective image representation in the classification and retrieval tasks. The accuracy is higher in comparison to tag-based image retrieval. The proposed system is an improvement from the user-tag only approach and is a promising step towards improving image retrieval using text

queries. The evaluation of sample dataset proves the effectiveness of the proposed system.

VII. FUTURE WORK

We have proposed an image ranking system that uses visual features of images for generating the content tags for images as well as later using these tags for the retrieval purposes. A hybrid approach is presented that uses both user and the generated content tags for the retrieval purposes. An evaluation study shows the usefulness of the results.

Although, the proposed system has improved the ranking of images, the algorithm still needs refinements so that it is able to retrieve more relevant results. Therefore, in future we plan to incorporate the concept of "visual keywords" which are the significant portion of images related to the user tags, to further improve the image retrieval process.

REFERENCES

- [1] Chunsheng Fang and Ryan Anderson, "A Parallel implementation of Content-based image Retrieval." Faculty of Computing, Viron University, Chicago, Technical Report cs15-0013, 2008.
- [2] V Rajakumar and Vipeen V Bopche, "Image Search Reranking," International Journal of Computer Trends and Technology (IJCTT), vol. 6, no. 5, pp. 242-247, 2013.
- [3] Bill Slawski. (2008, May) How Do Images Get Ranked In Image Search? [Online]. <http://www.seobythesea.com>
- [4] Aixin Sun, Sourav S. Bhowmick, Khanh Tran Nam Nguyen, and Ge Bai, "Tag-based social image retrieval: An empirical evaluation," Journal of the American Society for Information Science and Technology, vol. 62, no. 12, pp. 2364-2381, December 2011.
- [5] Jun Yang, Yu-Gang Jiang, Alexander Hauptmann, and Chong-Wah Ngo, "Evaluating Bag-of-Visual-Words Representations in Scene Classification," In International Workshop on Multimedia Information Retrieval, pp. 197-206, 2007.
- [6] Alceu Costa. (2012, July) Bag of words training and testing opencv, MATLAB. [Online]. <http://www.stackoverflow.com>
- [7] David G. Lowe, "Object recognition from local scale-invariant features," in International Conference on Computer Vision, Corfu, Greece, 1999, pp. 1150-1157.
- [8] Yan Ke and Rahul Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," Computer Vision and Pattern Recognition, pp. 66-75, 2004.
- [9] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in European Conference on Computer Vision, Graz, Austria, 2006.
- [10] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang, "Tag Ranking," in Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 351-360.
- [11] Luo Juan and Oubong Gwun, "A comparison of sift, PCA-sift and surf," Int. J. Image Process., vol. 3, no. 5, pp. 143-152, 2009.
- [12] Gil Levi. (2013, August) Bag of Words Models for visual categorization. [Online]. <https://www.gilscvblog.wordpress.com>
- [13] Hiroharu Kato and Tatsuya Harada, "Image Reconstruction from Bag-of-Visual-Words," in IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 955-962.
- [14] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
- [15] (2015, June) Bag of Words. [Online]. <http://www.cs.ubc.ca/research/flann/>
- [16] Marius Muja and David G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in International Conference on Computer Vision Theory and Application, 2009, pp. 331-340.
- [17] Fei-Fei Li, Marco Andreetto, and Marc'Aurelio Ranzato. (2003, September) Vision Lab. [Online]. <http://www.vision.caltech.edu>

A Disaster Document Classification Technique Using Domain Specific Ontologies

Qazi Mudassar Ilyas

College of Computer Sciences and Information Technology
King Faisal University
Al-Ahsa, 31982, Saudi Arabia

Abstract—Manual data collection and entry is one of the bottlenecks in conventional disaster management information systems. Time is a critical factor in emergency situations and timely data collection and processing may help in saving several lives. An effective disaster management system needs to collect data from World Wide Web automatically. A prerequisite for data collection process is document classification mechanism to classify a particular document into different categories. Ontologies are formal bodies of knowledge used to capture machine understandable semantics of a domain of interest and have been used successfully to support document classification in various domains. This paper presents an ontology-based document classification technique for automatic data collection in a disaster management system. A general ontology of disasters is used that contains the description of several natural and man-made disasters. The proposed technique augments the conventional classification measures with the ontological knowledge to improve the precision of classification. A preliminary implementation of the proposed technique shows promising results with up to 10% overall improvement in precision when compared with conventional classification methods.

Keywords—Disaster Management; Document Classification; Ontology; Supervised Learning; Information Retrieval

I. INTRODUCTION

EM-DAT International Disaster Database of the Centre for Research on the Epidemiology of Disasters¹ classifies disasters into two general categories, namely Natural Disasters and Technological Disasters. Some more specific subcategories of Natural Disasters include Earthquake, Mass Movement, Volcanic Activity, Extreme Temperature, Fog, Storm, Flood, Landslide, Wave Action, Drought, Glacial Lake Outburst, Wildfire Epidemic, Insect Infestation and Animal Accident. Similarly, Technological Disasters are subdivided into Chemical Spill, Collapse, Explosion, Fire, Gas Leak, Poisoning, Radiation, Air Accident, Road Accident, Rail Accident, Water Accident and Others. All disasters are humanitarian crisis of varying degrees and usually need some mitigation measure to minimize losses to lives and infrastructure. Information Technology can also play a vital role in disaster management. Conventional disaster management systems such as Sahana² depend on manual collection, entry, and management of database for disaster management. Ilyas and Ahmed propose SAHARA [1], a

semantic disaster management system to support disaster management. The proposed system comprises the following components:

- A knowledge base is used to formally capture knowledge about disasters and disaster management in the form of disaster ontologies. A base level disaster ontology is developed by Afzal et al. [2].
- A data collection components collects disaster-related information from various resources on World Wide Web such as blogs, social networks, wiki sites, news sites, government and non-government organizations etc [3]. Ontology developed during the previous phase may also be used to support data collection.
- A reasoner is used to perform reasoning on ontologies and the instance data collected by the data collection component. This process produces useful information to support disaster management such as location of disaster, intensity of disaster, information about inaccessible routes of affected area, services required in affected areas, infrastructure damage, number of casualties, livestock loss, services available and required in nearby hospitals.
- An alert management sub-system sends alerts to various stakeholders such as hospitals, government organizations, non-government organizations and volunteers to support decision making for effective disaster management.

This paper presents a document classification technique that can be used in data collection phase of SAHARA. The first step during data collection is to label a newly found document according to specified categories. A supervised learning approach is used because the categorization information is already available in the form of an ontology. These categories are formed by various concepts and properties in the domain of disaster management. A set of measures usually used in conventional classification techniques is supported with the ontological knowledge to improve the precision of classification process. The conventional measures include URL of a link, anchor text, inbound links, position & frequency of the target category and URL depth of the document being processed. Ontology computations involve ontology concepts, properties, relationships, annotations and instances. Rest of the paper is organized as follows:

¹ <http://www.emdat.be/>

² <http://sahanafoundation.org/>

Section 2 presents a review of use of ontologies in disaster management systems. Section 3 gives details of the proposed technique. Results are presented in section 4 followed by the conclusion and future directions in section 5.

II. RELATED WORK

To find relevance of a document with the target concept in a distributed environment like Internet, the traditional approaches in document classification focus on processing links in the document, popularity of the document through inbound links, frequency and position of the term in the document. More recently, the researchers have also used ontologies to support the classification process. As ontologies are used to capture domain knowledge in a formal and explicit way, they are a natural choice in document classification process. Ontologies have been used in a diverse range of domains from cultural heritage [4] to 3D modeling [5], e-commerce [6] to health services [7], human anatomy [8] to fraud detection [9] and cyber warfare [10] to agriculture [11]. Punitha et al. argue that ontology augmentation can improve the document classification process significantly [12].

Disaster management systems can also benefit from ontologies significantly in various phases and tasks of disaster management. Hristidis et al. have identified five phases in disaster management that need data analysis and management, namely information extraction, information retrieval, information filtering, data mining and decision support [13]. Each one of these phases has its own unique challenges and the researchers have explored the use of ontologies in all of them. Imran et al. have used ontologies to support information extraction process from micro blogging sites [14]. Their work is based on ontology proposed by Vieweg et al. that captures information about *Caution & Advice, Casualties & Damage, Donations of Money, Goods or Services, People Missing, Found, or Seen and Information Source* [15]. The proposed method achieved up to 93% accuracy and 64.5% recall for some concepts.

Fan and Zlatanova have used ontologies for semantic interoperability in disaster management [16]. The proposed methodology comprises two phases. In the first phase, ontologies are developed and evaluated for actors, static & dynamic data models, processes and task. In the second phase, several ontologies are matched together to identify and match common concepts in these ontologies. Ontologies are also updated if required. The authors have used a primitive case study to validate the proposed methodology.

Haghighi et al. have proposed Domain Ontology for Mass Gatherings (DO4MG); an ontology for intelligent decision support in medical emergency management for mass gatherings [17]. The top level concepts in the ontology include *Environmental Factors, Mass Gathering Plan, Gathering Type, Crowd Features* and *Event Venue*. Two evaluation approaches, namely criteria-based evaluation and application-based evaluation are used to evaluate the developed ontology. A prototype system is developed for application-based evaluation. The results are encouraging and prove that

DO4MG ontology can be used effectively to support the decision making process in mass gatherings. Amailef Lu have proposed a similar system and proved its effectiveness to support case-based reasoning in m-government emergency response services [18].

Chen et al. have proposed an ontology based decision support system for disaster management in typhoons [19]. The proposed system comprises three phases including feature extraction, damage prediction and risk analysis. An ontology is used to support these phases. The authors argue that the performance of the system depends on accuracy and completeness of the knowledge captured by ontologies.

Cabacas et al. have proposed an ontology-based messaging system to utilize social relations as a service [20]. The user query is analyzed by the system to “understand” the user’s social and physical environment. A service matching component finds the most suitable service based on several criteria such as location, time and situation. Finally, service messenger component broadcasts the message to the concerned stakeholders.

Hristoskova et al. have used a set of generic as well as domain specific ontologies to support the reasoning process in disaster management [21]. A data aggregator component collects data from various devices and sensors. This data is passed on to context engine which updates/queries a semantic model composed of ontologies. The context engine also interacts with a decision engine for updating, querying and evaluating the rules. The proposed approach is validated through implementation in two scenarios. A critical analysis of the related work strengthens the case and need of developing an ontology-based document classification method for disaster management system that can be used to categorize various kind of documents from World Wide Web.

III. PROPOSED METHODOLOGY

The proposed approach attempts to categorize a document with a target concept in the domain of disaster management. The process is divided into three phases, namely link relevance, page relevance and ontology relevance. Finally, these scores are combined into an overall document relevance score. The details of these three phases are as follows.

A. Link relevance computations

Link relevance is based on the measures commonly used in classical clustering methods. These include anchor text, URL text, and link popularity. A page will be assigned a higher relevance score if the target concept appears in the anchor text and URL text. Also, the relevance score will be higher for a popular page i.e., a page having more number of inbound links from external documents.

B. Page relevance computations

The structure and content of a document/webpage play important role in computing its relevance with a particular concept. Page computation is further divided into the following measures:

1) Term frequency-Inverse document frequency (TF-IDF)

TF-IDF score is a classical method of assigning more weight to a more frequent term in a document and a lower weight to unimportant terms in the entire document collection. Several variations exist and one of them is given below [22]:

$$P_{tf} = \log(f_{t,d}) + 1 \text{ if } f_{t,d} > 0; 0 \text{ otherwise} \quad (1)$$

Where $f_{t,d}$ represents frequency of term t in document d .

A commonly used formula for calculating inverse document frequency is:

$$P_{idf} = \log(N/N_t) \quad (2)$$

Where N is the total number of documents in the collection and N_t is the number of documents in which term t appears.

Finally, P_{tf-idf} can be calculated by simply multiplying P_{tf} and P_{idf} .

$$P_{tf-idf} = P_{tf} * P_{idf} \quad (3)$$

2) Attribute relevance

The position of a term appearing in a document plays an important role in classifying a document. If a term appears in title, first or second level heading, then the document is more relevant to that term as compared to another document in which the same term appears in a paragraph.

3) URL depth

URL depth refers to how deep a web page lies in a website. The closer a webpage is to site root; the more it is considered to be relevant to the target concept. A webpage located deeper in a site hierarchy is considered to be less important.

C. Ontology relevance computation

As mentioned above, Ontologies are an excellent source of document classification because they are formal bodies of knowledge developed for specific domains. In this work, the base level disaster ontology developed by Afzal et al. is used [2]. The top level concepts in the ontology include *Disaster*, *Disaster Location*, *Disaster Date*, *Losses*, *Services*, *Service Providers*, and *Relief Items*. A partial hierarchy of *Services* concept in the ontology is given in Fig. 1. Fig. 2 shows a detailed description of *Transportation Hazard* concept in the ontology. The details of ontology relevance computations are given below:

1) Ontology concepts

A positive match between concepts in a document with the ontological concept to be classified may serve as an important document classification measure. This measure is given the highest weight in our classification process because of the formal semantics captured in an ontology.

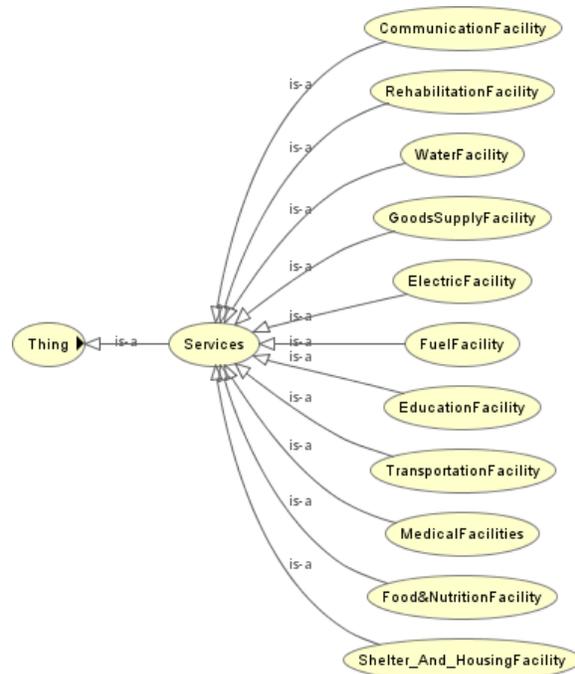


Fig. 1. A subconcept hierarchy of Service concept in the disaster management ontology

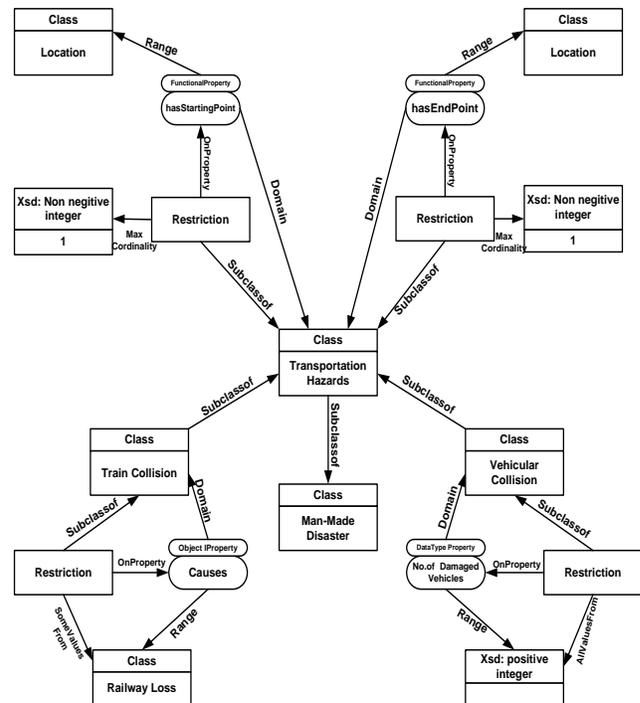


Fig. 2. A detailed visual description of Transportation Hazard concept in the disaster management ontology

2) Ontology properties

Ontology properties are used to define relationships of concepts with literals only such as *OccurredOn* is a property of *Disaster* concept to describe date and time of occurrence of disaster. Ontology properties can play an important role in document classification as they are used to define the concept unambiguously. Two cases may arise in this case. First, if an ontology concept is matched in a document and the properties are also similar, then the confidence of relevance is very high. On the other hand, if concepts are different but there is high similarity between the properties, then there are high chances of similarity and it is assumed that different synonyms are used for the same concept.

3) Ontology relationships

While ontology properties establish a link between ontology concepts and literals, ontology relationships are used to relate concepts with other concepts. Ontology relationships can give contextual and domain information such as *hasLocation* relates the *Disaster* concept with the *Location* concept. Relationships are important measure for document classification as they can help in reducing ambiguity with contextual information.

4) Ontology annotations

An ontology may have a number of annotation properties such as *SeeAlso* can be used to point to another source describing the same concept. Other examples include *Label*, *Comment*, *SeeAlso* and *IsDefinedBy*. These annotations may use to give synonyms of a term, refer to some other resources for further description or give human-readable labels.

5) Ontology instances

Instances relate concrete things to general class of concepts e.g., *Katrina*³ is an instance of *Hurricane disaster*. A document containing instance of the target concept is assigned a higher weight.

D. Proposed Algorithms

The algorithms for the three computational components mentioned above i.e., link relevance, page relevance and ontology relevance, are given below.

Algorithm LinkToConceptRelevance

Inputs: Source document, Target concept, Set of concepts from ontology, Weight of anchor text, Weight of URL text and Weight of link popularity

Output: Link relevance score

Let

Concept=Target concept in disaster domain

NumLinks = Total number of links in Page

Anchor= Anchor text of a link

S_a, S_u, S_{lp} = Temporary variables to store relevance scores for anchor text, URL and link popularity respectively

$Rel_a, Rel_u, Rel_{lp}, Rel_L$ =Relevance for anchor text, URL, link popularity and total link relevance with the target concept respectively

W_a, W_u, W_{lp} = Weight assigned to anchor text, URL and link popularity respectively

$S_a, S_u, S_{lp} \leftarrow 0$

$Rel_a, Rel_u, Rel_{lp}, Rel_L \leftarrow 0$

For all Links in the page

If target of Link is a valid page or an OWL/RDF file
Store Link in database

End if

For all Tokens in the Anchor

If Token contains Concept

$S_a \leftarrow S_a + 1$

End if

End for

For all Tokens in the URL

If Token contains Concept

$S_u \leftarrow S_u + 1$

End if

End for

End for

Get S_{lp} using Google API

Normalize S_a and S_u by length of document

$Rel_a \leftarrow S_a * W_a$

$Rel_u \leftarrow S_u * W_u$

$Rel_{lp} \leftarrow S_{lp} * W_{lp}$

$Rel_L = Rel_a + Rel_u + Rel_{lp}$

The algorithm for computing page relevance is given below.

Algorithm PageToConceptRelevance

Inputs: Source document, Target concept, Set of concepts from ontology, Weight of title tag, Weight of heading tag and Weight of TF-IDF

Output: Page relevance score

Let

Concept=Target concept in disaster domain

Title=Title of the page

TF=Term frequency

N=Total number of documents

N_t =Number of documents in which term t appears

W_t, W_h, W_{tf-idf} = Weight assigned to title, heading and tf-idf respectively

$S_t, S_h, S_{tf}, S_{idf}, S_{tf-idf}$ = Temporary variables to store relevance scores for title, heading, tf, idf and tf-idf respectively

$Rel_t, Rel_h, Rel_{tf-idf}, Rel_p$ = Relevance score for title,

heading, tf-idf, and total relevance for document with the

target concept respectively

$S_t, TF, S_h, S_{tf}, S_{idf}, S_{tf-idf} \leftarrow 0$

$Rel_t, Rel_h, Rel_{tf}, Rel_{idf}, Rel_{tf-idf}, Rel_p \leftarrow 0$

For all Tokens in Title Do

If Title contains Concept

$S_t \leftarrow S_t + 1$

End if

End For

For all Tokens in document Do

If Token contains Concept

$TF \leftarrow TF + 1$

If Token is in Heading 1

$S_h \leftarrow S_h + \log(TF)$

End if

If Token is Heading 2

³ <http://www.history.com/topics/hurricane-katrina>

```

                Sh ← Sh+ log (log (TF))
            End if
        End if
    End For
    If TF > 0
        Stf ← log (TF)
    End if
    Sidf ← log (N/Nt)
    Stf-idf ← Stf * Sidf
    Normalize St and Sh by length of document
    Relt ← St * Wt
    Reltf-idf ← Stf-idf * Wtf-idf
    Relh ← Sh * Wh
    Relp ← Relt + Reltf-idf + Relh

```

The algorithm for ontology relevance computation is given below.

Algorithm OntologyToConceptRelevance

Inputs: Word vector of document, Word vectors of ontology concepts, properties, annotations and instances, Weight assigned to concepts, properties, relations, annotations, instances and cosine similarity

Output: Ontology relevance score

Let

Concept=Target concept in disaster domain

S_c, S_p, S_r, S_a, S_i, S₀ = Temporary variables to store relevance scores for ontology concepts, properties, relations annotations, instances and ontology respectively

Rel_c, Rel_p, Rel_r, Rel_a, Rel_i, Rel₀ = Relevance of ontology concepts, properties, relations, assertions, instances and ontology with the target concept respectively

CS=Cosine similarity measure of the document

CS_c, CS_p, CS_r, CS_a, CS_i = Cosine similarity measures for concepts, properties, relations, annotations and instances respectively

W_c, W_p, W_r, W_a, W_i = Cosine similarity measure weights for concepts, properties, relations, annotations and instances respectively

\vec{D} = Word vector of document

\vec{C} , \vec{P} , \vec{R} , \vec{A} , \vec{I} = Word vector of concepts, properties, relations, annotations, and instances in the ontology respectively

W_c, W_p, W_r, W_a, W_i, W_{CS} = Weight assigned to concepts, properties, relations, annotations, instances and cosine similarity respectively

S_i, S_p, S_r, S_a, S_i, S₀ ← 0

Rel_c, Rel_p, Rel_r, Rel_a, Rel_i, Rel₀ ← 0

For all Tokens in document Do

 For all Concepts in ontology Do

 If Token contains Concept

 S_c ← S_c + 1

 End if

 End For

 For all Properties in ontology Do

 If Token contains Property

 S_p ← S_p + 1

 End if

End For

For all Relations in ontology Do

 If Token contains Relation

 S_r ← S_r + 1

 End if

End For

For all Annotations in ontology Do

 If Token contains Annotation

 S_a ← S_a + 1

 End if

End For

For all Instances in ontology Do

 If Token contains Instance

 S_i ← S_i + 1

 End if

End For

End For

$$CS_c \leftarrow \frac{\vec{D}}{|\vec{D}|} \cdot \frac{\vec{C}}{|\vec{C}|}$$

$$CS_p \leftarrow \frac{\vec{D}}{|\vec{D}|} \cdot \frac{\vec{P}}{|\vec{P}|}$$

$$CS_r \leftarrow \frac{\vec{D}}{|\vec{D}|} \cdot \frac{\vec{R}}{|\vec{R}|}$$

$$CS_a \leftarrow \frac{\vec{D}}{|\vec{D}|} \cdot \frac{\vec{A}}{|\vec{A}|}$$

$$CS_i \leftarrow \frac{\vec{D}}{|\vec{D}|} \cdot \frac{\vec{I}}{|\vec{I}|}$$

$$CS \leftarrow (CS_c + CS_p + CS_r + CS_a + CS_i) * W_{CS}$$

Normalize S_c, S_p, S_r, S_a, S_i and S₀ by length of document

$$Rel_c \leftarrow S_c * W_c$$

$$Rel_p \leftarrow S_p * W_p$$

$$Rel_r \leftarrow S_r * W_r$$

$$Rel_a \leftarrow S_a * W_a$$

$$Rel_i \leftarrow S_i * W_i$$

$$Rel_0 \leftarrow Rel_c + Rel_p + Rel_r + Rel_a + Rel_i + CS$$

Finally, the three algorithms given above are combined to compute the final relevance score of the document being processed.

Algorithm DocumentClassification

Input: Domain ontology of disaster, Set of documents, Weight assigned to link relevance, page relevance and ontology relevance

Output: Final relevance of a document with the target concept

Let

d=Document being processed

c= A concept in the ontology

W=Weight of a measure

W_L, W_P, W_O ← 0

Extract Concepts from disaster ontology

For all Concepts Do

 For all Documents Do

$$Relevance_L \leftarrow LinkToConceptRelevance (d_m, c_m,$$

$$\{c_1, c_2, \dots, c_n\}, W_a, W_l, W_{lp})$$

$$\text{Relevance}_P \leftarrow \text{PageToConceptRelevance}(d_m, c_m,$$

$$\{c_1, c_2, \dots, c_n\}, W_t, W_h, W_{\text{tf-edf}})$$

$$\text{Relevance}_O \leftarrow \text{OntologyToConceptRelevance}(\rightarrow_C, \rightarrow_P, \rightarrow_R, \rightarrow_A, \rightarrow_I, W_c, W_p, W_r, W_a, W_i)$$

$$\text{Relevance}_{\text{Total}} \leftarrow \text{Relevance}_L * W_L + \text{Relevance}_P * W_{P+} + \text{Relevance}_O * W_O$$

End For
End For

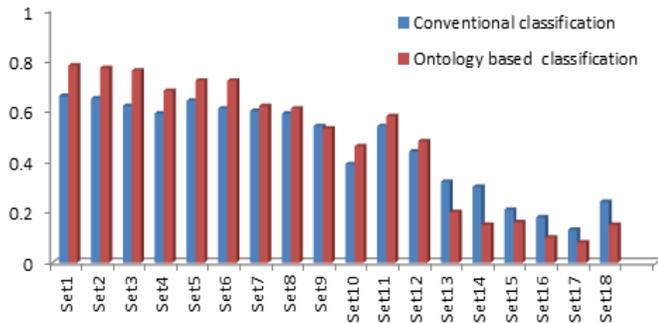


Fig. 3. A comparison of precision of conventional and ontology-based classification approaches

IV. RESULTS AND DISCUSSION

The proposed algorithm is tested on eighteen sets of documents related to various concepts in disaster management domain. These documents are categorized by human reviewers for their relevance with the target concepts. Then the results of conventional and ontology based classification are compared. Fig. 3 shows results of proposed algorithm on 18 sets of documents, each set consisting of 20 documents and the results are averaged for each set. The first six sets of documents (Set1 – Set6) were highly relevant to the target concept. The next six document sets (Set7 – Set12) were moderately related with the target concept. The last six sets (Set13 – Set18) were unrelated with the target concept. The results show that the ontology based classification performed better both for highly relevant and irrelevant documents. The proposed algorithm ranked relevant document higher than the conventional technique. The overall average gain achieved was 11%. For moderately relevant documents, the difference between proposed and traditional algorithm was marginal i.e., 3%. In case of unrelated documents, the proposed algorithm ranked the documents lower than the traditional algorithms. In this case, the average difference was 9%. Hence, the proposed algorithm achieved an overall improvement of about 10% because of use of ontologies.

V. CONCLUSION AND FUTURE WORK

The proposed ontology-based document classification technique outperforms the conventional methods because of formal semantics provided by the ontology. The initial evaluation on a selected set of documents showed up to 10% overall improvement in the precision of classification. However, the proposed techniques has some limitations. First, it depends on availability of ontologies. As there are no standard disaster ontologies available, the performance of a typical system depends on the quality and accuracy of

ontologies used. Another limitation is a lack of availability of instance data. Also, the ontological processing is computationally expensive as compared to traditional approaches.

The future work involves evaluation of the proposed technique in a distributed environment like World Wide Web. A real life implementation in a particular disaster situation is also required to evaluate the proposed methodology. Moreover, in this work, a general ontology of disaster management is used that covers several kinds of disasters. One may also consider using a specific ontology targeted to a particular kind of disaster to improve the effectiveness of the proposed approach, e.g., an earthquake ontology for classifying earthquake-related documents and an tsunami ontology for tsunami-related documents. More specific ontologies may also have added advantage of improved efficiency because of narrower coverage of domain. Another future direction may focus on the selection of ontologies in real time. In this case, the system is not given an initial ontology as input but the most suitable ontology is selected based on the first few documents. A system may also be designed to use different ontologies for different set of documents. The criteria might include level of granularity or specificity of the concepts in the documents being processed.

ACKNOWLEDGMENT

The support of Mr Raza Kashif, Mr Mehtab Afzal, Mr Hamad Ahmed, Mr M. Abdul Wahab and Mr Sohail Irshad in the implementation of this work is greatly appreciated.

REFERENCES

- [1] Q. M. Ilyas and I. Ahmad, "A conceptual architecture of SAHARA - a semantic disaster management system," World Appl. Sci. J, vol. 10, pp. 980-985, 2010.
- [2] M. Afzal, Q.M. Ilyas, I. Ahmad and J. Ajoon, " A base level ontology for disaster management," Journal of Internet Technology, 2017 (In press)
- [3] I. Ahmed, Q. M. Ilyas, J. Ajoon and M. Afzal, "Gleaning disaster related information from world wide web using GATE," Journal of Theoretical and Applied Information Technology, vol. 40, pp. 135-142, 2012.
- [4] E. Mäkelä, E. Hyvönen and T. Ruotsalo, "How to deal with massively heterogeneous cultural heritage data-lessons learned in CultureSampo." Semantic Web, vol. 3, pp. 85-109, 2012.
- [5] C. Metral, N. Ghoula and G. Falquet, "An ontology of 3D visualization techniques for enriched 3D city models," in Usage, Usability, and Utility of 3D City Models, final conference of the European COST Action TU0801, 2013.
- [6] A. I. La Paz, A. Ramaprasad, T. Syn and J. Vasquez, "Editorial: An Ontology of E-Commerce-Mapping a Relevant Corpus of Knowledge," Journal of Theoretical and Applied Electronic Commerce Research 2015.
- [7] L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng and W. A. Kibbe, "Disease Ontology: a backbone for disease semantic integration," Nucleic Acids Res., vol. 40, Jan, 2012.
- [8] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis and M. A. Haendel, "Uberon, an integrative multi-species anatomy ontology," Genome Biol., vol. 13, 2012.
- [9] R. N. Carvalho, S. Matsumoto, K. B. Laskey, da Costa, Paulo Cesar G, M. Ladeira and L. L. Santos, "Probabilistic ontology and knowledge fusion for procurement fraud detection in brazil." in URSW (LNCS Vol.), pp. 19-40, 2013.
- [10] R. Dipert, "The essential features of an ontology for cyberwarfare," Conflict and Cooperation in Cyberspace, pp. 35-48, 2013.

- [11] S. Lata, B. Sinha, E. Kumar, S. Chandra and R. Arora, "Semantic web query on e-Governance data and designing ontology for agriculture domain," *International Journal of Web & Semantic Technology (IJWesT)*, vol. 4, pp. 65-72, 2013.
- [12] S. Punitha, K. Mugunthadevi and M. Punithavalli, "Impact of ontology based approach on document clustering," *International Journal of Computer Applications*, 2011.
- [13] V. Hristidis, S. Chen, T. Li, S. Luis and Y. Deng, "Survey of data management and analysis in disaster situations," *J. Syst. Software*, vol. 83, pp. 1701-1714, 2010.
- [14] M. Imran, S. M. Elbassuoni, C. Castillo, F. Diaz and P. Meier, "Extracting information nuggets from disaster-related messages in social media," in *Proceedings of ISCRAM, Baden-Baden, Germany*, 2013.
- [15] S. Vieweg, A. L. Hughes, K. Starbird and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1079-1088, 2010.
- [16] Z. Fan and S. Zlatanova, "Exploring ontologies for semantic interoperability of data in emergency response," *Applied Geomatics*, vol. 3, pp. 109-122, 2011.
- [17] P. D. Haghighi, F. Burstein, A. Zaslavsky and P. Arbon, "Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings," *Decis. Support Syst.*, vol. 54, pp. 1192-1204, 2013.
- [18] K. Amailef and J. Lu, "Ontology-supported case-based reasoning approach for intelligent m-Government emergency response services," *Decis. Support Syst.*, vol. 55, pp. 79-97, 2013.
- [19] W. Chen, G. Sui and D. Tang, "A fuzzy intelligent decision support system for typhoon disaster management," in *2011 IEEE International Conference on Fuzzy Systems (FUZZ)*, pp. 364-367, 2011.
- [20] R. Cabacas, R. Sankar and I. Ra, "Context-aware emergency messaging system framework utilizing social relations as services," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, pp. 77-86, 2014.
- [21] A. Hristoskova, F. Ongenae and F. De Turck, "Semantic reasoning for intelligent emergency response applications," in *2013 11th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 547-554, 2013.
- [22] S. Büttcher, C. L. Clarke and G. V. Cormack, "Information Retrieval: Implementing and Evaluating Search Engines," MIT Press, 2010.

Intrusion Detection System in Wireless Sensor Networks: A Review

Anush Ananthakumar

Student of Electronics and
Telecommunication Engineering
Thadomal Shahani Engineering
College
Mumbai, India

Tanmay Ganediwal

Student of Electronics and
Telecommunication Engineering
Thadomal Shahani Engineering
College
Mumbai, India

Dr. Ashwini Kunte

HOD of Electronics and
Telecommunication Engineering
Vice Principal of Thadomal Shahani
Engineering College
Mumbai, India

Abstract—The security of wireless sensor networks is a topic that has been studied extensively in the literature. The intrusion detection system is used to detect various attacks occurring on sensor nodes of Wireless Sensor Networks that are placed in various hostile environments. As many innovative and efficient models have emerged in the last decade in this area, we mainly focus our work on Intrusion detection Systems. This paper reviews various intrusion detection systems which can be broadly classified based on certain traditional techniques, namely signature based, anomaly based and hybrid based. The models proposed by various researchers have been critically examined based on certain classification parameters, such as detection rate, false alarm, algorithms used, etc. This work contains a summarization study of various intrusion detection systems used particularly in Wireless Sensor Networks, and also highlights their distinct features.

Keywords—Wireless sensor networks; Intrusion Detection System; Signature based IDS; Anomaly based IDS; Hybrid based IDS; Algorithms

I. INTRODUCTION

Wireless Sensor Networks (WSN) are used for monitoring the environment or a given area by collection of data, such as temperature, sound, pressure, light, etc from various Sensor Nodes (SNs) and analyzing them at a Base Station [1, 2]. The WSN consists of hundreds of sensor nodes that are basically small sensors used for monitoring the environment. The advantage of these sensors is that they can be placed in any location where surveillance by humans is not possible, including harsh climatic conditions or underwater surveillance [3]. The WSNs are used in a variety of fields ranging from healthcare and area monitoring to environmental and industrial monitoring systems.

This paper focuses on one of the applications of Wireless Sensor Networks namely Intrusion Detection Systems (IDS) [5, 6]. Intrusion detection systems are used to detect intrusions in a certain network or an area under surveillance. Intrusion is defined as an unauthorized (unwanted) activity in a network. In [4], an efficient IDS has been proposed in the field of healthcare for prevention against intrusions. On the basis of detection methodology, IDS are traditionally classified into 3 models: Anomaly based, Signature based and Hybrid Based IDS. The signature based IDS have predefined set of rules that are designed on the basis of previously known security attacks

and the signatures of the attacks are stored in a database. The signature is a kind of pattern that describes a known attack. The incoming information is compared and checked with the previously identified signatures and hence protect against well known attacks and also have the advantage of low false alarm rate (FAR). A preliminary rule based approach to detect intrusions is developed in [7] that is based on comparison of the incoming packets with known signatures. On the other hand as it has been pointed out in [8, 9], the signature based model is similar to an anti-virus system that has a database and can detect known attacks but has problems when unknown attacks whose signatures are unknown are to be detected. To eliminate this particular drawback, the anomaly based IDS are used which works on the basis of a threshold [10]. This type of IDS defines what is called as a normal behaviour and an abnormal behaviour. Any new inbound information packet is verified against this normal behaviour and determined if it is an intrusion or not. As the detection mechanism is based on a threshold for normal traffic pattern, it has the capability to detect new intrusions, but on the other hand, it has a major disadvantage of missing out on well known attacks. The anomaly based model has a high detection rate and seldom classifies an actual intrusion as a normal packet, but it has a large false positive rate (FPR) i.e normal packets are defined as abnormal. Also as suggested in [11], there could be attacks due to hybrid anomaly which consists of multiple anomaly attacks, for which he proposes a model which has a detection technique based on K-means clustering. To improve on the disadvantages of these two conventional methods, a hybrid of the two IDS is usually incorporated known as a Hybrid Intrusion Detection System (HIDS). In this system, both the IDS are present, with the anomaly based IDS usually functioning as a filter and the signature based IDS as a second level of intrusion detection as it has low false positives and can accurately detect the intrusions. For example, [12] has proposed a hybrid intrusion detection model that integrates anomaly based IDS based on support vector mechanism (SVM) with a misuse detection based IDS to achieve a high detection rate of 98% and a low false positive rate. Apart from these, a developing area of intrusion detection is the cross layered IDS that can detect attacks on different OSI layers. A cross layer based IDS that integrates the Mac and Physical layer has been proposed by [13]. However in this paper, we focus only on the signature, anomaly and hybrid based IDS. This paper attempts to review the work carried out by various

researchers in the broad area of intrusion detection systems, which are traditionally classified as signature, anomaly and hybrid based IDS. It is of interest to see how various models perform with respect to certain critical parameters that help us in understanding the robustness and effectiveness of these models against various security threats. This will also help in drawing certain important insights about the algorithms used and the preferred detection techniques incorporated in different conditions.

In Section 2, various models of Signature, Anomaly and Hybrid based IDS proposed by various researchers has been

discussed. The subsequent section is on the analysis of these models based on eight parameters, namely the model used, algorithms used, the data set used for experiments and simulation, detection rate, false detection rate, attacks against which the IDS protects, adaptive/ learning nature of IDS and the distinct feature of the model. The last section is the conclusion.

II. LITERATURE SURVEY

A. Anomaly based Intrusion Detection System:

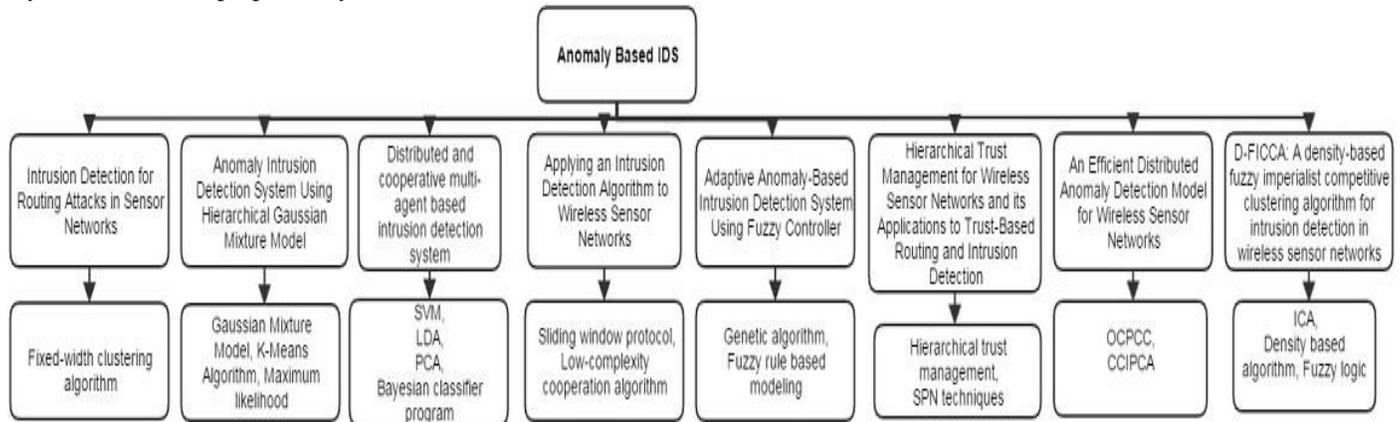


Fig. 1. Anomaly Based IDS

First the anomaly based intrusion detection systems have been discussed in detail. Chong eik loo et.al. [14] has designed an anomaly based IDS that collects information of normal traffic pattern which is then used to detect abnormal traffic patterns. In this technique no information is to be shared between the nodes and every node is equipped with an IDS which works independently without information from neighbouring nodes so as to conserve maximum energy. The anomaly based approach is based on a fixed width clustering algorithm which is used to model the distribution of training points. Using this model, 95% detection rate for a 5% false positive rate was achieved for periodic route attack. For passive sinkhole attack, the detection rate is 70% for a 5% false positive rate. For the active sinkhole (the most effective attack), detection rate is 100% with a 5% false positive rate. But in this method it is assumed that each node has sufficient power and resources so as to perform the computation required for proper functioning of the IDS. An anomaly based model incorporating Hierarchical Gaussian Mixture Model (HGMM) that classifies network attacks based on statistical pre-processing classification has been proposed in [15]. The normal and intrusive behaviours are learnt by Gaussian probability distribution functions and are used to classify observed system activities. The HGMM model proposed has also been compared with six other techniques: Gaussian Mixture, Radial Basis Function, Binary Tree Classifier, SOM, ART and LAMASTAR [34], and the results indicate that the proposed HGMM is able to achieve high accuracy, detection rate and low false positives. A major problem in WSNs is the availability of resources; hence the IDS must be resource efficient. The IDS presented in [16] uses mobile agents to collect data from the system and the classification of normal

behaviour of the nodes is based on a SVM classifier. The mobile agent gathers information from the local agents before allowing the system to send data. Whenever information is sent in the network to any another system, the mobile agent gathers information from the neighbouring node and then calls the SVM to detect if an attack has occurred. If no suspicious behaviour is encountered, the information is then sent on the network.

This type of model is able to stop intrusion in the network level, and promises high levels of detection rate compared to traditional security measures. Another IDS using the information shared between neighbouring nodes is developed in [17], which is based on a simple and resource constrained WSN. This WSN consists of various static sensor nodes which create a statistical model of normal behaviour of their neighbouring nodes. Once this statistical model is created for each node, then the neighbouring nodes analyze the incoming packets on various layers and classifies whether an intrusion has occurred or not. The statistical model of the neighbouring nodes is used to determine a maximum and minimum threshold of the power consumption per packet, so that incoming packets having a receive power less than or greater than the minimum and maximum thresholds respectively, are classified as abnormal packets. The use of the low complexity algorithm improves the detection and containment process. Bao et. al [19] proposes a cluster based hierarchical trust management protocol for wireless sensor networks(WSNs). This IDS based on trust management protocol [35, 36] detects selfish or malicious sensor nodes for intrusion tolerance and can dynamically learn from the past experiences and adapt to the environment. It maintains two levels of trust management:

at the sensor level and other at the cluster head. The false positive and negative probabilities are dependent on the trust threshold and weight of social trust. A variety of methods exist for classification of intrusions, such as statistical techniques, which we have already observed in the two initial papers, data mining methods, etc. A method that is widely used for intrusion detection is based on fuzzy rules, as proposed by [18] which uses fuzzy controller to increase system performance and accuracy based on Adaptive anomaly. Here detection model generator is used for generating a detection model while IDS engine classifies test records and stores them in Buffer which are monitored and reports it to Fuzzy model tuner which updates the confidence prediction ratio. The proposed model gives accuracy of 15% higher than other machine learning methods and static models. Using the fuzzy rules, a density based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks is proposed by [21]. It consists of the imperialist competitive algorithm (ICA) integrated with a density based algorithm and fuzzy logic for optimum

clustering in WSNs. This proposed model increases the accuracy of security attack detection compared with KMICA, Kmean, and DBSCAN. The results demonstrate that the proposed framework achieves higher detection accuracy of 87% and clustering quality 0.99 compared to existing approaches. There have also been innovative algorithms and methods to reduce the energy consumption in WSNs such as the model used by Rassam et. al.[20]. This paper introduces a distributed anomaly detection model based on one class Principal component classifier (OCPC) that uses the candid covariance free incremental principal component analysis (CCIPCA) algorithm so as to detect the intrusions as they occur. The sensor nodes classify every packet as either normal or abnormal according to the threshold specified in global normal model (GNM) that is formed during the training phase of the IDS. Various papers on anomaly based IDS that have been considered in this study are indicated in Fig I along with the respective algorithms used by each author.

B. Signature based Intrusion Detection System:

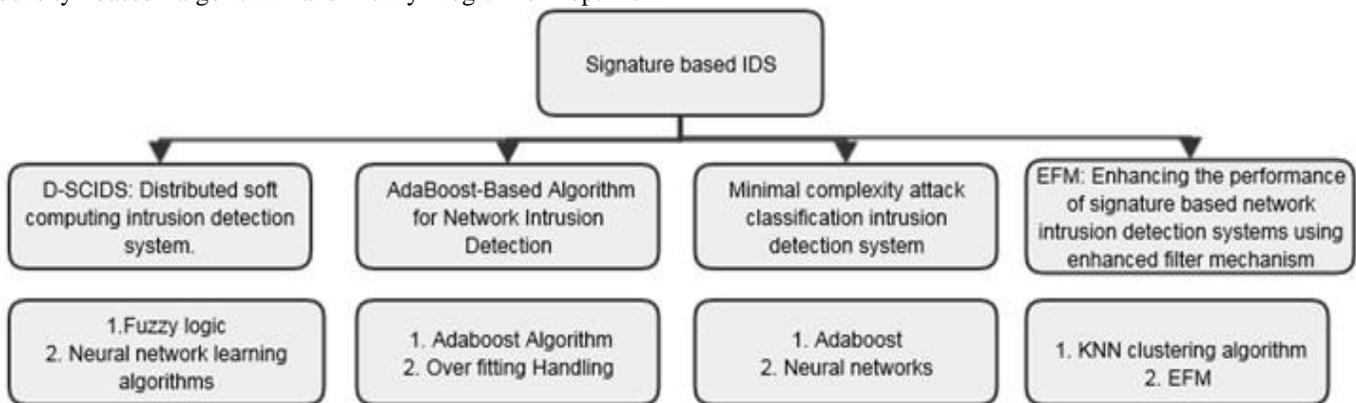


Fig. 2. Signature based IDS

The signature based IDS or misuse based IDS works on various set of rules and compares new information packets with already known signatures to detect intrusions. Abrahama et. al. [22] compared three fuzzy rule based approaches namely: 1:Rule generation based on the histogram of attribute values (FR1) 2:Rule generation based on partition of overlapping areas (FR2) 3:Neural learning of fuzzy rules (FR3). Since none of these approaches were able to single handedly get accurate results for all classes they proposed a new model which is a combination of different classifiers. The proposed heavy weight model was able to get 100% accuracy for all attacks and lightweight was able to get minimum accuracy of 94% for all attacks. A famous algorithm based on signature matching is the Adaboost algorithm and this algorithm has been incorporated in a network based IDS by [23]. The AdaBoost algorithm is a machine learning algorithm which corrects the misclassifications made by weak classifiers, which in this case are decision stumps [37].The decision rules are provided for both continuous and categorical features. Recognition performances of the AdaBoost based classifiers are fast and are generally encouraging. The following algorithm is compared against other algorithms such as SVM, SOM, RSSDSS, etc based on detection rate and false alarm rate. A simple overfitting

handling is used to improve the learning results. But the following adaboost algorithm cannot be applied for incremental learning and does not support offline learning. Using the concept of adaboost and neural network method, an innovative design has been proposed by [24] to lower computational complexity by incorporating rules learnt from the behaviour of the network. The rules have been made according to the data set of KDD99, which is analysed in this case. The proposed IDS has been compared with the adaboost and neural network method. Even though classification by adaboost is better than neural network method, the proposed rule based method provides higher classification rate and lower computational time and also has the capability to learn rules from the behaviour of the network. Statistical methods such as KNN, are being widely used to improve the performance and speed of the signature matching. W. Meng et. al. [25] has used the concept of enhanced filter mechanism (EFM) on a network based IDS which improves the performance of a signature based IDS such as Snort [44] and consists of a context-aware blacklist-based packet filter, an exclusive signature matching component and a KNN-based false alarm filter. The blacklist based packet filter reduces the work of NIDS as it filters out intrusions based on IP address. The signature matching performs the important function of

identifying the intrusion based on signatures and the KNN-based filter is used to reduce the false positives i.e false alarms. The average detection accuracy of this IDS is about 86%, but this is based on the training set, with appropriate training a detection accuracy of over 90% is possible. Also it promises a great reduction in the false alarms. Fig II contains

information about the various signature based IDS and algorithms which have been studied in this survey.

C. Hybrid based Intrusion Detection System:

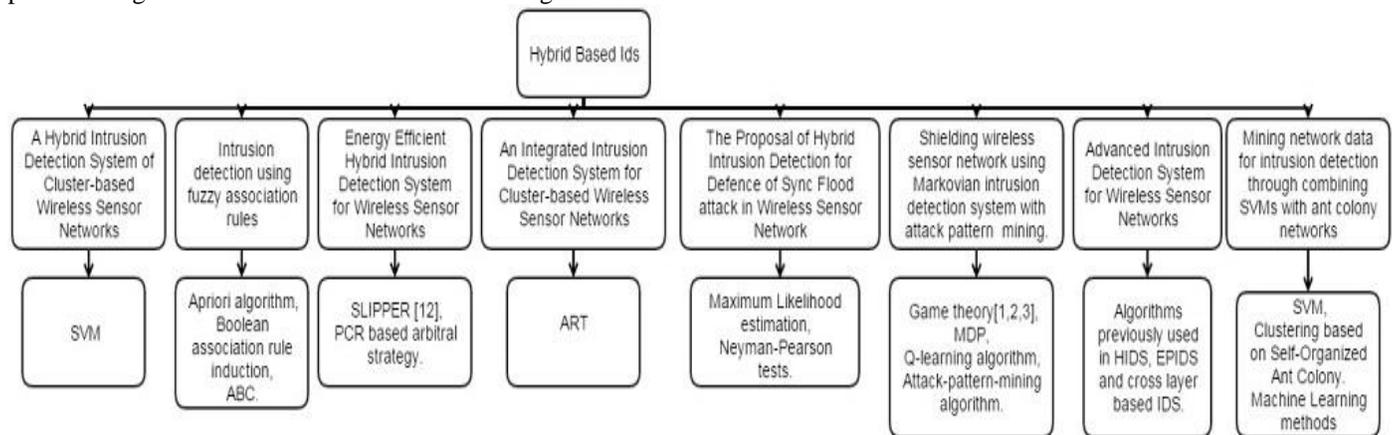


Fig. 3. Hybrid based IDS

The hybrid based IDS is a combination of both the signature based and anomaly based IDS and capitalizes on their advantages and results in a higher detection rate, false alarm, etc. Various data mining techniques like Association Based Classification (ABC) as incorporated by [27] are used to combine the two traditional detection methods. This paper was one of the early attempts that uses fuzzy association rulesets as descriptive models of different classes and combines anomaly based and signature based IDS using Association Based Classification (ABC) technique which is one of the well known approaches of data mining techniques. The fuzzy association rules are utilized to improve the time utilization of data mining technique. The performance of anomaly and misuse based IDS are evaluated separately and the proposed algorithm is shown to have a better performance than the two models independently. By combining the anomaly detection method with misuse detection method, the false positive error rate is very low and it also promises a good detection rate. Attacks on a WSN are usually on the Cluster Head (CH) as it collects data from different sensor nodes in a particular sensor and hence proper protection needs to be provided. K.Q. Yan [26] has proposed a hybrid based IDS for intrusion detection at the CH of a CWSN. The anomaly based model is used as a filter and a signature based IDS is used to detect the intrusion. It additionally consists of a decision making module that decides if an intrusion has occurred. The output of which is given to the administrator for the follow up work. In this model, the training sample must be sufficient to ensure high detection rates. A major difficulty in using Hybrid based models is the high consumption of resources and energy as indicated by Alrajeh [38]. To improve the energy efficiency, [28] has proposed a cluster based WSN (CWSN) so as to reduce communication costs and computational energy. CWSN helps to reduce the energy consumption and

increase the lifetime of the model. The following eHIDS has been compared with HIDS and eHIP. In this scheme, each node of eHIP consumes on an average 2,91J and HIDS consumes 2,58J for the total packet transmission process, whereas eHIDS uses only 1,93J. This model achieves high accuracy, high detection rates, low energy consumption and low computational costs. The intrusive attacks in a network may be unknown to the IDS many a times and using a learning mechanism will help in storing a signature of the particular attack for future prevention, as is the case in the model incorporated by [29]. In this paper a model is proposed which has 3 separate IDS for sink, cluster head (CH) and Sensor node. The model is a cluster based WSN (CWSN). The 3 proposed IDS are: Intelligent Hybrid Intrusion Detection System (IHIDS) for the sink that has learning ability, Hybrid Intrusion Detection System (HIDS) for the cluster Head (CH) and a misuse based IDS for the sensor nodes. The first level of filter is done by anomaly detection and the identified intrusions are sent to misuse detection for further analysis. If the intrusions are not identified by the misuse detection then it is sent to the learning mechanism of IHIDS. IHIDS decreases the energy consumption and also reduces the information efficiently. The proposed IHIDS can achieve a high detection rate and low false positive rate and it also learns about new attacks on the IDS using ART [39]. Based on incorporating the learning mechanism in IDS, [31] has also proposed a new model that uses Markovian IDS to protect sensor nodes from attack. It integrates Anomaly based, Misuse based and game theory to prevent malicious attacks. The Markov decision process is used in the self learning process of the IDS and determines the weakest nodes to be protected. The system is able to reveal the patterns from which it predicts future points of attack and devises appropriate defence strategies, and also has a high detection rate.

TABLE I. ANOMALY BASED IDS

Authors	Algorithms Used	Data Set Source	Adaptive	Detection Accuracy (DA) and False Detection Rate (FDR)	Protection against Attacks	Distinct Features and type of WSN
[14] Loo et al. (2006)	Fixed-width clustering algorithm	1. NRL 2. NS-2	No	DA= 1. 95% for routing attacks. 2. 100% for active sinkhole attack. FDR=5% (FN)	Periodic route error attack, Active sinkhole attacks	Routing protocol is AODV. No information exchange between neighbouring nodes. Ad hoc placement of sensors.
[15] Behroloolur and Khalegi (2008)	1. GMM 2. K-Means Algorithm 3. Maximum likelihood	MIT's Lincoln Lab[42]	No	DA=88.14% FDR=4.70	Probe, Dos, R2L, U2R	Uses statistical preprocessing classification. Classifies based on Gaussian probability distribution functions.
[16] Renjit and Shunmuganathan (2010)	1. SVM. 2. LDA. 3. PCA. 4. BCP	NA	Yes	DA=89%-98%. FDR= 5-9% (False Positive)	NA	Differentiates congestive packet loss from malicious packet loss. Anomaly detection result of neighbouring node is used.
[17] Wang et al. (2009)	1. SWP 2. Low-complexity cooperation algorithm	NA	No	DA > 90% FDR=Decreases with increase in intrusion buffer lengths.	Node Impersonation, Resource Depletion	Checks for anomalous packets from neighbouring nodes. Develops a statistical model of normal behaviour of these nodes.
[18] Abbaspour et al. (2012)	Genetic algorithm, Fuzzy rule based modeling	KDD Cup99	Yes	DA=86.71(TN) FDR= 13.29 (FN) 57.71(FP)	NA	Accuracy 78.6. Online Adaptation. CWSN
[19] Bao et al. (2012)	1.Hierarchical trust management 2. SPN	Self made	Yes	DA>90% when FP approaches zero FDR= Limited to 5%	BH,SH, Slandering attacks, Flooding-Based Routing	Hierarchical trust based IDS based on social trust and QoS trust. Learns from its past experiences and adapts to changes in network. CWSN
[20] Rassam et al. (2013)	1. OCPCC 2. CCIPCA	GSB	No	DA=96% FDR=7.2%	NA	High detection effectiveness. Utilizes network resources efficiently. Distributed online IDS.
[21] Shamshirband et al. (2014)	1. ICA 2. Density based algorithm. 3. Fuzzy logic.	1.IRL[41] 2.ARC[40]	Yes	DA> 87% FDR=15	DoS	Reinforces detection function against incoming DDoS attacks. Continuous self-learning from prior attacks. CWSN.

There have been IDS which are developed for protection against a specific attack, usually used for application specific IDS. One such model based on protection against sync flood attacks has been proposed by [30]. They propose a Hybrid Intrusion Detection System that works on Stream flow and state transition analysis by which the malicious nodes are effectively shut down. The main attack on which the model focusses is Sync Flood attack that establishes a number of TCP connections to use a large amount of resources on the affected nodes. The proposed hybrid detection approach is faster and effective in case of densely deployed sensor network and alarming the base station about the infected or abnormal behavior in the flow of the traffic.

To further improve on the range of attacks against which protection is provided and to enhance the detection rate considerably, Simenthy et.al. [32] proposes a new advanced intrusion detection system that consists of Hybrid Intrusion detection system(HIDS), Energy Prediction based Intrusion Detection System(EPIDS) and cross layer detection system in different stages to ensure maximum security. The Advanced intrusion Detection System has been compared with Energy

Prediction Model, HIDS and Cross Layer Model, and it was analyzed that the proposed model gave better attack detection, less false positives and better detection probability compared to the other 3 models. Also in this system, the energy efficiency and lifetime of the system increases. A recent model that works on the principle of Clustering based on Self Organized Ant Colony Network (CSOACN) and SVM has been proposed by [33] to develop a hybrid based IDS. The SVM is used to find support vectors and to generate hyperplane that separates normal and abnormal data while a CSOACN is used to find data added to active SVM training set and to finally generate models for normal data as well as for each class of abnormal data. An important aspect of this paper is that the processes of training and testing are done parallelly. The detection rate of this model is 94.86%, False positive is 6.01% and False negative is 1.00%. The paper highlights that the proposed CSVAC (Combining Support Vectors with Ant Colony) performs better than SVM and CSOACN applied independently. Hybrid based IDS which have been studied in this survey are depicted in Fig III along with the algorithms used in each study.

This paper attempts to review these three important techniques, namely anomaly, signature and hybrid based IDS. The need of such a research is to provide an insight into the recent developments in the area of intrusion detection and provide details about the different types of IDS required according to varying requirements of the wireless sensor network.

III. COMPARISON

Various papers of anomaly, signature and hybrid have been analyzed in this survey. Certain parameters, such as, algorithms used, detection accuracy, false alarm rate (Both FN and FP), protection against attacks, adaptive/ learning and the distinct feature of each model have been investigated. A number of algorithms are incorporated which can be classified based on three traditional methods, namely statistical methods, machine learning and optimization techniques. Some algorithms have been tailor made for particular applications and have been classified as ad-hoc procedures. The models have also been classified based on whether the IDS is adaptive or not. Adaptive signifies that the proposed model is capable of learning from previous attacks that have already occurred, and hence can detect it the next time it occurs.

The most important aspect being considered is the attacks against which considerable protection is provided by the proposed IDS, as the work of an IDS is to eliminate security threats in the network. We have also touched upon the distinct features in each model and also included any other miscellaneous parameter that may prove useful.

The surveyed anomaly based IDS's indicate that it has a detection rate of >87% largely and can reach a high detection rate of about 95%-96% in certain cases. But the false alarms generated in the IDS are large, i.e about 4-6%. Whereas the false alarms in a signature based IDS are very less, generally

around <1%. The hybrid based IDS ensures a high detection rate of >88%, and also has the advantage of low false alarms.

This indicates that the hybrid based IDS truly provides an improvement in terms of detection rate and false alarm reduction, than using signature and anomaly based IDS independently. A closer look on the various models proposed also suggests that the denial of service (DoS) attack is the most frequently detected intrusion, whereas the probe, U2R and R2L attacks have a lower detection rate. Hence there needs to be an improvement in detection of specifically the probe, R2L and U2R attacks.

A careful study of the comparison tables show that for the case of anomaly based IDS, statistical methods are preferred over the other algorithms. The statistical algorithms are being used in [14-16], [18-21], indicating that they are widely used in applications where a threshold has to be formed for the detection of intrusions in a network. The statistical algorithms used in various IDS include the fixed width clustering algorithm in [14], GMM, K-means and maximum likelihood algorithms in [15], Hierarchical trust management and SPN applied in [19] and OCPCC, CCIPCA incorporated in [20].

In [16], a mixture of both statistical and machine learning algorithms is incorporated that include SVM, LDA, PCA and BCP. The models proposed in [18, 21] incorporate statistical, machine learning and optimization algorithms simultaneously. They include genetic algorithms, fuzzy rule modeling, ICA and density based algorithm. The machine learning algorithm used in [17] includes SWP and low complexity cooperation algorithm. In the hybrid based IDS, a different scenario exists as the machine learning algorithms are the widely preferred methods, which includes ART, Q-learning, SVM, SLIPPER, CSOACN, etc.

TABLE II. SIGNATURE BASED IDS

Authors	Algorithms used	Data Set Source	Adaptive	Detection accuracy (DA) and False detection rate (FDR)	Protection against attacks	Distinct feature and types of WSN
[22] Abrahama et al (2007)	Fuzzy logic, Neural network learning algorithms	DARPA, 1998	Yes	DA= >94.11% FDR=NA	DoS, Probe, U2R, R2L	The detection accuracy can reach about 99.98 for R2L attack. Distributed IDS (DIDS).
[23] Hu et al. (2008)	Adaboost, Over fitting Handling	KDDCup99	Yes	DA= 90.04%-91%. FDR= 0.31%-1.79%	DoS, U2R, R2L, Probe	1. Decision stumps are used as weak classifiers, 2. Simple overfitting handling is used to improve the learning.
[24] Gowrisona et al (2013)	Adaboost [43], Neural networks	KDDCup99	Yes	DA= >99% FDR=0.1%	DoS, Probe, U2R, R2L	Can learn from network behaviour. High detection rate.
[25] Meng et al. (2014)	KNN clustering algorithm, Enhanced filter mechanism	1. DARPA, 1999 [49] 2. Real data set	No	DA= 86% - >90%. FDR= 85% less than snort.	IP Spoofing, Snort, algorithmic complexity attack	3 components: a context-aware blacklist-based packet filter, exclusive signature matching component and a KNN-based false alarm filter. Network based IDS

TABLE III. HYBRID BASED IDS

Authors	Algorithms used	Data Set Source	Adaptive	Detection accuracy (DR) and False detection rate (FDR)	Protection against attacks	Distinct Features and Type of WSN
[26] Yan et al.	SVM	KDDCup99	No	DA=99.81% FDR= 0.57% (FP)	DoS, U2R, R2L, Probe	High Accuracy of 99.75%. CWSN
[27] Tajbakhsh et al. (2009)	1.Apriori algorithm 2.Boolean association rule induction[9]. 3.Association Based Classification(ABC)	KDDCup99	No	DA= 88.5% FDR=6.9% (FP)	DoS, Probe, U2R, R2L.	1. Handling symbolic (categorical) attributes. 2. Efficient classification of large datasets.
[28] Abduvaliyey et al. (2010)	1. SLIPPER [48]. 2. PCR based arbitral strategy.	Self made	No	DA= 96% FDR=0.05%	NA	Low energy consumption: 1.93J/node, Low computational costs. CWSN
[29] Wang et al. (2011)	ART	KDDCup99	Yes	DA=90.96% FDR= 2.06% (FP)	Spoofed/Altered/ Replayed Routing Information, SF, SH, SY, WH, DoS,	1. Three IDS for Sink, CH and SN are proposed. 2. Learning mechanism. 3. Accuracy of 99.75%. CWSN
[30] Bhatnagar and Shankar (2012)	1. MLE. 2. Neyman-Pearson test.	Self made	No	DA= NA FDR= NA	DoS	1. Effective against SYNC flood attack. 2. Detection is faster & effective for densely deployed networks.
[31] Huang et al. (2013)	1. Game theory [45, 46, 47] 2. MDP. 3. Q-learning Algorithm. 4. Attack-pattern-mining algorithm.	Real world	Yes	DA= 1. 96.34% for high regularity attacks. 2. 79.75% for low regularity attacks. FDR= NA	Jamming, Blackhole, Flooding, De-synchronization capture attack	Reveals the patterns to predict future points of attack and devises defence strategies. Hierarchical clustered IDS.
[32] Simenthy et al. (2014)	Algorithms previously used in HIDS, EPIDS and cross layer based IDS.	Self made	Yes	DA >90% FP <0.175%	SF,WH,SY,SH,HF, DoS	Applicable to Small, medium and large sized networks. Integrates 3 types of IDS. CWSN.
[33] Feng et al. (2014)	1. SVM 2. CSOACN	KDDCup99	Yes	DA= 94.86% FDR= 6.01% (FP) 1.00% (FN)	DoS,U2R,R2L, Probe	1. The process of training & testing are done parallelly. 2. Combines both SVM and CSOACN. CWSN.

The proposed models incorporating machine learning methods are [26, 28, 29, 31] and [33]. In papers [26, 33], SVM is used for detecting intrusions and in [33], CSOACN is used along with SVM to provide a dual layer of intrusion detection. Whereas in [29], adaptive resonance theory is used extensively. In [28], both machine learning algorithms such as SLIPPER and optimization algorithms such as PCR based arbitral strategy are incorporated.

A combination of all the 3 techniques is used in [31] which comprises of statistical machine learning and optimization algorithms. In the hybrid based IDS, purely statistical based algorithms such as MLE and Neyman Pearson test are applied by [30]. The model [27] uses an ad-hoc methodology for efficient performance.

From tables [1, 2 and 3] it is clear that the data set used for experimentation is mainly based on KDDCup-99 data set. In the anomaly based models a wide variety of data sets are used. A couple of models [18, 15] are based on KDDCup-99 set, whereas GSB, IRL, ARC NRL data sets have been scarcely

used. The hybrid based IDS which have been reviewed in this paper, have majorly used only KDDCup-99. Four hybrid based models use KDDCup-99 and four hybrid models use the real data samples. On analyzing signature based IDS the KDDCup-99 is found to be the most widely used data set for training the sensor nodes.

A study of the literature reveals that the computation involved generally in an anomaly or signature based IDS is usually lower when compared to a hybrid based IDS. Also the energy consumption is higher in a hybrid based model than the signature or anomaly. But the higher consumption of resources by hybrid based IDS also ensures that the detection rate and protection against the attacks is enhanced and also the false alarms are greatly reduced in comparison to signature or anomaly based models.

This research provides an insight into the various recent developments in intrusion detection systems along with the types of algorithms which have been incorporated. It also provides the various merits and demerits of the models which

have been researched in this area by comparing them in a tabular format.

IV. CONCLUSION

This paper conclusively analyzes signature, anomaly and hybrid based intrusion detection systems. The models which have been proposed by various researchers, roughly in the past decade, have been reviewed on the basis of certain parameters. It indicates that the performance of IDS in detection of the attacks has been increasing consistently with time. There is an improvement in the detection rate, lesser false alarms generated and a considerable increase in the range of attacks being detected. It can be inferred from the analysis that the statistical algorithms are frequently used in anomaly based detection models and the machine learning algorithms are common in the hybrid based IDS. We have also observed that hybrid based models have a higher detection rate and lower false alarms compared to the two traditional methods namely, signature based and anomaly based IDS.

The protection against certain attacks such as R2L and U2R is usually low, and can be due to the skewed training data sets used, which contain fairly low number of data sets belonging to these attacks. Hence such attacks pose security concerns in some of the intrusion detection models. The presented information constitutes an important point for addressing future Research & Development in the field of IDS. As this paper essentially focuses on the traditional methods such as anomaly and signature based IDS, future work could include analysis of models based on cross layer or stack based IDS technologies. Techniques providing higher detection rate but utilising fewer resources are required so as to enhance WSNs. Countermeasures which are faster and more effective are needed to cope up with the ever-growing attacks to improve the protection of the networks under surveillance.

TABLE IV. ABBREVIATIONS

Name	Abbreviation	Name	Abbreviation
Cluster based WSN	CWSN	Adaptive Resistance Theory	ART
perialist competitive algorithm	ICA	Markov Decision Process	MDP
Intel Research Laboratories	IRL	Energy prediction based IDS	EPIDS
The Australian Research Council's research network	ARC	Clustering based on Self-Organised Ant Colony Network	CSOACN
Denial of Service	DoS	Maximum Likelihood Estimation	MLE
Stochastic Petri Net	SPN	Prediction Confidence Ratio	PCR
Support Vector Mechanism	SVM	Distributed Denial of Service	DDoS
Ad-hoc on demand distance vector	AODV	Black Hole	BH
Network Simulator-2	NS-2	Selective Forwarding Attack	SF
Naval Research Laboratories	NRL	Sink Hole Attack	SH
False Negative	FN	Sybil Attack	SY
False Positive	FP	Worm Hole Attack	WH
Gaussian Mixture Model	GMM	Hello Flood Attack	HF
Sliding window protocol	SWP	Linear discriminant analysis	LDA
Bayesian classifier program	BCP	Principal component analysis	PCA

REFERENCES

- [1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393-422, 2002.
- [2] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey", *Computer Networks*, Vol. 52, Issue 12, pp. 2292-2330, August 2008.
- [3] Muhammad Ayaz, Imran Baig, Azween Abdullah, Ibrahim Faye," A survey on routing techniques in underwater wireless sensor networks", *Journal of Network and Computer Applications*, Volume 34 Issue 6, pp.1908-1927,2011.
- [4] Jelena Mistic, Fereshteh Amini, Moazzam Khan,"Signature-based intrusion detection in healthcare wireless sensor networks implemented over IEEE 802.15.4 beacon enabled clusters".
- [5] Nabil Ali Alrajeh, S. Khan, Bilal Shams," Intrusion Detection Systems in Wireless Sensor Networks: A Review", *International Journal of Distributed Sensor Networks*, Volume 2013 (2013), Article ID 167575, 7 pages.
- [6] Robert Mitchell, Ing-Ray Chen," A survey of intrusion detection in wireless network applications", Elsevier, *Computer Communications* 42, pp.1-23,2014.
- [7] S Jha, M Hassan," Building agents for rule-based intrusion detection system", Elsevier, *Computer Communications*, Volume 25, Issue 15, 2002, pp.1366-1373.
- [8] T.S. Sobh, "Wired and wireless intrusion detection system: Classifications, good characteristics and state-of-the-art", Elsevier, *J. Computer Standards and Interfaces*, volume 28, number 6, pp.670-694, 2006.
- [9] C. Borgelt, 2005, *Association Rule Induction*, Available: <http://fuzzy.cs.uni-magdeburg.de/borgelt>
- [10] Miao Xie, Song Han, Biming Tian, Sazia Parvin," Anomaly detection in wireless sensor networks: A survey", *Journal of Network and Computer Applications*, Volume 34, Issue 4, July 2011, pp. 1302-1325.
- [11] Mohammad Wazid," Hybrid Anomaly Detection using K-Means Clustering in Wireless Sensor Networks".
- [12] Sedjelmaci, Hichem, Feham, Mohamed," Novel Hybrid Intrusion Detection System For Clustered Wireless Sensor Network", *Academic*

- Journal, International Journal of Network Security & Its Applications; Jul2011, Vol. 3 Issue 4, p1.
- [13] Djallel Eddine Boubiche and Azeddine Bilami, "Cross Layer Intrusion Detection System For Wireless Sensor Network", International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012, pp.35-52.
- [14] Chong Eik Loo And Mun Yong Ng, Christopher Leckie, Marimuthu Palaniswami, "Intrusion Detection for Routing Attacks in Sensor Networks", International Journal of Distributed Sensor Networks, 2006, pp.313-332.
- [15] M. Bahrololulom and M. Khaleghi, "Anomaly Intrusion Detection System Using Hierarchical Gaussian Mixture Model", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008, pp.264-271.
- [16] J. Arokia Renjit and K. L. Shunmuganathan, "Distributed and cooperative multi-agent based intrusion detection system", Vol.3 No.10, Indian Journal of Science and Technology, 2010, ISSN: 0974- 6846, pp. 1070-1074.
- [17] QiWang, ShuWang, ZhonglouMeng, "Applying an Intrusion Detection Algorithm to Wireless Sensor Networks", IEEE, Second International Workshop on Knowledge Discovery and Data Mining, 978-0-7695-3543-2/09 \$25.00,2009, pp.284-287.
- [18] Farzaneh Geramiraz, Amir Saman Memaripour, and Maghsoud Abbaspour (Corresponding author: Maghsoud Abbaspour), "Adaptive Anomaly-Based Intrusion Detection System Using Fuzzy Controller", International Journal of Network Security, Vol.14, No.6, Nov. 2012, pp.352-361.
- [19] Fenyao Bao, Ing-Ray Chen, Moon Jeong Chang, and Jin-Hee Cho, "Hierarchical Trust Management for Wireless Sensor Networks and its Applications to Trust-Based Routing and Intrusion Detection", IEEE Transaction on network and service management, Vol. 9, No. 2, June 2012, pp.169-183.
- [20] Murad A. Rassam, Anazida Zainala, Mohd Aizaini Maarof, "An Efficient Distributed Anomaly Detection Model for Wireless Sensor Networks", Elsevier, AASRI Procedia 5, 2013, pp.9 – 14.
- [21] Shahaboddin Shamshirband, Amineh Amini, Nor Badrul Anuar, Miss Laiha Mat Kiah, Ying Wah Teh, Steven Furnell, "D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks", Elsevier, Measurement 55,2014, pp.212–226.
- [22] Ajith Abraham, Ravi Jain, Johnson Thomas, Sang Yong Hana, "D-SCIDS: Distributed soft computing intrusion detection system", Elsevier, Journal of Network and Computer Applications 30, 2007, pp.81–98.
- [23] Weiming Hu, Wei Hu and Steve Maybank, "AdaBoost-Based Algorithm for Network Intrusion Detection", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 38, NO. 2, APRIL 2008, 1083-4419/\$25.00, pp.577-583.
- [24] G. Gowrisona, K. Ramar, K. Muneeswaran, T. Revathi, "Minimal complexity attack classification intrusion detection system", Elsevier, Applied Soft Computing 13, 2013, pp.921–927.
- [25] Weizhi Meng, Wenjuan Li, Lam-For Kwok, "EFM: Enhancing the performance of signature-based network intrusion detection systems using enhanced filter mechanism", Elsevier, computers & security 43,2014, pp.189-204.
- [26] K.Q. Yan, S.C. Wang, C.W. Liu, "A Hybrid Intrusion Detection System of Cluster-based Wireless Sensor Networks", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I, Hong Kong. ISBN: 978-988-17012-2-0
- [27] Arman Tajbakhsh, Mohammad Rahmati, Abdolreza Mirzaei, "Intrusion detection using fuzzy association rules", Elsevier, Applied Soft Computing 9,2009 , 1568-4946/\$,pp.462–469.
- [28] Abror Abduvaliyev, Sungyoung Lee, Young-Koo Lee, "Energy Efficient Hybrid Intrusion Detection System for Wireless Sensor Networks", 2010 International Conference on Electronics and Information Engineering (ICEIE 2010).
- [29] Shun-Sheng Wang, Kuo-Qin Yan, Shu-Ching Wang, Chia-Wei Liu, "An Integrated Intrusion Detection System for Cluster-based Wireless Sensor Networks", Elsevier, Expert Systems with Applications 38,2011, pp.15234–15243.
- [30] Ruchi Bhatnagar and Udai Shankar, "The Proposal of Hybrid Intrusion Detection For Defence Of Sync Flood Attack In Wireless Sensor Network", International Journal of Computer Science & Engineering Survey (IJCSSES) Vol.3, No.2, April 2012.
- [31] Jen-Yan Huang, I-En Liao, Yu-Fang Chung, Kuen-Tzung Chen, "Shielding wireless sensor network using Markovian intrusion detection system with attack pattern mining", Elsevier, Information Sciences 231 ,2013, pp.32–44.
- [32] Joseph Rish Simenthy CEng ,AMIE, K. Vijayan, "Advanced Intrusion Detection System for Wireless Sensor Networks", Vol. 3, Special Issue 3, International Conference on Signal Processing, Embedded System and Communication Technologies and their applications for Sustainable and Renewable Energy (ICSECSRE '14), April 2014.
- [33] Wenying Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiangji Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks", Elsevier, Future Generation Computer Systems 37,2014, pp. 127–140.
- [34] V. Venkatachalam, S. Selvan, "Intrusion Detection using Improved Competitive Learning Lamstar Neural Network", IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.2, February 2007.
- [35] J. H. Cho, A. Swami, and I. R. Chen, "A survey on trust management for mobile ad hoc networks," IEEE Commun. Surveys Tutorials, vol. 13, no. 4, pp. 562–583, 2011.
- [36] E. M. Daly and M. Haahr, "Social network analysis for information flow in disconnected delay-tolerant MANETs," IEEE Trans. Mobile Computing, vol. 8, no. 5, pp. 606–621, May 2009.
- [37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. Int. Conf. Comput. Vis. Pattern Recog., 2001, vol. 1, pp. I-511–I-518.
- [38] Nabil Ali Alrajeh, S. Khan, Bilal Shams, "Intrusion Detection Systems in Wireless Sensor Networks: A Review".
- [39] Carpenter, G. A., & Grossberg, S., "The ART of adaptive pattern recognition by a self-organizing neural network". IEEE, Computer 21(3), 1988, pp.77–88.
- [40] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks", Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2010 Sixth International Conference on 2010, pp. 269–274.
- [41] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, S. Madden, Intel lab data,
- [42] Lincoln Laboratory, Massachusetts Institute of Technology (MIT), 1998-2000. DARPA Intrusion Detection Evaluation.
- [43] Y. Freund, R.E. Schapire, "A short introduction to boosting", Journal of Japanese Society for Artificial Intelligence 14 (5), 1999, pp. 771–780.
- [44] Snort, The Open Source Network Intrusion Detection System. <http://www.snort.org/>.
- [45] A. Agah, S.K. Das, K. Basu, "A Noncooperative Game Approach for Intrusion Detection in Sensor Networks (VTC 2004)", 2004, pp. 2902–2906.
- [46] A. Agah, S.K. Das, K. Basu, "Intrusion detection in sensor networks: A noncooperative game approach", 3rd IEEE International Symposium on Network Computing and Applications (IEEE NCA04), 2004, pp. 1–4.
- [47] A. Agah, M. Asadi, S.K. Das, "Prevention of DoS attacks in sensor networks using repeated game theory", Proceedings of the International Conference on Wireless Networks, 2006.
- [48] W. Cohen and Y. Singer, "A Simple, Fast, and Effective Rule Learner", Proceedings of 6th national Conference on Artificial Intelligence and 11th Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida, pp.335342, July 1999.
- [49] DARPA, 1999; McHugh, 2000 was produced by MIT Lincoln Laboratory and Air Force Research Laboratory.

A Survey on the Internet of Things Software Architecture

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

¹Faculty of Electrical Engineering and Computer Science, ²Integrated Center 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—The Internet of Things (IoT) is a concept and a paradigm that considers the pervasive presence in the environment of a variety of things/objects through wired or wireless that are uniquely addressed and are able to interact with each other and cooperate with other things/objects in order to create new applications/services and to achieve common objectives. IoT defines a new world where the real, the digital and the virtual converge to create an environment that makes the energy, transport, city, and many other areas to become more intelligent. The IoT purposed is to validate the connection type: anytime, anywhere, and everything and everyone. IoT may be considered as a network of physical objects with embedded communication technologies that 'feel' or interact with internal or external environment. This paper presents a survey on the Internet of Things software architectures that meets the requirements listed above.

Keywords—middleware; Internet of Things; things; software architecture

I. INTRODUCTION

The Internet of Things is a paradigm that is included in the Internet of the Future. According to the International Telecommunication Union (ITU) [1], the Internet of Things will connect the world's objects, both in sensory and intelligent way. The ITU proposed an Internet of Things ecosystem that included all things from everyday live.

The Ecosystem proposed by ITU [1] can be represented according to the Fig. 1 [2]. The scanners are used to identify the things (by labels or RFID tags). These scanners can transmit the locations of the things to the others systems (upper layer). Middleware systems and development tools can be used to design applications and services that use the information from the things. This information can be stored in the cloud and can be accessed through the Internet providing greater flexibility of the services.

Currently, there is no definition for the Internet of things [3] accepted by the scientists. Because the terms Internet of Things is widely and increasingly used, in the specialized literature can be found several definitions of the IoT. A definition of the IoT is the following [4]: "global network of interconnected objects that are unique addressable based on the standard communication protocols." Another definition is provided by Atzori et al [5] that included the services provided by the things with virtual identity and the capability to communicate in the virtual environment. Other definitions and models can be found in [6]-[10].

The Internet of things includes the existing technologies such as Machine-to-Machine (M2M) [11], [12], wireless sensor networks (WSN), RFID, embedded systems, etc. The challenges of the Internet of Things are [6]-[10]: data confidentiality and encryption, security, safety, information privacy, standardization, naming, and identity.

The paper is organized as follows: Section II presents the IoT architectures presented in the specialized literature, and the conclusions are drawn in section III.

II. THE IOT ARCHITECTURES

This section will be an overview of the variously proposed architectures for IoT. Fig. 2 presents an IoT model which can have up to five layers and different names of the layers.

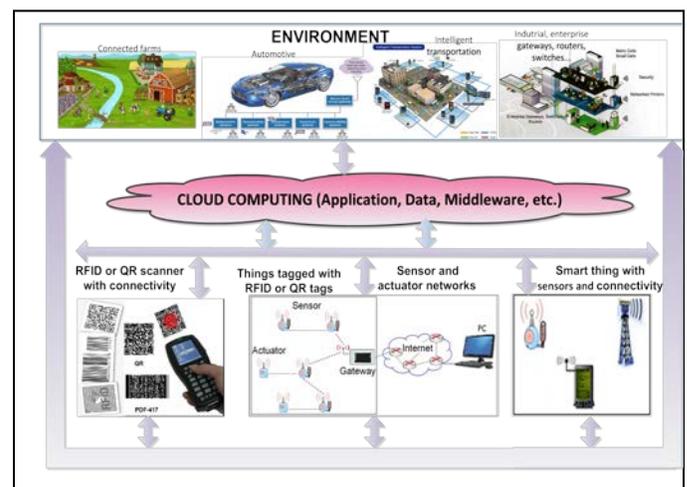


Fig. 1. The Internet of Things Ecosystem [1]

The perception layer (Layer 1) [14] represents the sense organs of the IoT and deals, mainly, with objects identification and data collection. The perception layer includes 2-D barcodes tag readers, RFID tags and appropriate readers/inscriptions, cameras, GPSs, sensors, terminals, sensor networks, etc. Its main task is to identify the object and collect information. In [15], Layer 1 is called the sensing layer and is similar to the perception layer in [14] but it is proposed as an innovative fusion between RFID and wireless sensor networks (WSN) called EPC sensor networks. It has the same meaning and name as the sensor layer in [16], indicating that it defines an additional base station. Another name for Layer 1 is given

in [19], namely the device layer with the two basic elements: gateway and device.

The next layer is Layer 2 called transport layer in [14]. Its main function is the transmission and processing of end-to-end information in a reliable or unreliable way. Another name given to this layer [14] is the network layer. The authors define this layer as a neural network that represents the brain of the IoT and includes a network convergence for communication, the Internet, a network management center, an information center and an intelligent processing, etc. Layer 2 is called the core layer in [16] which mainly includes the network access and the Internet. Another name of this layer can be found in [17] as the gateway layer. This layer establishes a communication channel for heterogeneous sensors and RFIDs, which is the next layer, namely the middleware.

The next layer, Layer 3, has been called the processing layer in [14] and, mainly, it stores, analyses and processes the information related to items received from Layer 2. In [17], Layer 3 is called the middleware layer. This is the layer where the IoT systems run. In order to modularize the physical objects, a proxy can map the messages of the objects to their logical components from the middleware.

Software components in execution are virtual representations of services and physical objects. The proxy is connected to servers for applications, ontology, lookup, database and management. A very similar model to the one proposed in [14] is proposed in [18]. This has all five layers, but the process layer is changed with the middleware layer.

Another name for Layer 3 is given in [19] where this layer is called the service support and application support layer that provides generic capabilities for all IoT applications (e.g. processing and data storage) and capabilities specific to various applications.

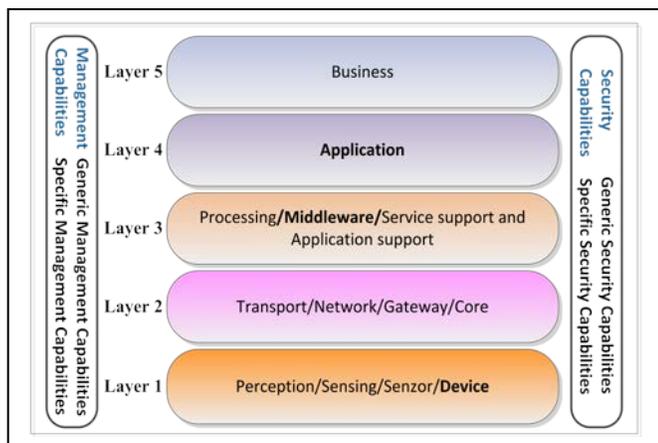


Fig. 2. Layered architecture of IoT [14]

Layer 4, the application layer, contains IoT applications [19]. Layer 5 is the business layer. As it is well known, the success of a technology depends not only on the priority of technology, but also on reasonable innovation in the business model [14]. The models presented so far are a little vague and do not yet provide a complete standardization (concrete implementations are not shown). The architecture presented in

Fig. 2 can be a starting point for the standardization process. Below, we make a brief presentation of other approaches related to the IoT architecture.

In [20], the authors highlight the need for a transparent and standardized end-to-end architecture in order to replace proprietary approaches. Thus, for the physical layer, they opted for IEEE 802.15.4-2006 PHY. From the MAC perspective, for the MAC layer, the IEEE 802.15.4 was used.

The MAC protocol of this new family is tailored for multihop/mesh industrial applications under extreme interference and attenuation (fading). From the network perspective, the introduction of the IETF 6LoWPAN family of protocols has an essential role in connecting low-power radio devices to the Internet and the working group from IETF ROLL introduced appropriate routing protocols to achieve universal connectivity.

Indeed, the two working groups worked for IPv6 connectivity, which is a great advantage in ensuring worldwide accessibility, true scalability and reliable security. From an application perspective, the introduction of the CoAP IETF protocols family had an essential role in ensuring that the application layer and the applications themselves must be redesigned to run on networks with low power consumption. A similar architecture to that described in [20] (that is based on WSN - Wireless Sensor Network, CoAP and REST) is found in [21]. Here, the authors implement and evaluate the model, which includes Linux, Contiki, as well as Linux service to integrate with the Hadoop HBAs data store.

In the vision of IoT-A project [22], the Internet of Things is based on the fact that the interoperability of the solutions for both the communication and the services must be provided on various platforms. This justifies, firstly, the creation of a reference model for the IoT domain, in order to promote a common approach. Secondly, companies that want to create their own IoT compatible solutions must be sustained by a reference architecture that describes the essential constituents and the choices related to designing support in order to meet contradictory requirements in terms of functionality, performance and security implementation. The central choice for the IoT-A project was to base the work on actual "state of art" techniques, rather than on the use of new technologies. In [23] the authors provide a brief description throughout the state of art in IoT, with a special focus on the concepts and technologies related to mobility, communication and wireless networking. The authors concluded that although the concept of IoT has some years, there are still many technical problems that were not solved such as heterogeneity, scalability, security, connectivity, energy, management, naming and identification. The complexity of these issues, especially concerning the nature of resource constraints in most IoT components and the use of wireless communication requires a unified architectural vision able to be solved in a consistent manner. In the article, the authors describe briefly a recent case study of architecture and protocol suite that were used to implement the IoT. In their vision, the basic elements of the architecture are the Wireless Sensors Networks (WSN) and 6LoWPAN [24] used to connect to the Internet by IPv6 that has enough Internet addresses and web services. The article refers to the model proposed in the

IoT-A project, but it is not supported by substantial discussion in relation to this model.

An effective IoT implementation used to monitor normal domestic conditions through a pervasive (ubiquitous) system of low-cost sensors is presented in [25]. The proposed model is based on a wireless sensor network based on the ZigBee protocol. End devices collect and send data on a ZigBee coordinator, after which the specific data of the ZigBee protocol format are translated for the Internet IPv6 protocol implemented on a gateway layer. In conclusion, there could be highlighted three layers, namely: smart metering devices, IoT gateways and Internet servers. The domestic application seems to be exciting, but it is interoperable with other IoT models. It looks more like a silage model type or an Intranet of Things.

A more sustained architecture with implementation and a practice test is presented in [26]. In this article, the authors present a new architecture called Sensor Networks for an All-IP World (SNAIL). This architecture includes four major technologies - mobility, web, time synchronization and security in architecture for adaptation to IP. Afterwards, the authors describe how they have verified the feasibility and interoperability of the architecture by implementing a SNAIL platform and testing it on a Korea Advanced Research Network (Koren) national model. The model is more complete but the research continues.

An interesting new concept was introduced in [27]. It was called the Social Internet of Things (SIoT) and it is based on a type of relationships between objects, similar to the relationships between human beings. The authors analyzed statistically the SIoT network structure through simulations that modeled the mobility of objects and the relations between them. Preliminary results have shown that most SIoT features are these observed in social networks of people. Based on the results of these analyses, the authors investigated whether the navigability can be reached in SIoT and identified techniques in setting the social networking that can improve the navigability. The proposed model has three layers: the base layer that contains a database for storing and managing data with relevant descriptors, a database of ontology and engines for semantics and communications. Another approach of the SIoT is presented in [28] and an original way to approach the future IoT is presented in [29] and [30]. The authors introduce two aspects: Unit IoT and Ubiquitous IoT. Unit IoT refers to the basic IoT unit that focuses on providing solutions for special applications and the architecture is the man-like nervous (MLN) model type. At the same time, their vision of the future Internet and especially the global IoT is about ubiquity in the sense that "everything must be connected, intelligently controlled and covered from everywhere." The model was called Ubiquitous IoT which refers to the global, national, industrial or local IoT and represents the integration of multiple IoT units (Unit IoT) with a "ubiquitous" character. The Ubiquitous IoT architecture looks like the social organization framework (SOF) model.

The architecture proposed in [31] starts from the open EPCglobal Network architecture. The authors emphasize that there are many approaches regarding the IoT; they claim to be followers of the architectural approach based on the EPCglobal

Network. However, the IoT requires a more holistic architecture. It can be built on design principles such as the EPCglobal Architecture Framework [32].

An interesting architecture that is based on the EPCglobal architecture and the IEEE 1451 is presented in [33]. Both are integrated into the IoT architecture framework and the EPCglobal and IEEE 1451 standard framework, in order to form an open environment. It simulated a scenario after which, finally, the authors conclude that the proposed IoT is feasible.

The industrial environment is made explicit in [34]. It proposed an architecture called IoT@Work whose main component is ENS (Event Notification Server) which aims to collect, organize and deliver, in a controlled way, the production data from the shop floor. The ENS middleware provides a communication model based on events like publish/subscribe communication to support templates such as one-to-many and many-to-many and the dynamic coupling of the services, processes and devices. The model uses the AMQP (Advanced Message Queuing Protocol) protocol and architecture [35].

In order to address to the specific challenges of the IoT, in [36] both the VIRTUS architecture (as an event-driven middleware built on existing standards such as XMPP and OSGi) and security issues are discussed. The VIRTUS architecture is a middleware solution for the management of IoT applications. Using the paradigm of "publish&subscribe" and XMPP native security facilities, VIRTUS simplifies the IoT application development.

III. CONCLUSIONS

In this paper, we were presented the main IoT architecture presented in the literature. From these architectures can see that the most include a middleware level to distribute the data in the Internet. However, at this time there is not a middleware standard that is accepted by all in the deployment of IoT systems. In fact, the most IoT architectures include existing technologies that are used in order to meet the requirements for the IoT systems in terms of the interaction of things via the Internet. Furthermore, the majority of the IoT architectures are organized on five layers, according with Fig. 2.

ACKNOWLEDGMENT

This paper was supported by the project "Increasing the competitiveness of the EURONEST ICT&Hub Regional Innovation Cluster and stimulating interactions between members to develop high tech products and services" - Contract no.: 1CLT/800.020/19.05.2014, project co-funded from European Social Fund through Sectorial Operational Program Increase of Economic Competitiveness 2007-2013.

REFERENCES

- [1] International Telecommunications Union, ITU Internet Reports 2005: The Internet of Things. Executive Summary, Geneva: ITU, 2005.
- [2] Louis COETZEE, Johan EKSTEEN, The Internet of Things – Promise for the Future? An Introduction, IST-Africa 2011 Conference Proceedings Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, 2011 ISBN: 978-1-905824-26-7.

- [3] Y. Huang and G. Li. Descriptive Models for Internet of Things. In Proc. of Int. Conf. on Intelligent Control and Information Processing (ICICIP), Dalian, China, Aug. 2010.
- [4] INFSO D.4 Networked Enterprise RFID INFSO G.2 Micro Nanosystems in Co-operation with the Working Group RFID of the ETP EPOSS. Internet of Things in 2020, Roadmap for the Future, Version 1.1. Technical report, 27 May 2008.
- [5] L. Atzoria, A. Ierab, and G. Morabito. The Internet of Things: A survey. *Computer Networks*, 54(15):2787–2805, Oct. 2010.
- [6] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi. From Today's INTRANet of Things to a Future INTERNet of Things: A Wireless- and Mobility-Related View. *IEEE Wireless Communications*, 17(6):44 – 51, Dec. 2010.
- [7] E. Fleisch. What is the Internet of Things? - An Economic Perspective. Auto-ID Labs, 2010.
- [8] European Research Cluster on Internet of Things (IERC). Internet of Things - Pan European Research and Innovation Vision. IERC, Available online: <http://www.internet-of-thingsresearch.eu/documents.htm>, Oct. 2011.
- [9] L. Mainetti, L. Patrono, and A. Vilei. Evolution of Wireless Sensor Networks towards the Internet of Things: A survey. In Proc. of 19th Int. Conf. on Software, Telecommunications and Computer Networks (SoftCOM), Split, Dubrovnik, Sept. 2011.
- [10] O. Hersent, D. Boswarthick, and O. Elloumi. The Internet of Things: Key Applications and Protocols. Wiley, 2012.
- [11] G. Lawton. Machine-to-Machine Technology Gears up for growth. *Computer*, 37(9):12 – 15, 2004.
- [12] ETSI TS 102 689 v1.1.1. Machine-to-Machine communications (M2M): M2M service requirements, Aug. 2010.
- [13] Li, S.; Xu, L.; Wang, X., "Compressed Sensing Signal and Data Acquisition in Wireless Sensor Networks and Internet of Things," *Industrial Informatics, IEEE Transactions on* / , vol.PP, no.99, pp.1,1, doi: 10.1109/II.2012.2189222
- [14] Miao Wu; Ting-Jie Lu; Fei-Yang Ling; Jing Sun; Hui-Ying Du, "Research on the architecture of Internet of Things," *Advanced Computer Theory and Engineering (ICACTE)*, 2010 3rd International Conference on / , vol.5, no., pp.V5-484,V5-487, 20-22 Aug. 2010, doi: 10.1109/ICACTE.2010.5579493.
- [15] Handong Zhang; Lin Zhu, "Internet of Things: Key technology, architecture and challenging problems," *Computer Science and Automation Engineering (CSAE)*, 2011 IEEE International Conference on / , vol.4, no., pp.507-512, 10-12 June 2011, doi: 10.1109/CSAE.2011.5952899.
- [16] Jing Pei Wang, Sun Bin, Yang Yu, Xin Xin Niu, Distributed Trust Management Mechanism for the Internet of Things, *Applied Mechanics and Materials (Volumes 347 - 350)*, Instruments, Measurement, Electronics and Information Engineering, pp. 2463-2467, doi: 10.4028/www.scientific.net/AMM.347-350.2463.
- [17] Wei Wang; Lee, K.; Murray, D., "Building a generic architecture for the Internet of Things," *Intelligent Sensors, Sensor Networks and Information Processing*, 2013 IEEE Eighth International Conference on / , vol., no., pp.333,338, 2-5 April 2013, doi: 10.1109/ISSNIP.2013.6529812.
- [18] Khan, R.; Khan, S.U.; Zaheer, R.; Khan, S., "Future Internet: The Internet of Things Architecture, Possible Applications and Key Challenges," *Frontiers of Information Technology (FIT)*, 2012 10th International Conference on / , vol., no., pp.257,260, 17-19 Dec. 2012, doi: 10.1109/FIT.2012.53.K. Ashton, "Internet of Things," *RFID Journal*, June 22 2009.
- [19] International Telecommunications Union, ITU-T Y.2060, Overview of the Internet of things, 2012.
- [20] 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.
- [21] Tracey, D.; Sreenan, C., "A Holistic Architecture for the Internet of Things, Sensing Services and Big Data," *Cluster, Cloud and Grid Computing (CCGrid)*, 2013 13th IEEE/ACM International Symposium on / , vol., no., pp.546,553, 13-16 May 2013. doi: 10.1109/CCGrid.2013.100.
- [22] http://www.iot-a.eu/public/public-documents/copy_of_d1.2, Introduction to Architectural Reference Model for the Internet of Things.
- [23] Zorzi, M.; Gluhak, A.; Lange, S.; Bassi, A., "From today's INTRANet of things to a future INTERNet of things: a wireless- and mobility-related view," *Wireless Communications, IEEE* / , vol.17, no.6, pp.44,51, December 2010, doi: 10.1109/MWC.2010.5675777.
- [24] Z. Shelby and C. Borman, *6LoWPAN: The Wireless Embedded Internet*, Wiley, 2009, ISBN: 978-0-470-74799-5.
- [25] Kelly, S.D.T.; Suryadevara, N.K.; Mukhopadhyay, S.C., "Towards the Implementation of IoT for Environmental Condition Monitoring in Homes," *Sensors Journal, IEEE* / , vol.13, no.10, pp.3846,3853, Oct. 2013, doi: 10.1109/JSEN.2013.2263379.
- [26] Sungmin Hong; Daeyoung Kim; Minkeun Ha; Sungho Bae; Sang Jun Park; Wooyoung Jung; Jae-Eon Kim, "SNAIL: an IP-based wireless sensor network approach to the internet of things," *Wireless Communications, IEEE* / , vol.17, no.6, pp.34,42, December 2010 doi: 10.1109/MWC.2010.5675776
- [27] Atzori, L.; Iera, A.; Morabito, G., "IIoT: Giving a Social Structure to the Internet of Things," *Communications Letters, IEEE* / , vol.15, no.11, pp.1193,1195, November 2011, doi: 10.1109/LCOMM.2011.090911.111340.
- [28] Turcu, C.; Turcu, C., "The Social Internet of Things and the RFID-based robots," *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2012 4th International Congress on / , vol., no., pp.77,83, 3-5 Oct. 2012
- [29] Huansheng Ning; Ziou Wang, "Future Internet of Things Architecture: Like Mankind Neural System or Social Organization Framework?," *Communications Letters, IEEE* / , vol.15, no.4, pp.461,463, April 2011, doi: 10.1109/LCOMM.2011.022411.110120
- [30] Huansheng Ning; Hong Liu; Yang, L.T., "Cyberentity Security in the Internet of Things," *Computer* / , vol.46, no.4, pp.46,53, April 2013, doi: 10.1109/MC.2013.74.
- [31] Dieter Uckelmann, Mark Harrison, Florian Michahelles, book chapter - An Architectural Approach Towards the Future Internet of Things, *Architecting the Internet of Things*, pp 1-24, ISBN 978-3-642-19156-5, Springer 2011.
- [32] EPCglobal (2007) The EPCglobal Architecture Framework, Standard Specification. www.epcglobalinc.org/standards/architecture/architecture_1_2-framework-20070910.pdf.
- [33] Chao-Wen Tseng; Chih-Ming Chang; Chua-Huang Huang, "Complex sensing event process of IoT application based on epcglobal architecture and IEEE 1451," *Internet of Things (IOT)*, 2012 3rd International Conference on / , vol., no., pp.92,98, 24-26 Oct. 2012, doi: 10.1109/IOT.2012.6402309.
- [34] Gusmeroli, S.; Piccione, S.; Rotondi, D., "IoT@Work automation middleware system design and architecture," *Emerging Technologies & Factory Automation (ETFA)*, 2012 IEEE 17th Conference on / , vol., no., pp.1,8, 17-21 Sept. 2012, doi: 10.1109/ETFA.2012.6489652.
- [35] AMQP Working Group, "Advanced Message Queuing Protocol – Specification". Available: <http://www.amqp.org/confluence/download/attachments/720900/amqp.pdf?version=1&modificationDate=1318011006000>
- [36] Conzon, D.; Bolognesi, T.; Brizzi, P.; Lotito, A.; Tomasi, R.; Spirito, M.A., "The VIRTUS Middleware: An XMPP Based Architecture for Secure IoT Communications," *Computer Communications and Networks (ICCCN)*, 2012 21st International Conference on / , vol., no., pp.1,6, July 30 2012-Aug. 2 2012, doi: 10.1109/ICCCN.2012.6289309.

A Carrier Signal Approach for Intermittent Fault Detection and Health Monitoring for Electronics Interconnections System

Syed Wakil Ahmad

EPSRC Centre for Innovative Manufacturing in Through-life Engineering Services, Cranfield University, MK 43 0AL, UK

Dr. Suresh Perinpanayagam

Integrated Vehicle Health Management Centre, Cranfield University, MK43 0QF, UK

Prof. Ian Jennions

Integrated Vehicle Health Management Centre, Cranfield University, MK43 0QF, UK

Dr. Mohammad Samie

Integrated Vehicle Health Management Centre, Cranfield University, MK43 0QF, UK

Abstract—Intermittent faults are completely missed out by traditional monitoring and detection techniques due to non-stationary nature of signals. These are the incipient events of a precursor of permanent faults to come. Intermittent faults in electrical interconnection are short duration transients which could be detected by some specific techniques but these do not provide enough information to understand the root cause of it. Due to random and non-predictable nature, the intermittent faults are the most frustrating, elusive, and expensive faults to detect in interconnection system. The novel approach of the author injects a fixed frequency sinusoidal signal into electronics interconnection system that modulates intermittent fault if persist. Intermittent faults and other channel effects are computed from received signal by demodulation and spectrum analysis. This paper describes technology for intermittent fault detection, and classification of intermittent fault, and channel characterization. The paper also reports the functionally tests of computational system of the proposed methods. This algorithm has been tested using experimental setup. It generate an intermittent signal by external vibration stress on connector and intermittency is detected by acquiring and processing propagating signal. The results demonstrate to detect and classify intermittent interconnection and noise variations due to intermittency. Monitoring the channel in-situ with low amplitude, and narrow band signal over electronics interconnection between a transmitter and a receiver provides the most effective tool for continuously watching the wire system for the random, unpredictable intermittent faults, the precursor of failure.

Keywords—NFF; Intermittent; Intermittency; Fault detection; Health Monitoring

I. INTRODUCTION

An intermittent fault (IF) is an electrical spike that develops from ageing of electric interconnects, cuts, rubs, or loose contacts, and manifests itself intermittently in an unpredictable manner. If these are not detected on time or at the early stage, it would gradually lead to permanent fault and are also safety critical [1]. This also lead to, many other problems for example delayed or cancellation of flights,

electrical arc or spark that could lead disaster that progressed from IFs.

Manufacturing imperfections, poor design and system degradations are main causes of intermittent faults [2] Although Sheng et al are disagree that intermittent faults are precursor of permanent failure [3] but S. Bryan et al says that intermittent faults are precursor of hard failure [4]. These both statements could be true, depends upon causes of intermittence. IF due to system/component degradation are precursor of permanent faults but marginal design or manufacturing imperfection are not signs of hard failure. Irrespective of causes; IFs are random and non-reproducible incidents, and are most frustrating, elusive, and expensive faults to detect and locate in wiring / interconnection systems.

IFs are identified by visual or traditional instruments for electronic/electrical interconnects. It has also been reported that conventional test equipment, which is required to carry out the fault investigation, are not always successful. This can be due to the fact that the necessary levels of confidence and efficiency are inappropriate in the many industries which are suffering No Fault Found (NFF) failures [5]. If testability as a design characteristic was successful, perhaps NFF would not be so problematic. This is particularly evident in the case of attempting to detect and isolate intermittent faults at a test station the ability to test for short duration non-stationary intermittency at the very moment that it re-occurs using conventional methods is so remote that it will almost certainly result in a NFF. The one major issue with designing component testability is that the focus is on functionality and integrity of the system.

There are many test equipment that are used to detected anomalies in electrical interconnection systems. The more common ones include multi-meters that detect steady or invariant signals. On the other hand, digital oscilloscopes, and spectrum analyzers are used to monitor time domain and frequency domain time invariant signals. Problem with an intermittent fault is that it occurs for only a short duration and it is time variant, making it difficult to detect unless a very

high sample rate it used. This goes beyond the capabilities of typical test equipment. The current state-of-the-art in intermittent fault detection during maintenance testing includes latching continuity testing, analogue neural network technology and time domain reflectometry.

There are various disadvantages of these techniques: to halt operation for inspection, hard to capture or watch on oscilloscope or voltmeter as well as ineffectiveness due to many inspection points and some time being in the location frequently hard to reach or observe. These are unable to detect the fault in many cases since the duration of the fault was often short and not consistent. System would behave normally and it would find the interconnection/wire system normal or NFF status. Therefore, it is easy for the observer or instrument to miss the occurrence of intermittent fault.

Much research has been done on reflectometry wiring fault detection and that is used for high power electrical wirings and could not be used for interfaces and loose solder joints or for other electronics circuits. The concept of reflectometry relies on transmitting electromagnetic waves across the wire and observe the reflections. These reflections depends upon the variation of impedance in the wire system as $\frac{Z_1 - Z_2}{Z_1 + Z_2}$, where Z_1 and Z_2 are impedances of two electrical mediums [6] Time between the incident and the reflected wave is used to locate the fault. Magnitude of reflections are used to determine if it is a potential fault or not. These techniques have drawbacks for modern electronics / electrical system that any change in the wire material (e.g., connection in circuit) reflects the incident waves resulting in incorrect fault determination. These techniques usually requires high voltage pulses.

Recently, direct-sequence spread-spectrum (DS-SS) signals are used instead of high voltage signals employing digital signal processing techniques to find and locate electrical faults [7]. Taylor and Faulkner proposed direct-sequence spread spectrum modulation on power line carrier, and outlined optimal signal processing techniques and frequency domain correlation techniques for the on-line test in high voltage line [8]. Lately, slightly different use of spread spectrum was reported from the research result of on detecting live wire problems [9]. These techniques work on reflectometry, and it solves the need to use low voltage signal, that does not interfere with online signals and could be used in-situ, but still there is a problem of reflection occurring at all points of interconnections in the circuit. So this technique is inadequate for interconnecting system, where there are many interfaces and connectors. This is also not suitable to use for electronic circuits i.e. for PCBs, solder joints, interfaces, and similar interconnecting systems. Otherwise, the injected signal would be reflected from both ends and result in a combined, distorted, and reflected false signal due to impedance mismatch.

The novel approach of IF detecting and characterization has been developed by the author to overcome above mentioned issues and it is very different from traditional diagnostic methods. Novelty of the proposed new technique is the fact that signs of IF intrinsically modulated on a carrier signal, in compare with healthy wired communication channel

and interconnection system. In healthy communication link carrier signal propagates without any changes that affect amplitude/phase/frequency of signal but with Additive White Gaussian Noise (AWGN). The proposed technique aims to look at signature of intermittency as a modulated message on carrier, and employ demodulation techniques to explore behavior of aged channel/interconnections.

The new approaches of the author send a sinusoidal carrier to interconnecting system and demodulate the received signal from interconnection channel for IF detection and feature extraction to find the root cause of problem. This could be used for multipoint of electrical/ electronic interconnection system and diagnose the health status of the wire after demodulation to retrieve an intermittent signature of channel. The essence of this approach is using communication modulation techniques to detect and electrical interconnection system. The transient caused by the intermittent fault in the wire would disrupt the signal sent over interconnection from a transmitter, and thus arriving signal at the receiver would contain intermittent signal information. When intermittent signals are found it will extract IF information by demodulation algorithms. The features of amplitude, phase, and frequency are computed by AM (Amplitude Modulation), PM (Phase Modulation) and FM (Frequency Modulation) demodulation schemes. The benefits of computing phase, amplitude, and frequency of IF could be used to classify intermittent signal for root cause analysis, and degradation monitor.

In the next section, we describe the communication technology and its devised method for detection and computation of fault's information in terms of duration, occurrence frequency, and channel noise. Then, third section describes, devised communication approach for IF detection using demodulation computations. Fourth section describes the test rig and application. Following section describes the results and validations of algorithm then last section concludes this paper.

II. COMMUNICATION APPROACH FOR INTERMITTENT FAULT DETECTION

Related to fault detection, author has used radar communication approach where it sends blank carrier signal and extras desire information from received signal. Intermittent characteristics of channel will change the propagating signal and these intermittent signature could be computed by removing original signal. Carrier modulation / demodulation concept is being used to as sounding techniques to extract IF signature.

There are many carrier modulation schemes but fundamentally there are three modulations schemes called amplitude modulation (AM), frequency modulation (FM) and phase modulation (PM). In AM, the amplitude of carrier signal changes according to input signal and this concept is being used that if there is an intermittent open/close it changes the amplitude of carrier signal. Similarly phase and frequency changes could be computed by using PM and FM demodulation concept.

A. Theory and formulation

Any AM, PM or FM signal $x(t)$ can be written as shown in equation 1

$$x(t) = R(t)\cos(\omega t + \varphi(t)) \dots (1)$$

In equation (1) $R(t)$ is the envelope of signal (amplitude of signal as function of time), ω is angular frequency, and $\varphi(t)$ is a phase of signal at t time.

For AM $\varphi(t)$, and ω are constant only envelope $R(t)$ is time variant, thus equation (1) can be written as below

$$x(t) = (C + m(t))\cos(\omega t) \dots (2)$$

In equation (2) $R(t)$ envelope is replaced to $C + m(t)$ where $m(t)$ is amplitude of base signal, in our case this is an IF signal, and "C" is carrier amplitude.

The IF signal $m(t)$ could be extracted by simple diode rectification and low or band pass filtration for analogue circuits and could be compute digital filtering / modulation algorithms. Filter band must be according to the band range of IF signal else information of IF will be lost.

For PM and FM the amplitude envelope will remain constant but it varies the phase/frequency. For FM/PM demodulation, the signal is fed into a Phase Loop Lock (PLL) and the error signal is used as the demodulated signal.

III. NOVEL FAULT DETECTION ALGORITHM

In wireless communication, to model channel behavior they measure its properties by sending and receiving wireless signal, are called channel sounding techniques [10]. Author

has adapter similar method to measure an intermittency in electric/electronics interconnection systems. To measure IF and its properties it sends and receives suitable signal through interconnection system. A novel algorithm has been developed to compute intermittency for IF detection and classification. Its features of amplitude, frequency, and phase are computed using AM, PM, and FM demodulation algorithms while Fast Fourier Transform (FFT) computes its spectrum. This algorithm has been shown in Fig. 1, it consists of signal source (carrier frequency), intermittent channel (test rig), demodulating unit, digital filter, IF detection using AM, FM and PM algorithms. It counts an intermittency and IF fault detection turn on. Each fault duration and frequency of occurrence are stored in output buffer.

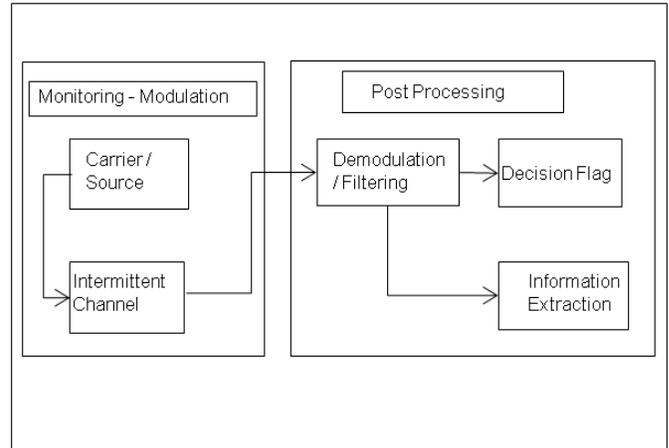


Fig. 1. Block Diagram of IF detection Algorithm

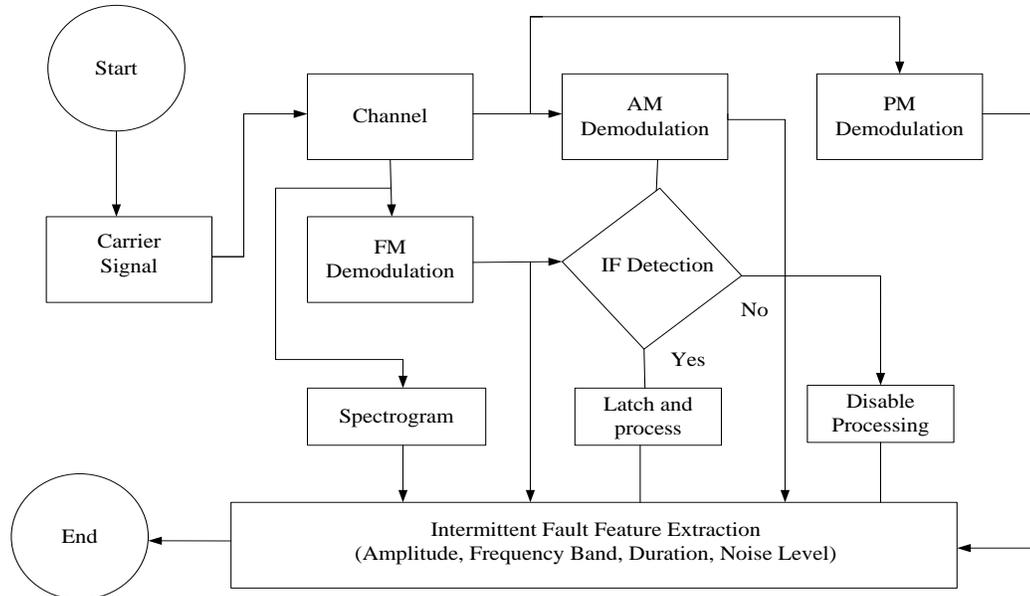


Fig. 2. Intermittent Fault Detection Algorithm

Figure 2 shows the flow diagram of this algorithm. It starts with suitable selected carrier signal that fulfil the required resolution; 1 k Hz sine wave is selected to give one millisecond resolution that is suitable for our repetitively producing IFs test rig. The advantage of using one millimetre resolution, is that it will eliminate debouching harmonics but

if high resolution is required for less frequent IF, carrier frequency could be increased accordingly i.e. resolution is inversely proportional to carrier frequency. Carrier signal propagates through interconnection system to terminating point to complete a circuit. IF detection unit constantly process carrier signal to compute IF and dynamics. Processing

unit demodulates using amplitude, frequency, and phase demodulation schemes. The spectrogram is also computed to check the bandwidth and noise level. Frequency, amplitude and bandwidth information are used to detect IF.

To make IF detection decision AM and FM demodulation techniques are used, if there is not any IF then it will disable the feature extraction and memory but if IF is detected it latches the signal and extracts its feature.

This also save computation power and memory. It also computes the amplitude, bandwidth, noise level, and time information of signal when decision flag is on.

IV. APPLICATION & CASE STUDY

RJ45 Ethernet socket with Ethernet cable/plug under external vibration is used to generate intermittence in the connection. A Female RJ45 Ethernet socket is used to hold it with assembly on shaker that Connector can vibrate as shown in Fig. 3. The grid has been installed on the shaker by screws and a metal plate as shown Fig. 3. This Ethernet connection assembly is used to produce the intermittent fault under vibration. Other ends of Ethernet cable are connected to a circuit and data acquisition system. A complete circuit setup is shown in Fig. 4. It consists of a test rig, oscilloscope, and data acquisition system.

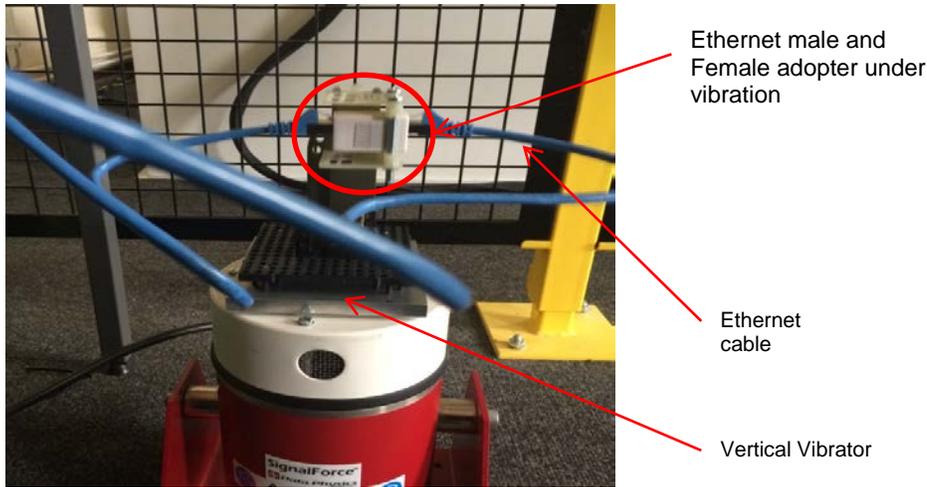


Fig. 3. Ethernet Male and Female Socket with Cable Connection as an Intermittent Test Rig

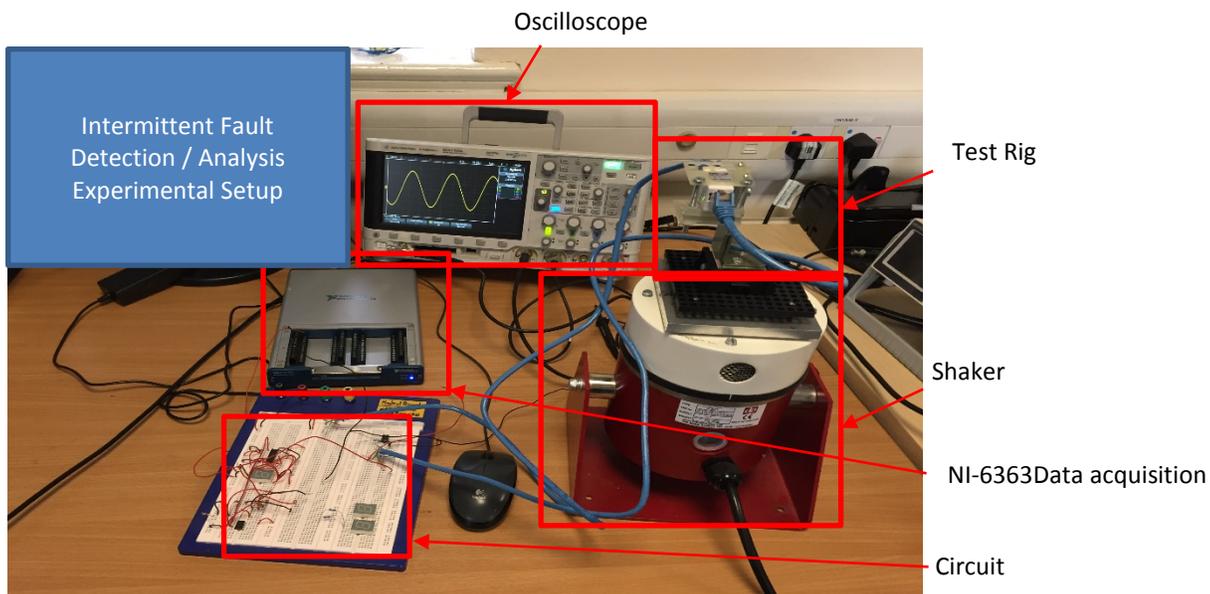


Fig. 4. Experimental Setup for Intermittent Fault Detection

This oscilloscope has four channels, 4 G bits/second sample rate, 200MHz bandwidth and built-in function generator that can output variety of signals but we used 1.00K Hz 3v peak to peak sinusoidal signal as voltage source to voltage divider circuit. The NI-6363 data acquisition card can acquire up to 2 mega samples per second.

The input sine wave of one kilo hertz is propagates through test rig to receiver. To detect an IF and other information, the data is acquired using NI data acquisition card. Received data is being processed using FFT, AM, FM and PM demodulation algorithms. The decision has been taken if there is an IF fault or not; if there is an IF then its noise level, duration and frequency is calculated for IF classification or analysis.

V. SIMULATION AND VALIDATION

The algorithm has been validated by acquiring data from above mention experimental setup and processed in matlab using algorithm described in section 2.

Input carrier signal at 1 k Hz frequency, to electronic interconnection system is shown in 5. This propagated through a test rig under vibration as shown in Fig. 4

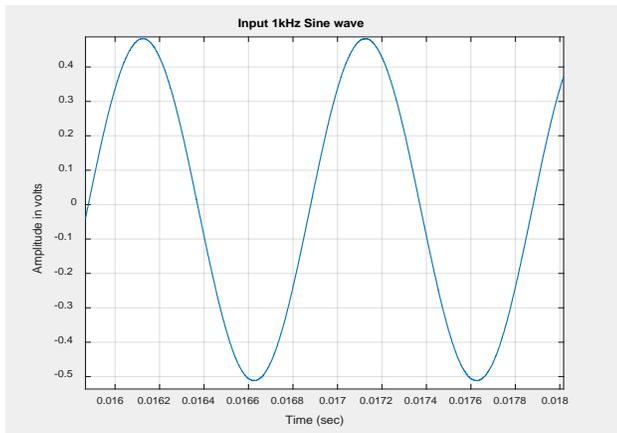


Fig. 5. Input sine wave to unit under test

Shaking test rig adds an intermittency and other noises to a carrier signal due to lose electrical / electronic circuit. Received signal is shown in Fig. 6. This shows that how IF effect on propagating signal. This is output of channel as described in Figure 4.

To detect IF and to extract its feature, it has been demodulated with respect to amplitude, frequency and phase. Amplitude demodulation gives information where amplitude of signal drops due to intermittent discontinuity while change in frequency can be calculated using frequency demodulation. Intermittent fault also changes the phase of signal due to nonlinear discontinuities and could be calculated using phase demodulation.

Fig. 7 shows AM demodulated signal that gives an intermittent signal with twenty spikes of an intermittent fault of a connection shaking at 20 Hz. The amplitude of these spikes shows the change in the amplitude with respect to carrier signal at that instant.

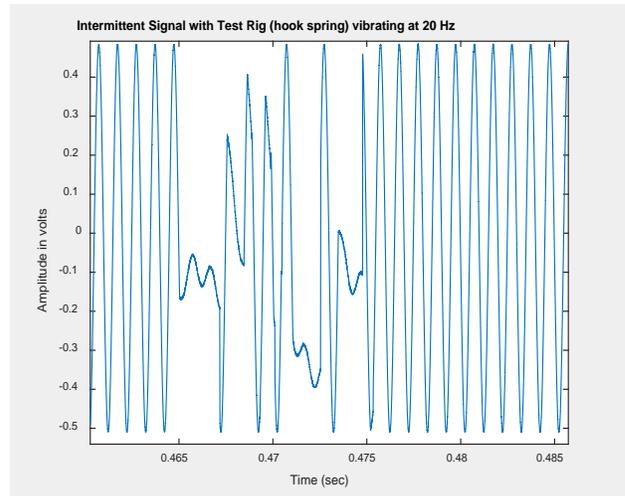


Fig. 6. Received noisy signal with IF information

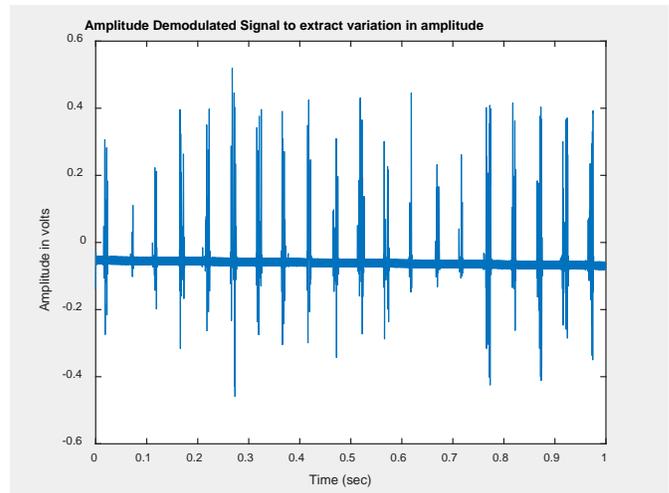


Fig. 7. AM Demodulated Signal

Fig. 8 shows frequency changes with respect to carrier signal. The magnitude indicates changes in the frequency at that instant. The feature of change in frequency are used to calculate the duration of an intermittent interval by subtracting it from carrier frequency and taking inverse. The IF detection decisions are made by comparing both AM and FM demodulated signals and these are also used to calculate its duration and frequency of intermittent fault. It only enable processing unit then there is an intermittent interval as described in Figure 2. The phase change is calculated by phase demodulation as shown in Fig. 9. It gives an information that how phase of intermittent signal has changed. This could be used to study that how an IF effect the signal and change the phase of transmission and adds noise to signal.

The power spectrum of IF signal is shown in Fig. 10. The carrier frequency and intermittent signal are shown in this figure at different frequencies. The normalized frequency has peaks at 0.05 and 0.001; these corresponds 1000 Hz and 20 Hz frequencies when samples at 20k sample/second sampling frequencies. This power spectrum shows the power spectrum of its signals at carrier and around 20Hz intermittent signal's spectrum.

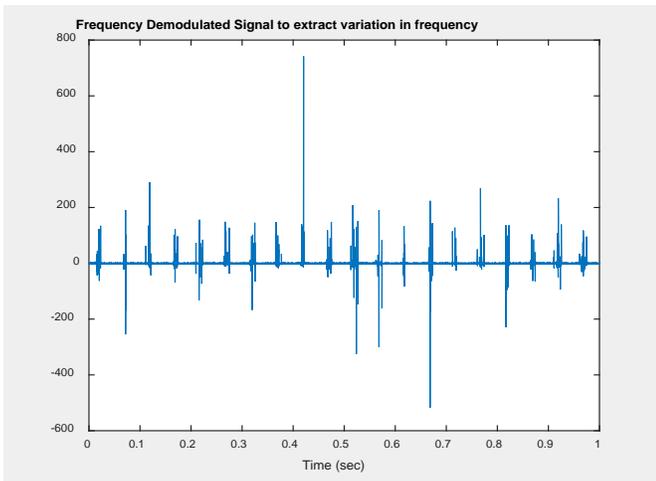


Fig. 8. Frequency Demodulated Signal with 20 Hz Shaking Connector

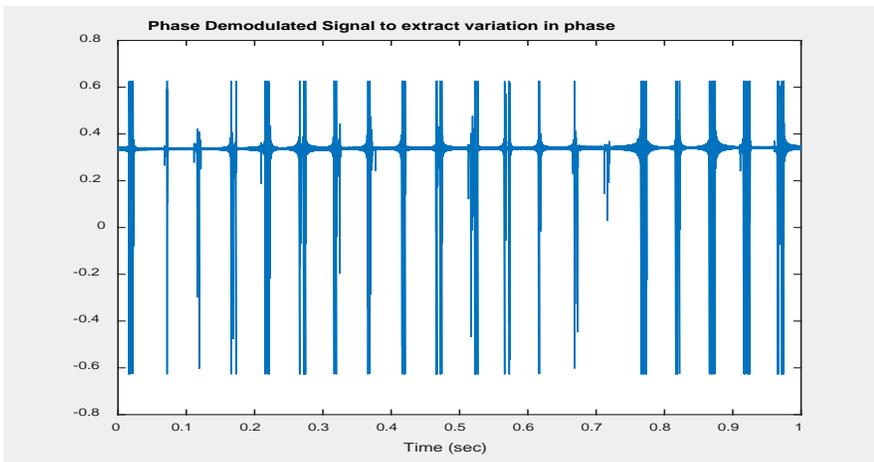


Fig. 9. Phase Demodulated Signal with shaking 20 Hz External Vibration to a connector

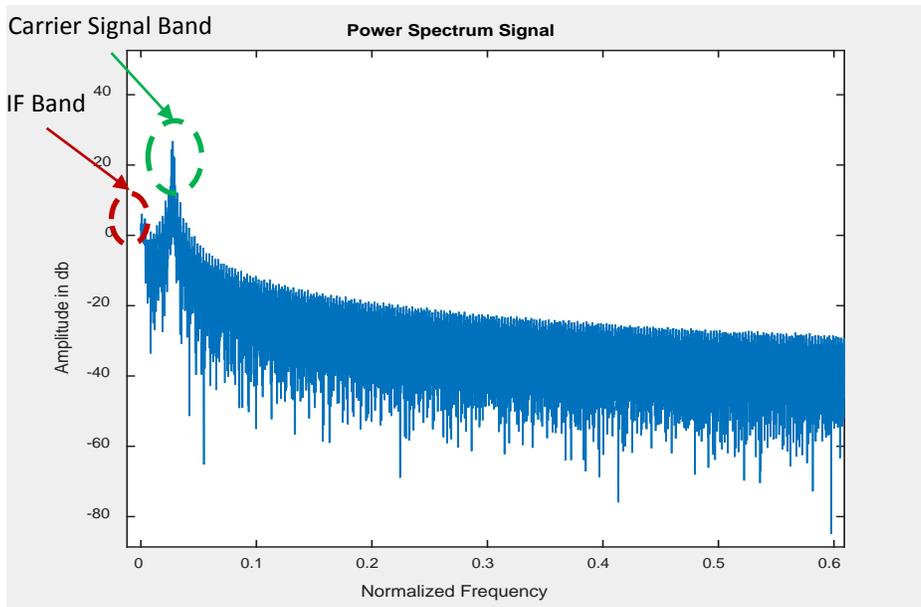


Fig. 10. Spectrum of Intermittent signal

The algorithm described in section 3 has been verified and it gives IF detection and its feature. In this experiment IF correlates with external vibration as we have seen that at 20 hertz shaking lose connection gives us 20 hertz intermittency but the duration and magnitude of IF are not identical for all faults.

VI. CONCLUSION

NFF can be overcome by using in-situ health monitoring algorithm to check it for abnormalities in the interconnection system. An intermittent signal can be detected and classified by using classical demodulation schemes. Amplitude, frequency, and phase variation has been extracted by AM, FM and PM modulation schemes, which gives intermittent channel information.

We have seen that amplitude and frequency variations are very useful for root cause analysis. It also highlights that phase information are not very helpful to understand the exact cause. This algorithm also gives IF signal's characteristics of amplitude, and frequency variation that could help to understand its effects on the system performance. Power spectrum is a very useful tool and that could be used to determine the frequency of intermittency as in experiment IF could be seen at 20 Hz because test rig was shaken at this frequency. Although it is not necessary that spectrum exactly relates to external vibration but could be used to understand the root cause; if it is due to vibration or due to other noise.

Further work could be carried out using filter banks to segment different possible bands of spectrum and this could be used to detect intermittence, and to find root cause.

REFERENCES

- [1] A. Correcher, E. García, F. Morant, E. Quiles and L. Rodríguez, "Intermittent Failure Dynamics Characterization," IEEE TRANSACTIONS ON RELIABILITY, vol. 61, no. 3, pp. 649-658, 2012.
- [2] W. Syed, S. Khan and P. Philips, "Intermittent fault finding strategies," in CIRP, 2013.
- [3] S. Sheng, X. Mingqing, L. Zhao and R. Tingting, "Misunderstandings on intermittent failures and LCP-based description of intermittent failures," *Zhangjiajie*, August 2014.
- [4] S. Bryan, B. Floyd, O. Nathan and S. Brent, "Intermittent Fault Detection and Isolation System," Salt Lake City, 2008.
- [5] H. Qi, S. Ganesan and M. Pecht, "No-Fault Found and Intermittent Failures in Electronic Products," *Microelectronics Reliability Elsevier*, vol. 48, no. 5, pp. 663-674, 2008.
- [6] S. Paul, F. Cynthia and et-al, "Spread Spectrum Sensors for Critical Fault Location on Live Wires," *IEEE Sensors Journal*, June 2005.
- [7] P. Smith, F. Cynthia and J. Gunther, "Analysis of Spread Spectrum Time Domain Reflectometry for Wire Fault Location," *IEEE SENSORS JOURNAL*, vol. 5, no. 6, pp. 1469-1478, December 2005.
- [8] C. Kim, "Detection and Location of Intermittent Faults by Monitoring Carrier Signal Channel Behaviour of Electrical Interconnection System," in *Electric Ship Technologies Symposium, 2009. ESTS 2009. IEEE*, Baltimore, MD, 20-22 April 2009.
- [9] F. Cynthia, S. Paul, L. Chet and et-al, "Spread Spectrum for Critical Fault Location on Live Wire Networks," *Structural Control Health Monitoring*, Wiley InterScience, pp. 257-267, 2005.
- [10] A. F. Molisch, "Andreas F. Molisch," in *Wireless Communications*, 2nd Edition, Wiley, 2012, p. 884.

A Synchronous Stream Cipher Generator Based on Quadratic Fields (SSCQF)

Younes ASIMI

LabSiv, Equipe SCAM

Faculty of sciences, Ibn Zohr

University B.P 8106, City Dakhla, Agadir, Morocco

Ahmed ASIMI

LabSiv, Equipe SCAM

Faculty of sciences, Ibn Zohr

University B.P 8106, City Dakhla, Agadir, Morocco

Abstract—In this paper, we propose a new synchronous stream cipher called SSCQF whose secret-key is $K_S = (z_1, \dots, z_N)$ where z_i is a positive integer. Let d_1, d_2, \dots, d_N be N positive integers in $\{0, 1, \dots, 2^m - 1\}$ such that $d_i \equiv z_i \pmod{2^m}$ with $m \in \mathbb{N}$ and $m \geq 8$. Our purpose is to combine a linear feedback shift registers LFSRs, the arithmetic of quadratic fields: more precisely the unit group of quadratic fields, and Boolean functions [14]. Encryption and decryption are done by XOR'ing the output pseudorandom number generator with the plaintext and ciphertext respectively. The basic ingredients of this proposal stream generator SSCQF rely on the three following processes:

In process *I*, we constructed the initial vectors $IV = \{X_1, \dots, X_N\}$ from the secret-key $K_S = (z_1, \dots, z_N)$ by using the fundamental unit of $\mathbb{Q}(\sqrt{d_i})$ if d_i is a square free integer otherwise by splitting d_i , and in process *II*, we regenerate, from the vectors X_i , the vectors Y_i having the same length L , that is divisible by 8 (equations (2) and (3)). In process *III*, for each Y_i , we assign $L/8$ linear feedback shift registers, each of length eight. We then obtain $N \times L/8$ linear feedback shift registers that are initialized by the binary sequence regenerated by process *II*, filtered by primitive polynomials, and the combine the binary sequence output with $L/8$ Boolean functions. The keystream generator, denoted K , is a concatenation of the output binary sequences of all Boolean functions.

Keywords—Synchronous stream cipher SSCQF; linear feedback shift registers LFSRs; arithmetic of quadratic fields; Boolean functions; pseudorandom number generator and keystream generator

I. INTRODUCTION

The proposed stream cipher SSCQF is a binary addition stream cipher [14]. In a binary addition stream cipher, the plaintext is given as a string m_1, m_2, \dots of elements of the finite field $\mathbf{k}_2 = \{0, 1\}$. The keystream z_1, z_2, \dots is a binary pseudorandom sequence [13]. The sender encrypts the plaintext message according to the rule $c_t = m_t \oplus z_t$ for all

$t \geq 0$. The ciphertext c_1, c_2, \dots is decrypted by the receiver by adding bitwise the keystream z_1, z_2, \dots to the received ciphertext sequence c_1, c_2, \dots . Sender and receiver produce the keystream z_1, z_2, \dots via identical copies of the stream generator.

Let z_1, z_2, \dots, z_N be N positive integers, d_1, d_2, \dots, d_N be N positive integers in $\{0, 1, \dots, 2^m - 1\}$ such that $d_i \equiv z_i \pmod{2^m}$ with $m \in \mathbb{N}$ and $m \geq 8$, and ε_i be a fundamental unit of a quadratic field $\mathbb{Q}(\sqrt{d_i})$, if d_i is a square free integer.

In this paper, we propose a new synchronous stream cipher called SSCQF whose secret-key is $K_S = (z_1, \dots, z_N)$ where z_i are positive integers, based upon the combination of a linear feedback shift registers LFSRs [14], the congruence modulo 2^m with $m \in \mathbb{N}$ and $m \geq 8$, the arithmetic of quadratic fields: more precisely the unit group of quadratic fields, and the $L/8$ combining functions. The basic ingredients of this proposal stream cipher generator SSCQF rely on the following three processes:

In process *I*, we construct the initial vectors $IV = \{X_1, \dots, X_N\}$ from the secret-key K_S by using the fundamental unit of $\mathbb{Q}(\sqrt{d_i})$ if d_i is a square free integer otherwise by splitting d_i , and in process *II*, we regenerate, from the vectors X_i , the vectors Y_i having the same length L , more precisely the length L must be divisible by eight (Equations (2) and (3)). In process *III*, for each Y_i , we assign $L/8$ linear feedback shift registers of length eight filtered by primitive polynomials of degree eight. They are $\frac{\varphi(2^8 - 1)}{8} = 25$ primitive polynomials [12]. We then obtain $N \times L/8$ linear feedback shift registers that are initialized by the binary sequence regenerated by process *II*. And we combine the output binary sequence of all linear feedback

shift registers, namely, $LFSR_{ij}$ with $L/8$ Booleans functions $R_1, \dots, R_{L/8}$. The Boolean function R_j combines the output bits of $LFSR_{ij}$ for all $i \in \{1, \dots, N\}$. The keystream generator denoted K , is a concatenation of the output binary sequences of all Boolean functions R_j .

The output function of our stream cipher is parameterized only by the secret-key K_s . As the keystream bits are produced independently of the plaintext, the proposed stream cipher SSCQF belongs to the category of synchronous stream ciphers.

In this section, we introduce the notations that will be used throughout this paper in TABLE I.

TABLE I. NOTATIONS

K_s	: Input secret-key.
keystream	: Output secret-key.
\oplus	: XOR operation.
\parallel	: Concatenation.
$LFSR_{ij}$: Linear feedback shift registers.
R_j	: Boolean functions.
F	: Feedback function.
x^{-2}	: Binary sequence of any integer x.
IV	: Initial Vector.
\mathbf{k}_2	: Binary finite field of characteristic two.
\mathbf{k}_2^m	: \mathbf{k}_2 -vector space of dimension m .
$Lmc(k, k')$: Lowest common multiple of positive integers $k; k'$.
Γ	: Set of periodic binary functions not necessarily the same period.
\mathbb{N}	: Set of natural numbers.
$\sqrt{\quad}$: Square root.
L_{Bi}	: Length of i^{th} binary sequence.
$L_{1/2Bi}$: Half-length of i^{th} binary sequence.

II. PRELIMINARY

Stream cipher [14] is a secret-key cryptosystem constructed for improve secrecy of transmitted data. It is a lightweight and efficient cryptographic primitive for ensure confidentiality of transmitted data between two communicated pairs. It proves its robustness by its ability to resist against attacks [3][4] [7][14]. It has a wide application area especially in mobile devices and embedded systems. In this section we introduce the notation and terminology that will be used throughout the proposal. We use the symbol $\mathbf{k}_2 = \{0,1\}$ to denote the binary finite field of characteristic two, \oplus to denote logical XOR (OR exclusive), $\mathbf{k}_2^m = \{0,1\}^m$ to denote

the \mathbf{k}_2 -vector space of dimension m , n^{-2} to denote the binary sequence of any integer $n \in \mathbb{N}^*$ and \parallel denotes concatenation of two bits sequences. Bit sequence means a sequence built from 0 and 1.

Definition 2.1: Let $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$ be two vectors of $\mathbf{k}_2^n = \{0,1\}^n$.

- 1) $X = Y$ if only if $x_i = y_i$ for all $i \in \{1, \dots, n\}$.
- 2) $X \oplus Y = (x_1 \oplus y_1, \dots, x_n \oplus y_n)$.

Theorem 2.1: Let X , Y and Z be three vectors of $\mathbf{k}_2^n = \{0,1\}^n$.

$X = Y$ if and only if $X \oplus Z = Y \oplus Z$.

Proof : Let $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_n)$ and $Z = (z_1, \dots, z_n)$ be three vectors of $\mathbf{k}_2^n = \{0,1\}^n$. $X \oplus Z = Y \oplus Z$ if and only if $x_i \oplus z_i = y_i \oplus z_i$ for all $i \in \{1, \dots, n\}$

(Definition 2.1), if and only if $(x_i \oplus z_i) \oplus z_i = (y_i \oplus z_i) \oplus z_i$ if and only if $x_i \oplus (z_i \oplus z_i) = y_i \oplus (z_i \oplus z_i)$ if and only if $x_i \oplus 0 = y_i \oplus 0$ if and only if $x_i = y_i$ if and only if $X = Y$.

Let m be a positive integer. A binary feedback shift register (FSR) of length m is uniquely determined by its feedback function $F : \{0,1\}^m \rightarrow \{0,1\}$.

Definition 2.2 (see [20]): A feedback function $F : \{0,1\}^m \rightarrow \{0,1\}$ is nonsingular if and only if the algebraic normal form of F has the form $F(x_0, \dots, x_{m-1}) = x_0 + G(x_1, \dots, x_{m-1})$, where $G : \{0,1\}^{m-1} \rightarrow \{0,1\}$ is a polynomial in the variables x_1, \dots, x_{m-1} .

If the feedback function F of an m -stage feedback shift register is linear, one speaks of a linear feedback shift registers (LFSR). Otherwise one speaks of a nonlinear feedback shift register (NLFSR). All feedback shift registers used in this paper are nonsingular and linear. In this case, $F(x_0, \dots, x_{m-1}) = x_0 + a_1x_1 + \dots + a_{m-1}x_{m-1}$ modulo 2 where the a_i 's are either 0 or 1 for all $i \in \{1, \dots, m-1\}$ and its linear recursion is of the form: $x_{n+m} = x_n + \sum_{i=1}^{i=m-1} a_i x_{n+i}$

modulo 2 for all $n \geq 0$ [6][11][17]. An alternative way to describe this recursion is to specify the m^{th} degree binary characteristic polynomial [16]: $f(x) = x^m + \sum_{i=1}^{m-1} a_i x^i + 1$.

To obtain the maximal period of $2^m - 1$, a sufficient condition is that $f(x)$ be a primitive m^{th} degree polynomial modulo two.

Definition 2.3 (see [12]): Let $f(x) \in \mathbf{k}_2[x]$ be a polynomial of degree at least l . Then $f(x)$ is said to be irreducible over \mathbf{k}_2 if it cannot be written as a product of two polynomials in $\mathbf{k}_2[x]$, each of positive degree.

Definition 2.4 (see [12]): Let $f(x) \in \mathbf{k}_2[x]$ be an irreducible polynomial of degree N . Then $\mathbf{k}_2[x]/(f(x))$; the set of polynomials in $\mathbf{k}_2[x]$ of degree less than N , is a field of order 2^N . Addition and multiplication are performed modulo $f(x)$. Therefore $\mathbf{k}_{2^N} = \mathbf{k}_2[x]/(f(x))$. In this case, \mathbf{k}_{2^N} is called the splitting field of $f(x)$.

Definition 2.5 (see [12]): A polynomial $f(x) \in \mathbf{k}_2[x]$ of degree N is called a primitive polynomial over \mathbf{k}_2 if it is the minimal polynomial over \mathbf{k}_2 of a primitive element of \mathbf{k}_{2^N} .

Definition 2.6 : We call a Boolean function upon $\{0,1\}^N$, all function defined from $\{0,1\}^N$ into $\{0,1\}$. They are 2^{2^N} Boolean functions upon $\{0,1\}^N$.

III. A BRIEF DESCRIPTION OF SSCQF ALGORITHM

Stream cipher encrypts the plaintext by using a key stream generator. The latter can be a synchronous or an asynchronous stream cipher. This property is related to regenerate a nature of secret-key. A generator is qualified as a synchronous stream cipher if the regeneration of the secret-keys carries out independently of the plaintext and ciphertext messages. By contrast, an asynchronous stream cipher products the keystreams as a function of the input secret-key and previous ciphertexts [14]. Our synchronous algorithm SSCQF can briefly be described as follows:

It takes a secret-key constructed by a sequence of positive integers z_1, \dots, z_N and let $d_i \equiv z_i \pmod{2^m}$ with $m \in \mathbf{N}$ and $m \geq 8$.

For each d_i we assign them only two positive integers n_i and m_i as follows:

- If $d_i = s_i^2 r_i$ where $r_i = 1$ or r_i is a square free integer, then $n_i = r_i$ and $m_i = s_i^2$.
- If d_i is a square free integer, then we assign only one fundamental unit ε_i of the quadratic field

$\mathbb{Q}(\sqrt{d_i})$ [2] [5] where

$$\varepsilon_i = \begin{cases} n_i + m_i \sqrt{d_i} & \text{if } d \equiv 2 \text{ or } 3 \pmod{4} \\ \frac{n_i + m_i \sqrt{d_i}}{2} & \text{if } d \equiv 1 \pmod{4} \end{cases} \quad (1)$$

We then construct the initial vectors $IV = \{X_1, \dots, X_N\}$ where $X_i = \overline{n_i} \parallel \overline{d_i} \parallel \overline{m_i}^{-2}$ for all $i \in \{1, \dots, N\}$. Since the vectors X_i do not have the same length, then we regenerate the vectors Y_i , from the vectors X_i , having the same length L . The number L is divisible by eight via equations 2 and 3. Each binary standard sequence is subdivided into $L/8$ binary sequences of length eight, each of them initializes one linear feedback shift register of length eight. We then obtain $L/8$ LFSRs for each Y_i , namely, $\text{LFSR}_{i1}, \dots, \text{LFSR}_{iL/8}$ filtering by primitive polynomials of degree eight. And we combine the output binary sequence of all LFSR_{ij} with $L/8$ Boolean functions $R_1, \dots, R_{L/8} : \{0,1\}^N \rightarrow \{0,1\}$ defined as follows: For each $j \in \{1, \dots, L/8\}$, the Boolean function R_j combines the output bits of LFSR_{ij} for all $i \in \{1, \dots, N\}$. The keystream digit is obtained by concatenation of the output binary sequences of all Boolean functions R_j .

IV. DETAILED DESCRIPTION OF SSCQF ALGORITHM

The overall structure of the keystream generator SSCQF is depicted in the following figure.

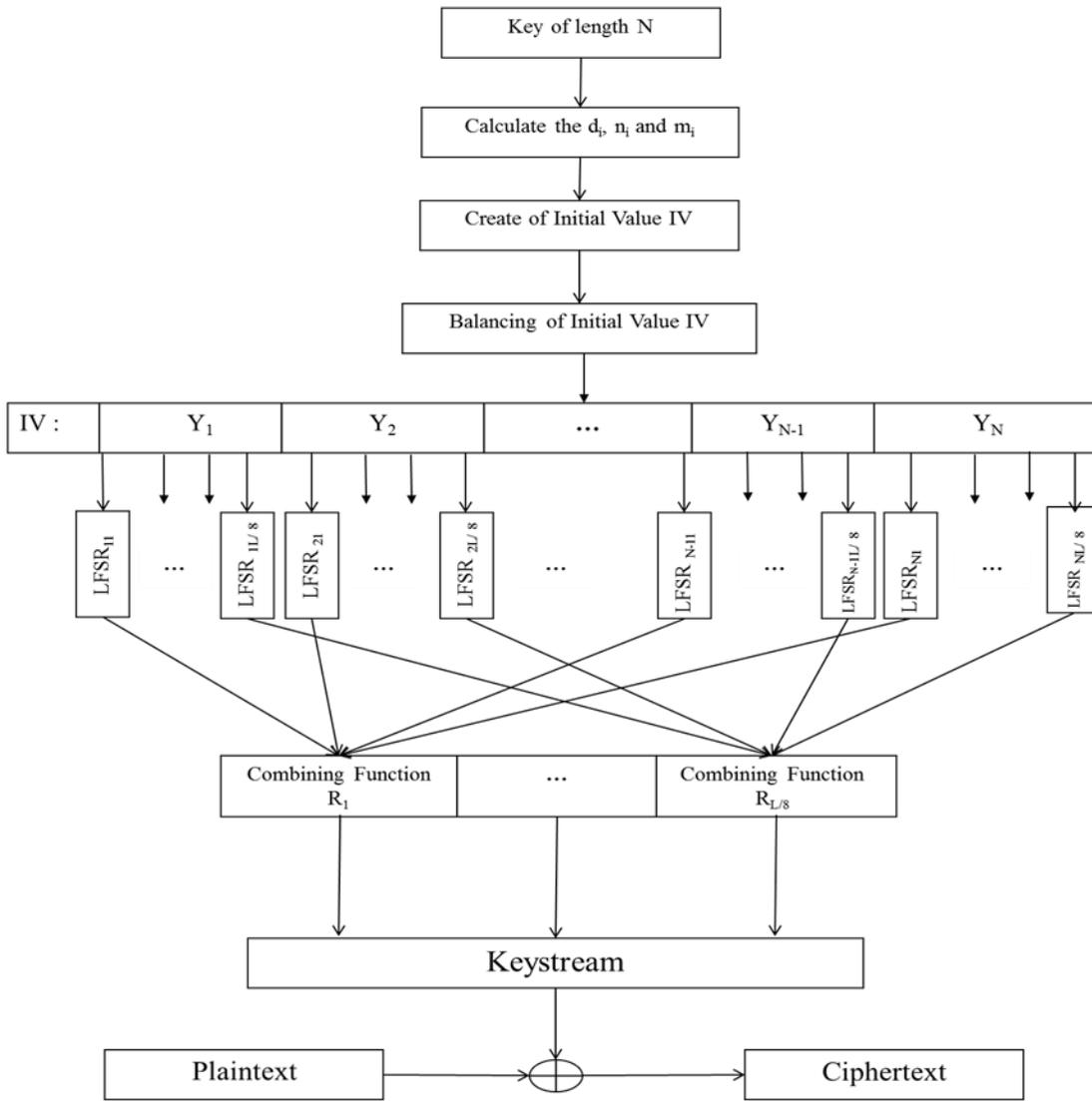


Fig. 1. Detailed description of ASCGQF algorithm

The basic ingredients of the keystream generator SSCQF rely on the following three processes:

A. Process I

The main goal of this process is to generate the initial vectors $IV = \{X_1, \dots, X_N\}$ from a secret-key $K_S = (z_1, \dots, z_N)$ where z_i are positive integers for all $i \in \{1, \dots, N\}$. We then proceed as follows:

- We compute the positive integers d_i such that $d_i = z_i \bmod 2^m$ with $m \in \mathbb{N}$ and $m \geq 8$ for all $i \in \{1, \dots, N\}$.
- For each d_i we assign only two positive integers n_i and m_i :

- Assume that $d_i = s_i^2 r_i$ where $r_i = 1$ or r_i is a square free integer, we then get $n_i = r_i$ and $m_i = s_i^2$.

- Assume that d_i is a square free integer, we assign only one fundamental unit ε_i of the quadratic field $\mathbb{Q}(\sqrt{d_i})$ [2] [5] together with

$$\varepsilon_i = \begin{cases} n_i + m_i \sqrt{d_i} & \text{if } d \equiv 2 \text{ or } 3 \pmod{4} \\ \frac{n_i + m_i \sqrt{d_i}}{2} & \text{if } d \equiv 1 \pmod{4} \end{cases}$$

- For all $i \in \{1, \dots, N\}$, $X_i = \overline{n_i}^{-2} \parallel \overline{d_i}^{-2} \parallel \overline{m_i}^{-2}$.

B. Process II

The vectors X_i for all $i \in \{1, \dots, N\}$ are not necessarily of the same length. The goal of this process is to balancing those vectors. For that, we then choose a vector of a maximal length, for example X_k of length $l_k = L'$, and we proceed as follows :

For each vector $X_i = (x_{i1}, \dots, x_{il_i})$ one assigns the only vector $Y_i = (y_{i1}, \dots, y_{il_i})$ defined as follows:

If $L \equiv 0 \pmod 8$, $L = L'$, we get:

$$\begin{cases} y_{ij} = x_{ij} & \text{for all } 0 \leq j \leq l_i \\ y_{i(l_i+t)} = x_{i(t \bmod l_i)} \oplus x_{kt} & \text{for all } 0 \leq t \leq L - l_i \end{cases} \quad (2)$$

Otherwise, $L = L' + (8 - L' \bmod 8)$, we get:

$$\begin{cases} y_{ij} = x_{ij} & \text{for all } 0 \leq j \leq l_i \\ y_{i(l_i+t)} = x_{i(t \bmod l_i)} \oplus x_{kt} & \text{for all } 0 \leq t \leq L' - l_i \\ y_{i(L'+s)} = \sum_{t=0}^s x_{it} \oplus x_{ks} & \text{for all } 0 \leq s \leq 8 - (L' \bmod 8) \end{cases} \quad (3)$$

C. Process III

The vectors Y_i for all $i \in \{1, \dots, N\}$ generated in the process II, are of the same length L divisible by eight. We subdivide it into $L/8$ binary sequences of length eight; each initializes a linear feedback shift register filtered by the primitive polynomial of degree eight. We then obtain, for each Y_i , $L/8$ linear feedback shift registers, namely, $LFSR_{i1}, \dots, LFSR_{iL/8}$. And we combine the output binary sequence of all $LFSR_{ij}$ with $L/8$ Boolean functions $R_1, \dots, R_{L/8} : \{0,1\}^N \rightarrow \{0,1\}$ defined as follows: For each $j \in \{1, \dots, L/8\}$, the Boolean function R_j combines the output bits of $LFSR_{ij}$ for all $i \in \{1, \dots, N\}$, together with $R_j(x_1, \dots, x_N) = R(x_1, \dots, x_{j-1}, 1, x_{j+1}, \dots, x_N)$ and $R(x_1, \dots, x_N) = \sum_{i=1}^{i=N} x_i + \sum_{i < j=1}^N x_i x_j \pmod 2$. The keystream is obtained by concatenation of the output binary sequences of all Boolean functions.

V. BEHAVIORAL STUDY

After presenting and explaining the principle components of our SSCQF algorithm, in this section, we focus a behavioral study for all elements constituting our regenerator in order to highlight its internal characteristics. We begin by studying the complexity of the output binary sequences of all Boolean

functions R_j related to their lengths for a given password. Effectively, our goal, in this subsection, is to appear the cryptographic nature of the internal states of our regenerator of binary sequences. Then, we pass to analysis the keystream regenerated by our system after the minimal perturbations on the initial condition. Finally, we present an analytical study simulating the human system.

A. Correlation and normalized distance of periodic binary strings

For the binary sequences, we must exploit the Hamming principle to make sure their nature distribution. It aids in estimating the complexity of binary strings that have the same period. However, the testing of the keystreams regenerated by our regenerator show that not necessarily of the same period. Hence, we should use an extension of a Hamming distance as we defined in [1] [21]:

Let S and S' be two elements of Γ of periods k and k' respectively and $K = Lmc(k, k')$.

The function $D' : \Gamma \times \Gamma \rightarrow [0,1]$ defined by:

$$D'(S, S') = \frac{\sum_{i=0}^{K-1} ((S(i) + S'(i)) \% 2)}{K} \quad (4)$$

is a normalized distance of Γ .

Also in [21], we defined another interesting property allowing to more ensure the nature of binary sequences: uncorrelation of the binary strings. Thus, for all S and S' in Γ , we say that two binary strings are weakly correlated if:

$$D'(S, S') \simeq 0.5 \quad (5)$$

This property allows us to prove the complexity of the binary sequences not necessarily of the same period. More precisely, the obtained values of a normalized distance are used to make sure about the uncorrelation or the correlation of the sets of periodic binary strings.

B. Impact of the lengths on the output binary sequences of all Boolean functions

Firstly, we propose an analysis study of each output binary sequences of all Boolean functions R_j related to their lengths for a given password. In this case, we change the length of output binary sequences of all Boolean functions R_j in order to ensure the internal nature of our regenerator. For this object, we propose a fixed secret-key $K_S = (z_1, \dots, z_N)$ where z_i are positive integers and N equal to 50 as follows:

$$K_S = \{12, 3, 6, 77, 80, 81, 90, 95, 44, 54, 56, 47, 2, 8, 10, 15, 18, 16, 28, 99, 29, 55, 60, 67, 86, 84, 26, 37, 35, 34, 311, 57, 41, 5, 13, 11, 512, 73, 92, 40, 42, 47, 19, 388, 39, 71, 73, 79, 188, 115\}$$

For each case, for same secret-key K_s , we adapt our program to regenerate the primitive signals not have the same length. Then, we obtain:

- In first case (Fig.2), the length of a binary sequence is: $L_{B1}=2005$ bits.
- In second case (Fig. 3), the length of a binary sequence is: $L_{B2}=4005$ bits.
- In third case (Fig.4), the length of a binary sequence is: $L_{B3}=6005$ bits.

From [14], we say the binary sequences X_1, \dots, X_N of same lengths are independent if each taking on the values 0 or 1 with probability $\frac{1}{2}$. Then, we talk about the unpredictable and uncorrelated primitive signals if the distribution of hamming distance accumulates near to half-length ($L_{1/2Bi}$) of this binary sequence. This means that almost half the bits in same position of two set of the binary sequence are different.

- $L_{1/2B1} \approx 1002$ bits.
- $L_{1/2B2} \approx 2002$ bits.
- $L_{1/2B3} \approx 3002$ bits.

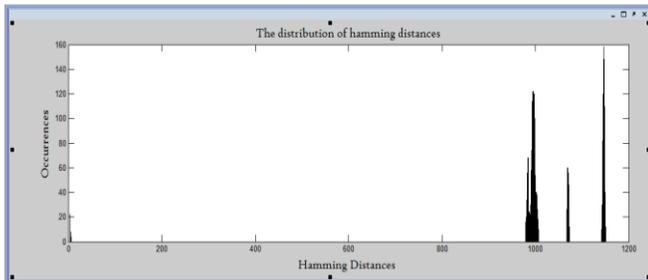


Fig. 2. The distribution of hamming distances for $L_{B1}=2005$ bits

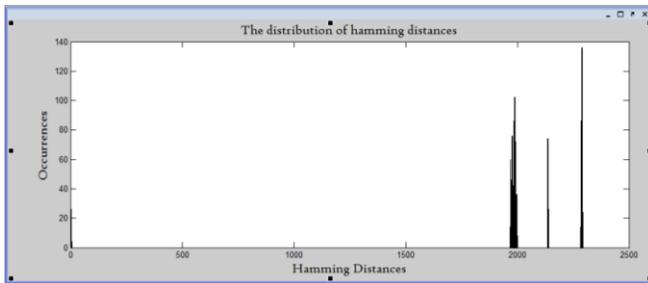


Fig. 3. The distribution of hamming distances for $L_{B2}= 4005$ bits

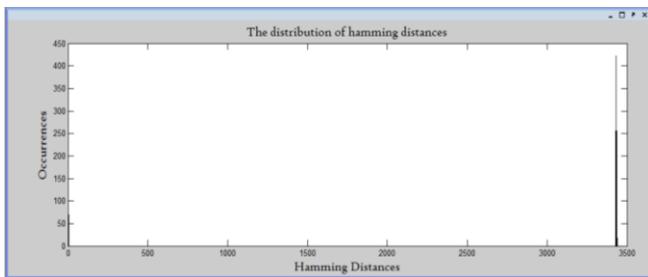


Fig. 4. The distribution of hamming distances for $L_{B3}= 6005$ bits

From these histograms, we notice, for a same secret-key, the distribution of hamming distances in these three cases accumulates in the vicinity of half-length of each output binary sequences of all Boolean functions. In addition, the obtain results are almost identical in all three histograms. In two first cases, we have three accumulations regions nearest to half-length. But, in third case, we have only a peak nearest to half-length. Accordingly, the cryptographic nature of each primitive signal in any internal state is not only related to the length of the regenerated a binary sequence. Effectively, these results are strongly linked to Boolean Functions and linear feedback shift registers filtered by the primitive polynomials of degree eight integrated in our system. Hence, our purpose has unpredictable internal characteristics [1][21], which is recommended in order to resist against attack periodic sequences [5][10]. This enables us to ensure the cryptographic nature of SSCQF algorithm. Finally, for each internal state, we can summarize these features as follows:

- The length of each block regenerated has a positive effect on the cryptographic quality of the regenerated primitive signals.
- The distribution of lengths and periods are random.
- The primitive signals are unpredictable or cryptographically strong.
- When we increase the period length of the internal states, their regenerated the primitive signals became more uncorrelated. Then, long period has a positive impact on the cryptographic nature of internal primitive signs. This property is more desirable for an efficient stream cipher generator.
- The cryptographic quality of each regenerated primitive signals is strongly related to Boolean Functions and linear feedback shift registers filtered by the primitive polynomials of degree eight integrated in our system.

C. Impact of Minimal Perturbations

After introducing an analytical study of the internal states of our system, in this subsection, we concentrate to the behavioral study of external states Keystream of our system. The benefit is to interpret the responses of our proposed system in the minimal conditions. Objectively, for each iterations, we choose the secret-keys the same length $K_s = (z_1, \dots, z_N)$ where z_i are a positive integer in an interval $[2, \dots, 50]$, N equal to 6, the first secret-key is $K_s = (2, 2, 2, 2, 2, 2)$ and the last secret-key is $K_s = (50, 50, 50, 50, 50, 50)$. Also, we perform the minimal perturbations on the input secret-key in order to examine their impact on the lengths and the nature of primitive signals of the associated keystreams. We increment, in each iteration, an integer number z_i of input secret-key in a given position progressively. The importance is to show if the linearity of input secret keys has an effect on the cryptographic quality of output secret-keys.

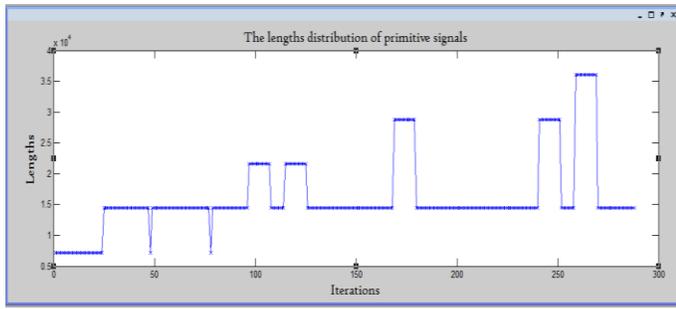


Fig. 5. The lengths distribution of primitive signals

From this histogram (Fig.5), we observe, for the minimal perturbations, that the lengths distribution of primitive signals does not admit a probabilistic law. That means, it hard to an attack to infer the input length according to the lengths of output secret-keys. Its period represents an important benefit to distinguish a good stream cipher regenerator. This dynamite confirms another robustness factor of our regenerator of binary sequences.

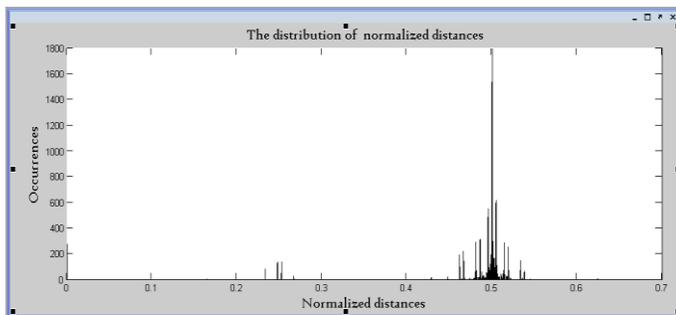


Fig. 6. The distribution of normalized distances

In this histogram (Fig.6), it appears clearly the accumulation of normalized distances nearest to 0.5 followed by small peaks and a large peak exactly in 0.5. This result of normalized distances reassures another significant property filled by our proposed system: unpredictable of each binary sequence. Therefore, we confirm the uncorrelation of generated primitive signals able to withstand the collision and correlation attacks [5][8] [9][10][14][18] [19] [21].

D. Simulating a human system

In reality, Man has a chaotic mind. It is hard to control an user during the choice its input secret-key K_s . But, we can -

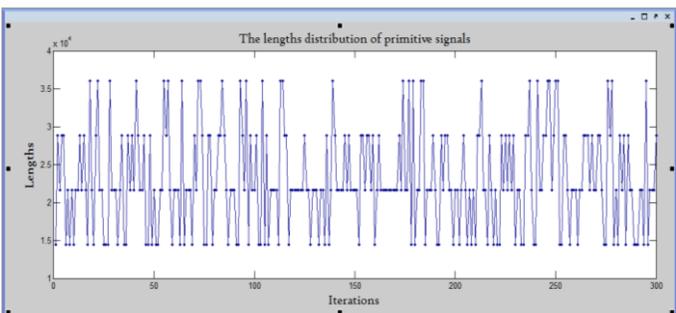


Fig. 7. The lengths distribution of primitive signals

simulate a human system for regenerate the inputs secret-keys the same length ($N=6$). For this work, we adapt a Rand function in order to product the integer numbers z_i in interval $[1, \dots, 200]$ randomly. The aim, in this emulation, is to study the dynamic nature and the cryptographic quality of regenerated primitive signals in the real situations.

This dynamite (Fig.7) reconfirms the random nature of the lengths distribution of regenerated primitive signals for the inputs secret-keys of same length. It is random and unpredictable over time. This result is highly dependent on calculated positive integers d_i such that $d_i = z_i \bmod 2^m$ with $m \in \mathbb{N}$ and $m \geq 8$. More specifically, it depends on the quadratic structure (square-free integer or integer with square factor) of the calculated positive integers d_i . Because, the binary representations of positive integers d_i , n_i and m_i , have an impact on the balancing results. Wherefore, our system inspires its robustness.

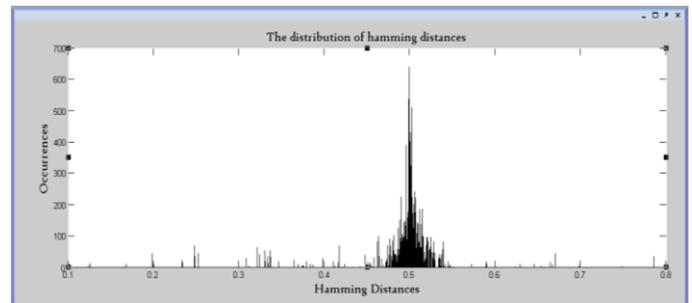


Fig. 8. The distribution of normalized distances

This outcome (Fig.8) is identical to the result obtained in figure 6. It proves, in the minimal conditions, the cryptographic nature of SSCQF algorithm [21]. In effect, our algorithm is efficient and able to resist against attack periodic sequences [5][10]. Likewise, the keystreams are cryptographically strong. This stream ciphers design generate the keystream digits pseudo-randomly from smaller inputs secret-keys without lessening security. They are also able to withstand against to correlation, collision and exhaustive search attacks on stream ciphers [3][4][7][8][9][14][15][18][19][21]. We aim, by this work, to evolve and improve at the symmetric-key encryption scheme.

VI. IMPLEMENTATION

This SSCQF regenerator of binary sequences can be executed in different types of symmetric cryptosystem. We aim, in this work, to evolve the cryptographic quality secret-keys against various types of attacks [3][4][7][9][10][14][18] [19]. Thus, according to behavioral study, this property of the primitive signals regenerated is assured. In this section, we itemize practically different execution stages of our proposed system.

A. Implementation of process I

The first aim of this process is to generate the integer numbers d_i , n_i and m_i for each element z_i of a secret-key

K_s , then, their binary representations. In each iteration, the binary representations of d_i, n_i and m_i will be combined in order to create an initial vector as follows $X_i = \overline{n_i}^{-2} \parallel \overline{d_i}^{-2} \parallel \overline{m_i}^{-2}$.

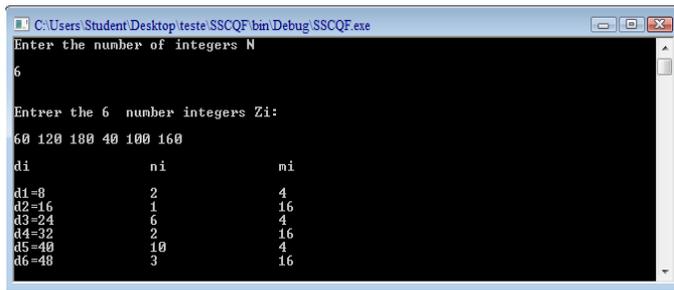


Fig. 9. Regeneration of the d_i, n_i and m_i for a secret-key

From this figure (Fig.9), we show that the values of n_i and m_i don't depend on the values of d_i , but, these are strongly related to its quadratic structure. In reality, it gives more complexity and dynamite of our proposed system. It suffices to behold here that any added bit has an impact on the balancing results of initial binary vectors X_i .

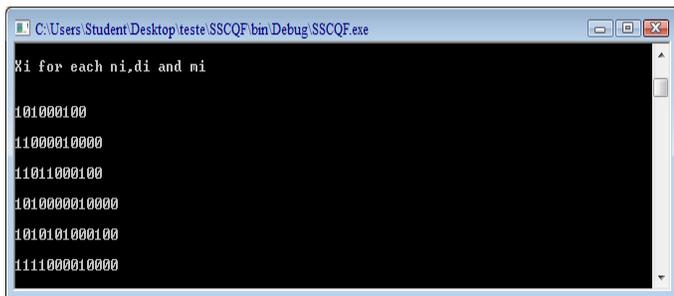


Fig. 10. Binary representation of each initial vector X_i

From this outcome (Fig.10), the binary representations of each initial vector X_i don't have the same length. But, in our

proposal, we want to get the binary sequences which have the same length L divisible by eight. This is the object of the following process.

B. Implementation of process II

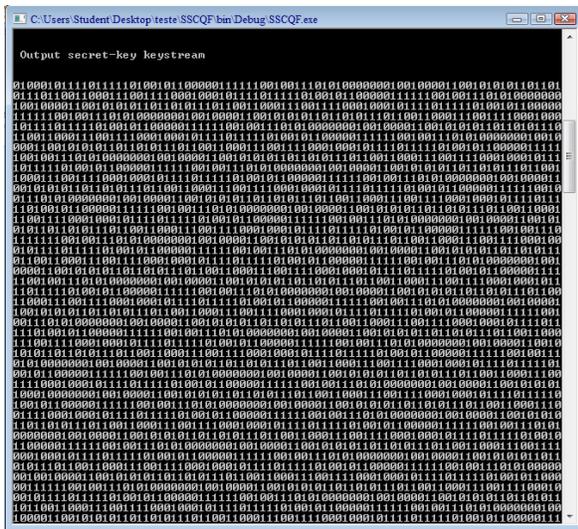
As we have previously explained, we dedicate this process to balancing the binary sequences generated in previous process. The aim is to obtain initial binary vectors X_i that have a length multiple to eight. Because, in these situation, we use a linear feedback shift register filtered by the primitive polynomial of degree eight. So, if we change the degree of primitive polynomial, in this case, we should adapt this process to regenerate the initial vectors that have a length of its degree. The results of this process are presented in following figure (Fig.11).



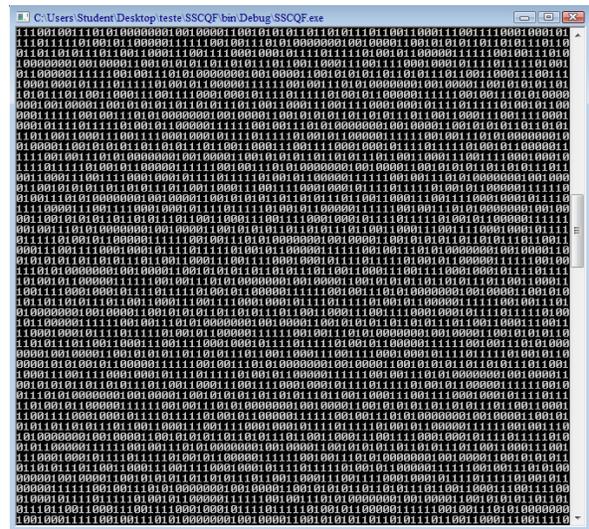
Fig. 11. Balancing of each X_i

C. Implementation of process III

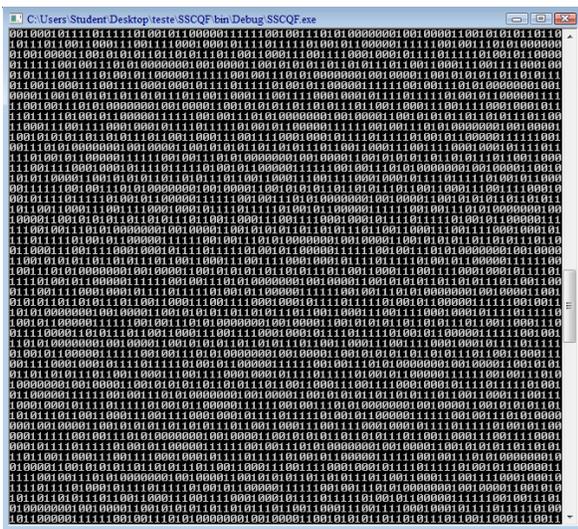
After balancing each initial vectors comes this important process. We implement this process for create an output Keystream digit specific to each input secret-key K_s . In first time, for each Y_i , we construct $L/8$ linear feedback shift registers, namely, $LFSR_{i1}, \dots, LFSR_{iL/8}$. Then, we exercise the Boolean functions R_j on all $LFSR_{ij}$ as defined in process III. The output keystream is obtained by concatenation of the output binary sequences of all Boolean functions. This following figure presents an embodiment of this process (Fig.12).



Binary sequence (1)



Binary sequence (2)



Binary sequence (3)



Binary sequence (4)

Fig. 12. Binary sequence of Keystream digit

Note: The Keystream is obtained by concatenation of all binary sequences (1, 2, 3, 4).

In this work, we innovate a quick, dynamic and complex generator of the binary sequences. We are combined a large theory concept for product a pseudorandom stream cipher. It will be used as a symmetric key cipher for avoid the serious security problems. This synchronous generator products primitive signals uncorrelated, unpredictable and independents of the same input secret-keys lengths. Moreover, it ensures the cryptographic quality of internals states in order to avoid correlation attacks [9][14][18] [19].

VII. CONCLUSION

We introduced, in this paper, a new synchronous stream generator cipher named SSCQF. Our proposed symmetric key system is founded on quadratic fields. We aim by this work to

improve the confidentiality of transmitted data between two communicated pairs. A behavioral study, in the minimal conditions, appears the cryptographic nature of our construction. It also confirms the concrete security of the internal and external states, more, its ability to conserve the unpredictable nature of each regenerated primitive signals. In addition, the output secret-key length is not related to the input secret-key length, but, is strongly linked to quadratic nature of each element constructing an input secret-key. Idem, these dynamite and robustness are clearly proved in implementation section.

REFERENCE

- [1] A. Sabour, A. Asimi, and A. Lbekkouri, "The three states functions: Theoretical foundations and estimated complexity", in The 3rd International Conference on Information Technology, pp. 1{9, 2007).
- [2] Cassels. J.W.S and Frohlich. A, "Algebraic number theory", Academic Press, 1967.

- [3] Courtois N. and Meier W.: "Algebraic attacks on stream ciphers with linear feedback", Advances in Cryptology Eurocrypt 2003, LNCS 2656, Springer-Verlag, pp. 345-359, 2003.
- [4] Courtois N.: "Algebraic Attacks on Combiners with Memory and Several Out-puts", ICISC 2004, LNCS 3506, pp. 320, 2005.
- [5] E.R. Berlekamp. "Algebraic coding theory". McGraw-Hill, 1967.
- [6] Golomb. S.W, "Shift register sequences", revised edition, Aegean park press, laguna hills, California, 1982.
- [7] Hawkes P. and Rose G.: "Rewriting variables: the complexity of fast algebraic attacks on stream ciphers", Advances in Cryptology Crypto 2004, LNCS 3152, SpringerVerlag, pp.390-406, 2004.
- [8] J. D. Golic. "Cryptanalysis of Alleged A5 Stream Cipher". In Advances in Cryptology { Eurocrypt'97, LNCS 1233, pp. 239-255, Springer-Verlag, 1997.
- [9] J. D. Golic. "Towards Fast Correlation Attacks on Irregularly Clocked Shift Registers." In Advances in Cryptography { Eurocrypt'95, pp. 248-262, Springer-Verlag, 1995.
- [10] J.L. Massey. "Shift-register synthesis and BCH decoding". IEEE Transactions on Information Theory, vol. 15, pp. 122-127, 1969.
- [11] Lewis. T.G and Payne, W.H, "Generalized feedback shift register pseudo-random number algorithms", Journal of the ACM, 20: 456-468, 1973.
- [12] Lidl. R and Neiderreiter. H, Introduction to finite fields and their applications, Cambridge University Press: Cambridge, London, New York, 1968.
- [13] Menezes A.J., Oorschot P.C., Vanstone S.A.: "Handbook of Applied Cryptography", Chapter 5: Pseudorandom Bits and Sequences, CRC Press, 1996.
- [14] Menezes A.J., Oorschot P.C., Vanstone S.A.: Handbook of Applied Cryptography, Chapter 6: Stream Ciphers, CRC Press, 1996.
- [15] S. Babbage. "A Space/Time Tradeoff in Exhaustive Search Attacks on Stream Ciphers". European Convention on Security and Detection, IEE Conference publication, No. 408, May 1995.
- [16] Samuel. P, "Théorie algébrique des nombres", Hermann, Paris 1971.
- [17] Tausworthe. R. C, Random numbers generated by linear recurrence modulo two, Mathematics of Computation, 19: 201-209, 1965.
- [18] V.V. Chepyzhov, T. Johansson and B. Smeets. "A Simple Algorithm for Fast Correlation Attacks on Stream Ciphers". In Fast Software Encryption { FSE 2000, LNCS 1978, pp. 181-195, Springer-Verlag, 2000.
- [19] W. Meier and O. Staffelbach. "Fast Correlation Attacks on Certain Stream Ciphers". Journal of Cryptography, 1(3):159-176, 1989.
- [20] Walker. E. A, "Nonlinear recursive sequences can". J. Math 11, 370-378, 1959.
- [21] Younes Asimi, Abdallah Amghar, Ahmed Asimi, and Yassine Sadqi, "New Random Generator of a Safe Cryptographic Salt Per Session", International Journal of Network Security, Vol.18, No.3, PP.445-453, May 2016.

Pneumatic Launcher Based Precise Placement Model for Large-Scale Deployment in Wireless Sensor Networks

Vikrant Sharma

Dept. of Computer Science and Engineering
GB Pant Engineering College
Pauri, Uttarakhand, India

H S Bhadauria

Dept. of Computer Science and Engineering
GB Pant Engineering College
Pauri, Uttarakhand, India

R B Patel

Dept. of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Chandigarh, India

D Prasad

Dept. of Information Technology
M M Engineering College
Mullana, Haryana, India

Abstract—Sensor nodes (SNs) are small sized, low cost devices used to facilitate automation, remote controlling and monitoring. Wireless sensor network (WSN) is an environment monitoring network formed by the number of SNs connected by a wireless medium. Deployment of SNs is an essential phase in the life of a WSN as all the other performance matrices such as connectivity, life and coverage directly depends on it. Moreover, the task of deployment becomes challenging when the WSN is to be established in a large scale candidate region within a limited time interval in order to deal with emergency conditions. In this paper a model for time efficient and precise placement of SNs in large-scale candidate region has been proposed. It constitute of two sets of pneumatic launchers (PLs), one on either side of a deployment helicopter. Each PL is governed by software which determines the launch time and velocity of a SN for its precise placement on the predetermined positions. Simulation results show that the proposed scheme is more time efficient, feasible and cost effective in comparison to the existing state of art models of deployment and can be opted as an effective alternative to deal with emergency conditions.

Keywords—WSN; deployment; placement; aerial; coverage

I. INTRODUCTION

Sensors are being used over the years to facilitate automation and remote monitoring [1]. Wireless sensors have relieved from the mesh of connecting wires used so far and significantly extended the application domain of sensors. WSNs are widely used for the purpose of disaster management, military, health care, industrial and agricultural monitoring and automation [2][3][4][5]. Deployment is a prime phase in the life of any wireless sensor network (WSN) and the performance of any WSN largely depends on it. It becomes more challenging and difficult when the candidate region is extremely large and unreachable. The matter becomes even more sensitive when it is all about disaster management and life rescue. In such cases a quick, effective and generic technique is required to optimally place the SNs within a candidate region in order to handle the situation.

Deployment can be broadly classified as indoor or open area which is further classified as blanket type, border type and point of interest type [6] [7]. There are scenarios where entire candidate region need to be monitored such as forest fire detection, in such cases SNs are positioned such that complete candidate region is covered and this type of deployment is called blanket type. In many cases particular region need to be isolated from intruders, thus a boundary is made around it by placing the SNs, which detects the movement of intruders, such a deployment is called border type. Even there are cases where only few point within a candidate region need to be monitored such a deployment is called point of interest based deployment.

Various researchers suggested the techniques for the uniform distribution of SNs within a candidate region but none of them considered the emergency conditions raised by natural calamity. In such cases the size of a candidate region is generally large and time is the major constraint to deal with a situation.

In this paper a model for precise and time efficient placement of SNs has been proposed. It is a pneumatic launcher based precise placement model (PLM) which uses a number of SN-launchers powered by the pressure of air. It is a generic model and can be used for the deployment of mobile or static SNs and can be used to effectively deploy SNs for any kind of deployment, such as barrier, blanket or point of interest based.

Rest of the paper is organized as follows. Section II outlines the related work. Preliminary is described in Section III. Section IV constitute of the proposed model. Simulation results are discussed in Section V followed by a conclusion in Section VI.

II. RELATED WORK

A lot has been done in the field of deployment of SNs. Initially, Andrew et al. [8] proposed a potential field based

method of uniform distribution of SNs which assumes that each SNs and obstacles possess a charge due to which they exert a repulsive force on each other. Thus the SNs relocate themselves in order to balance the force exerted by their neighbors thereby distributing themselves uniformly within a candidate region. A Virtual force driven deployment model was proposed by Zou et al. [9], which considers that both repulsive and attractive forces exist between the SNs, If the distance is less than the threshold then there exists a repulsive force if the distance is greater than the threshold then there exists an attractive force. Thus to balance the forces exerted by the neighbors the SNs relocate thereby distributing themselves uniformly within a candidate region. Both the schemes use mobile SNs (MSNs) and focuses on the uniform distribution of the SNs but least concern was given to the connectivity with the base station (BS). Connectivity Preserved virtual force (CPVF) and FLOOR based schemes were proposed by Guang et al. [10] to deal with the BS connectivity issues.

Corke et al. [11][12] proposed deployment model which uses a robot helicopter equipped with screw groove assembly to carry and precisely drop the SNs to predefined locations. Although, the model precisely deploys the SNs within a candidate region, it is not feasible to be used for large scale deployments, due to limited battery life and carrying capacity of the robot helicopter. Yoshiaki et al.[13], proposed a uniform aerial deployment (UAD) model to deploy SNs from air in large scale candidate region. It uses special parachute to carry a SN. The Parachutes are assumed to have a capability to switch between “gliding” and “falling” states in order to achieve the required density level. Although, UAD is an improved mechanism for aerial scattering, but it can only work for the SNs falling at the same level (altitude) and the design of a parachute is also not defined.

In [14] authors proposed a Centrifugal Cannon based Sprinkler (CCS) to randomly scatter the SNs within a candidate region. CCS is an assembly of variable sized cannons rotated by a motor at specified RPM. It alone cannot yield the coverage equivalent to optimal, as it is a random scattering model, but it provides an effective and time efficient method for random scattering of SNs over large unreachable regions.

In order to achieve the blanket coverage over a large-scale candidate region, most of the previously proposed state of art models either randomly scatters the SNs from the air or use MSNs, which are programmed to relocate to the optimal locations after random dropping. While, other uses robot helicopter to precisely place the SNs on the pre-computed, optimal locations within a candidate region. Among these models, although the random scattering model of SN deployment is the simplest, but it cannot yield optimal

coverage. Usage of MSNs In place of static SNs emerged as an effective solution for the optimal deployment problem of randomly scattered SNs but, MSNs are relatively costlier and have their own limitations of mobility in uneven and diverse terrain.

III. PRELIMINARY

A. Deployment helicopter

It traverses the entire candidate region in order to aerially deploy the SNs.

B. Deployment path

It is a virtual track on which the deployment helicopter moves while traversing the candidate region.

C. Path width

It is width of a strip on the candidate region covered by ECCS. It is equal to the twice of the horizontal distance covered by a SN launched from the longest cannon.

D. Base line

It is a virtual path above which the deployment helicopter moves, while traversing the candidate region.

E. Optimal deployment

Motivated from the cellular architecture of mobile networks, the entire candidate region is divided into hexagonal cells and center of these cells forms the optimal deployment locations (shown in Fig. 1).

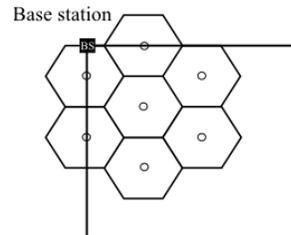


Fig. 1. Optimal deployment pattern

IV. PROPOSED MODEL

A. Model Assumptions

It is assumed that SNs are encapsulated within a spherical casing, in order to ensure the evenness in shape and to protect them from any kind of physical damage while landing. The deployment helicopter is equipped with precise positioning system. The density of air ρ is assumed to be constant, i.e., 1.255 Kg/m^3 . Mass, M of SNs is 0.250 Kg .

Prior information such as buildings, water bodies or any other structure where deployment is not required is available either in the form of digital map or satellite image.

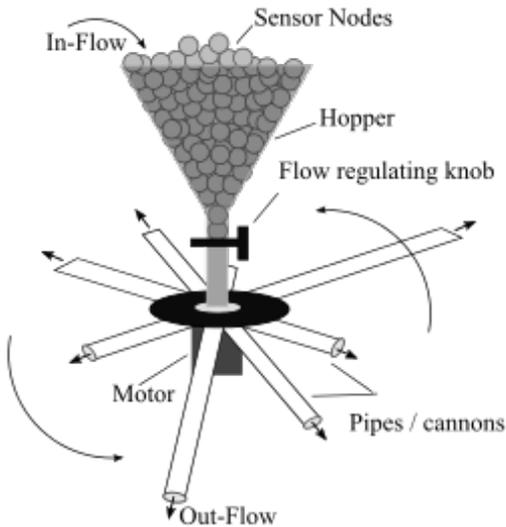


Fig. 2. Centrifugal cannon based sprinkler (CCS)

B. Centrifugal Cannon based Sprinkler (CCS)

CCS was designed as a time efficient and effective method to randomly scatter the SNs within a large scale candidate region. It constitute of assembly of variable sized cannons rotated by a motor to sprinkle the SNs within a candidate region (see Fig. 2). CCS is mounted on a deployment helicopter, which traverses the entire candidate region while following a predefined scan-path.

C. Pneumatic launcher based model (PLM)

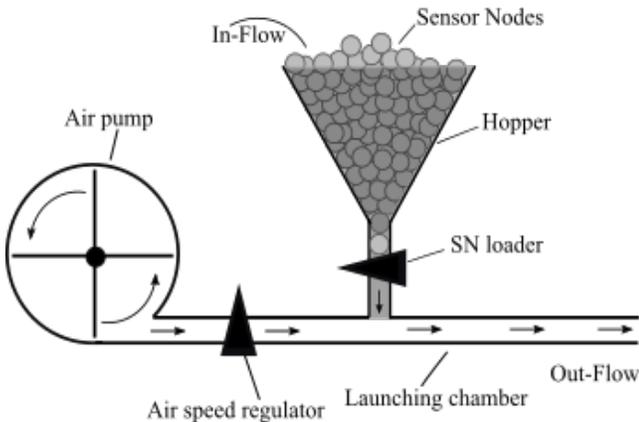


Fig. 3. Pneumatic launcher

Following are the prime components of a pneumatic launcher:

- **Hopper:** SNs are collectively held inside the hopper, from where they are sequentially loaded into the launching chamber.
- **Air pump:** It is a centrifugal pump, driven by a high speed DC motor. It blows the air through the launching chamber.
- **Launching chamber:** It is a pipe, connected to the hopper and air pump as shown in Fig. 3. It launched the SNs with the thrust of air pumped by the air pump.

- **Air speed regulator:** It controls the flow of air through the launching chamber, so as to ascertain the required launch velocity of a SN.
- **SN loader:** Loads the SN into the launching chamber.
- **SN capsule:** The SN is placed within a spherical shell, so as to ensure the evenness and alike shape of each SN. It is made up of two concentric spheres. The inner part is made up of shock absorbing material (e.g. sponge or thermocol) in order to cushion the SN and absorb the shocks generated while landing. The outer part is made up of a thin layer of hard and brittle material. The capsule is divided into two hemispheres (i.e., upper and lower), both containing a groove to pack a SN within. Bottom of a lower-half is filled with a sticky gel, in order to keep the bottom heavy for ensuring the landing position, absorb the landing shock and minimize the post landing movements of SN.

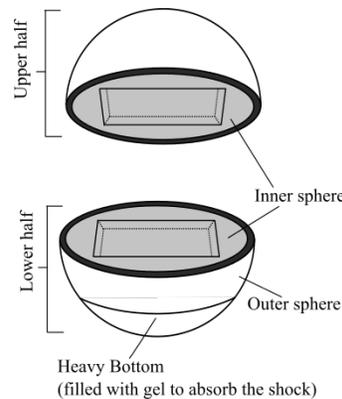


Fig. 4. SN capsule

1) **Pre-deployment configuration and computation:** The Entire candidate region is divided into hexagonal cells (regular hexagons with each side equal to r_s) to achieve the optimal coverage and center of these hexagons form the DLs for the placement of SNs. The cellular division of candidate region is motivated from cellular networks [15]. The relation between r_c and r_s is given by equation (1).

$$r_c = \sqrt{3} * r_s \quad (1)$$

The helicopter follows a pre-defined path in order to traverse the entire candidate region. The candidate region is marked by the master grid consisting of vertical and horizontal lines formed by joining the adjacent DLs as shown in Fig 5.

The horizontal lines are labeled as:

$$HB_{-n}^0, HB_{-n+1}^0, HB_{-n+2}^0, \dots, HB_0^0, HB_1^0, HB_2^0, \dots, HB_n^0, \\ HB_{-n}^1, HB_{-n+1}^1, HB_{-n+2}^1, \dots, HB_0^1, HB_1^1, HB_2^1, \dots, HB_n^1, \\ \dots, \\ HB_{-n}^k, HB_{-n+1}^k, HB_{-n+2}^k, \dots, HB_0^k, HB_1^k, HB_2^k, \dots, HB_n^k,$$

where n is the number of PLs in each set of PLM and k is the total number of parallel scan lines. The distance between adjacent horizontal lines d_h and adjacent parallel scan path d_p is given by equation (2) and (3) respectively. However, the total number of parallel-line scan paths is given by equation (4).

$$d_h = \frac{3r_s}{2} \quad (2)$$

$$d_p = 2 * n * \frac{3r_s}{2} \quad (3)$$

$$k = \frac{w}{ds} \quad (4)$$

where w is the width of a candidate region.

The helicopter moves above the parallel-line scan path labeled as HB_0^x . Vertical lines are labeled as B_1, B_2, \dots, B_m , where m is the total number of vertical lines and its value depends on the width of a candidate region. The distance d_v between adjacent vertical lines is given by equation (5).

$$d_v = \frac{\sqrt{3}r_s}{2} \quad (5)$$

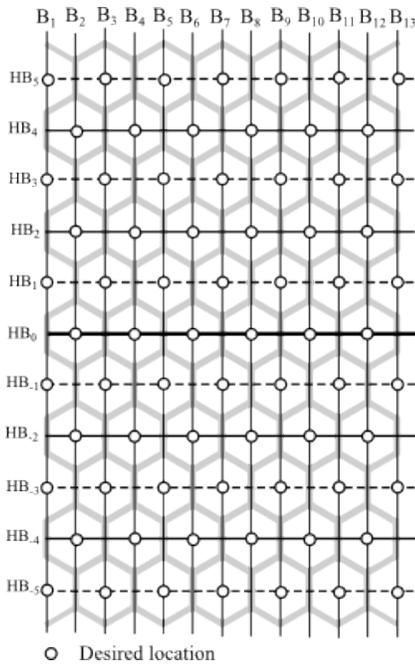


Fig. 5. Logical division of a candidate region

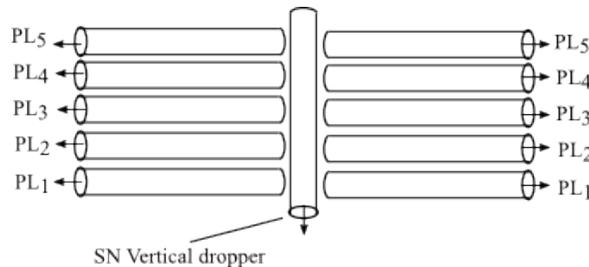


Fig. 6. Sets of PLs used in PLM

PLM consist of two sets of pneumatic launchers, $PL = \{PL_1, PL_2, PL_3, \dots, PL_n\}$, one on each side of a helicopter, all the pneumatic launchers are of equal lengths l but have variable range of launch velocities. Launch velocities VL_i of

any pneumatic launcher PL_i is given by the corresponding elements set VL .

$VL = \{VL_1, VL_2, VL_3, \dots, VL_n\}$, such that, $VL_1 < VL_2 < VL_3 < \dots < VL_n$.

The computation of value of VL_i is further discussed in this section.

2) Without air resistance: The time of flight, t of SN fired from PL_i is given by equation (2).

$$t = \sqrt{\frac{2H}{g}} \quad (6)$$

Where, H is a dropping height and g is the acceleration due to gravity ($g = 9.8$). Horizontal distance D_h covered by SN fired from PL_i is given by equation (3)

$$D_h = VL_i * \sqrt{\frac{2H}{g}} \quad (7)$$

3) With air resistance: In real scenarios, the air plays an important role in determining the trajectory of a launched SN. The launched SN stops accelerating vertically after achieving its terminal velocity v_t , due to the resistance offered by the air. Terminal velocity is a function of weight, radius of spherical SN and density of air (given in equation (8)). The relation between weight, radius and terminal velocity of a spherical SN is shown in shown in Fig. 7.

$$v_t = \sqrt{\frac{2Mg}{C_d \rho A}} \quad (8)$$

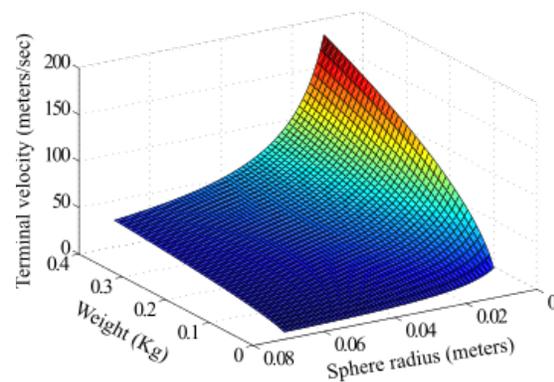


Fig. 7. Relation between radius, weight and terminal velocity of SN

Resistive force F_d is exerted by air on every SN fired from PLM, which reduces the horizontal distance covered by it. The value of F_d is given by equation (9).

$$F_d = \frac{1}{2} v^2 \rho A C_d \quad (9)$$

Where, v is the current speed of SN, A is the area of cross-section of SN's capsule, ρ is the density of air (i.e., 1.255 Kg/m^3) and C_d is a drag-coefficient for sphere (0.5). Fig. 8 (a) and (b) represents the trajectory formed by the SNs fired from PLM (with and without air resistance, respectively).

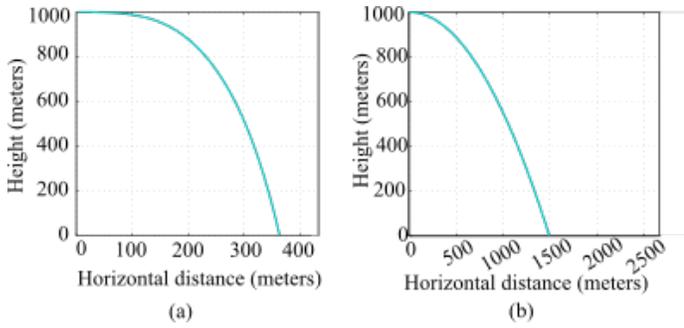


Fig. 8. (a). Trajectory formed by SN fired from PLM (With air). (b). Trajectory formed by SN fired from PLM (Without air)

Algorithm 1: Computation of launch distance of SNs (with air resistance).

```

getDistWithAir(H, M, A, vinit)
While H >= vdist
    Fv ← (ρ * A * Cd * vv2)/2
    av ← Fv/M
    vv ← vv + (g - av) * dt
    vdist ← vdist + dt * vv

    Fh ← (ρ * A * Cd * vinit2)/2
    ah ← Fh/M
    vinit ← vinit - ah * dt
    hdist ← hdist + vinit * dt

```

End

Return h_{dist}

End

4) *Computation of launch velocity*: Launch velocity is a speed with which the SN is launched from the PL. Each PL in a set is assigned different magnitude of launch velocity. It is adjusted such that the projectile formed by the launch from each PL ends on the corresponding HB.

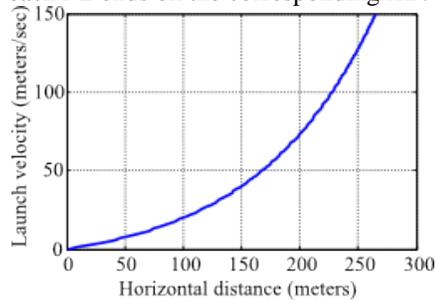


Fig. 9. Relation between VL and horizontal distance covered

Graph in Fig. 9 represents the relation between VL and horizontal distance covered by a SN. The deployment height considered to be 200 m. VL_i for particular PL_i is adjusted such that the SN launched by PL_i covers the horizontal distance D_i before hitting the ground. D_i is the distance between baseline and the horizontal line HB_i.

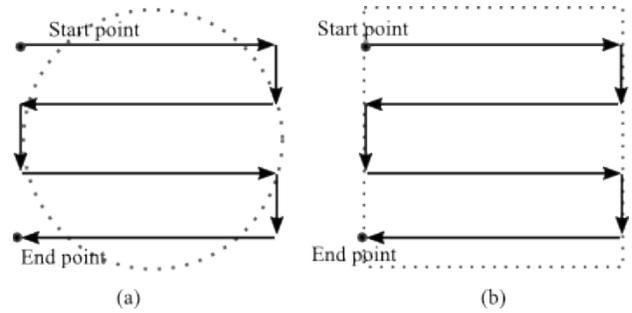


Fig. 10. (a) Scan path for circular candidate region. (b) Scan path for rectangular candidate region

5) *Scan Path*: It is a path defined for a deployment helicopter to traverse the candidate region while deploying the SNs [14]. In this paper we have considered the scan path in the form of horizontal parallel lines (see Fig. 10). SNs are precisely launched by the PLM while moving in the scan path. Operation of PLM is given by Algorithm 2.

6) *SN horizontal-launch regulation*: Launch regulation is an important task performed by PLM. It depends on the length of the PL_i and its VL_i. Each loaded SN passes through the PL_i in time interval

Each PL_i consumes a specific time T_i to launch a SN, called dispense time (given in equation (10)).

$$T_i = T_L + t_i \quad (10)$$

where T_L is the time taken by the actuator to load the SN into the cannon and t_i (given in equation (11)) is the time taken by the SN to pass through PL_i.

$$t_i = \frac{2l}{VL_i} \quad (11)$$

Vertical lines are categorized as real-vertical lines (RB_j) and virtual-vertical lines (VB_j). RB_j constitutes of actual lines which are plotted on the ground on the basis of computed DLs, while the VB_j is the copy of RB_j above the ground at height H, with a shift of distance D_s. The shift is opposite to the direction of movement of the helicopter as shown in Fig. 11 and its value is given by equation (12). It is done in order to compensate the displacement caused due inertia induced in the SN by the movement of the helicopter.

$$D_s = \text{getDistWithAir}(H, M, A, V_H) \quad (12)$$

where H is the altitude of a deployment helicopter, M is the mass of SNs, A is the cross-sectional area of the SN and V_H is the velocity with which the helicopter.

The errors are introduced while deployment of SNs due to various unavoidable factors such as environmental winds, humidity and temperature. These errors are called uncertainty errors E_u and their magnitude largely depends on the height of deployment. In this paper the value of E_u is considered as 7.5% of H.

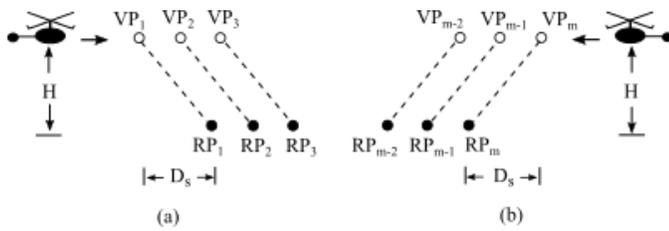


Fig. 11. (a). Vertical line-shift on left to right movement of deployment helicopter. (b). Vertical line-shift on right to left movement of deployment helicopter

7) *Exception handling*: Exceptions are the selected regions within a candidate region where deployment is not required these may include water bodies, buildings or any other structure or selected region. The DLs overlapping these regions are located on the digital map of a candidate region and removed from the set of DLs before feeding DLs to the PLM.

Algorithm 2: PLM operation

- *moveOnPath(speed, altitude)*: Moves the deployment helicopter above the scan path (HP_0^X) at specific altitude with specific speed.
- *getNextVPOnPath()*: Returns the next vertical line (VP_X) on the scan path.
- *getVPNumber()*: Returns the reference number of a vertical line.
- *dropSN()*: Simply drops the SN vertically without using the cannons.
- *loadEvenCannons()*: Loads cannons labeled with even number.
- *loadOddCannons()*: Loads cannons labeled with even number.
- *currentX*: X-coordinate at current position of deployment helicopter.

Thread 1

1. *startMovingOnPath(speed, altitude)*;
2. *setRPM(rpm)*;

Thread 2

```

Till endOfPath do
    If VP != getNextVPOnPath() Then
        VP ← getNextVPOnPath();
        If |currentX - VP.X| = speed * Ti Then
            If VP.getVPNumber%2=0 Then
                loadEvenCannons(VP);
                dropSNonPath();
            Else
                loadOddCannons();
            End
        End
    End
End
    
```

```

dropSNonPath(VP)
    While currentX != VP.X do
        Wait();
    End
    dropSN();
End
    
```

V. SIMULATION RESULTS AND DISCUSSION

The simulation of the proposed model has been performed Quorum Comm (our own simulator developed in java). Simulation is repeated 500 times and the average values are presented as the results. The values of variables used while simulation is given in TABLE I. Since this model is a unique of its kind, not much is available for comparison

Fig. 12 represents the coverage pattern of PLM as well as CCS. It is observed that the coverage achieved by PLM is very close to that of optimal. It uniformly covers the maximum part of a candidate region (see Fig. 12 (a)). However the coverage pattern achieved by CCS comparatively non-uniform (see Fig. 12 (b)).

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Communication range (r_c)	70m
Sensing range (r_s)	40m
Sprinkler RPM	1000
Height of deployment (H)	200m
SN radius	0.05m
SN weight	0.25Kg
Helicopter speed (V_H)	27.7 m/s (100 Km/h)
Uncertainty error (e_u)	7.5 % of H
Area of candidate region	1000 m X 1000 m
Number of SNs	250
Number of PLs in a set	5

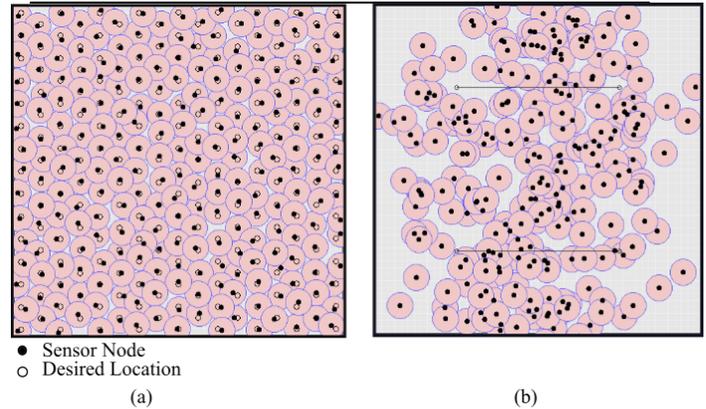


Fig. 12. (a) Coverage by PLM (b) Coverage by CCS

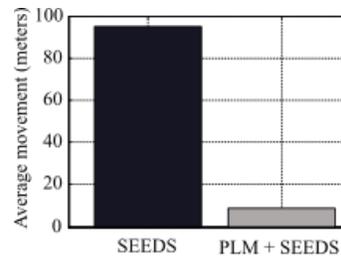


Fig. 13. Enhanced performance of SEEDS with PLM

Fig. 13 demonstrates the impact of PLM on SEEDS. SEEDS is a relocation based deployment scheme which uses MSNs to facilitate mobility. The average movement of MSNs in SEEDS is reduced from 95 m to 9.5 m when PLM is used to scatter the MSNs.

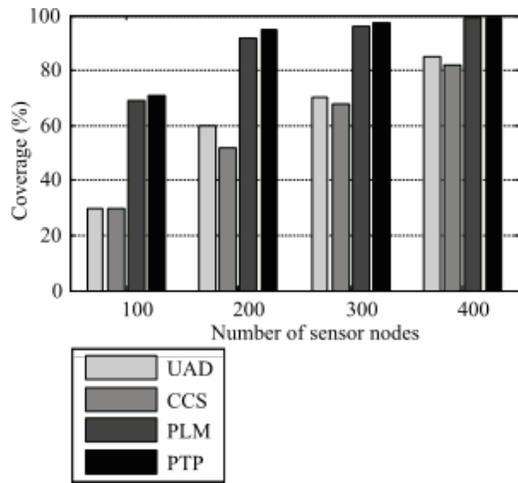


Fig. 14. Coverage achieved by various deployment models

The coverage achieved by PLM is very close to the optimal. Comparison of coverage achieved by various aerial deployment schemes is shown in Fig. 14.

Fig. 15 represents the comparison between the time taken by PTP and PLM to deploy the SNs. It is observed that PLM is 5.2 (approx) times faster than PTP deployment model and yields approximately same coverage. It is due to the fact that PLM covers the wide band on the candidate region in a single scan, thus minimizing the number of scans to deploy the SNs.

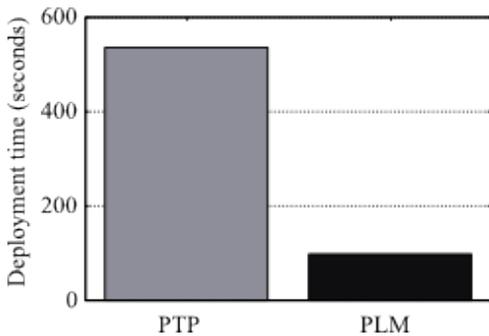


Fig. 15. Comparison between deployment time taken by PTP and PLM

VI. CONCLUSION

In this paper a model for time efficient and precise placement of SNs in large-scale candidate region has been proposed. It constitute of two sets of PLs, one on either side of a deployment helicopter. SNs are launched from these PLs with controlled velocity and time such that they land on the pre-computed locations. Each PL is governed by software which determines the launch time and velocity of a SN for its precise placement. Simulation results show that the proposed scheme is more time efficient, feasible and cost effective in

comparison to the existing state of art models of deployment. We are designing a hardware model of PLM for its hardware level testing and implementation. Moreover, the future enhancements of PLM include its generalization for all types of terrain.

REFERENCES

- [1] Kun Yang, "Wireless sensor networks," Principles, Design and Applications, 2014.
- [2] Mainwaring Alan, Polastre Joseph, Szewczyk Robert, and Culler David, "Wireless sensor networks for habitat monitoring," in WSN '02 Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications, New York, 2002, pp. 88-97.
- [3] G W Allen et al., "Deploying a Wireless Sensor Network on an Active Volcano," IEEE Internet Computing, vol. 10, no. 2, pp. 18-25, April 2006.
- [4] M Felamban, B Shihada, and K Jamshaid, "Optimal node placement in underwater wireless sensor networks," in 27th International Conference on Advanced Information Networking and Applications (AINA), 2013 IEEE, 2013, pp. 492-499.
- [5] Zhi Sun et al., "BorderSense: Border patrol through advanced wireless sensor networks," Ad Hoc Networks, vol. 9, no. 3, pp. 468-477, May 2011.
- [6] Vikrant Sharma, RB Patel, HS Bhadauria, and D Prasad, "Deployment schemes in wireless sensor network to achieve blanket coverage in large-scale open area: A review," Egyptian Informatics Journal, p. in press, 2015.
- [7] Shigeaki Tanabe, Kei Sawai, and Tsuyoshi Suzuki, "Sensor node deployment strategy for maintaining wireless sensor network communication connectivity," International Journal of Advanced Computer Sciences and Applications, vol. 2, no. 12, pp. 140-146, 2011.
- [8] Andrew Howard, Maja J Matari'c, and Gaurav S Sukhatme, "An incremental self-deployment algorithm for mobile sensor networks," Autonomous Robots, vol. 13, no. 2, pp. 113-126, 2002.
- [9] Yi Zou and Krishnendu Chakrabarty, "Sensor deployment and target localization based on virtual forces," in INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies, San Francisco, 2003, pp. 1293-1303.
- [10] Guang Tan, Stephen A Jarvis, and A-M Kermarrec, "Connectivity-guaranteed and obstacle-adaptive deployment schemes for mobile sensor networks," IEEE Transactions on Mobile Computing, vol. 8, no. 6, pp. 836-848, 2009.
- [11] P. Corke, S. Hrabar, R. Peterson, D. and Saripalli, S. Rus, and G. Sukhatme, "Autonomous deployment and repair of a sensor network using an unmanned aerial vehicle," in IEEE International Conference on Robotics and Automation, 2004, pp. 3602-3608.
- [12] Peter Corke et al., "Deployment and connectivity repair of a sensor net with a flying robot," Experimental robotics IX, pp. 333-343, 2006.
- [13] Yoshiaki Taniguchi, Tomoya Kitani, and Kenji Leibnitz, "A uniform airdrop deployment method for large-scale wireless sensor networks," International Journal of Sensor Networks, Inderscience, vol. 9, no. 3/4, pp. 182-191, 2011.
- [14] V Sharma, R B Patel, H S Bhaduria, and D Prasad, "Policy for random aerial deployment in large scale Wireless Sensor Networks.," in International Conference on Computing, Communication Automation (ICCCA), 2015, Noida, 2015, pp. 367-373.
- [15] Stuber and L Gordon, Principles of mobile communication, 3rd ed. Atlanta, USA: Springer, 2011.

Tree-Combined Trie: A Compressed Data Structure for Fast IP Address Lookup

Muhammad Tahir

Department of Computer Engineering,
Sir Syed University of Engineering and Technology,
Karachi

Shakil Ahmed

Department of Computer Engineering,
Sir Syed University of Engineering and Technology,
Karachi

Abstract—For meeting the requirements of the high-speed Internet and satisfying the Internet users, building fast routers with high-speed IP address lookup engine is inevitable. Regarding the unpredictable variations occurred in the forwarding information during the time and space, the IP lookup algorithm should be able to customize itself with temporal and spatial conditions. This paper proposes a new dynamic data structure for fast IP address lookup. This novel data structure is a dynamic mixture of trees and tries which is called Tree-Combined Trie or simply TC-Trie. Binary sorted trees are more advantageous than tries for representing a sparse population while multibit tries have better performance than trees when a population is dense. TC-trie combines advantages of binary sorted trees and multibit tries to achieve maximum compression of the forwarding information. Dynamic reconfiguration of TC-trie, made it capable of customizing itself along the time and scaling to support more prefixes or longer IPv6 prefixes. TC-trie provides a smooth transition from current large IPv4 databases to the large IPv6 databases of the future Internet.

Keywords—IP address lookup; compression; dynamic data structure; IPv6

I. INTRODUCTION

Improvement of Internet-based multimedia applications in recent years drives new demands for high-speed Internet. It seems that the demand for achieving higher bit-rates never saturates. Having the fast optical fiber technology for data transmission, data processing elements, i.e. routers, became main bottleneck of the current Internet speed. Inside a router, components that limit its speed are IP address lookup and classification engines. The main role of router is to forward millions of packets per second on each of its destination by finding address of next-hop router or the egress port through which packet should be forwarded. This forwarding decision is limiting the speed as there are millions of addresses and finding destination IP from millions of IPs is not an easy task. There is a need to have an algorithm for efficient IP lookup. Before we go to details, let's see how IP addressing architecture works and evolving. Reviewing it will help us to understand the address lookup problem. IP addressing architecture can be divided into two schemes; classful IP addressing scheme and classless IP addressing scheme. Classful IP scheme has two main issues; first, large number of IP addresses is wasted because of using IP address classes, second, the routing tables become very large. The growth of the forwarding tables resulted in higher lookup times and higher memory requirements in the routers and threatened to

impact their forwarding capacity. In order to resolve two main issues there are two possible solutions one is IPv6 IP addressing scheme and second is Classless Inter-domain Routing or CIDR.

Finding a high-speed, memory-efficient and scalable IP address lookup method has been a great challenge especially in the last decade (i.e. after introducing Classless Inter-Domain Routing, CIDR, in 1994). In this paper, we will discuss only CIDR. In addition to these desirable features, reconfigurability is also of great importance; true because different points of this huge heterogeneous structure of Internet have different traffic shapes and network topology changes along the time at each point as well.

This paper proposes a new cost-efficient data structure for fast IP address lookup that dynamically reconfigures itself. This novel data structure combines binary sorted trees with variable-stride multibit tries and put advantages of them all together in itself.

A. Paper Organization

Section 1 explains importance and related work. Section 2 explores the idea of TC-trie by examples and explains how to build TC-trie from a binary trie. Section 3 shows the experimental results of the TC-trie implementation and finally Section 4 concludes the paper and reveals our future works.

B. IP Address Lookup & Forwarding Tables

IP address lookup is a special search problem in a database of hundreds thousands of network addresses. Routers keep network addresses in their forwarding tables. For each incoming packet, the router finds a network address inside its table that matches with the destination IP address of that packet. Joint with each network address, there is a result field. The result could be simply an egress port of the router that packet should be exited from router via it to reach its destination network. Actually more information than an out port is required for forwarding a packet properly. This information includes next-hop layer-2 address, next hop layer-2 MTU (Maximum Transfer Unit), out port and so on. This information is kept in another table. This table can be called NHT (Next-Hop Table). Having the NHT, the result of the lookup could be a short length pointer to an entry of NHT. Fig. 1 shows an example of a forwarding table and an NHT. As this figure depicts, the forwarding table holds an 8-bit pointer corresponding to each network address.

A network address is a prefix of the 32-bit IP address. After introduction of CIDR (Classless Inter-Domain Routing) in 1994, network prefixes can be of any arbitrary length. In the classless routing, more than one network prefix may match with the destination IP address; in this case, the router must

choose the longest prefix that matches; so the IP lookup problem is known as a Longest Prefix Match (LPM) problem.

C. Trees & Tries

During the last ten years, many solutions have been proposed that issue the LPM problem. Simply, the IP

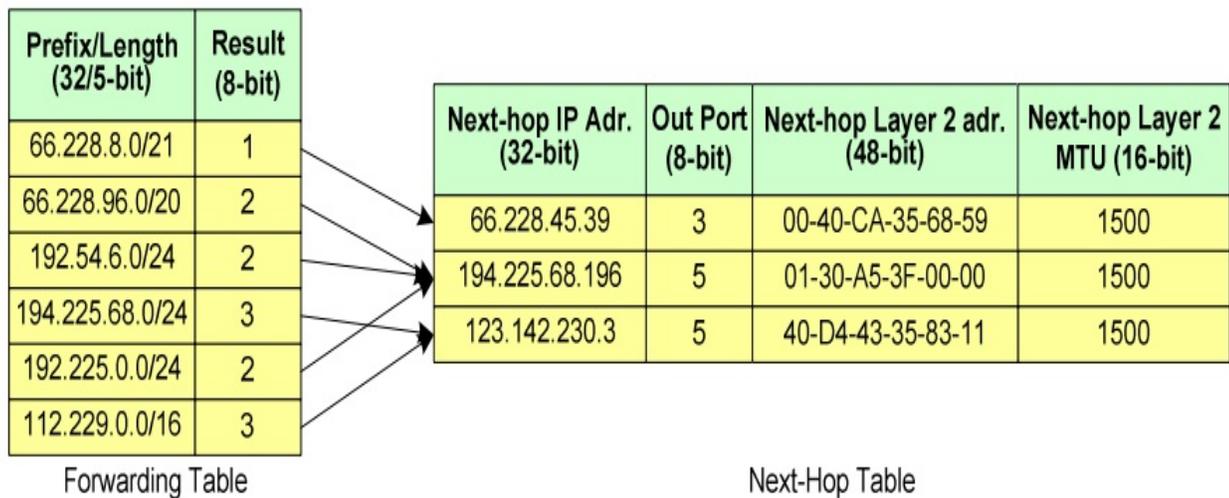


Fig. 1. simple forwarding table pointing to a Next-Hop Table (NHT)

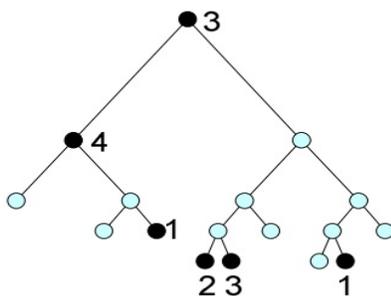
lookup solutions can be categorized as trie-based and tree-based methods. Fig. 2 shows trie and tree representations of a forwarding table in a supposedly 4-bit address space. In trees, information are hold explicitly in the nodes; so number of nodes is equal to the number of the network prefixes. If the binary tree to be balanced, its depth is ceiling $\lceil \log_2 N \rceil$ while N is the number of prefixes. In the opposite case, in tries, information are distributed on the edges. In a binary trie, each edge implicitly holds one bit of information. Left edges mean a zero bit and right edges mean a one bit. A path from the root to each node is a bit-string that corresponds to a network prefix. If this prefix exists in the table, the node should be labeled with corresponding result. Depth of a trie is equal to the length of the longest prefix that exists in the table.

binary tries, they are faster than binary tries but they suffer from rebalancing overhead.

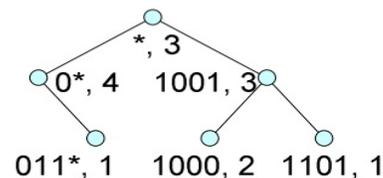
Memory consumption of trees and tries can be calculated by considering number of nodes in the data structure and size of each node. Number of nodes in tries is more than it in trees, because some nodes of the tries do not correspond to any valid prefix while in trees each node exactly keeps one prefix. The situation is different when considering the size of nodes. Since in the trees, prefixes are explicitly kept in the nodes, tree nodes are bigger than trie ones. Suppose that trie nodes are 32-bit while tree nodes are 64-bit. In the example of Fig. 2, memory consumption of the trie is 68 bytes while the memory consumption of the tree is 48 bytes; this means that in this example the tree is not only faster but also more compact than the trie. Fig. 3 shows another example for comparing trees with tries. In this example, the memory consumption of the trie is less than the one of the tree.

Address Prefix	Result
*	3
0*	4
011*	1
1000	2
1001	3
1101	1

A. Forwarding Table



B. Trie



C. Tree

Fig. 2. Trie and tree representations of a forwarding table. In this example, memory consumption of tree is less than trie

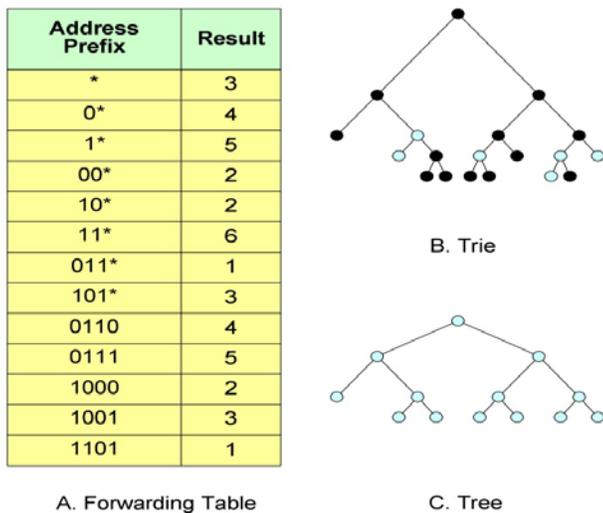


Fig. 3. Trie and tree representations of a forwarding table. In this example, memory consumption of tree is more than trie

When number of nodes in a trie is less than twice of number of tree nodes, trie consumes less memory than tree; in the other case, tree is more memory-efficient. As a rule of thumb, it can be concluded that trees are more compact than tries for sparse population of prefixes while when the prefixes are dense, tries are more compact.

D. Binary tries & Multibit Tries

Since a binary trie needs 32 memory accesses at the worst case for each address lookup, many solutions use multibit trie concept to accelerate the lookup search. In multibit tries, degree of nodes may be more than two. It means that an edge can hold more than one bit of information implicitly. Obviously, depths of multibit tries are less than binary tries and hence they are faster than binary tries. In fact, in multibit tries, the IP address segments into some strides. For example, the 32 bit IPv4 address can be segmented as 16-4-4-8. This segmentation corresponds to a four strides (or four levels) multibit trie with strides of length 16, 4, 4 and 8. The depth of a multibit trie is equal to the number of its strides. The memory consumption of a multibit trie depends to the number of strides and the length of each stride. Fig. 4 shows tree examples for comparing a tree, a binary trie and a one level multibit trie in a 4-bit address space. In the first example, the tree, in the second example, the trie, and in the third example, the multibit trie is the most memory efficient data structure. These examples demonstrate that the proper data structure should be chosen regarding the sparseness and the distribution shape of the prefixes in the address space. In general, it could not be said that trees are always more compact than tries or multibit tries consumes more memory than binary tries; the memory consumptions of trees, tries and multibit tries vary case by case. TC-trie is a dynamic mixture of trees, tries and multibit tries. Suppose a large 32-bit IP address space in your mind; in each part of this space, the TC-trie acts in a way that the data structure reaches the maximum compression ratio. It means that in a situation like

Fig. 4-A, the TC-trie would be a tree and in situations like Fig. 4-B and 4-C, the TC-trie would be a trie and a multibit trie respectively. Table 1 summarizes the results of the speed and memory-consumption comparison between trees, binary tries, and multibit tries of Fig. 4.

For a moment, forget combination of trees with tries and just consider the combination of tries with different multibit tries. This structure is a variable stride multibit trie. A variable stride multibit trie is a multibit trie that the number of strides and length of each stride vary in different paths from the root to the nodes. It could be said that the TC-trie is a variable stride multibit trie that its sparse parts are dynamically represented by binary sorted tree structures.

E. Related Works

Several IP address lookup solutions have been proposed previously which are based on tries. Some of them compress the trie for decreasing the memory consumption and improving the speed [1-2], [19-21]. Since a binary trie needs lots of comparisons and memory accesses, many solutions use multibit trie concept to accelerate the lookup search [3-13]; however, using a multi-bit trie instead of binary trie normally increases the memory consumption. Some papers have proposed compression methods for solving this problem [8]-[13]. Compression methods usually suffer from update overhead. Besides methods that work based on trie, other methods are also proposed that some of them use hash tables [14] while some others use CAMs (Content Addressable Memories) or TCAMs (Tertiary CAMs) for solving the LPM problem [15-17]. Some of these methods are not fast enough and some others consume lots of memory and those ones which have good speed and memory consumption suffer from heavy update overhead. A comparison between some of these methods comes in [18]. It seems that finding a method that meets all of the requirements of high-speed search, low memory consumption, and high-speed update is an endless challenge.

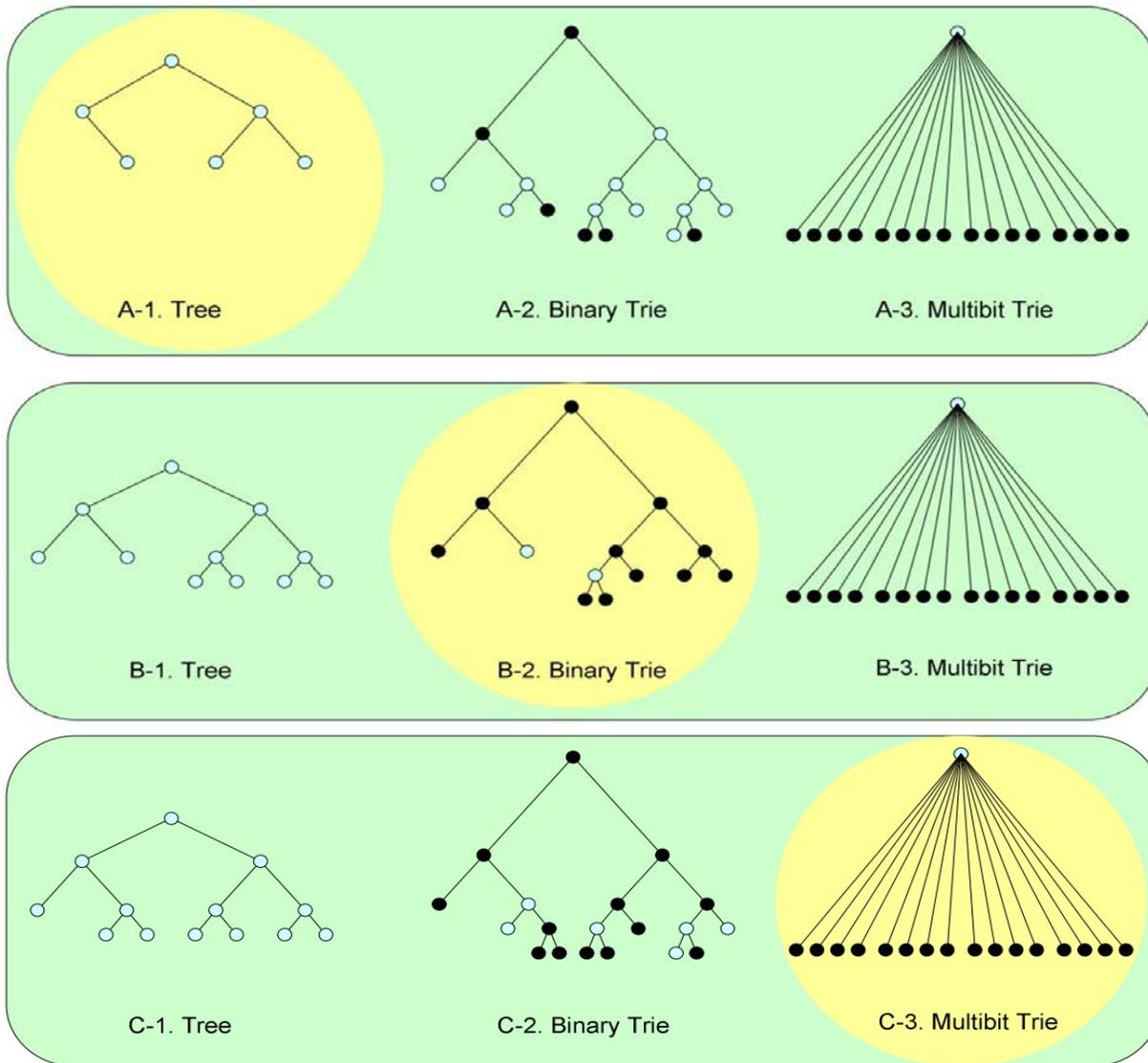


Fig. 4. Comparing the memory consumption of trees, binary tries and multibit tries in a 4-bit address space. A. Tree consumes less memory than trie and multibit trie. B. Binary trie consumes less memory than tree and multibit trie. C. Multibit trie consumes less memory than tree and binary trie

TABLE I. SPEED AND MEMORY CONSUMPTION COMPARISON BETWEEN TREES, BINARY TRIES AND MULTIBIT TRIES OF FIG. 4

		Number of Nodes	Size of Each Node (Byte)	Memory Consumption (Byte)	Maximum Depth
Example 1 (Fig. 4-A)	Tree	6	8	48	2
	Binary Trie	17	4	68	4
	MultibitTrie	16	4	64	1
Example 2 (Fig. 4-B)	Tree	11	8	88	3
	Binary Trie	13	4	52	4
	MultibitTrie	16	4	64	1
Example 3 (Fig. 4-C)	Tree	13	8	104	3
	Binary Trie	19	4	76	4
	MultibitTrie	16	4	64	1

II. TREE-COMBINED TRIE (TC-TRIE)

TC-trie is a flexible data structure that combines multibit tries with trees. Fig. 5 shows a binary trie representing the network prefixes in the 6-bit address space. We want to convert this structure to the TC-trie of Fig. 6. This conversion should be done in order to reduce the depth and memory consumption.

In our implementation for 32-bit IPv4 address space, the TC-trie starts with a stride of length 16-bit. Doing this, most significant 16 bits of address are considered at the first step and hence the depth of the structure never exceeds 17. Since 16 bits of each prefix are implicitly encoded in the first stride, all tree nodes should keep just the remaining 16 bits for each prefix. Therefore 16 bits from the prefix and 1 bit from the prefix length field will be saved.

A. Trie & Tree Nodes

For measuring the memory consumption, size of trie and tree nodes are required. Fig. 7 illustrates the trie and tree nodes. As this figure demonstrates a tree node in our architecture is exactly two times bigger than a trie node.

A trie node is composed of the following fields: Pointer: a pointer to a table in the next level that contains its children

nodes. Len: a value between 0 and 15 that shows the length of stride minus one. For a binary node, len equals to 0 and for a node with degree 16 (a node at the head of a stride with length 4), len equals to 3.

Result: for the nodes that contain a valid prefix, result is a pointer to the NHT (Next-Hop Table); for other nodes, it has the reserved value of "11111111"₂ that means no network prefix exists for this node.

The following fields consist in a tree node: Trie Pointer / Result: if a tree node to be at the head of a multibit trie cone, this field is a pointer to a table in the next level that contains its trie children. Otherwise, eight least significant bits of this field compose a pointer to the NHT (Next-Hop Table) while all the other bits are set to one. Len: if a tree node to be at the head of a multibit trie cone, this field shows the length of stride minus one. Tree Pointer: a pointer to a two-entry table in the next level that contains its tree children nodes. Prefix: the remaining least significant 16 bits of the prefix that corresponds to the tree node. Plen: length of the remaining part of the prefix minus one.

The following sub-sections explain how to build TC-trie from a binary trie and how to search it. The explanations about the incremental update and IPv6 implementation are ignored due to the lack of space.

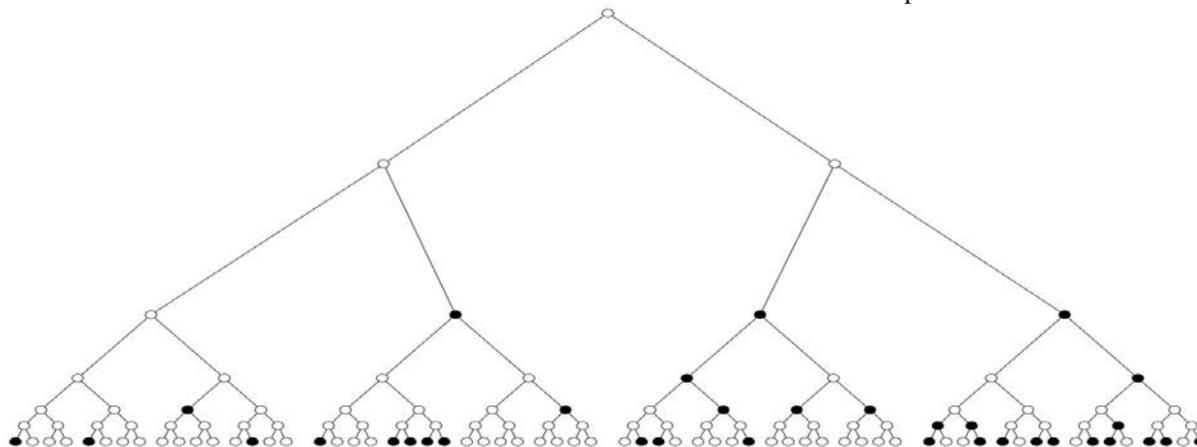


Fig. 5. A binary trie representing the network prefixes in a 6-bit address space

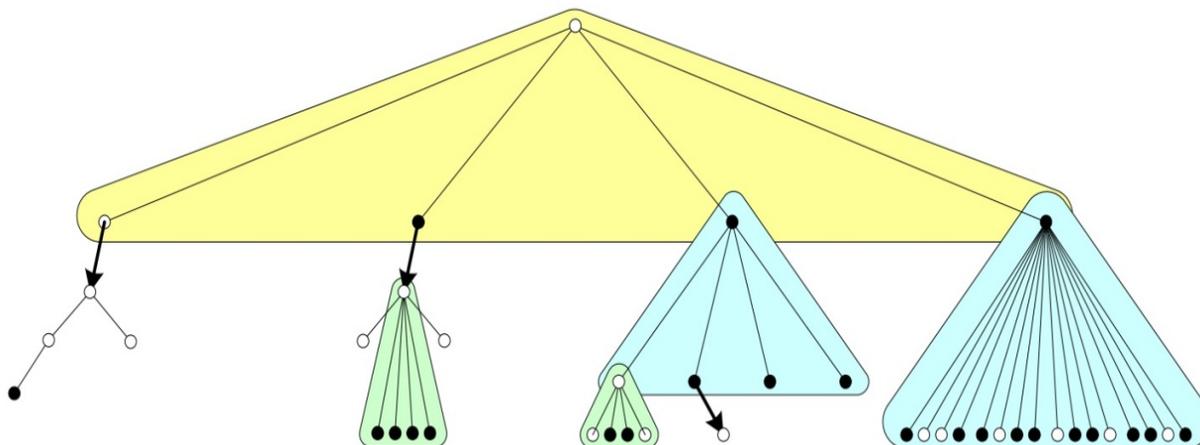


Fig. 6. A TC-trie equivalent to the binary trie of Fig. 5

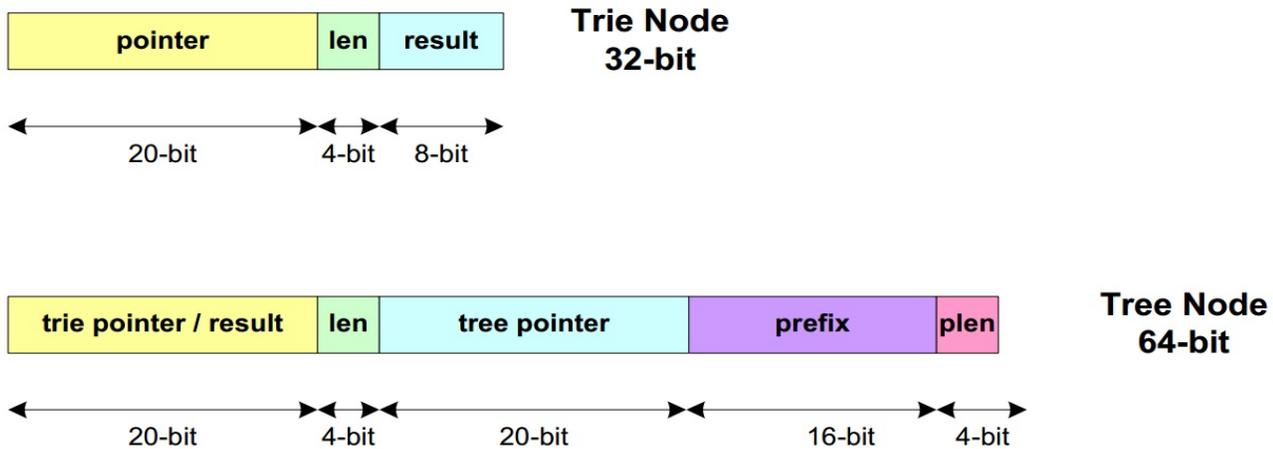


Fig. 7. Trie and tree nodes

B. Building TC-trie

For building a TC-trie, the starting point is a binary trie which is kept in the control unit of the router. The control unit (sometimes called slow path) of a router usually is a GPP (General Purpose Processor) that runs a routing protocol like RIP, OSPF or BGP. The control unit is responsible for updating the forwarding table of the lookup engine. Since CPE (Controlled Prefix Expansion) [6] in multibit tries and many compression techniques used in other IP lookup methods removes parts of information, the control unit must have an original copy of information inside itself to do the update operation properly. The control unit uses DRAM and update doesn't occur very frequently, so the size and speed of the original structure in the control unit is not critical. In our implementation, the control unit keeps a binary trie that contains the original non-scratched information. For building a TC-trie at the first time or for incrementally updating it, the control unit uses its binary trie structure.

To build a TC-trie from a binary trie, two main steps should be followed. The first step is finding dense regions and representing them with multibit tries. In the second step, prefixes which are not covered by multibit tries must be represented by binary sorted trees. When searching for dense regions, two issues should be considered. The first issue is the search resolution. Resolution equal to one means the maximum resolution that yields the most accurate results. Resolution equal to two means that searching the binary trie is being done with step size two. So the resulting multibit tries would be of depth 2, 4, 6 and etc. In general, if resolution equals to r , the depth of all multibit tries which are obtained would be a multiple of r .

The second issue is the threshold between denseness and sparseness concepts. How many prefixes have to be in a region of a trie to call it a dense region? To answer this question both memory consumption and lookup speed should be considered. Suppose that the resolution is four and we want to find out whether a cone with depth four in the original binary trie has the essential condition for being a stride of depth four or not. A stride of depth four needs 16 trie nodes that consumes $16 * 4 = 64$ bytes of memory. On the other hand, 64 bytes is equal to 8 tree nodes. So, if the number of prefixes is less than eight, tree representation is more compact; otherwise, a stride of depth four is better. Therefore, by considering only the memory constraint, it could be said that a binary trie of depth d is dense if it contains at least 2^{d-1} prefixes. We refer to this threshold as 50% threshold.

It's clear that a single stride is faster than any tree structures. So, if speed to be considered in addition to the memory consumption, the threshold should be less than 50%. Since compressing the higher parts of the trie improves the lookup speed of more prefixes, it's wise to apply different threshold for different heights of the trie. In other words, it's reasonable to increase the threshold from top to the bottom of the trie up to 50%.

C. Lookup Search

Fig. 8 illustrates an example of a lookup search on a TC-trie in a 6-bit memory space. This TC-trie is the one that was shown in Fig. 6. Notice how tree and trie nodes filled the memory space. Each word of the memory can be filled with one tree node or two trie nodes. Since each stride of multibit trie always has even number of nodes, no part of the memory space would be dissipated.

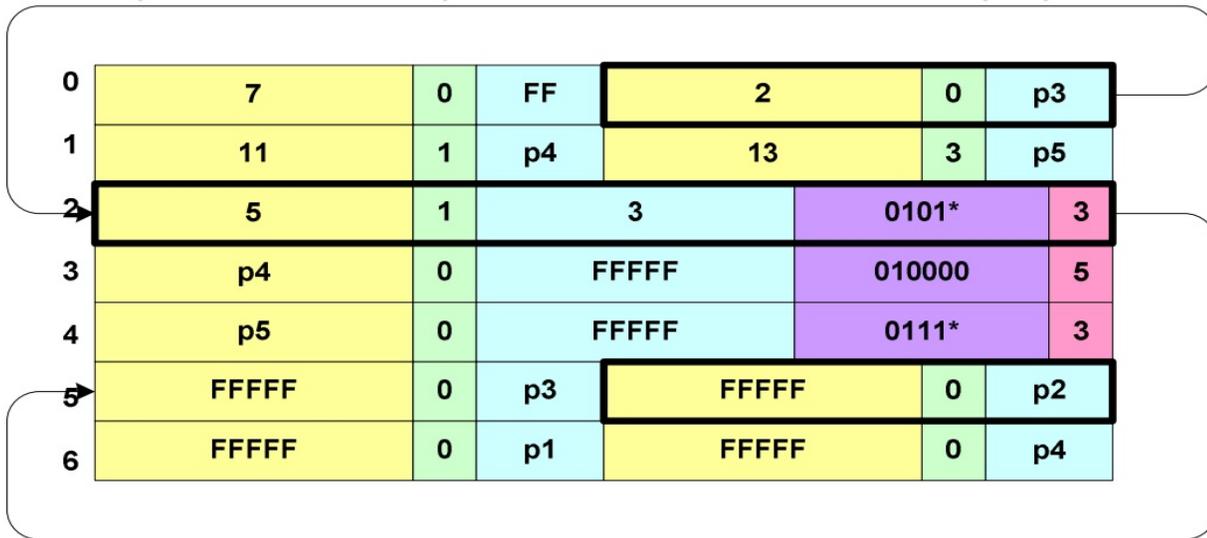


Fig. 8. An example of a lookup search in the memory architecture of a TC-trie structure

TABLE II. MEMORY CONSUMPTION, AVERAGE DEPTH, AND MAXIMUM DEPTH OF FIVE DIFFERENT FORWARDING TABLES¹. THESE RESULTS ARE OBTAINED WITH RESOLUTION 2 AND 50% THRESHOLD VALUE

Table Name	Table Size	Memory Consumption (Byte)	Maximum Depth	Average Depth
AS1221	156535	1303024	11	3.98
AS267	134024	1136144	8	3.50
AS286	134236	1137232	8	3.49
AS3333	139033	1170904	10	3.53
AS3549	133869	1135136	8	3.51

TABLE III. MEMORY CONSUMPTION, AVERAGE DEPTH, AND MAXIMUM DEPTH OF AS1221¹ FOR RESOLUTIONS. THESE RESULTS ARE OBTAINED WITH 50% THRESHOLD VALUE

Resolution Step	Memory Consumption (Byte)	Maximum Depth	Average Depth
2	1303024	11	3.98
4	1357288	11	3.92
6	1392992	8	3.95

TABLE IV. MEMORY CONSUMPTION, AVERAGE DEPTH, AND MAXIMUM DEPTH OF AS1221¹ FOR DIFFERENT THRESHOLD VALUES. THESE RESULTS ARE OBTAINED WITH RESOLUTION 2

Threshold value	Memory Consumption (Byte)	Maximum Depth	Average Depth
50%	1303024	11	3.98
25%	1969456	11	3.27
Variable	1890496	12	3.32

In this example, the query address is "010101". Fig. 6 shows how trie and tree nodes must be traversed for searching this address. In Fig. 8, nodes that must be read are highlighted. The final result is the value p2.

III. EXPERIMENTAL RESULTS

Different experiments have been done based on different forwarding tables, different resolutions, and different threshold values. Table 2 shows the memory consumption, average depth, and maximum depth of TC-trie structure achieved for five forwarding tables. These forwarding tables are obtained via potaroo website [22].

Table 3 shows the effect of changing the resolution in the memory consumption and depth of the forwarding table¹ BGP routing table analysis reports: <http://bgp.potaroo.net/>, retrieved on January 2000 & December 2014. AS1221 [22].

Table 3 shows the effect of changing the resolution in the memory consumption and depth of the forwarding table AS1221 [22].

In Table 4, the effects of changing the tree-trie threshold

are illustrated. This table shows that higher threshold (up to 50 %) yields lower memory consumption, while lower threshold yields smaller depth. In this table the last row stands for a variable threshold. This variable threshold increases from the root of the trie (original binary trie) to the leaves. The variable threshold causes higher parts of the trie have more chance of being compressed.

IV. CONCLUSION AND FUTURE WORKS

A new data structure for fast and memory-efficient IP address lookup was presented. This structure, which is a variable stride multibit trie combined with binary sorted tree was called Tree-Combined Trie (TC-trie). TC-trie collects benefits of multibit tries and binary sorted trees in itself. Dynamic reconfigurability of TC-trie made it a very flexible data structure that is scalable to the number of prefix, prefix distribution and prefix length. The proposed data structure prepares a smooth transition from IPv4 toward IPv6. Different aspects of this new data structure were considered and the building procedure and lookup search in this structure were explained. Examples and experiments demonstrated that our method consumes less memory than trie-based methods and is faster than tree-based methods. The flexibility of TC-trie is more than other methods and it better fulfills the requirements of current unsteady Internet. Our future work can be outlined as follows:

- Finding a better tree structure for combining with multibit tries by considering the following issues:
 - i. Multiway trees
 - ii. The sorting mechanism of the tree nodes
- Doing more theoretical and experimental studies about a threshold point between trees and tries.
 - i. Best static threshold conditions regarding the memory consumption and speed
 - ii. Dynamic threshold conditions
- Adding more intelligent to the system during the TC-trie build up.
 - i. How to assign priorities to the tree nodes to sort them in a way that the average TC-trie depth to be minimum.
- Simulating a scenario of growing IPv6 tables in an actual condition that may occur in the future.

ACKNOWLEDGMENTS

This work has been supported by the Saeed Shamshiri.

REFERENCES

- [1] D. R. Morrison, "PATRICIA - Practical algorithm to retrieve information coded in alphanumeric," *J. ACM*, vol. 15, no. 4, pp. 514–34, Oct. 1968.
- [2] K. Sklower, "A tree-based packet routing table for Berkeley UNIX," *Proc. 1991 Winter Usenix Conf.*, pp. 93–99, 1991.
- [3] P. Gupta, S. Lin, and N. McKeown, "Routing lookups in hardware at memory access speeds," *Proc. IEEE INFOCOM '98*, pp. 1240–47, Apr. 1998.
- [4] Tomas Henriksson, Ingrid Verbauehede, "Fast IP address lookup engine for SOC integration," *Proc. of Design and Dignostics of Electronic Cricuits and Systems, Brno, Czeck Republic*, pp. 200-210, Apr 2002.
- [5] Chen, W.E.; Tsai, C.J. "A fast and scalable IP lookup scheme for high-speed networks," *Proc. IEEE ICON99*, pp. 211-218, 1999.
- [6] V. Srinivasan and G. Varghese. "Fast address lookups using controlled prefix expansion," *ACM Transactions on Computer Systems*, vol. 17, no. 1, pp. 1-40, Feb. 1999.
- [7] T. Chiueh and P. Pradhan, "High performance IP routing table lookup using CPU caching," *Proc. IEEE INFOCOM'99, New York, NY, USA*, pp. 1421-1428, April 1999.
- [8] KARI SEPPANEN, "Novel IP address lookup algorithm for inexpensive hardware implementation", *WSEAS Transactions on Communications*, vol. 1, no. 1, pp. 76-84, 2002.
- [9] Nen-Fu Huang, Shi-Ming Zhao, Jen-Yi Pan, and Chi-An Su, "A fast IP routing lookup scheme for gigabit switching routers", *Proc. IEEE INFOCOM*, pp. 1429-1436, Mar. 1999.
- [10] Stefan Nilsson, Gunnar Karlsson, "Fast address lookup for internet routers", *Proc. IFIP 4th International Conference on Broadband Communications*, pp. 11-22, 1998.
- [11] S. Nilsson and G. Karlsson "IP-address lookup using LC-tries," *IEEE JSAC*, vol. 17, no. 6, pp. 1083–92, June 1999.
- [12] M. DegerMark, et al., "Small forwarding tables for fast routing lookups," *Proc. ACM SIGCOMM 97*, pp. 3-14, 1997.
- [13] Derek Pao, Cutson Liu, Angus Wu, Lawrence Yeung and K. S. Chan, "Efficient hardware architecture for fast IP address lookup," *IEE Proceedings-Computers and Digital Techniques*, vol. 150, no. 1, pp. 43-52, Jan. 2003.
- [14] M. Waldvogel, G. Varghese, J. Turner, B. Plattner, "Scalable high speed IP routing lookups", *Proc. ACM SIGCOMM '97*, pp. 25–36, Sept. 1997.
- [15] Huan Liu, "Routing table compaction in ternary CAM", *IEEE Micro*, vol. 22, no. 1, pp.58-64, January 2002.
- [16] Francis Zane, Girija Narlikar, Anindya Basu, "CoolCAMs: power-efficient TCAMs for forwarding engines", *IEEE INFOCOM*, vol. 1, pp. 42-52, 2003.
- [17] Anthony J. McAuley, Paul Francis, "Fast routing table lookup using CAMs", *Proc. IEEE INFOCOM*, pp. 1382-1391, March/April 1993.
- [18] Miguel Á. Ruiz-Sánchez, Ernst W. Biersack and Walid Dabbous, "Survey and taxonomy of IP address lookup algorithms," *IEEE Network Magazine*, vol. 15 no. 2, pp. 8-23, March/April 2001.
- [19] Yi-Mao Hsiao , Yuan-Sun Chu, Jeng-Farn Lee, Jinn-Shyan Wang, "A high-throughput and high-capacity IPv6 routing lookup system" , *Computer Networks*, Elsevier, 2013.
- [20] KunHuang , GaogangXie , YanbiaoLi , DafangZhang, "Memory-efficient IP lookup using trie merging for scalable virtual routers", *Computer Networks*, Elsevier, 2014.
- [21] Hyuntae Park, Hyejeong Hong, Sungho Kang, "An efficient IP address lookup algorithm based on a small balanced tree using entry reduction", *Computer Networks*, Elsevier, 2012.

Performance Evaluation of K-Mean and Fuzzy C-Mean Image Segmentation Based Clustering Classifier

Hind R.M Shaaban / Professor
Faculty of Computer Science and
Mathematics
University of Kufa
Najaf, Iraq

Farah Abbas Obaid
Faculty of Computer Science and
Mathematics
University of Kufa
Najaf, Iraq

Ali Abdulkarem Habib / Assistant
Lecture
Faculty of Computer Science and
Mathematics
University of Kufa
Najaf, Iraq

Abstract—This paper presents Evaluation K-mean and Fuzzy c-mean image segmentation based Clustering classifier. It is followed by thresholding and level set segmentation stages to provide an accurate region segment. The proposed stay can get benefits of the K-means clustering

The performance and Evaluation of the proposed image segmentation approach was evaluated by comparing K-mean and Fuzzy c-mean algorithms in case of accuracy, processing time, Clustering classifier, and Features and accurate results performance.

The database consists of 40 images executed by K-mean and Fuzzy c-mean image segmentation based Clustering classifier. The experimental results confirm the effectiveness of the proposed Fuzzy c-mean image segmentation based Clustering classifier. The statistical significance Measures of mean values of Peak signal-to-noise ratio (PSNR) and Mean Square error (MSE) and discrepancy used to Performance Evaluation of K-mean and Fuzzy c-mean image segmentation.

The algorithm higher accuracy can be found by increasing number of Clustering classifier and with Fuzzy c-mean image segmentation.

Keywords—Segmentation; image segmentation; Evaluation image Segmentation; K-means clustering; Fuzzy C-means

I. INTRODUCTION

Segmentation plays an integral part in partitioning an image into sub regions with respect to a particular application. The image might be having certain characteristics like gray level gray level, color intensity, texture information, depth or motion based on the measurement. The traditional methods used for the medical image segmentation are Clustering, threshold, region based Segmentation, edge based methods and ANN image Segmentation [1].

Image segmentation methods can be classified into three categories: edge based methods, region based methods and pixel based methods .K-Means clustering is a key technique in pixel based methods [2].

Fuzzy K-Means (also called Fuzzy C-Means) is an extension of K-Means , the popular simple clustering technique. While K-Means discovers hard clusters (a point belong to only one cluster), Fuzzy K-Means is a more

statistically formalized method and discovers soft clusters where a particular point can belong to more than one cluster with certain probability[3].

The goal of image segmentation is to cluster pixels into salient image regions such as individual surfaces, objects, natural parts of objects. Clustering technique can be used for image segmentation. Clustering in image segmentation is the process of identifying groups of related images. To achieve the super pixel formation, many clustering techniques can be classified. The purpose of using clustering technique is to get proper result with high efficiency effective storage image [4].

The paper is organized as follows; Section 2 deals with K-Means and Fuzzy C-Means, section3 the proposed method with results is introduced, in section4 Experimental Results, and the conclusion of this study is given in section 5.

II. K-MEANS AND FUZZY C-MEANS [5,6,7,8]

The clustering Algorithms groups a sample set of feature vectors into K clusters via an appropriate similarity or dissimilarity criterion.

The k-means algorithm assigns feature vectors to clusters by the minimum distance assignment principle, which assigns a new feature vector $\mathbf{x}^{(n)}$ to the cluster $\mathbf{c}^{(k)}$ such that the distance from $\mathbf{x}^{(n)}$ to the center of $\mathbf{c}^{(k)}$ is the minimum over all K clusters. The basic k-means algorithm is as follows:

- Put the first K feature vectors as initial centers
- Assign each sample vector to the cluster with minimum distance assignment principle.
- Compute new average as new center for each cluster
- If any center has changed, then go to step 2, else terminate.

Fuzzy clustering plays an important role in solving problems in the areas of pattern recognition and fuzzy model identification. A variety of fuzzy clustering methods have been proposed and most of them are based upon distance criteria. One widely used algorithm is the fuzzy c-means (FCM) algorithm. It uses reciprocal distance to compute fuzzy weights.

Fuzzy C-means Clustering (FCM) is an clustering method which is separated from k-means that employs hard partitioning (FCM is an iterative algorithm). The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. The aim of FCM is to find cluster centers that minimize a dissimilarity function.

To accommodate the introduction of fuzzy partitioning, the membership matrix(U) is randomly initialized according to Equation (1)

$$\sum_{i=1}^c u_{ij} = 1, \forall j = 1, \dots, n \quad (1)$$

The dissimilarity function which is used in FCM is given Equation (2)

$$J(U, c_1, c_2, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 \quad (2)$$

u_{ij} is between 0 and 1

c_i is the centroid of cluster i ;

d_{ij} is the Euclidian distance between i th centroid(c_i) and j th data point;

$m \in [1, \infty]$ is a weighting exponent.

To reach a minimum of dissimilarity function must find two conditions These are given in Equation (3) and Equation

$$c_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (3)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (4)$$

Detailed algorithm of fuzzy c-means proposed by Bezek in 1973[5]. This algorithm determines the following steps:

Step 1. Randomly initialize the membership matrix (U) that has constraints in Equation (1).

Step 2. Calculate centroids(c_i) by using Equation (3).

Step 3. Compute dissimilarity between centroids and data points using equation (2). Stop if its improvement over previous iteration is below a threshold.

Step 4. Compute a new U using Equation (4). Go to Step 2.

Performance depends on initial centroids Because of cluster centers (centroids) are initialize using U that randomly initialized.(Equation 3) the FCM does not ensure that it converges to an optimal solution.

III. THE PROPOSED METHOD

The proposed method has been applied using gray scale images size (256*256) ,format are(.tiff and .png) a detailed experimental comparison of the above stated study has been presented. We have used gray image databases. Figure (1) shows sample data base for astronomical images, which are used in this paper.

Data base for paper contain 40 brain images applied for all the K-mean and Fuzzy c-mean image segmentation based Clustering classifier.

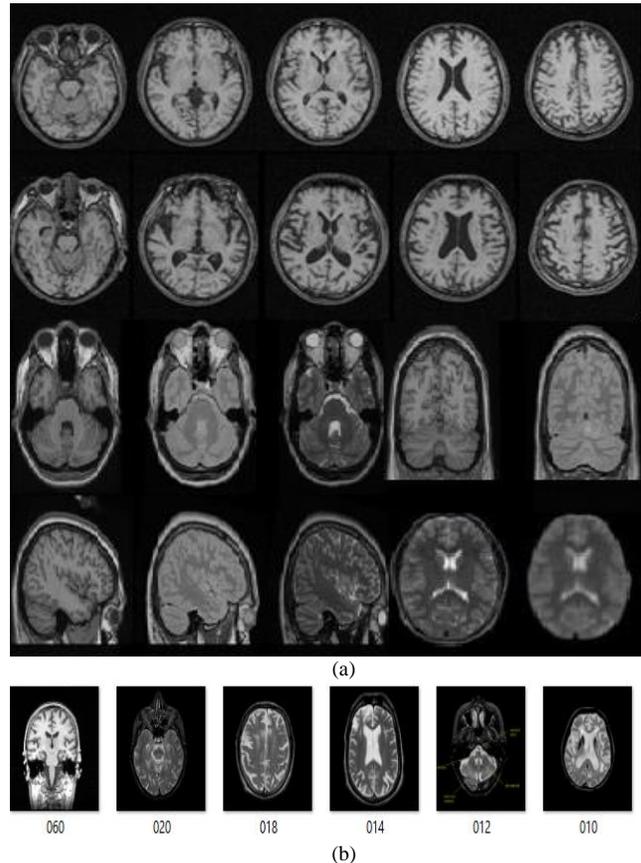


Fig.1. Sample Data Base for brain Images (b) image with its number Appling in paper

The flowchart for system show in figure (2)

The following flowchart showed the sub-key work processes through which they are determined to best work distinctive characteristics and the way.

In end of flowchart Analysis of the results and determine the best algorithm and ask if there is another image to testing.

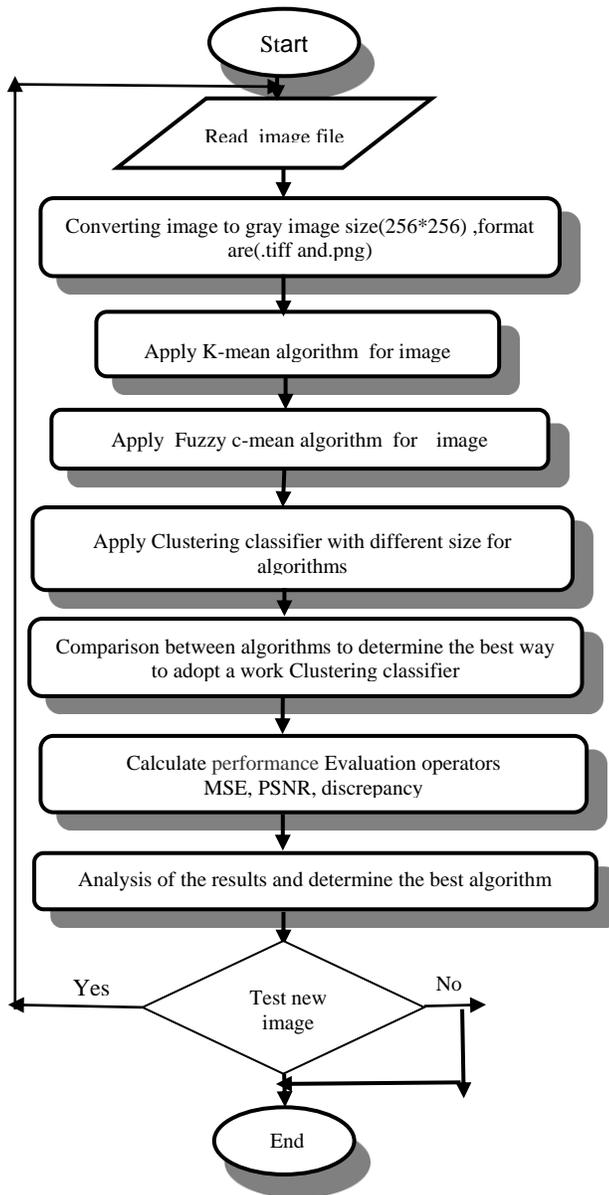


Fig. 2. flowchart for system

IV. EXPERIMENTAL RESULTS

In this section, the results are presented which are obtained by applying and evaluation K-mean and Fuzzy c-mean image segmentation.

The statistical significance Measures of mean values of Peak signal-to-noise ratio (PSNR) and Mean Square error (MSE) and discrepancy use to Performance Evaluation of K-mean and Fuzzy c-mean image segmentation based Clustering classifier

Peak signal-to-noise ratio, often abbreviated PSNR, is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale.

$$PSNR = 10 \log_{10} \frac{(L-1)^2}{\frac{1}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} [f(x,y) - f^*(x,y)]^2} \quad (5)$$

where L is the number of gray levels(e.g., for 8 bits L=256).

$f(x,y)$: the original image, $f^*(x,y)$: the decompressed image, x, y : row and column[10].

Mean Squared Error (MSE) of an estimator measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss[11].

suppose that we measure the quality of t , as a measure of the center of the distribution, in terms of the *mean square error*

$$MSE(t) = \frac{1}{\#} \sum_{i=1}^k f_i (x_i - t)^2 = \sum_{i=1}^k p_i (x_i - t)^2 \quad (6)$$

MSE(t) is a weighted average of the squares of the distances between t and the class marks with the relative frequencies as the weight factors. Thus, the best measure of the center, relative to this measure of error, is the value of t that minimizes MSE[10].

Calculate Discrepancy by Equation (7)

$$Discrepancy = \sum_i^{I_n} \sum_j^{I_w} (C_{gt}(i,j) - L(i,j)) \quad (7)$$

Where $C_{gi}(I_i,j)$ is the gray level value of pixel $p(I_i,j)$ on original image and $L(I_i,j)$ is the gray level value of pixel on the image after thresholding[13].

$$E_{intra} = \frac{\sum_{p \in I} \mu (\| C_x^0(p) - C_x^s(p) \|_{L^*a*b} - TH)}{S_I}$$

Where $C_x^0(p)$ and $C_x^s(p)$ are pixel feature value(color components in CIEL*a*b space) for pixel p on original and segmented image respectively, TH is the threshold to judge significn diffrence, and $\mu(t) = 1$ when $t > 0$, otherwise $\mu(t) = 0$.

From the experiments results, which they illustrated in table (1) showed MSE & PSNR with K-Means Clustering for five images,

TABLE I. SHOWED MSE & PSNR FOR K-MEANS CLUSTERING

IMAGE	MSE	PSNR
010	21.1371	29.5471
012	13.4794	33.0530
014	33.2907	26.4325
018	23.8972	27.2781
020	17.6975	26.3113
060	33.1643	30.4140

Figure(3) showed Recurring planned rates for MSE & PSNR with K-Means Clustering for five images,



Fig. 3. Recurring planned rates for MSE & PSNR with K-Means Clustering

table (2) showed MSE & PSNR with Fuzzy C-Means for five images,

TABLE II. SHOWED MSE & PSNR FOR FUZZY C-MEANS

IMAGE	MSE	PSNR
010	0.7151	1.4564
012	59.5771	1.295
014	63.3936	1.6431
018	0.6960	1.5740
020	0.7174	1.4426
060	0.6441	1.9105

Figure(4) showed Recurring planned rates for MSE & PSNR with Fuzzy C-Means for five images.

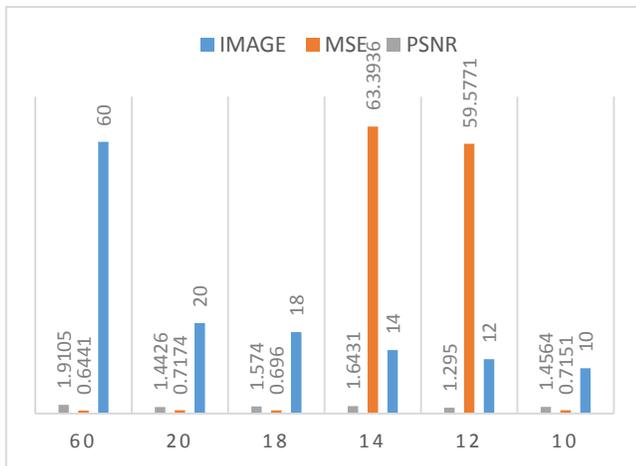


Fig. 4. Recurring planned rates for MSE & PSNR with Fuzzy C-Means for five images

Table (3) showed Discrepancy with K-Means Clustering for five images,

TABLE III. DISCREPANCY WITH K-MEANS CLUSTERING

Image	Disc
010	24980
012	-61750
014	-23515
018	13958
020	3495
060	18242

Figure (5) showed Recurring planned rates Discrepancy K-Means Clustering algorithm

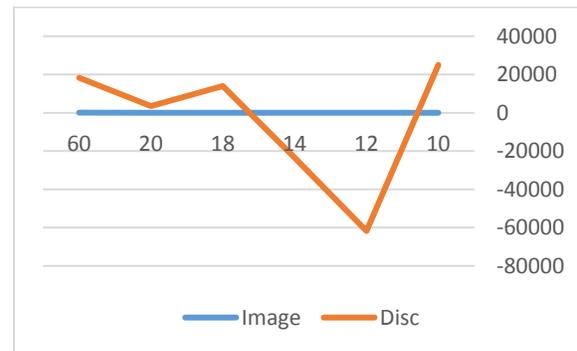


Fig. 5. Recurring planned rates Discrepancy K-Means Clustering algorithm

Table (4) showed Discrepancy with Fuzzy C-Means for five images.

TABLE IV. DISCREPANCY WITH FUZZY C-MEANS

Image	Disc
010	-4.4481
012	-1.9027
014	-4.2590
018	-4.3668
020	-4.5621
060	-3.8431

Figure (6) showed Recurring planned rates Discrepancy Fuzzy C-Means algorithm

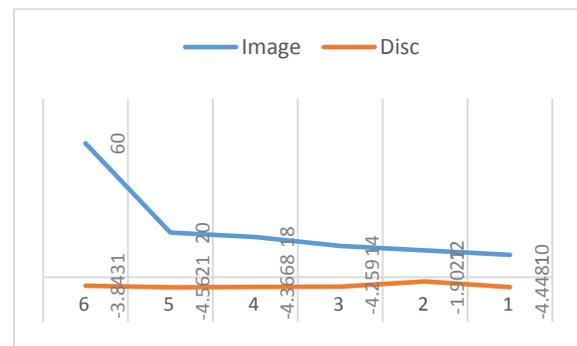


Fig. 6. Recurring planned rates Discrepancy Fuzzy C-Means algorithm

Table(5) showed rate of E-Intra and Threshold for K-Means Clustering to five images.

TABLE V. E-INTRA AND THRESHOLD FOR K-MEANS CLUSTERING TO FIVE IMAGES

Image	E-Intra	Thresholding
010	69.5432	69.7977
012	59.2619	58.8513
014	66.7928	62.7437
018	59.3109	57.9370
020	46.2355	44.8562
060	102.2822	102.5606

Figure (7): show Recurring planned rates E-Intra and Threshold for K-Means Clustering

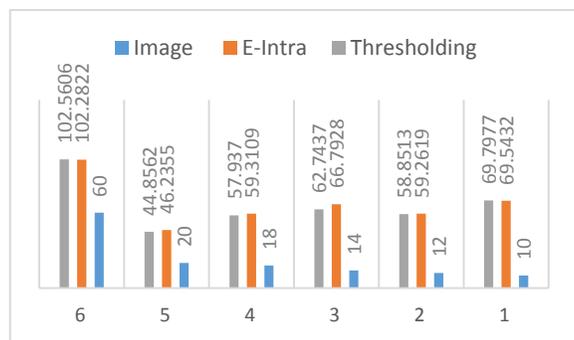


Fig. 7. E-Intra and Threshold for Fuzzy C-Means to five images

TABLE VI. E-INTRA AND THRESHOLD FOR FUZZY C-MEANS TO FIVE IMAGES

Image	Disc
010	24980
012	-61750
014	-23515
018	13958
020	3495
060	18242

Figure (8) showed Recurring planned rates Discrepancy K-Means Clustering

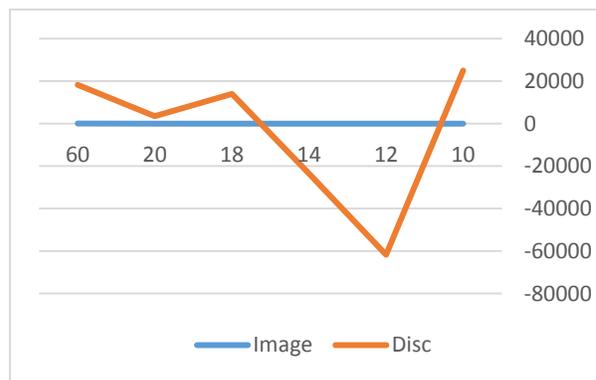


Fig. 8. showed Recurring planned rates Discrepancy K-Means Clustering

TABLE VII. FUZZY C-MEANS WITH DIFFERENT CLASSES

Original Image	C=2	C=3	C=8	C=16
	 Dis = -4.94 E = 70.4764 MSE = 0.7151 PSNR = 1.4564	 Dis = -4.39 E = 69.7878 MSE = 0.3169 PSNR = 4.9913	 Dis = -1.04 E = 69.9602 MSE = 0.2437 PSNR = 6.1313	 Dis = -9.55 E = 69.6448 MSE = 0.3180 PSNR = -10.4160
	 E = 59.5771 MSE = 0.7420 PSNR = 1.295	 Dis = -9.5414 E = 58.8149 MSE = 0.1234 PSNR = 9.0866	 Dis = -4730 E = 59.0982 MSE = 0.3387 PSNR = -4.7014	 Dis = 2.1765 E = 58.7683 MSE = 0.2029 PSNR = -8.4644

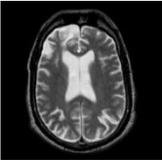
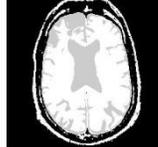
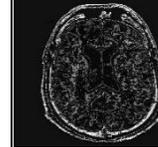
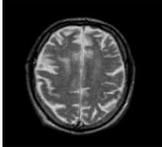
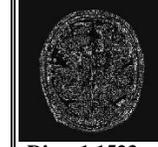
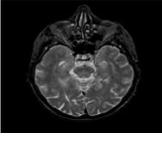
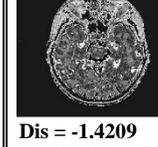
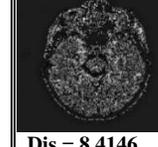
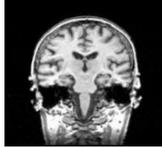
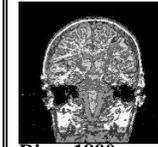
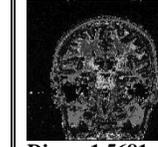
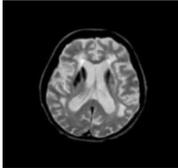
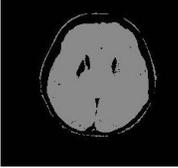
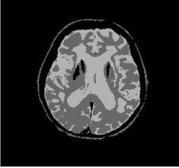
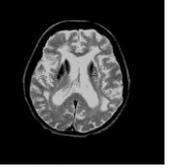
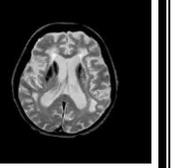
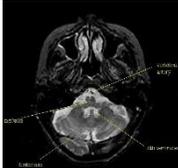
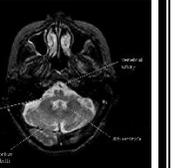
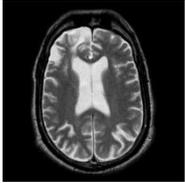
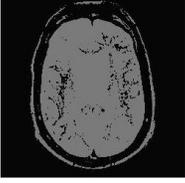
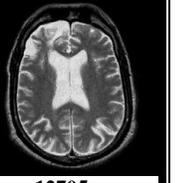
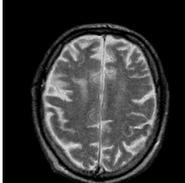
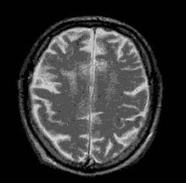
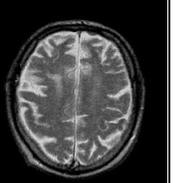
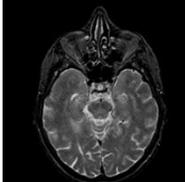
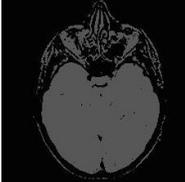
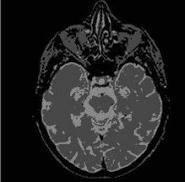
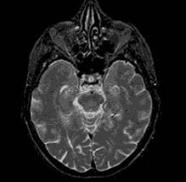
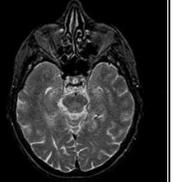
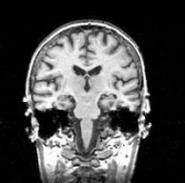
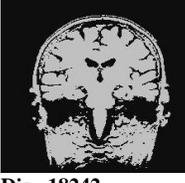
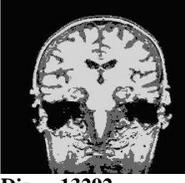
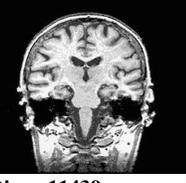
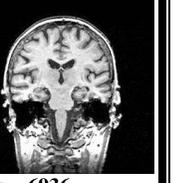
	 Dis = -4.2590 E = 63.3936 MSE = 0.6850 PSNR = 1.6431	 Dis = -8.7354 E = 62.8770 MSE = 0.4466 PSNR = 3.5008	 Dis = -1.3892 E = 62.9557 MSE = 0.2989 PSNR = 5.2454	 Dis = 1.1933 E = 62.5616 MSE = 0.3282 PSNR = -10.5519
	 Dis = -4.3668 E = 58.6033 MSE = 0.6960 PSNR = 1.5740	 Dis = 5.1774 E = 57.8580 MSE = 0.1958 PSNR = 7.0829	 Dis = -1.3290 E = 58.1398 MSE = 0.2732 PSNR = 5.6347	 Dis = 1.1523 E = 57.7612 MSE = 0.3206 PSNR = -10.4503
	 Dis = -4.5621 E = 45.5523 MSE = 0.7174 PSNR = 1.4426	 Dis = -4.6756 E = 44.9275 MSE = 0.3306 PSNR = 4.8068	 Dis = -1.4209 E = 45.0730 MSE = 0.2830 PSNR = 5.4827	 Dis = 8.4146 E = 44.7278 MSE = 0.2586 PSNR = -9.5166
	 Dis = -3.8431 E = 103.1470 MSE = 0.6441 PSNR = 1.9105	 Dis = -611.0118 E = 102.5699 MSE = 0.2164 PSNR = 6.6479	 Dis = -.1890 E = 102.7008 MSE = 0.2958 PSNR = 5.2899	 Dis = 1.5681 E = 102.3213 MSE = 0.3870 PSNR = -11.2686

TABLE VIII. K-MEANS CLUSTERING WITH DIFFERENT NUMBER OF CLUSTER

Original Image	C=2	C=3	C=8	C=16
	 Dis = 24980 E = 69.5432 MSE = 21.1371 PSNR = 29.5471	 Dis = -15254 E = 70.0316 MSE = 10.8361 PSNR = 34.8048	 Dis = 9470 E = 69.6532 MSE = 4.5071 PSNR = 40.1501	 Dis = 2747 E = 69.7558 MSE = 2.5187 PSNR = 43.5198
	 Dis = -61750 E = 59.2619 MSE = 13.4794 PSNR = 33.0530	 Dis = 74161 E = 58.5684 MSE = 6.1268 PSNR = 39.9464	 Dis = 945 E = 44.8598 MSE = 5.2216 PSNR = 39.8656	 Dis = 66273 E = 58.5985 MSE = 4.0058 PSNR = 42.0355

	 Dis = -23515 E = 66.7928 MSE = 33.2907 PSNR = 26.4325	 Dis = 4921 E = 62.6686 MSE = 16.0652 PSNR = 33.6072	 Dis = 15225 E = 62.5114 MSE = 8.0751 PSNR = 37.657	 Dis = -12705 E = 62.9376 MSE = 3.5400 PSNR = 42.1142
	 Dis = 13958 E = 59.3109 MSE = 23.8972 PSNR = 27.2781	 Dis = 14657 E = 57.7159 MSE = 13.5720 PSNR = 33.0232	 Dis = 14773 E = 57.7116 MSE = 5.1849 PSNR = 38.8732	 Dis = 19544 E = 57.6388 MSE = 3.8630 PSNR = 41.5521
	 Dis = 3495 E = 46.2355 MSE = 17.6975 PSNR = 26.3113	 Dis = 7725 E = 44.8649 MSE = 11.5880 PSNR = 32.0943	 Dis = -8302 E = 44.9829 MSE = 5.0357 PSNR = 38.9565	 Dis = 8054 E = 44.7333 MSE = 3.0978 PSNR = 42.4738
	 Dis = 18242 E = 102.2822 MSE = 33.1643 PSNR = 30.4140	 Dis = -13292 E = 102.7634 MSE = 20.3983 PSNR = 33.5528	 Dis = -11439 E = 102.7351 MSE = 8.4418 PSNR = 38.2672	 Dis = 6936 E = 102.4548 MSE = 4.5379 PSNR = 41.4939

From above experiments we notes the characteristic of K-means clustering changed depend on number of clusters, this proved when analysis results of experiments as bellow.

1) Whenever number of clusters increased the discrepancy will be reduced, if see when Cluster number = 2 , discrepancy= 24980

and discrepancy= 2747

2) From another side high percentage for number of generated regions refer to bad segmentation for original image and low value for regions means suffusion segmentation.

About fuzzy C-means we see different characteristic as follow

1) Discrepancy gradually will be increase with increase

number of clusters, moreover this create vast differences between original image and segmented image.

2) Also Fuzzy C-means dell with K-means clustering on measurement

$E_{\text{intra region}}$

It is worth mentioning the means square error increased when there is difference among original image and segmented image, this means whenever original image dramatically segmented, MSE became high.

While peak signal to noise rate measure the quality of segmented image then if its high this means the segmented image nearly to original image and this Insufficient segmentation , and if PSNR value is low then the original image segmented In order to be sufficiently clear vision.

As seen before the PSNR depend on cluster numbers in K-means clustering, also this with Fuzzy c-means.

V. CONCLUSION

We propose an algorithm to Performance Evaluation of K-mean and Fuzzy c-mean image segmentation based Clustering classifier.

The concluded that all of K-means and Fuzzy C-means approximately generate same number of regions in all selected cluster, from another side we note K-means generate a large proportion of error (MSE) with high PSNR Compared with the Fuzzy C-means generate low small percentage of error with low PSNR.

The algorithm higher accuracy can be found by increasing number of Clustering classifier and with Fuzzy c-mean image segmentation.

REFERENCES

- [1] Paresh Chandra et.al., "MRI image Segmentation using level set method and implement an medical diagnosis system", Computer science & Engineering: An international journal(CSEIJ).Vol.1, No.5,2011.
- [2] Tsai,C.S.,Chang,C.C.,"An Improvement to image segment based on Human Visual system for object based coding",Fundamentainformaticae,Vol.58,No.2,2004.
- [3] Dweepna Garg, Khushboo Trivedi, B.B.Panchal , "A Comparative study of Clustering Algorithms using MapReduce in Hadoop", International Journal of Engineering Research & Technology , Vol.2 - Issue 10 (October - 2013)
- [4] Hongyuan Zhu,Fanman Meng,Jianfei Cai and Shijian Lu,"Beyond pixels: A comprehensive survey from bottom up to semantic image segmentation and segmentation ",Preprint submitted to Elsevier ,February ,2015.
- [5] Rafael C. Gonzalez and Richard E.Woods , " Digital Image Processing ", Prentice Hall,3rd. Edition, 2008.
- [6] Uri Kroszynski and Jianjun Zhou, Fuzzy Clustering Principles, Methods and Examples, IKS, December 1998
- [7] J.-S. R. Jang, C.-T. Sun, E.Mizutani, Neuro-Fuzzy and Soft Computing, p (426-427)Prentice Hall, 1997
- [8] Brundha B.Nagendra Kumar M," MIR image segmentation of brain to detect brain tumor and its area calculation using K-means clustering and fuzzy c-means algorithm", International journal for Technological Research In Engineering,Vol.2,Issue.9,2015.
- [9] Hind Rustum Mohammed, Dr. Ali Hassan Nasse and Raghad Saaheb, "CT Angiography Image Segmentation by Mean Shift Algorithm and Contour with Connected Components Image", International Journal of Scientific & Engineering Research, Vol.3,Issue.8,2012.
- [10] Soumi Ghosh,Sanjay Kumar Dubey , " comparative analysis of means and Fuzzy c-means algorithms, International journal of advanced computer science and applications, Vol.4,Issue.4,2013.
- [11] Yusra A. Y.Al_Najjar,Der Chen Soong, "comparison of image quality assessment:PSNR,HVS,SSIM,UIQI", International Journal of Scientific & Engineering Research,Vol.3, Issue 8,2012.
- [12] YogendraKumar Jain,Garima Silakari", Performance Evaluation of filters for enhancement of Image in different application areas", Journal of computer Engineering, Vol.10, Issue.5 ,2013.
- [13] Hui Zhang ,Jason E.Fritts and Sally A.Goldman,"Image segmentation evaluation :a survey of unsupervised methods", computer vision and image understanding, Issue.110,2008.

Identifying Cancer Biomarkers Via Node Classification within a Mapreduce Framework

Taysir Hassan A. Soliman

Associate Professor, Information Systems Department, Faculty of Computers & Information, Assiut University, Egypt

Abstract—Big data are giving new research challenges in the life sciences domain because of their variety, volume, veracity, velocity, and value. Predicting gene biomarkers is one of the vital research issues in bioinformatics field, where microarray gene expression and network based methods can be used. These datasets suffer from the huge data voluminous, causing main memory problems. In this paper, a Random Committee Node Classifier algorithm (RCNC) is proposed for identifying cancer biomarkers, which is based on microarray gene expression data and Protein-Protein Interaction (PPI) data. Data are enriched from other public databases, such as IntACT1 and UniProt2 and Gene Ontology3 (GO). Cancer Biomarkers are identified when applied to different datasets with an accuracy rate an accuracy rate 99.16%, 99.96% precision, 99.24% recall, 99.16% F1-measure and 99.6 ROC. To speed up the performance, it is run within a MapReduce framework, where RCNC MapReduce algorithm is much faster than RCNC sequential algorithm when having large datasets.

Keywords—Big data; cancer biomarkers; MapReduce; node classification

I. INTRODUCTION

Bioinformatics is one of the main applications that adopt big data through microarray gene expression analysis, next generation sequencing, text mining of literature publications, and large graph analysis of biological networks, such as metabolic networks, signal pathways, and protein-protein interaction networks. Bioinformatics researchers have an excellent opportunity to achieve scalable efficient and reliable computing performance on Linux clusters and within cloud computing environment [1]. However, scalable and efficient data mining algorithms are needed to perform different tasks in bioinformatics. Biomarkers play an important role in diagnosing, assessing prognosis and directing treatment of cancer. A cancer biomarker refers to a substance or process that is indicative of the presence of cancer in the body. A biomarker may be a molecule secreted by a tumor or a specific response of the body to the presence of cancer. Genetic, epigenetic, proteomic, glycomic, and imaging biomarkers can be used for cancer diagnosis, prognosis, and epidemiology⁴. Biologists can now quickly identify hundreds, and even thousands of candidate genes associated with a target disease or functionality. One of the main traditional techniques to find interactions and similar structure is applying text mining techniques to literature abstracts, i.e. through PubMed⁵ [2,3]. However, this is a very time consuming issue because of

the tremendous high volume of current literature reviews.

Other techniques fall into two main categories: Microarray gene expression analysis and biological networks. Microarray gene expression analysis can measure thousands of gene expressions which make it a good chance to identify biomarkers through microarray technology [4-6]. However, better prediction accuracy is required since the accuracy of applying network techniques is relatively low. Identifying significant gene sets or pathways involved in diseases or biological processes by incorporating some prior biological knowledge, such as gene set enrichment analysis or pathway enrichment analysis are proposed via several methods [7-9]. In addition, PPIs, protein-DNA interactions, or regulatory pathways algorithms are developed. For instance, Chuang et al. [10] identified biomarkers of metastasis using breast cancer gene expression data, based on protein-protein interaction networks. Li et al. [11] introduced a network-constrained term based on L1-norm of regression coefficients of microarray data. Jahid and Ruan [12] identified a small number of intermediate genes containing important information about the pathways involved in metastasis genes, using a randomized steiner tree. Zhu et al. [13] recently built binary classifiers as prediction models, using support vector machines. In addition, Wei and Li [14] developed a Markov Random Field Model for network-based Analysis. Furthermore, Chen et al. [15] developed network-constrained Support Vector Machine (netSVM) for cancer biomarker identification with an improved prediction performance. Hwang et al. [16] applied the network propagation algorithm to study three large-scale breast cancer datasets, achieving competitive classification performance. Xia et al [17] have developed Network Analyst, enabling high performance network analysis with rich user experience in order to identify genes/ proteins of interest in biological networks.

One of the main computational challenges have become increasingly important is using High Performance Computing (HPC) in bioinformatics data analysis [18]. Another computer architecture / service model is cloud computing [19-21], where it is used to scale up the performance of the required service. Recently, biomarker prediction based on large-scale feature selection and MapReduce has been discussed in [22], where Kmeans clustering and Signal to Noise Ratio have been combined with optimization technique as Binary Particle Swarm Optimization. A key problem arises when using hybrid approaches of microarray gene expression and network-based methods is handling very large networks which require high performance time.

¹<http://www.ebi.ac.uk/intact/>

²<http://www.uniprot.org/>, ³<http://geneontology.org/>,

⁴<https://en.wikipedia.org> ⁵<http://www.ncbi.nlm.nih.gov/>

In this paper, a node classification algorithm is suggested in order to identify biomarkers, which is considered one of the main problems in the bioinformatics domain. This algorithm is applied and compared to other machine learning algorithms, such as naïve bayes and random forest. In addition, the RCNC algorithm is applied within MapReduce framework, as one of the open source Apache Hadoop project. Node classification has been previously introduced in dynamic content-based networks [23]. The main contributions of this paper are:

- 1) A hybrid approach of microarray gene expression and PPI networks is proposed to predict protein biomarkers via Random Committee Node Classifier algorithm (RCNC).
- 2) Speeding up the performance of the algorithm via MapReduce.
- 3) Developing an information topological PPI network

The organization of this paper as follows: section two explains materials and methods and section three illustrates results and discussion. Finally, section four concludes the work and gives insights into future work.

II. MATERIALS AND METHODS

In this section, identifying biomarkers based on node classification within a MapReduce framework is proposed, as illustrated in Fig. 1. This framework depends on a hybrid approach of microarray gene expression data and PPI network. The framework consists of two main phases: data preprocessing and biomarker identification, which will be discussed in details in the following subsections. Data preprocessing phase has two main goals, which are 1) Computing Differentially Expressed Genes (DEGs) and 2) Integrating data. The goal of biomarker identification phase is to identify biomarkers for different types of cancer (Breast, colon, ovarian and hepatocellular carcinoma), using the proposed RCNC algorithm.

A. Phase i: data preprocessing

The objectives of this phase are to a) Compute Differentially Expressed Genes (DEGs) and b) Integrate Data.

1) Computing deg:

Microarray technologies now enable the simultaneous interrogation of the expression level of thousands of genes to obtain a quantitative assessment of their differential activity in a given tissue or cell. Microarray analysis has enabled the identification of gene signatures for diagnosis, molecular characterization, prognosis and treatment prediction. Microarray gene expressions data are obtained from GEO⁴ database for Breast, colon, liver (hepatocellular carcinoma), and ovarian cancer. For each type of cancer, five series are used, which are illustrated in Table I, where both healthy and unhealthy microarray gene expression series are downloaded (Affymetrix experiments). Differentially Expressed Genes (DEGs) are computed for all downloaded samples, using R statistical language⁴; in addition, p value < 0.05 is set as the threshold for DEGs and t-test [23] is applied.

2) Integrating data

Data integration is one of the vital tasks in bioinformatics, where many diverse public databases' formats exist, such as

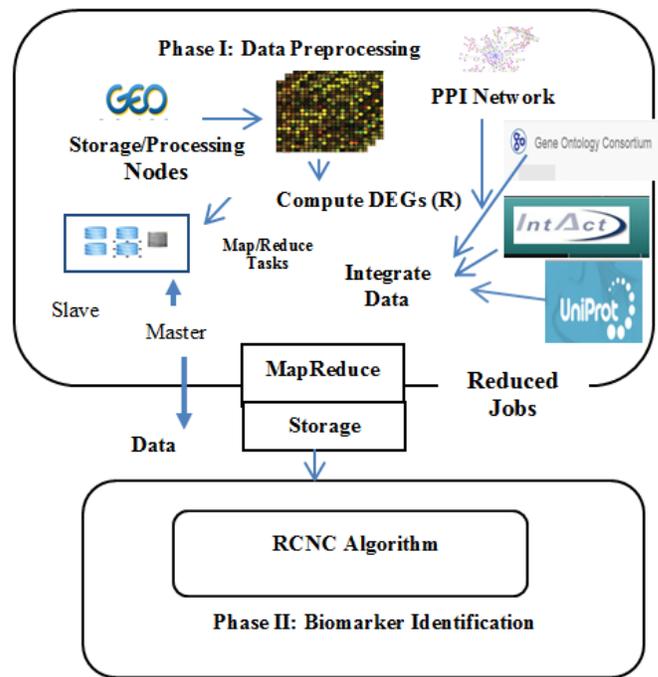


Fig. 1. Biomarker Identification Framework using RCNC Algorithm

TABLE I. GEO CANCER SERIES

Cancer Type	Series (Samples)	# of Gene Instances
Breast	GSE44024 (4)	22,278
	GSE53394 (8)	22,278
	GSE38376 (18)	48,804
	GSE45804 (12)	33,298
	GSE41816 (36)	33,298
Liver (Hepatocellular carcinoma)	GSE41804 (40)	54,676
	GSE49515 (26)	54,676
	GSE21955 (22)	24,527
	GSE29084 (4)	54,676
Ovarian	GSE32474 (174)	54,676
	GSE31432 (23)	48,804
	GSE51373 (28)	54,676
	GSE22600 (15)	54,676
	GSE23616 (15)	20,603
Colon	GSE13525 (12)	54,589
	GSE14773 (4)	54,676
	GSE34299 (4)	
	GSE18088 (53)	
GSE18560 (12)		

XML, csv, and RDF. PPI data sometimes are not enough to identify biomarkers. As a result, in this approach data are integrated from heterogeneous resources: IntAct (release 2.5) and UniProt (August 2015) in addition to the DEGS results of microarray gene expressions, computed at step 2.1.a.

In this work, cancer interaction datasets are downloaded from IntAct, which contain the target types of cancer discussed here: breast cancer, ovarian cancer, hepatocellular carcinoma, and colon cancer. The following preprocessing steps are accomplished for IntAct and UniProt data:

- 1) Removing missing values
- 2) Deleting irrelevant attributes
- 3) Extracting data
- 4) Mapping attributes

To illustrate the idea, downloaded cancer interaction data contain UniProtkb identifiers of interacting proteins, alternative identifiers for each protein at IntAct database European Bioinformatics Institute identifier, aliases, interaction detection method (two hybrid, pull down, etc), publication date of each, taxonomy identifier, interaction type (physical association, colocalization, direct interaction, and association), database source, interaction identifier, and confidence. Some of the GO ontologies are missing so the corresponding values are deleted. In addition, irrelevant attributes (attributes not used as parameters for determining biomarkers) are deleted: the publication date, taxonomy identifier, interaction detection method, interaction identifier and source database.

Gene name is extracted from attribute (Alias), and mapped to the DEGs found in microarray experiments. For example protein A: uniprotkb: P35125-3 Ubiquitin carboxyl-terminal hydrolase 6 (alternative identifier: intact:EBI-954590), interacts with protein B uniprotkb:P10916 (alternative identifier: intact:EBI-725770|uniprotkb:Q16123). In addition, alias of P35125-3 is psi-mi:p35125-3(display_long)|uniprotkb:"210(ORF1)"(isoform synonym)|uniprotkb: oncTre210p (isoform synonym)| uniprotkb: USP6(gene name)|psi-mi: USP6 (display_short)|uniprotkb:HRP1 (gene name synonym)|uniprotkb:TRE2(gene name synonym) |uniprotkb: Deubiquitinating enzyme 6(gene name synonym) |uniprotkb: Proto-oncogene TRE-2(gene name synonym)| uniprotkb: Ubiquitin-specific- processing protease 6 (gene name synonym)| uniprotkb:Ubiquitin thioesterase 6 (gene name synonym), alias of protein B is psi-mi:mlrv_human (display_long) |uniprotkb: MYL2(gene name)|psi-mi:MYL2(display_short). In addition, other attributes are interaction detection method (psi-mi:"MI:0018"(two hybrid)), publication 1st author (Dechamps et al. (2006)), publication identifier (pubmed:16555005), Taxid interactorA (taxid:9606(human)|taxid:9606(Homo sapiens), Taxid interactorB (taxid: 9606 (human) | taxid:9606(Homo sapiens)), interaction type (psi-mi:"MI:0915"(physical association)), source database(s) (psi-mi:"MI:0469"(IntAct)), interaction identifier (intact:EBI-1225898), and confidence value (intact-miscore:0.61).

For each protein, each UniProtkb identifier is mapped into its corresponding Uniprotkb identifier in UniProtkb database. Other included information from UniProtkb is protein function, Gene Ontology (GO) molecular function, biological process, and cellular component. In addition, DisGeNet database has been used as for validation of biomarkers' prediction results.

B. Phase II: Biomarker Identification

To identify biomarkers, RCNC algorithm is proposed, which depends on topological node classification algorithm in

an ensemble learning manner. The problem of node classification has been addressed in a number of applications, such as social network analysis [25]. In this section, RCNC algorithm of biomarkers identification is explained in details. RCNC uses a random committee technique, which is an ensemble tree classifiers based. Ensemble methods like combine the decisions of multiple hypotheses are some of the strongest existing machine learning methods [26-28]. Ensemble classifiers gather randomizable base classifiers, where each base classifier is built using a different random number seed. A random committee algorithm is an ensemble of random tree classifiers, where it predicts a class label by averaging probability estimates over these classification trees. This algorithm produces better overall accuracy for all testing cases than any individual committee member. In this paper, a random committee technique is used to handle: 1) too large data volume, 2) inadequate data, and 3) complexity of decision boundary. The learning procedure for ensemble algorithms can be divided into the following two parts:

1) *Constructing base classifiers/base models*: In this part, data preprocessing is performed first where noisy data are removed then base classifier are constructed. Data preprocessing step is already at the data integration phase, as previously explained.

2) *Voting*: The main objective of this part is to combine the base classifiers models built in the previous step into the final ensemble model. There are several kinds of voting but the most used ones are the weighted and un-weighted voting. Voting includes the weighted average (of each base classifier holds) when using regression problem and majority voting when doing classification and the weighted-majority output is given by, which is used in this paper:

$$\text{Argmax} \left[\sum_{i=1}^k p_i(x), w_i \right] \quad (1)$$

$P_i(x)$ is the results of the prediction of i th prediction model and $P_i(x, w)$ is indicator function defined as:

$$P_i(x, w) = \begin{cases} 1 & x = w \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Problem Definition: given a graph, which is represented as $G = \{V, E, W\}$, where V is a set of nodes, E is the set of Edges, and W is the edge weight matrix $n \times n$; $W = [w_{ij}]$ and $n = |V|$. L is the set of labels $L = \{l_1, l_2, \dots, l_q\}$ for the set of q attributes associated with each node V .

Homophily: is a term used in social networks and defined as a link between individuals (i.e. friendship or other social connection) when they are being similar in nature. When applying "homophily" to PPI information network, two protein nodes are connected based on "homophily" property if they interact with each other and have similar characteristics. These characteristics include:

- Sequence similarity scores.
- GO relations where two nodes are GO related if there is a semantic relation holding between those proteins. This semantic relation between two proteins is divided into the following:

- If functions are connected through ontology
- If cellular components relations exist.
- If Biological process relations exist.

For example, for the protein P35125 which is a biomarker for ovarian cancer interacts with protein Q8N8A2. P35125 has gene molecular functions: calmodulin binding, cysteine-type endopeptidase activity, nucleic acid binding, ubiquitin-specific protease activity. Q8N8A2 has a protein binding molecular function, where calmodulin binding is a protein binding type. P35125 and Q8N8A2 proteins have 84.3% sequence similarity. Sequence similarity scores are taken into consideration when >70%, as shown in Fig. 2. Table II explains the steps of graph construction algorithm.

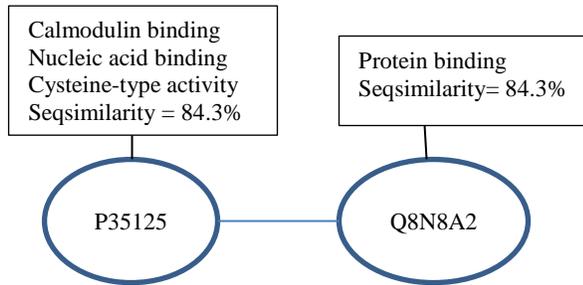


Fig. 2. An example of Breast Cancer PPI Information Network

TABLE II. GRAPH CONSTRUCTION ALGORITHM

Algorithm 1: Graph Construction

```

map(key, value):
begin
edge = 1;
Node V(edge);
If homophily exists
Emit(V.id, V);
end
reduce(key, values):
begin
Emit(key, serialize(values));
End
    
```

Machine learning algorithms have the advantage of making use of Hadoop distributed computing platform and the MapReduce programming model to process data in parallel. Many machine learning algorithms have been investigated to be transformed into the MapReduce paradigm in order to make use of the Hadoop Distributed File System (HDFS). In the current work, RCNC is run under the MapReduce framework and is evaluated on four datasets in order to evaluate scalability comparisons of using RCNC sequentially and RCNC under the MapReduce environment (RCNC MapReduce). The proposed MapReduce architecture used for this classifier is clarified in Fig. 3.

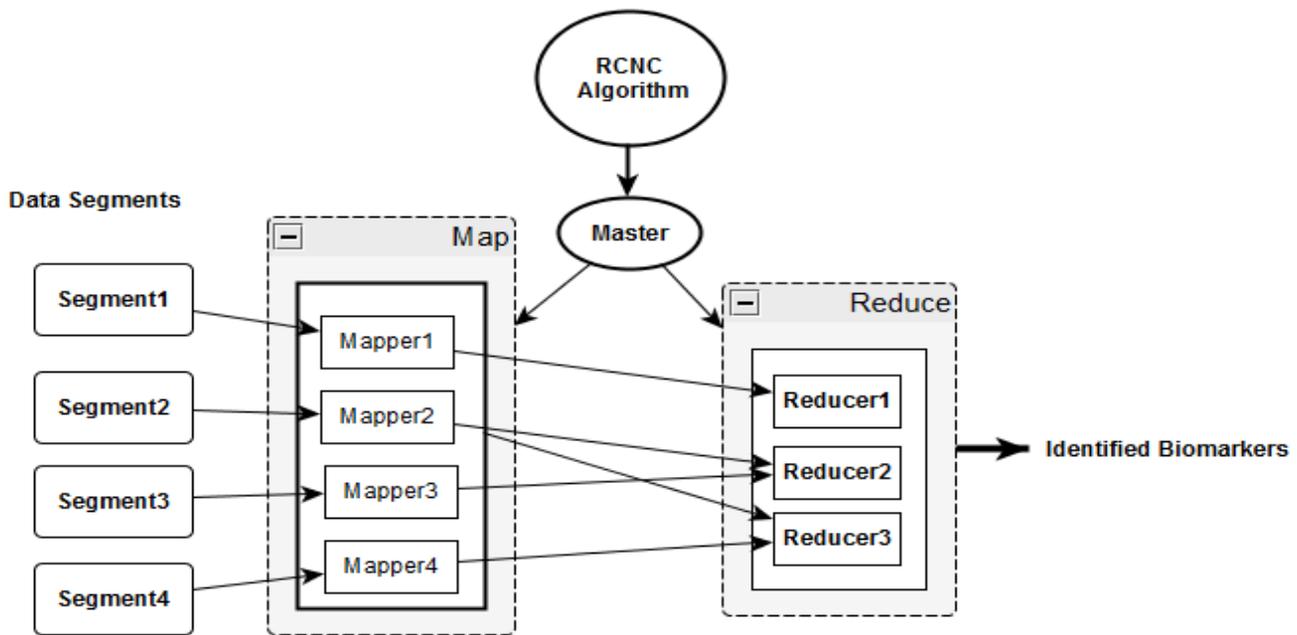


Fig. 3. Workflow of the proposed MapReduce framework for RCNC Node Classifier

Through this architecture, the number of occurrences of an attribute with a specific value given a certain class is obtained. The Hadoop uses Input Data Format to divide the big file into small input files which record Key and Value. In this case, the key will be the feature of the data (i.e. interaction type). Then, the Map process defines the data structure (key, value) on the Map operation. The Map process is applied to each input

dataset in parallel. With the result from the MapReduce task, one can assign the instance to a class after training each segment via a random committee algorithm. Finally, the ensemble of classifiers is computed via equation (2). Table III illustrates the steps of RCNC MapReduce algorithm. RCNC sequential is the same idea but without dividing the algorithm into Map & Reduce functions.

TABLE III. MAPREDUCE RCNC ALGORITHM

Algorithm 2: Random Committee Node Classifier (RCNC) MapReduce	
Input	Graph G=(V, E,W), T= ensemble size; Max= Maximum number of nodes
Output	f = (f ⁽¹⁾ ... f ^(T)) (ensemble of classifiers)
Process	Map(Vertexid V.id, Vertex V) Begin For E ∈ n.adjancylist do emit(E.neighbor, <V.label, E.EdgeWeight>) End Emit(Vertexid V.id, Vertex V) End Reduce(V.id, W) Begin For i = 1 to Max do Begin f(i) <- (p _i (V.id, W)) End V.label ← (f ⁽¹⁾ ... f ^(T)) Emit(Vertexid V.id, Vertex V) End

III. RESULTS

In this paper, four kinds of cancer are used: breast, colon, liver (hepatocellular carcinoma) and ovarian interaction datasets. Data are split into 66% for training and the rest for testing within a 10-Fold validation on the training dataset to select the optimal value of parameters. Experiments have been performed using Java JDK version 1.7 and for MapReduce implementation Hadoop version 2.4.1. MapReduce implementation is tested in a cluster of 4 data nodes running Linux. Each node is an Intel® Core™ i7-3770 CPU @3.4 GHZ, and 32GB RAM. Several comparisons are performed: 1) the proposed RCNC algorithm for node classification in a sequential manner versus naïve bayes, random forest classifiers, proposed method in [22], and [29], as shown in Table IV. In [29], an approach based on Neighborhood Rough Set and Probabilistic Neural Networks Ensemble is proposed for the classification of Gene Expression Profiles. Comparison contains the precision, recall, F1-measure, and ROC.

As summarized in Table IV, RCNC is always higher than Random Forest and naïve bayes classifiers when for all datasets. For example, for breast cancer dataset, RCNC has shown an accuracy of 99.72% , a recall of 99.7%, ROC of 100%, where the True positive rate is 99.7% and False Positive rate is 0.05% with F1-measure 99.7% for breast cancer datasets. For ovarian datasets, both datasets 15,154 and 54,675 are tested for all algorithms: RCNC, Random Forest, naïve bayes, BSMO, and [34]. In the first case, RCNC is higher than BSPO and [34], where in the second case RCNC and BSMO give the same accuracy rate. However, RCNC gives more information regarding related biomarkers from the PPI information network. Furthermore, datasets are enlarged to 4GB each synthetically and the accuracy is the same but performance time is very fast.

The second testing of RCNC MapReduce is its time performance versus RCNC MapReduce, as illustrated in Fig. 6, where the time of RCNC MapReduce is faster than RCNC sequential.

TABLE IV. COMPARISONS OF RCNC WITH OTHER CLASSIFIERS

# Genes	Classifiers	P %	Rec %	ROC %	F1-	Acc. %
Breast 22,278	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	98.9	98.9	99.9	98.7	98.8
	Naïve Bayes	98.3	98.3	100	98.2	98.3
Colon 15,154	RCNC	96	97.4	98	97	97
	Random Forest	83	84	84	95	84.1
	Naïve Bayes	81.8	82.8	81.8	95	82
Hepato 24,527	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	80.1	38.8	83.6	88.6	75.7
	Naïve Bayes	76	76.4	76.1	81.6	75.7
Ovarian 15,154	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	81.5	79.1	90.6	81.5	81.4
	Naïve Bayes	96	97.4	98	97.1	97
	BSPO[23]					99
	[34]					96
Ovarian 54,675	RCNC	99.7	99.7	100	99.7	99.7
	Random Forest	81.5	79.1	90.6	81.5	81.4
	Naïve Bayes	96	97.4	98	97.1	97
	BSPO[23]					100
	[34]					96

Finally, Fig. 7 clarifies the runtime of RCNC MapReduce having one, two, and four nodes for each dataset. Experiments for different size of data chunk and different number of maps are performed to evaluate impact of MapReduce parallelism. One can notice that having two nodes, the time performance is reduced to near half of the time required when having one node only. In addition, having four nodes, the runtime of the algorithm is reduced. The accuracy rate of RCNC sequential versus RCNC MapReduce is also tested when having four nodes, where the accuracy remains the same.

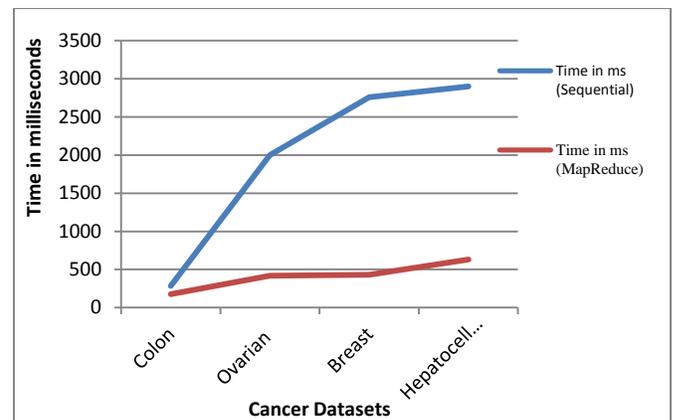


Fig. 4. Comparison of RCNC Sequential and RCNC MapReduce

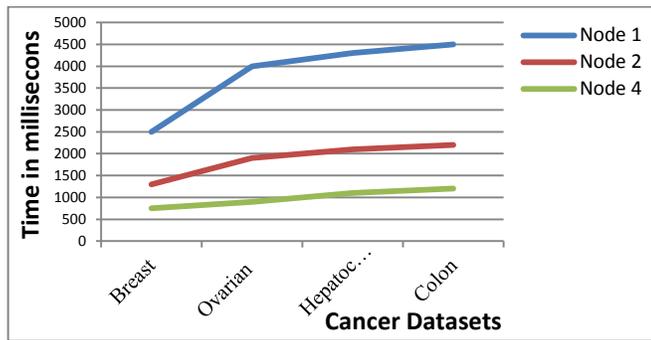


Fig. 5. Time Comparisons of RCNC MapReduce for Different Number of Nodes

Identified genes are evaluated against the DisGeNet database, where the relation between genes as biomarkers can be downloaded for cancer datasets. Examples of cancer detected biomarkers are: HSP60 (ovaries), HSPD1 (ovaries), FANCD2 (breast), FANCD3 (breast), FANCD4 (breast), MYL2 (breast), FANCD1 (ovaries), FACD (ovaries), XRCC9 (breast), DGKI (breast), APCS (colon), STK11 (colon), PTEN (colon), MLH1 (colon), MLH6 (colon), POLE (colon), EPCAM (colon), and MYH (colon)

IV. CONCLUSIONS

In this paper, a Random Committee Node Classifier algorithm (RCNC) was proposed to predict cancer biomarkers, where microarray gene expression and network based methods were used. These datasets had a very large volume, which caused main memory problems. Compared with other classifiers, RCNC had proven high accuracy. Biomarker genes were identified when applied to different datasets with an accuracy rate 99.16%, 99.96% precision, 99.24% recall, 99.16% F1-measure and 99.6 ROC. To speed up the performance, it was run within a MapReduce framework, where RCNC MapReduce were much more faster than RCNC sequential when having large datasets. Future work includes taking RNAseq data into consideration and enlarging the datasets into multiple types of cancer. In addition, more ontologies will be added as ChEBI and disease ontologies. Furthermore, more enhancements can be performed to RCNC for covering multi-dimensional graphs.

REFERENCES

- [1] Q. Zou, X. Li, W. Jiang, Z. Lin, G. Li, and K. Chen, "Survey of mapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, pp. 1-12, 2013.
- [2] H. Li and C. Liu, "Biomarker identification using text mining," *Computational and Mathematical Methods in Medicine*, pp. 1-4, 2012.
- [3] W. Fleuren, et al., "Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining," *BioDataMining*, Vol. 6, 2, pp.1-15, 2013.
- [4] T. Golub, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*," Vol. 286, issue 5439, pp.531-537, 1999.
- [5] C. Sotiriou, and L. Pusztai, "Gene-expression Signatures in breast cancer," *The New England Journal of Medicine*, 360, 8, pp. 790-800, 2009.
- [6] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl Acad Sci USA*, Vol. 98, 9, pp.5116-5121, 2001.

- [7] T. Bo, and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biology*, Vol. 3, issue 4, pp. 1-11, 2002.
- [8] A. Subramanian, et al., "Gene set enrichment analysis: a Knowledge-based Approach for interpreting genome-wide expression profiles," *Proceedings of National Academy Science*, 102, 43, pp.15545-15550, 2005.
- [9] R., Curtis, M. Oresic, and A. Vidal-Puig, "Pathways to the analysis of microarray data," *Trends Biotechnology*, 23, 8, pp. 429-435, 2005.
- [10] H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker, "T: network-based classification of breast cancer metastasis," *Molecular Systems Biology*, 3, 140, 2007.
- [11] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, 24, 9, pp. 1175-1182, 2008.
- [12] M. Jahid and J. Ruan, "A steiner tree-based method for biomarker discovery and classification in breast cancer metastasis," *BMC Genomics*, 13 (Suppl 6):S8, pp. 1-9, 2012.
- [13] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," *BMC Bioinformatics*, 10 (Suppl 1):S21, 2009.
- [14] Z. Wei and H. Li, "A markov random field model for network-based analysis of genomic data," *Bioinformatics*, 23, 12, 1537-1544, 2007.
- [15] L. Chen, J. Xuan, R. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines," *BMC Systems Biology*, 5, 16, 2011.
- [16] Hwang et al., "Robust and efficient identification of biomarkers by classifying features on graphs," *Bioinformatics*, Vol. 24, 18, pp.2023-2029, 2008.
- [17] J. Xia, M. Benner, and R. Hancock, "NetworkAnalyst - integrative approaches for protein-protein interaction network analysis and visual exploration," *Nucleic Acids Research*, 42, 167-174, 2014.
- [18] R. Taylor, "An overview of the hadoop/mapReduce/HBase framework and its current applications in bioinformatics," *BMC Bioinformatics*, 11(Suppl 12):S, 1-6, 2010.
- [19] C. Sansom, "Up in a cloud?," *Nature Biotechnology*, 28, 1, pp.13-15, 2010.
- [20] L. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, 11:207, 2011.
- [21] M. Schatz, B. Langmead, and S. Salzberg, "Cloud computing and the DNA data race," *Nature Biotechnology*, 28, pp.691-693, 2011.
- [22] A. Kourid, and M. Batouche, "Biomarker discovery based on large-scale feature selection and mapreduce," *Proceedings of the 5th IFIP TC 5 International Conference, CIIA 2015, Saida, Algeria, May 20-21, pp.81-92, 2015.*
- [23] C. Aggarwal, and N. Li, "On node classification in dynamic content-based networks," *Proc. of the 2011 SIAM International Conference on Data Mining (SDM'11)*, Phoenix, AZ, USA, Apr. 28-30, pp.355-366, 2011.
- [24] X. Cui and A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology*, Vol. 4, No. 210, pp.1-10, 2008.
- [25] S. Bhagat, G. Cormode, and S. Muthukrishnan, "Node classification in social networks," in: Aggarwal, C. (Eds), *Social Network Data Analytics*. Springer Science+Business Media, LLC., US, pp.115-148, 2011.
- [26] P. Melville, "Creating diverse ensemble classifiers. Technical Report, University of Texas," 2003.
- [27] T.G., Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, 40, pp.139-157, 2000.
- [28] G. Biau, L. Devroye, G. Lugosi, "Consistency of random forests and other averaging classifiers," *Journal of Machine Learning Research*, 9, pp. 2015-2033, 2008.
- [29] J. Yun, X. Guocheng, C. Na, C. Shan, "A New gene expression profiles classifying approach based on neighborhood rough set and probabilistic neural networks Ensemble," In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) *ICONIP 2013, Part II. LNCS*, 8227, pp. 484-489, 2013.

Intelligent Mobility Management Model for Heterogeneous Wireless Networks

Sanjeev Prakash

¹Department of Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, India

R B Patel

Dept. of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Chandigarh, India

V. K. Jain

Department of Electrical and Instrumentation Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, India

Abstract—Growing consumer demands for access of communication services in a ubiquitous environment is a driving force behind the development of new technologies. The rapid development in communication technology permits the end users to access heterogeneous wireless networks to utilize the swerve range of data rate service “anywhere any time”. These forces to technology developers to integrate different wireless access technologies which is known as fourth generation (4G). It is become possible to reduce the size of mobile nodes (MNs) with manifold network interfaces and development in IP-based applications. 4G mobile/wireless computing and communication heterogeneous environment consist of various access technologies that differ in bandwidth, network conditions, service type, latency and cost. A major challenge of the 4G wireless network is seamless vertical handoff across the heterogeneous wireless access network as the users are roaming in the heterogeneous wireless network environment. Today communication devices are portable and equipped with manifold interfaces and are capable to roam seamlessly among the various access technology networks for maintaining the network connectivity, since no single-interface technology provides ubiquitous coverage and quality-of-service (QoS).

This paper reports a mobile agent based heterogeneous wireless network management system. In this system agent’s decision focuses on multi parameter system (MPS). This system works on the parameters- network delay, received signal strength, network latency and study of the collected information about adjoining network cells viz., accessible channel. System is simulated and a comparative study is also made. From results it is observed that system improves the performance of wireless network.

Keywords—FNS; MNS; MN; WLAN; Mobile Agent

I. INTRODUCTION

The adoption of wireless technology is increasing rapidly from few decades and has become popular means for computing and communication to access the information even the users are away from their home network. The driving force behind this development is basically portability of hand held devices viz., Smart mobile phones, tablets, Laptops and Personal Digital Assistances (PDA), etc. These devices are equipped with multiple interfaces due to availability of several wireless technologies and mobile applications. To meet the increasing demand of computing and communication services in ubiquitous environment there is need to integrate different emerging wireless access technologies together which is

known as Fourth Generation (4G) wireless computing and communication system [1, 12]. This wireless system is popular due to its heterogeneity. A heterogeneous wireless network generally is an integration of fixed backbone network and wireless networks. Mainly three distinct sets of entities play key role in the system, viz., mobile nodes (MNs), Mobile Network Stations (MNSs) and fixed network stations (FNSs). A moving network station (NS) which retains its network connection is called a MNS. A fixed network consists of FNSs and communication links among FNSs and none moving nodes or devices. Some of the FNSs are designated as base stations (BSs). A BS is augmented with wireless interfaces and works as a gateway for communication among the fixed and wireless networks. Communication range of wireless transceivers limits the MNS/MN to communicate with a BS. The average covering area of a region is of the order of 1KM in radius [2].

A handoff management is required when a user moves from one wireless cell to another, abandoning the connection with one point of attachment to another [15]. When a handoff occurs within the domain of a homogeneous wireless access technology occurring event is known as horizontal handoff and when this event occurs among heterogeneous wireless access network technologies is known vertical handoff. Horizontal handoff occurs when the MNs are moving far from point of attachment and enter into the low signal strength area in a homogeneous wireless network. In a heterogeneous environment, users have an opportunity to access the different technologies networks. A user may be benefited from different network characteristics (coverage, bandwidth, latency, power consumption, cost, etc.) which are not comparable directly [18]. MN mobility is supported by vertical handoff as the communication technology and access supporting infrastructure change. Sometime vertical handoff takes place due to user’s convenience rather than unavailability of connection [17][24]. The handoff process becomes more complex in such an environment compared to the homogeneous one.

Vertical handoff event comprises three handoff steps: 1) initiation 2) decision 3) execution. Network and user related information are collected during the handoff initiation, this step is also known as system discovery, system detection, handover information gathering, and handoff initiation. Handoff decision stage plays an important role and is one of

critical process of handoff. It is also known as network selection or system selection. Handoff execution is either hard handoff or soft handoff.

The constantly changing environment is demanding for more and more services. In such a condition to maintain the required grade of services in a region is a driving force for the technology developer to split a region for handling the traffic increased without increasing the bandwidth of the system [3]. A BS is accountable for forwarding information and voice packets among a MNS/MN and a fixed network. To achieve the required goal a MNS/MN may cross the boundary among two regions while it is in conversation. The job of forwarding information/voice packets among the fixed network and the MNS/MN may be routed through the new regions for facilitating the end-to-end links in the dynamically network topological changing environment. Heterogeneous wireless network facilitates users to access the diverse range of access network technologies which are differing in coverage range, bandwidth, throughput, latency and cost [14]. To provide a seamless connectivity among these heterogeneous technologies an efficient mobility management framework is needed that facilitates roaming to users from one network to another [4][5][6][7].

Thus, researchers are continuously making their efforts for developing a common platform for the heterogeneous wireless computing and communication networks to offer secure, seamless, and required bandwidth connectivity to users [8][9][11][16]. The rest of this paper is organized as follows. In the section II system architecture is described. Registration management protocol and handoff management protocol are presented in the section III and IV respectively. simulation and performance study is presented in section V. Finally, conclusion and future work is described in the section VI.

II. SYSTEM ARCHITECTURE

It is assumed that global wireless mobile communication network is divided into network domains (mobile switching centre-MSC), regions (sub-networks, i.e., base switching centre (BSC) controlled network) and Mobile Network Stations (MNSs) as shown in Figure 1. A MSC works as a network management server (NMS) in each network domain and keeps information other existing NMSs in the global wireless mobile communication system. A NMS behaves like a Home Subscriber Server (HSS) which houses subscribers profiles. It is also known as subscriber profile repository and formally known as home location register (HLR). It maintains the current position of all the MNs which are registered in that network domain or transited through. This component work as a carried forward for UMTS and GSM and is a central database which is consisting of information regarding all the available MSCs (network operator's subscribers). Other component of a NMS is Packet Data Network Gateway (PDNG) which communicates with the outside world, it works similar to packet data networks (PDNs), using SGi interface [10]. Each data packet network is identified by BS and MAC address of the MN. The PDNG act as a GPRS support node (GGSN) and providing GPRS support node (SGSN) for UMTS and GSM. A BSC is also known as serving gateway (SGW) and play similar role as a router, and forwards

information packets among a BS and a PDNG. The function of mobility management entity (MME) is to control the sophisticated process of the mobility using signalling messages and HSS. NMS also implements the Policy Control and Charging Rules (PCRs) and is accountable for policy control and management. It implements flow-based charging functionalities.

NMS is also equipped with Policy Traffic Switch (PTS) for identifying the location at which it intersects the traffic. It embeds a subscriber policy broker (SPB) and a service delivery engine (SDE) in the data plane of any network. It may be physical or virtual, with any combination of access technologies. Embedded within the PTS, the SDE makes policy decisions locally and prevents unnecessary signalling, reduces the load on NMS, and delivers faster decisions. SPB works like as single data warehouse for subscriber's information. SDE makes system to operate in heterogeneous environment.

A NMS also maintains information about all the available BSCs. A BSC defines a boundary of a region. It maintains unique name of each BSCs (regions). It identifies the region/BSC in/under which a MNS/MN is currently available.

A region (BSC) maintains information about all the available BSs/MNSs in a region. A MN may be a member of an available MSC/BSC covered area/region or may register in a new region (MSC/BSC). In a region, an authentication authorization access server (AAAS) is used to maintain database of users presently available at a BS. It works like a gateway (BSC) of a subnetwork. It contains information of location of each MNS/MN which are registered in that region or transited through it. It works like a visitor location register (VLR). This network station (NS) also acts as the Mobile Node Name Server (MN2S). It maintains unique name of all MNSs/MNs, registered in a particular region. When a new MNS/MN is registered it details are registered in the AAAS of its birth region.

III. REGISTRATION MANAGEMENT PROTOCOL

HSS uses a tuple of three attributes in the format (MN, FD, r) to stores the information in its database. In this tuple FD stands for foreign network domain and r stands for a region. It signifies that a MN may be present in the r of the FD or transited through it. A tuple of three attributes (MN, r, Nil) entered in AAAS shows that MN is available in region r or transited through it. A tuple in form (MN, Nil, MNS) signifies that a MN exists in that region and is at a particular BS/MNS. The name/id of MN is used as a primary key for NMS and MN2S.

A MN movement among network domains is always achieved through the NMS. An inter domain movement of a MN updates location in HSS of the current network domain and registers in the HSS of the destination network domain.

In intra region movement, a MN updates its current position in the AAAS of the region which is called as an Intra Region Location Update. When inter region movement takes

place. A MN updates the position information in the AAAS of current region and registers in AAAS of the destination region. It specifies a BS/MNS in that region to which it is travelling. This protocol uses three processes to deal with MNs: Identification (Id) attachment, movement and location update in the lifetime of a MN. In total this operation is called a particular phase for a MN movement. The protocol defines four atomic operations on home location register SPR (HSPR) and visitor location register SPR (VSPR). (a) Id attachment process is executed for naming a newly registered MN, whose birth location is also stored. This process signifies the insertion of a new tuple in the database. This process fails if a tuple with the same Id already exists in the database. (b) newloc process is executed when a MN changes its position, by moving to a new location. It updates the tuple already available in the database. (c) Find process is executed when interaction with a MN is required. For a given MN name, it returns the current location of the MN. (d) Id detachment process is executed when a MN id is no longer used (i.e., the MN has been disposed off). This process deletes the related tuple from the database.

Each NMS, BSC, BS and MNS are equipped an intelligent agent and a mobile agent (MA)[25] for maintaining network topology and current status of its neighbouring BSs. Normally agents observe technique for order preference by similarity to the ideal solution (TOPSIS) for getting the

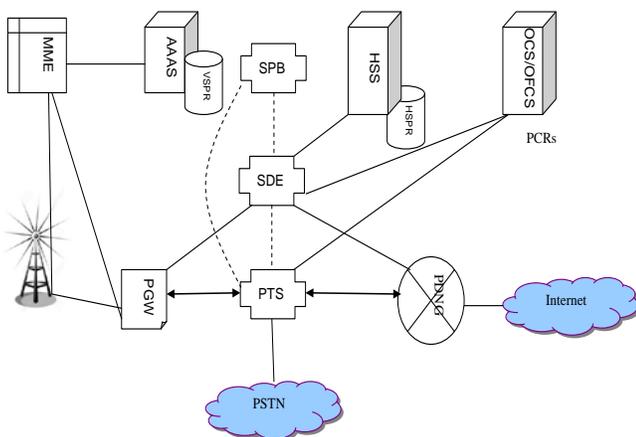


Fig. 1. Heterogeneous Wireless Network Environment Distribution Policy

services and network path from one point to another. In TOPSIS the best alternative is a shortest distance from the positive ideal key and the farthest distance from the negative ideal key. A matrix is constructed on the basis of set of alternative and corresponding attributes (criteria). In this method it is assumed that there are A_m alternatives and C_m criteria [23].

IV. HANDOFF MANAGEMENT PROTOCOL

In a heterogeneous wireless networks there are diverse access technology and each have different transmission power, bandwidth, security and cost. Thus, a single attribute is not sufficient to take handoff decision to provide a desired QoS. In heterogeneous wireless network environment, a dynamic negotiation helps to improve its performance of the network

for dynamic negotiation mobile agents technology play the key role. The goal of negotiation is to maximize the utility of a future decision. Each cell acquires a free channel for establishing a connection. In a given limit it requires to achieve a target. Negotiation stops as soon as defined limit is attained. A cell goes through this stage, when all channels are occupied except reserved channels. During this stage a mobile agent uses message exchange protocol for gathering the information about the channels status. In this stage it concludes for avoiding the handoff call blocking probability. In decision agent focuses on multi parameter systems (MPS). This system works on the parameters- network delay, received signal strength, and study of the collected information about adjoining network cells viz., accessible channel. Mobile agent sitting on the MSC/BSC/BS/MNS executes the following algorithm to update the candidate's network in database.

```
if (RSScurrent - RSSTh) < 0
  search for another network
endif
if (RSSnew - RSSTh) > 0
  update the network database
endif
```

A mobile execute the algorithm shown in Figure 2. After certain interval a MN reads the available network database in the area. If a network is there in the area then MN stay connected with available network and it checks regularly the network database within coverage area. If a network is not available in the coverage area then it terminates the process. If there is more than one available network then TOPSIS is used to arrange the network in an order. If the application priority is higher and bandwidth requirement is higher and low cost network is required then a MN chooses handoff to the alternative-1. If application priority is low and bandwidth requirement is higher and low cost network is required the MN goes handoff to the alternative-2. And if low bandwidth or low cost is required then the MN selects handoff to the alternative-3.

V. SIMULATION AND PERFORMANCE STUDY

For the deployment of agents in the presented model, a MATLAB and its Java features, MAC-SF [23] and PMADE agent framework [25][26] are used to study the agent migration under different network load conditions. To study the agent migration, three networks each having twenty one nodes are used. These networks are connected through one gateway. Main control frames are modified for simulating the essential measurement techniques and included the suitable information. The parameters communication cost, network bandwidth requirement of each MN, etc. are taken into consideration. This alteration does not modify the functioning of the system. A heterogeneous wireless network environment setup is created using the following technologies- WLAN, WiMAX [13], and GSM. Figure 3 shows the simulation scenario of the network used to simulate the agent guided distance based scheme and RSS scheme. The simulation network consists of minimum of 4 BS (GSM), 3 MNSs, 21 WLAN APs and 1 WiMAX point against 1 BS/1 MNS and 3 APs. A minimum number of 200 MNs are considered. This

number may increase or decrease because of the dynamic nature of the MNs. BS's/MNSs signals will be in overlapping fashion. In simulation cell radius is fixed to 1 KM. It is assumed that cells have equal bandwidth capacity and it is 10 MHz.

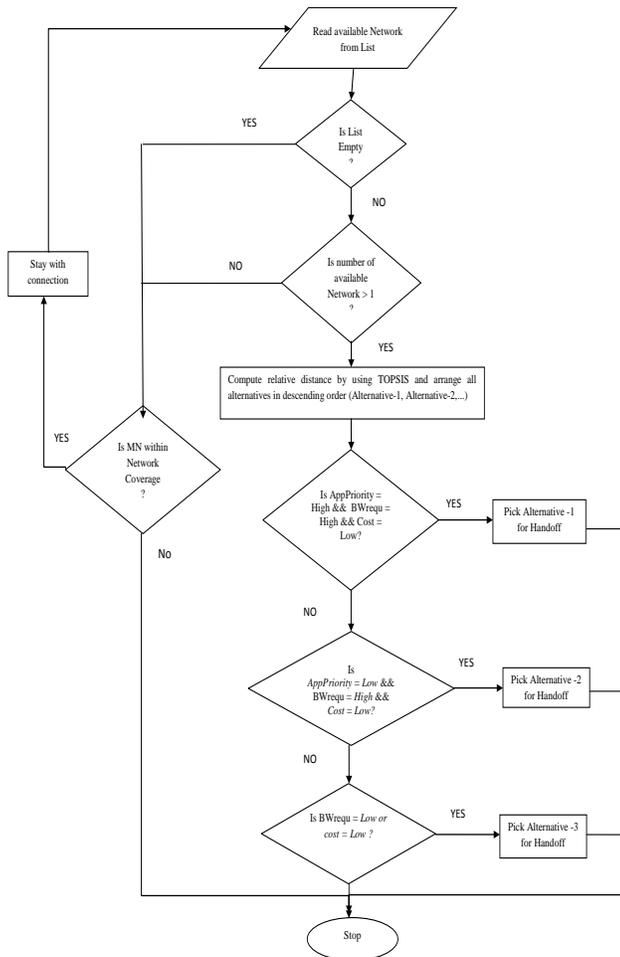


Fig. 2. A Multi Parameter Algorithm for Handoff

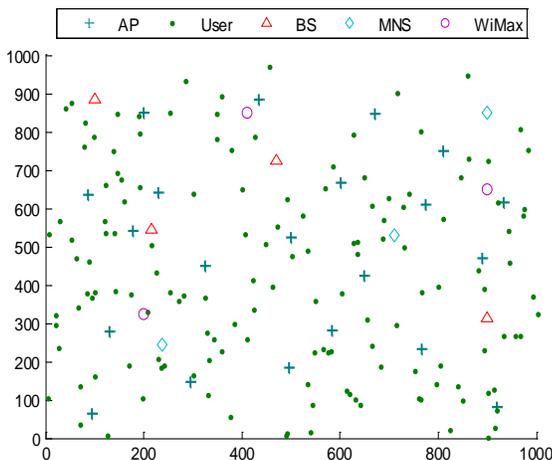


Fig. 3. Simulation Scenario

To avoid the boundary effect of a network size a wraparound edge cells are used. This ensures that the number of adjoining cells for a cell varies between three and six. This guarantees that there is always a network available for the communication. In the system there may be some areas which will be accessible by two or more networks. This covering helps the user to understand the network choice procedure in the presence of customer preferences.

The parameters given in Table 1 are used in the simulation. It is assumed that only an authenticated MN in the foreign networks (FNs) has a free access. It is also not taken into account authorization authentication and access (A³) and admission control mechanisms. A MN is moving with a fixed speed of 1.5 m/s. In simulation it is assumed that MN is executing real-time as well as non-real time applications, viz., audio/video as well as textual data. A priority is set as per the importance of applications to allow system to create single/multiple sessions.

Throughput of a network is directly related to the number of customers utilizing the link. Data transfer rates may differ upon accessible access network category, configuration, and load. A successful handoff decreases the number of conflicts on heavily loaded network. A conflict always regenerates the network traffic. Because packets are retransmitted and if this retransmission of packets reaches to the threshold value then packets are discarded. Figure 4 presents throughput in handoff mechanism for four systems. One is the conventional handoff system [19], second is cost based system in 3rd system [20] author is using intelligent agent and 4th is using intelligent stationary and mobile agents. In the past research algorithm when a MN is moving and a new network is detected whose RSS is higher and MN switched to higher RSS network. In MPS algorithm unnecessary handoffs are avoided and selection of network is not only based on the RSS but also on considered the application types, cost and peak-hour time. Agents (stationary intelligent and mobile agents) play the key role in successful handoff process when MNs are interested in priority based services. It decreases the number of conflicts at peak hour use of system and in heavy traffic load. The probability of loss of packets increases when the conflicts occur. To avoid such situation agent technology is used and better result is observed as shown in Figure 4.

A MN requirement always depends on its capability to process the received information. Power of a MN is measured in terms of communication & computing speed, memory, and power consumption against the throughput. Based on these parameters a node is being in the position to compute the requirement of the resources for the completion of the task which it wants to initiate in a network after the handoff. Further the same MN may change its expectation depending on the present context, or eventually as the network bandwidth improves and become available and/or access cost decreases. A network always announces about its usage and the available bandwidth. This information is intelligently maintained in the form of intelligent clustered database. System throughput is observed at a MN after the handoff completion.

TABLE I. ASSUMPTION AND VALUES OF PARAMETERS

Parameter	Values
Agent Platform	PMADE 1.1 and MAC-SF
Simulator	MLAB 15a
Simulation Area	1000x1000 m ²
Simulation time	200 seconds
Mobile Nodes	200
Access Points	21
BS	4
MNS	3
Number of WiMax	3
Users	128
Threshold(WLAN to cellular network)	-85 dBm
WLAN range	200 m
Access Point Transmitter Power (min)	30 dBm
Access Point Transmitter Power (max)	100 dBm
BS Transmitter Power	33 dBm
Cable loss	1.7 dB
Threshold (cellular network to WLAN)	-80 dBm
Channel gain power	33 dBm
Antenna height of BS	30 m
BS Operating Frequency	894 MHz
Bandwidth for GSM	10 MHz
Maximum output power of GSM	39dBm
Data rate	2 Mbps
Communication Channels	16
Traffic type	Audio/Video/Text
Transmit power	5 mW
AP capacity	40 sessions

Figure 4 gives a comparative study of different approaches. After the handoff execution classic and cost based approaches are showing the average throughput for transferring 1 MB of data. This handoff takes place between GSM and WiMAX networks. The result also shows that the presented schemes alter the networks between 1 and 6 seconds. Further, it is also observed that handoff period varies between 1 to 3 seconds. The professed throughput of MPS approach is steady till the completion of the task. The intelligent clustered database is key factor behind a handoff toward the best available network. MPS is an intelligent choice method gives the priority to network bandwidth, communication cost and continued with its activity till the end of the initiated task. Mobile agents working in the network are periodically updating the database for the selected network parameters. Thus, every network is always being with the

updated and above the threshold value of the RSS. Further this property of the system removes RSS issue of connectivity. MPS improves the system throughput in comparison to cost based, classic handoff and approach given in [20]. A MN moves arbitrarily and channel fading is a function of distance. It may vary with the changes in distance among the MN and BS. It also showed that the RSS of two BS change up and down.

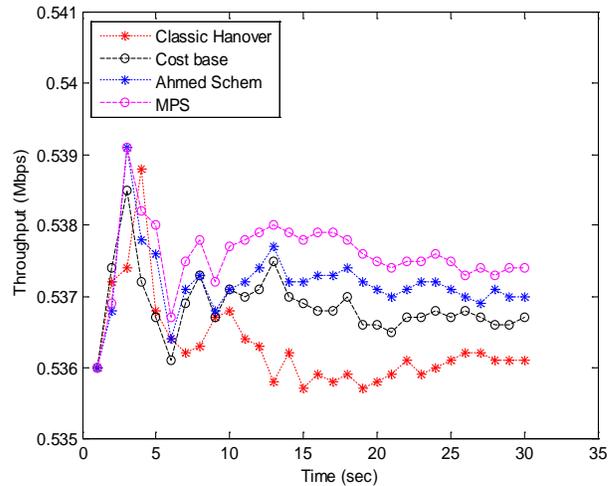


Fig. 4. Average Throughput

Figure 5 shows that the average handoffs delay for voice packet transmission while performing handoff. Handoff delay depends on the routing delay. And same is considered in the simulation while switching from one network to another. The make before-break scheme helps MPS to achieve low handoff delay. MPS improves network performance approximately by 6%, 13% and 21% in comparison to approach given in [20], cost-based and classic handoff schemes, respectively.

Another parameter which is important in handoff process is to compute the handoff blocking rate. It is percentage of calls which are not able finish their services. It may be due selection of an un-appropriate wireless access network or due to unavailability of the list of available wireless access networks. When a BS fails to assign a free channel an incoming call process is automatically blocked. The overall system service stability always depends on blocking rate. If a system is reflecting low blocking rate means system is efficient. TOPSIS method is used for selection of the target network which reduces the probability of blocking of mobility management protocols. This model takes into consideration updated database maintained by the mobile agents. The network topology parameter plays major role in the analysis because it gives the mobility patterns of a MN.

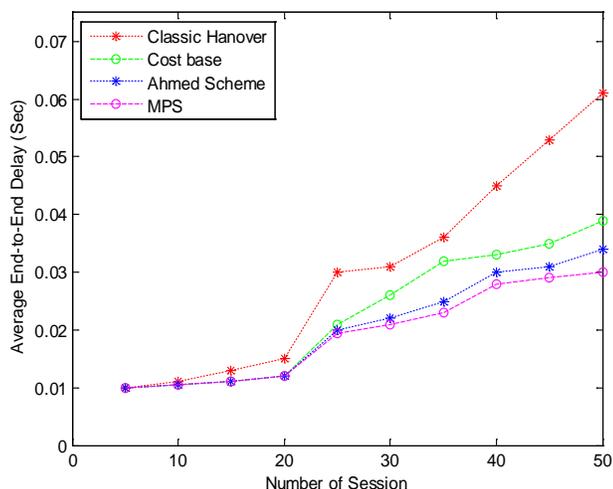


Fig. 5. Average end-to-end transmission delay

Figure 6 shows that MPS decreases handoff blocking rate in comparison to [20], classic scheme [22] and cost-based scheme [21]. In simulation it is assumed that 64 MNs are working simultaneously. It is seen that when the number of MN are less than 25–27, all the systems behave similarly. But when this active number of MNs increases, MPS approach outperforms in comparison to classic and cost-based systems but it performs much better with [20]. There may be reason that the handoff resources, in these two (classic and cost-based) systems are not adequate to satisfy all the handoff processes and better list of resources available in case of [20] to satisfy all the handoff requests. MPS steadies even after the access of more MNs.

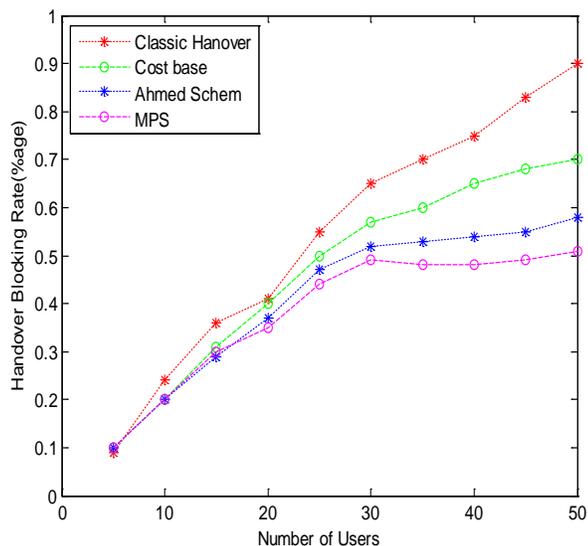


Fig. 6. Handoff blocking rate comparison

VI. CONCLUSION AND FUTURE WORK

In this paper a mobile agent based heterogeneous wireless network environment distribution policy is presented for 4G

networks. This is MPS based system in which handoff initiations and decisions are delegated by intelligent mobile agents. A clone of agent is required to serve a MN. A handoff is initiated by a client agent which is executed either in a MN or at the network side, depending on the basis of handoff generating actions. Normally it is observed that MN devices having limited battery power, computing power so, it is useful if a handoff decision takes place at the network side. MPS permits both type of access facility and core carrier services for handoff decision for better adjustment to client and demand requirements. Proposed system maintains an intelligent clustered of database of carrier service and agent for carrier selection. The system is evaluated through a simulation is carried out focusing on its impact on handoff. The results are satisfactory for the feasibility of the presented system. Future work will be focused on mobility management for high speed users.

REFERENCES

- [1] Anna Marie Vegni, Gabriele Tamea, Tiziano Inzorilli and Roberto Causani, "A combined Vertical Handover Decision Metric for QoS Enhancement in next Generation Network," IEEE international conference on wireless and mobile computing, Networking and Communication, 2009.
- [2] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical Modelling and computing techniques," Wireless Communications, IEEE, 12(3), 42-48, 2005.
- [3] A. H. Zahran, B. Liang, and A. Saleh, "Signal threshold adaptation for vertical handoff in heterogeneous wireless networks," Mob. Netw. Appl., vol. 11, no. 4, pp. 625–640, 2006.
- [4] Ben-Jye Chang and Jun-Fu Chen, "Cross-Layer Based Adaptive Vertical Handoff with Predictive RSS in Heterogeneous wireless Network," IEEE Transaction on Vehicular Technology, Vol. 57, No. 6, November 2008.
- [5] Q. Yu, W. Jiang, and Z. Xiao, "3G and WLAN heterogeneous network handover based on the location information," International Conference on Communications, Circuits and Systems (ICCCAS), Nov. 2013, Vol. 2, pp. 50-54.
- [6] X. Li and R. Chen, "Adaptive vertical handover algorithm based on user experience for heterogeneous network," 6th International Congress on Image and Signal Processing (CISP), December 2013, Vol. 3, pp. 1540-1544.
- [7] A. Çalhan and C. Çeken, "Case study on handoff strategies for wireless overlay networks," Computer Standards & Interfaces, 35(1), 170-178, 2013
- [8] Sang-Jo Yoo, D. Cypher, and N. Golmie, "Timely effective handover mechanism in heterogeneous wireless networks," Wireless Personal Communications, 52(3), 449-475, 2010.
- [9] R. Verma and N.P. Singh, "GRA based network selection in heterogeneous wireless networks. Wireless personal communications, 72(2), 1437-1452, 2013.
- [10] F. Bari and V. Leung, "Automated network selection in a heterogeneous wireless network environment," Network, IEEE, 21(1), 34-40, 2007.
- [11] X. Yan, N. Mani, and Y.A. Şekerioğlu, "A travelling distance prediction based method to minimize unnecessary handovers from cellular networks to WLANs. Communications Letters, IEEE, 12(1), 14-16, 2008.
- [12] S. Mohanty and I. Akyildiz, "A cross-layer (layer 2 + 3) handoff management protocol for next-generation wireless systems," IEEE Trans. Mobile Computing, 2006.
- [13] Z. Becvar, P. Mach, and B. Simak, "Improvement of handover prediction in mobile WiMAX by using two thresholds," Computer Network, vol. 55, no. 16, pp. 3759–3773, 2011.

- [14] H. Wang, R. Katz, J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks," In Proceedings of 2nd IEEE Workshop on Mobile Computing Systems and Applications, (WMCSA'99), 1999, pp. 51–60.
- [15] W. Chen, J. Liu, H. Huang, "An adaptive scheme for vertical handoff in wireless overlay networks," in: Proceedings on the 10th International Conference on Parallel and Distributed Systems (ICPADS 2004), 2004, pp. 541–548.
- [16] O. Ormond, J. Murphy, G. Muntean, "Utility-based intelligent network selection in beyond 3G systems," IEEE International Conference on Communications (ICC 2006), vol. 4, pp. 1831– 1836.
- [17] B.-J. Chang and J.-F. Chen, "Cross-layer-based adaptive vertical handoff with predictive RSS in heterogeneous wireless networks," IEEE Trans. Veh. Technol., 2008.
- [18] B.-J. Chang, J.-F. Chen, C.-H. Hsieh, and Y.-H. Liang, "Markov decision process-based adaptive vertical handoff with RSS prediction in heterogeneous wireless networks," in IEEE Wireless Communications and Networking Conference, 2009, pp. 1–6.
- [19] J. Raiyn, "A Novel Handover Scheme Based on Adaptive Agent for Reducing Real-Time Communication Latency in Automation Environment," EUROSIM, pp. 555-561, April, 2008.
- [20] Atiq Ahmed, Leïla Merghem-Boulahia, and Dominique Gaïti. "An intelligent agent-based scheme for vertical handover management across heterogeneous networks," annals of telecommunications-Annales des telecommunications 66, no. 9-10, pp. 583-602, 2011.
- [21] W. Yifei, L. Xiaowei, S. Mei, S. Junde, "Cooperation radio resource management and adaptive vertical handover in heterogeneous wireless networks". In: International conference on natural computation, vol 5, 2008, IEEE Computer Society, Los Alamitos, pp 197–201.
- [22] Ylianttila M, Mäkelä J, Pahlavan K, "Analysis of handoff in a location-aware vertical multi-access network," Computer Network, 47:185–201, 2005.
- [23] Sanjeev Prakash, R. B. Patel and V. K. Jain, "Movement Assisted Component Based Scalable Framework For Distributed Wireless Networks", International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5, pp. 71-86, 2015.
- [24] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks", Computer Communications, 31(10), pp.2607-2620, 2008.
- [25] R. B. Patel, K. Garg, "PMADE - A Platform for Mobile Agent Distribution & Execution," in Proceedings of 5th World MultiConference on Systemics, Cybernetics and Informatics (SCI2001) and 7th International Conference on Information System Analysis and Synthesis (ISAS 2001), Orlando, Florida, USA, July 22-25, 2001, Vol. IV, pp. 287-292.
- [26] R. B. Patel and K. Garg, "A New Paradigm for Mobile Agent Computing," WSEAS Transaction on Computers, 1(3): 57-64, Jan. 2004.

Development of Adaptive Mobile Learning (AML) on Information System Courses

I Made Agus Wirawan¹

Department of Informatics Technology Education,
Faculty of Engineering and Vocational,
Ganesha University of Education, Bali, Indonesia

Made Santo Gitakarna²

Department of Electronics Technnics Education,
Faculty of Engineering and Vocational,
Ganesha University of Education

Abstract—In general, the learning process is done conventionally, where the learning process is done face to face between teachers with learners in the classroom. Teachers have a very important role in determining the quantity and quality of the implementation study. Therefore, teachers must think and plan carefully to improve learning opportunities for learners and improve the quality of teaching. Along with the development of mobile technology and communication is rapidly increasing, enabling the learning process is not only done in the classroom, but can be done anywhere and anytime. Based on the analysis of the results of observations in the class conducted by a researcher and as a teacher in the learning courses of Information Systems, found some obstacles encountered during the learning process

This research is to develop an Adaptive Mobile Learning on Information Systems courses. The method used in this research is the development of research methods (research and development), which selected the design development using System Development Life Cycle model. Adaptive Mobile Learning will be validated and tested through three phases of testing are: (1) Product technical test as a software. (2) Testing of the product as a medium of learning, through expert review by a media expert, (3) Field test to evaluate the response of the students that learned Adaptive Mobile Learning.

The results show that Adaptive Mobile Learning software is can present the material in the course of Information Systems. Media Adaptive Mobile Learning can be used as an alternative medium (supplement) of learning Information Systems courses. The response of students to the development and use of software for Adaptive Mobile Learning Information Systems courses is likely to very positive, which is at 67.7% very positive and 32.3% is positive.

Keywords—*Mobile Learning; Information System Course; Learning Media; Adaptive Learning; Learners Response; Research and Development*

I. INTRODUCTION

In general, the learning process is done conventionally, where the learning process is done face to face between teachers with learners in the classroom. Teachers have a very important role in determining the quantity and quality of the implementation study. Therefore, teachers must think and plan carefully to improve learning opportunities for learners and improve the quality of teaching. Along with the development of mobile technology and communication is rapidly increasing, enabling the learning process is not only done in the classroom, but can be done anywhere and anytime.

Based on the analysis of the results of observations in the

class conducted by a researcher and as a teacher in the learning courses of Information Systems, found some obstacles encountered during the learning process, such: 1) Students are less active in taking the time to learn the material outside of class Information Systems. That is because the learning tools that are used less flexible. 2) The less effective learning process in the Department of Informatics Technology Education especially on Information System course, caused by some national holidays and religious holidays. 3) The learning process is given by the lecturer in the classroom of a general nature, which is considered the same level of students ability by lecturers. E-Learning can be viewed as an innovative approach for delivering well-designed, learner-centered, interactive, and facilitated learning environment to anyone, anytime by utilizing the attributes and resources of various digital technologies along with other forms of learning materials suited for open, flexible, and distributed learning environment” [1]. However, eLearning courses have witnessed high drop out rates as learners become increasingly dissatisfied with courses that do not engage them [2]; [3].

Each student has a different cognitive abilities [4]. In [5]; [6]; [7]; [8]; [9] those problems can be overcome with the use of adaptive learning system. Many studies have been conducted on adaptive learning. Among them, studies that provide the most universal method of adaptability offer courseware by considering learner styles [10]; [11]; [12]; [13]; [14]; [15]; [16]. However, there are also theories that assert that a learning strategy created according to either a task or content is much more effective than the learning style [17]; [18]; [19]. It is thus necessary to provide adaptability according to the learning content along with the learning style.

In [20] the previous study researchers have developed a smartphone application on the SQL Advanced Database for the same student. This material is still static, and the system was not able to present the material by the ability level of each learner.

Based on the description of the problem and the other researches, researchers looked at the need for the development of Adaptive Mobile Learning (AML) to support the learning process of Information Systems. This research is the development of technology-based adaptive learning media on Information Systems material. Application of this smartphone can be used as a supplement in the learning process of Information System.

II. LITERATURE REVIEW

A. Introduction of Adaptive Learning

Adaptive Learning based on constructivist theory and the theory of cognitive flexibility. Adaptive learning is a specific way of learning in the process of solving a particular problem, learners acquire knowledge and skills through positive thinking and operating. Adaptive learning is active learning. Learners can monitor their own learning process, and choose the most appropriate learning content to their actual needs [21].

B. Features of Adaptive Learning

The rapid development of the Internet and new technologies, which are related to the distance learning system in the network environment has been greatly advanced. Adaptive learning system emerged in response to the characteristics of learners. The above systems have common features as follows [21]:

1) A personalized learning system with the learner as the main body. In accordance with the learning needs, abilities and learning styles of learners. The system actively adjusts the learning content, learning styles, learning strategies, learning flow and learning support, and present learning materials in accordance with the level of ability of learners, whole learning process centered on the learner to meet the learning process of students, in which the dominant position of learners is fully realized.

2) The self-construction of knowledge. Learners are actively interacting with adaptive learning system and analyzing feedback information, which aims to build up their knowledge. By recording system adaptive learning of the learning process, learners can control with timely and adjust their own learning process to achieve the learning objectives required.

3) The adaptive learning system is intelligent. Intelligence is a basic guarantee for the adaptation of the system to realize independently. This allows the system comprehensively in diagnosing the extent and actual psychological condition of students. So that the learning content is presented and learning support in accordance with the pretest and tracking the learning process.

C. Learner Profile

In adaptive system, learner profile components use to obtain student information. This information is stored without making changes, and does not close the possibility of changing information. Changes occur because learner profiles information such as: level of motivation, learning style, and others also change. The learner profile has four categories of information that can be used as benchmarks, namely: [15].

1) Student's behavior, consists of information: level of motivation, learning style and learning materials.

2) Student's knowledge, the information about the knowledge levels of students. There are two approaches that can be used, namely: the test automatically (auto-evaluation) by an adaptive system and the test manual (manual-

evaluation) by a teacher. Levels of knowledge students can be categories: new, beginner, medium, advance and expert.

3) Student's achievement, the information relate to student achievement results.

4) Student's preferences, which explain the concept of information preferences, such as: cognitive preferences: (introduction, content, exercise, etc.), preferences physical support (text, video, images, etc.).

III. METHODOLOGY

A. Research setting and procedures

The method used in this research is the development of research methods (research and development. Because the media developed in the research development will produce the final product in software simulation program, then software development method used is the System Development Life Cycle (SDLC) Model. SDLC method is a method of software development is structured.

B. Adaptive Mobile Learning Concept

Functional design/learning process flow of Adaptive Mobile Learning in this study looks at the figure 1.

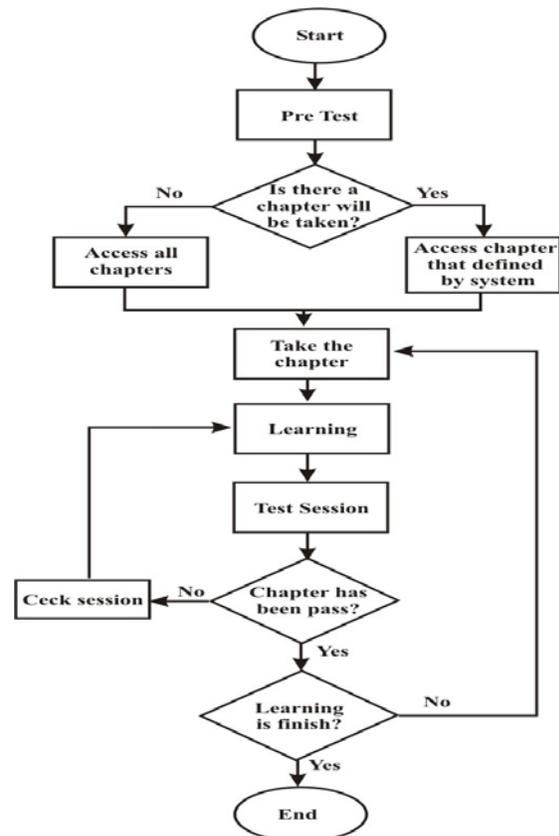


Fig. 1. Flow chart of Adaptive Mobile Learning [9]

The stages of the learning process adaptive mobile learning, as follows:

1) Perform pre-test trials on learners before doing the learning of a chapter (courses), learners will acquire a pre-test with the following provisions.

- a) Questions were taken representing each chapter.
- b) Teachers determine the questions selected for the pre-test.
- c) The results passed or not passed are determined by the percentage of the chapter of correct answers to the questions in each chapter, where the teacher determines the percentage of pass for each chap Questions are taken that represent each chapter.
- d) The selection of question is determined through the analysis of test item such as content validity, item analysis and internal consistency of item and test.
- e) Teacher determines number of questions for pre test.
- f) Results of pass or not are determined by the percentage of chapter number of correct answers on the questions in each chapter, wherein teachers determine the percentage of pass for each chapter.

2) The purpose of the pre-test is to determine the ability of early learners. After testing the pre-test, the system will provide a chapter that can be accessed in accordance with the level of understanding of learners. Some possible after pre-tests conducted.

a) *There is no chapter that pass*

If all chapters do not pass, the learners may only access the lowest chapter that has not been passed.

b) *Some or all of the chapters pass*

If there are several chapters that pass, then only chapter passed that can be accessed. If all the chapters pass then all the chapters can be accessed.

3) Taking a chapter. There are two conditions in taking the matter:

- a) The conditions in which learners are free to choose the material.
- b) The conditions in which learners must take the material determined by the system. This happens because there are chapters that do not pass the pre-test.

4) Taking the test session. After the learning process, learners are required to take a test session. Provisions in the test session is as follows:

- a) Questions used from chapters that have been taken.
- b) Standard passing score is determined by the teacher
- c) If learners do not take the test session, learners are not able to continue the learning process.
- d) The results obtained are passed or not passed on each chapter.

There are several possibilities in this test session:

- 1) If learners do not pass the test session, the learners will repeat the learning process for the chapters that did not pass. Repetition of the learning process will be stored by the system as a learning session.
- 2) If learners pass, then learners can continue to the next chapter.

3) If all the chapters have been completed, learners can complete the learning process.

C. *Research Location*

The location study was conducted in the Department of Informatics Technology Education, Faculty of Engineering and Vocational, Ganesha Education of University.

D. *Variables*

The independent variable in this study is the tool of learning interaction on subjects Information Systems are used, namely by Adaptive Mobile Learning that was developed in this study. The dependent variable were measured in the study is a learners response in the Department of Informatics Technology Education are use of Adaptive Mobile Learning as a tool for learning interaction on subjects Information Systems.

E. *Samples/Subject of research*

This research was conducted at the Department of Informatics Technology Education. Which will be samples/subjects in this study were learners at the Department of Informatics Technology Education who took a course of Information Systems In the odd semester of academic year 2014/2015.

F. *Sampling Techniques*

In the odd semester of academic year 2014/2015, the number of learners that as many as 31 people (1 class) and own a smartphone. The data taken from this research is learners response using the Adaptive Mobile Learning as a tool (supplement) in the learning process in the subject of Information Systems.

G. *Data Analysis*

Learners responses were analyzed using a questionnaire with Likert scales of 5 (the value of 1 to 5) were analyzed descriptively. Conversion learners response rates can be seen in Table I below:

TABLE I. CONVERSION TABLES LIKERT SCALE LEARNERS RESPONSE [22]

The range of values	Response categories
$M_i + 1,5 S_i \leq x$	Very Positively
$M_i + 0,5 S_i \leq x < M_i + 1,5 S_i$	Positive
$M_i - 0,5 S_i \leq x < M_i + 0,5 S_i$	Hesitant
$M_i - 1,5 S_i \leq x < M_i - 0,5 S_i$	Negative
$x < M_i - 1,5 S_i$	very negative

$M_i = 1/2$ (highest score ideal + lowest score ideal)
 $S_i = 1/6$ (highest score ideal - lowest score ideal)

H. *Criteria for the success of this research*

Development and the use of Adaptive Mobile Learning as interaction tools Information Systems course is considered successful if it meets the criteria of the research are: 1) The presence of Adaptive Mobile Learning as a tool in teaching Information Systems. 2) Learners response in the Department of Informatics Technology Education to use of Adaptive Mobile Learning as a tool for learning interaction Information Systems achieve positive category or more

IV. RESULTS

A. System Implementation

This research is the development of Adaptive Mobile Learning is applied to the Information Systems course. This research was conducted for 8 months starting from March to October. Here are excerpts of the program that has been developed.



The screenshot shows a mobile application window titled 'Login'. It contains an icon of a person with a key, a text input field for 'User Name', and another for 'Password'. At the bottom, there are two buttons: 'Batal' on the left and 'Login' on the right.

Fig. 2. Login form

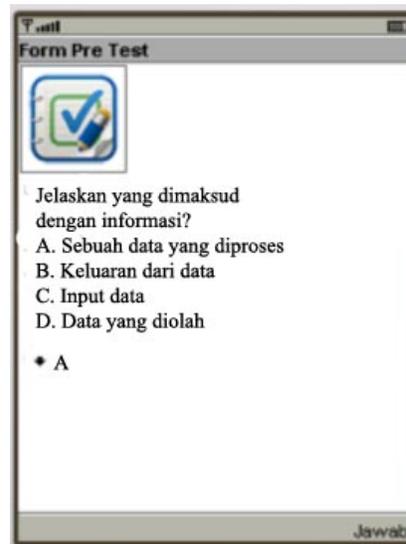
This form is used by the learner to the login process. Login process using username (Student ID Number) and password.



The screenshot shows a mobile application window titled 'Form Petunjuk'. It features a blue character icon holding a magnifying glass and a book. Below the icon, it displays the text 'Selamat Datang' followed by user details: 'NIM : >>12', 'Nama : >>Agus Wirawan', and 'Alamat : >>Bangli'. At the bottom, there are two buttons: 'LogOut' on the left and 'Test' on the right.

Fig. 3. Main form

The main form displays information of learners who successfully login. Information such as identification numbers, names and addresses. In the main form, learners can perform the logout process or Test. Test button was used to measure the ability of learners before the learning process begins.



The screenshot shows a mobile application window titled 'Form Pre Test'. It contains an icon of a checkmark and a question: 'Jelaskan yang dimaksud dengan informasi?'. Below the question are four multiple-choice options: 'A. Sebuah data yang diproses', 'B. Keluaran dari data', 'C. Input data', and 'D. Data yang diolah'. Option 'A' is selected with a radio button. At the bottom right, there is a 'Jawab' button.

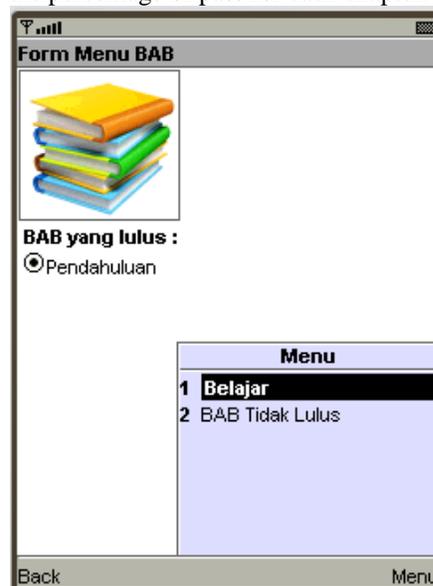
Fig. 4. Pre test form

Form pre-test is used to measure the initial capabilities of learners. There are several provisions in the pre-test processes such as:

1) The questions are presented is the early material of each chapter.

2) The teacher can determine the number or type of questions selected for the pre-test.

3) The results passed or not is determined by the percentage of the number of chapters of the correct answers to the questions in each chapter, where the teacher determines the percentage of pass for each chapter.



The screenshot shows a mobile application window titled 'Form Menu BAB'. It features an icon of a stack of books. Below the icon, it says 'BAB yang lulus :'. Underneath, there is a radio button selected for 'Pendahuluan'. At the bottom, there is a table with the following content:

Menu	
1	Belajar
2	BAB Tidak Lulus

At the bottom of the window, there are two buttons: 'Back' on the left and 'Menu' on the right.

Fig. 5. Chapter form

This form is used to display the chapter material that pass or not pass from the pre test. This form is used to display the chapter material that pass or not pass from the pretest. There are two conditions in the learning process:

- 1) The learner can access the chapter that has passed freely
- 2) learners can access the chapter which does not pass (pre-test) in accordance with the directives of the system.

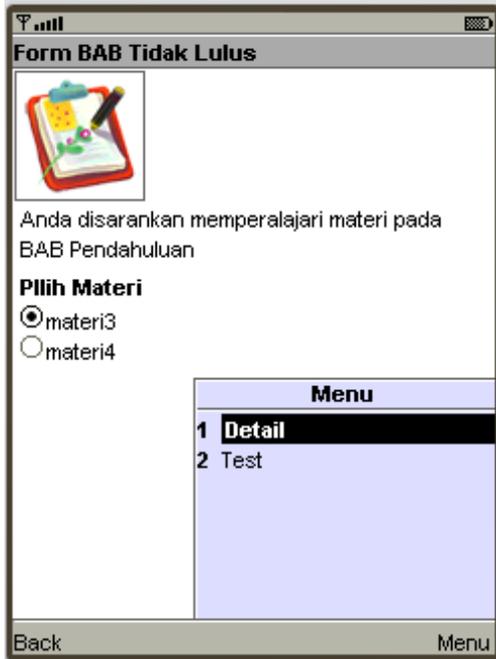


Fig. 6. Content form

This form are used to display the detail of material of each chapter.

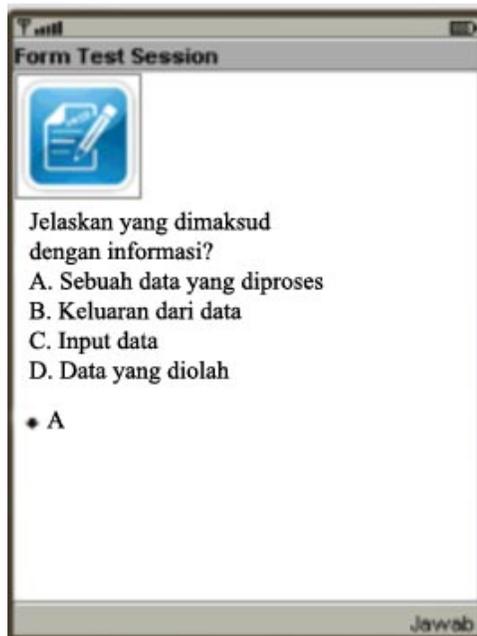


Fig. 7. Test session form

This form is used to display the questions of the test session

B. System testing process

In this study conducted three testing phases, namely:

1) *Technical testing, the test was conducted by researchers themselves to functional test of software (Black box testing) and examine the structure of software algorithms (White box testing).*

2) *Software testing as a medium of learning, carried out through expert review content, and media experts.*

3) *Field testing process, testing is given to learners in Department of Informatics Technology Education, who took a course on Information Systems. Results of learners response that consists of a 10-point declaration was measured with a Likert scale of 1-5. Ideal highest score and the lowest score is the ideal respectively 50 and 10. The calculation of the mean and standard deviation are as follows.:*

$$\text{The mean of ideal} = \frac{1}{2} (\text{highest score ideal} + \text{lowest score ideal})$$

$$= \frac{1}{2} (50 + 10)$$
$$= 30$$

$$\text{Deviation standard of ideal} = \frac{1}{6} (\text{highest score ideal} - \text{lowest score ideal})$$

$$= \frac{1}{6} (50 - 10)$$
$$= 6,7$$

Response categories of students who use Adaptive Mobile Learning in Information Systems courses shows that students that learned responses Adaptive Mobile Learning in Information Systems course is very positive for 67.7% and 32.3% positive. It can be concluded that the response of the learners tend to be very positive.

C. Discussion

At this stage of the software development using J2ME programming language and MySQL database. For the connection between mobile phone with a database requires server using the PHP language. Utilization of mobile devices on the learning process of Information Systems can be done anywhere and anytime. Besides, the process of learning can be adjust the level of ability of the students, this is because the concept of adaptive learning.

At the time limited testing in a class, to determine the response of students to use the software in Information Systems course, visible a very positive response who use Adaptive Mobile Learning. With the very positive response from the students on the course Information System using Adaptive Mobile Learning software, it is hoped able to increase the level of student ability of Information Systems, and is expected to improve student-learning outcomes. For this year, the research was conducted only the software development and limited testing to determine the response of students using Adaptive Mobile Learning.

V. CONCLUSIONS AND FUTURE WORK

Based on this research, suggestions is can be presented as:

1) *The design of Adaptive Mobile Learning can model of Information System material.*

2) *Implementation of Adaptive Mobile Learning software can present material of information system and as an alternative media in the process of learning materials of information system anywhere and anytime.*

3) *The student response to the development and use of Adaptive Mobile Learning software for learning Information System is very positive, with 67.7% is very positive, and 32.3% is positive.*

Based on this research, suggestions is can be presented as follows:

1) Need more mobile learning media to help students in the learning process anywhere and anytime and supports the concept of go green campus.

2) For further research can continue the experimental study of this study, so that the learning outcomes of the use of Adaptive Mobile Learning of Information Systems can be known.

3) For further research can be combined with adaptive learning models so that the learning process will be more interactive.

ACKNOWLEDGMENT

The authors express their gratitude to the faculty, staff and students of Department of Information Technology Education, Faculty of Technical and Vocational Education Ganesha University who has helped and supported the implementation of this research, especially on the data of student responses.

REFERENCES

- [1] Khan. B, "Managing E-Learning Strategies: Design, Delivery, Implementation and Evaluation", Hershey, PA, USA: Idea Group Inc, 2001.
- [2] Meister. J, Pillars of e-learning success, New York, USA: Corporate University Exchange, 2002.
- [3] Frankola. K, Why online learners dropout. Workforce, 10, 53-63, 2001.
- [4] Hernawati. K., "*E-Learning Adaptif Berbasis Karakteristik Peserta Didik*", Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA, Fakultas MIPA, Universitas Negeri Yogyakarta, 14 Mei 2011, ISBN: 978-979-99314-5-0, 2011.
- [5] Sfenrianto, "*A Model Of Adaptive E-Learning SystemBased On Student's Motivation*", Proceedings International Conference on Creative Communication and Innovative Technology (ICCIT), 8 Agustus 2009, Tangerang-Indonesia , ISSN 1978-8282, 2009.
- [6] Surjono. H. D, "Pemanfaatan Teknologi E-Learning Adaptif untuk Mengatasi Keragaman Gaya Belajar", Jurnal Penelitian Saintek, Vol 8, No 1, April 2013.
- [7] Dantes. G. R, Suarni. N. K and Sujaya. I. G, "*Model Dynamic Intellectual Learning (DIL): Pergeseran Paradigma E-Learning Menuju Adaptive Learning*", Konferensi Nasional Sistem dan Informatika 2010; Bali, November 13, 2010
- [8] Esichaikul. V, Lamnoi. S and Bechter. C, "*Student Modelling in Adaptive E-Learning Systems*", Knowledge Management & E-Learning: An International Journal, Vol.3, No.3, 2011.
- [9] Wirawan. I. M. A and Wahyuni. D. S, "*Adaptive Mobile Learning Concept*", Proceedings of the International Mobile Learning Festival, Bali - Indonesia June 2 – 4, 2014.
- [10] C. A. Carver, R. A. Howard and E. Lavelle, "Enhancing student learning by incorporating learning styles into adaptive hypermedia", Proceedings of ED-MEDIA '96 World Conf. on Educational Multimedia and Hypermedia, (1996) June, 17-22, Boston, USA.
- [11] J. E. Gilbert and C. Y. Han, "Adapting instruction in search of a significant difference", Journal of Network and Computer Applications, vol. 22, (1999), pp. 1-12.
- [12] M. Stern and P. Woolf, "Adaptive content in an online lecture system", Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web based System, (2000) August 28-30, Trento, Italy, pp. 29 1-300.
- [13] M. Grigoriadou, K. Papanikolaou, H. Kornilakis and G. Magoulas, "INSPIRE: An intelligent system for personalized instruction in a remote environment", Proceedings of 3rd Workshop on Adaptive Hypertext and Hypermedia, (2001) July 13-17, Sonthofen, Germany, pp. 13-24.
- [14] P. Paredes, and P. Rodrigues, "Considering sensing-intuitive dimension to exposition-exemplification in adaptive sequencing", Proceedings of AH2002 Workshop, Second International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, (2002) May 29-31, Malaga, Spain, pp. 556-559.
- [15] N. Stash, A. Cristea and P. D. Bra, "Authoring of learning styles in adaptive hypermedia: problems and solutions", Proceedings of the WWW 2004 Conference, (2004) May 17-22, New York, USA, pp. 114-123.
- [16] J. Lee, "Adaptive Courseware Using Kolb's Learning Style", International Magazine on Advances in Computer Science and Telecommunication, vol. 3, no. 1, (2012), pp. 45-59.
- [17] G. K. Tallmudge and J. W. Shearer, "Study of Training Equipment and Individual Differences: the Effects of Subject Matter Variables", Supplementary Report, American Institutes for Research, (1968).
- [18] D .C. Berliner and L. S. Cahen, "Trait-Treatment Interaction and Learning", Review of Research in Education, vol. 1, (1973), pp. 58-94.
- [19] D. H. Jonassen, "Aptitude-versus content-treatment interactions, implication for instructional designs", Journal of Instructional Development, vol. 5, no. 4, (1982), pp. 15-27.
- [20] Wirawan, "Development Learning Media for Mobile Phone Based Materials Basic SQL Syntax in Subjects Advanced Database (Case Study on IT Educational Programs Semester III)", Prosiding Seminar Internasional Teknologi Informasi dan Pendidikan, Bridging ICT and Education, ISSN 1907-3739, DIN EN ISO 9001:2008, Cert. No.01 100 086042, Fakultas Teknik Universitas Negeri Padang, 2011.
- [21] Qing L, Shaochun Z, Peng W, Xiaozhuo G, and Xiaolin Q, "*Learner Model in Adaptive Learning System*", Journal of Information & Computational Science 7: 5 (2010) 1137-1145
- [22] Sugiono, Metode Penelitian Kuantitatif, Kualitatif dan R & D, Alfabeta, Bandung, 2012.

AUTHOR PROFILE



I Made Agus Wirawan, S.Kom., M.Cs he is a lecturer in Ganesha University of Education. He earned her Master degree from Gadjah Mada University. Her research interests include artificial intelligence, adaptive learning, case based reasoning, and mobile phone technology. He has published papers in International Journal of Computer Applications, Journal of Educational Development Research, and National journal of information engineering education and among others.



Made Santo Gintakarna, S.T., M.T he is a lecturer in Ganesha University of Education. He earned her Master degree from ITS. Her research interests include artificial intelligence, E-learning, and robotic.

Ontology-Based Clinical Decision Support System for Predicting High-Risk Pregnant Woman

Umar Manzoor

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, KSA

Muhammad Usman

Department of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan

Mohammed A. Balubaid

Industrial Engineering Department, Engineering Faculty
King Abdulaziz University
Jeddah, KSA

Ahmed Mueen

Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, KSA

Abstract—According to Pakistan Medical and Dental Council (PMDC), Pakistan is facing a shortage of approximately 182,000 medical doctors. Due to the shortage of doctors; a large number of lives are in danger especially pregnant woman. A large number of pregnant women die every year due to pregnancy complications, and usually the reason behind their death is that the complications are not timely handled. In this paper, we proposed ontology-based clinical decision support system that diagnoses high-risk pregnant women and refer them to the qualified medical doctors for timely treatment. The Ontology of the proposed system is built automatically and enhanced afterward using doctor's feedback. The proposed framework has been tested on a large number of test cases; experimental results are satisfactory and support the implementation of the solution.

Keywords—High-risk patient; Pregnant woman; Ontology-based CDSS; Clinical Decision Support System

I. INTRODUCTION

The world has a shortage of professional medical doctors; even in the most developed countries have inadequate position regarding the availability of medical doctors. According to Association of American Medical Colleges, U.S is facing a shortage of approximately 20,000 medical doctors [1]. In Pakistan, the situation is even much worse. According to Pakistan Medical and Dental Councils (PMDC), Pakistan is facing a shortage of approximately 182,000 medical doctors [2]. Because of this shortage, pregnant women are also affected due to lack of proper and timely treatment, which increased the mortality rate of pregnant women over the years. According to world health organization (WHO), almost 500,000 women die every year from pregnancy-related complications [3]. Especially focusing on Maternal Mortality Rate, it was observed that most of the deaths occur because of few basic complications. Moreover, these complications can be easily treated once the reasons are diagnosed; therefore, the major problem is the unavailability of proper diagnosis because of the shortage of medical doctors.

The four high risk pregnancy complications handled in this work are Hypertension, Obstructed Labor, Septicemia and Hemorrhage. Increase in blood pressure during pregnancy

indicates Hypertension, Obstructed Labor is an anomaly that may arise during the process of labor, Septicemia pollutes the patient's blood and may occur due to infections caused by bacteria and Hemorrhage occurs due to excessive loss of blood from the patient's body.

To overcome this problem, we have proposed ontology-based clinical decision support system which can partially work in place of doctors to diagnose high-risk pregnant woman and refer them to the qualified medical doctors for treatment. This way, the high-risk patients will get proper treatment well in time, and many lives can be saved. The main focus of this system is to build a diagnostic procedure which can work independently of qualified doctors and identify high risk patients; Once these patients are identified, they can be treated by medical doctors. So our system will help in reducing the workload of doctors as well as providing basic health care to more and more patients. The framework is composed of three components: 1) Automatic Ontology Construction, 2) Feedback System and 3) Ontology Enhancement Component.

The rest of the paper is organized as follows. In section 2, existing work in ontology based clinical decision support system and automatic construction of ontology is discussed. In section 3, the proposed approach is discussed. Experimental results are presented in section 4. Finally, the conclusion is drawn in section 5.

II. LITERATURE REVIEW

A. Ontology Based CDSS

There are many mistakes made on regular basis by humans in clinical environments. Hazmy Iman Abas et al in [16] have identified the three common mistakes made by clinicians that are; they failed to meet guidelines, they are not educated on regular basis and they are not aware of their responsibilities. According to the authors, these mistakes can easily be avoided by the use of ontology based clinical decision support system. The early detection of Alzheimer Disease is a challenging task in medical domain. Eider Sanchez et al in [17] proposed an ontological CDSS approach to detect Alzheimer in early stages. In this system multidisciplinary knowledge is used (i.e.

the system uses three ontologies that are SWAN, SNOMED CT and MIND). SWAN is used for the diagnosis of Alzheimer Disease, SNOMED CT's purpose is standardization and MIND ontology is used to carry out patient tests. Farahidayah Bt. Mahmud et al in [15] designed a CDSS that finds the right time of weaning a patient from ventilator. According to the authors, the proposed Ontology based CDSS is very helpful because of the ontology's ability of presenting complex concepts, reusability and specification of shared conceptualization. Similarly, Adnan et al in [18] proposed ontology based clinical decision support system to assist electronic discharge summary (EDS) while prescribing the patient's medications. Matt et al in [20] developed an ontology based CDSS for preoperative risk assessment. The proposed system takes the patient's data as input, stores it in database and also passes it to Ontology Modeler for conversion into OWL format (i.e. the same record is stored in two different formats). Afterwards, inference is performed on both the database and ontology. Rule engine is used on the database to calculate the numeric scores (i.e. cardiac scores) whereas classification algorithm is run on the ontology to assign category to the patient, the results of both are then combined to calculate the patient preoperative risk assessment.

B. Automatic Construction of Ontology

Abd-Elrahman Elsayed et al in [19] used data mining technique (c4.5 decision tree) on structured data to construct ontology automatically. In this approach, the authors proposed decision tree to ontology mapping where the tree decision nodes are mapped to ontology classes and leaf nodes are mapped to individuals. The authors tested the proposed approach on the soybean disease dataset and showed the efficiency of the same. Seongwook Youn et al in [20] proposed architecture for the classification of emails as spam or legitimate using ontology based approach. The authors created a dataset D based on the features of the email (i.e. spam email) and used WEKA (J48 decision tree) to generate decision tree which afterwards is passed to JENA for conversion into RDF ontology format. Authors divided the original dataset into two parts (i.e. training and testing dataset) and tested the proposed approach on testing dataset; according to the authors the results are satisfactory. Amit Bhagat et al [21] used association rule mining to construct ontology from large transaction databases. In the proposed approach, multiple level association rule mining is used to extract more specific and relevant knowledge as compared to single level rule mining. Patrick Clerkin et al in [22] proposed automatic construction of ontology using the COBWEB algorithm. COBWEB is a clustering algorithm which creates different clusters of the data in a hierarchical manner. In the proposed technique, the hierarchy of the clusters given by COBWEB is mapped into ontology classes in such a way that parent cluster(s) is mapped as parent or super class(es) in the ontology whereas the sub or child cluster(s) is mapped as sub-class(es) in the ontology. Henrihs Gorskis et al in [23] reviewed 1) the work done in the field of ontology building using data mining techniques and 2) the potential of different techniques in the construction of ontology. According to the authors, the ontologies created with data mining technique(s) may be inferior to those constructed manually.

III. PROPOSED TECHNIQUE

In this paper, we have proposed a framework for predicting high risk woman using ontology based CDSS. The framework is composed of three main modules: 1) Automatic Ontology Construction 2) Feedback System 3) Ontology Enhancing Process.

A. Automatic Ontology Construction

This module automatically constructs the ontology of high risk pregnant woman using pregnant woman dataset and is composed of two sub-components namely Rules extractor and Rules to Ontology Mapper. Rules extractor extracts the rules from the pregnant woman dataset using WEKA and is developed in Java. WEKA API is used to call WEKA functions from the program. The dataset is given as input to this component which uses WEKA API to extract rules from the dataset. The dataset is given in the ARFF (Attribute-Relation File Format) format which is compatible with WEKA. ARFF format has two sections, the first section is Header which contains the name of relation, the attributes of the relation and the attributes' data types whereas the second section has Data which contains the real instances of the relations. Example of partial IRIS dataset in ARFF format is given below.

Header Section:

```
@RELATION iris
  @ATTRIBUTE sepallength NUMERIC
  @ATTRIBUTE sepalwidth NUMERIC
  @ATTRIBUTE petallength NUMERIC
  @ATTRIBUTE petalwidth NUMERIC
  @ATTRIBUTE class {Iris-setosa,
Iris-versicolor, Iris-virginica}
```

Data Section:

```
@DATA
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
4.7, 3.2, 1.3, 0.2, Iris-setosa
4.6, 3.1, 1.5, 0.2, Iris-setosa
5.0, 3.6, 1.4, 0.2, Iris-setosa
```

Once the dataset is loaded, attribute selection algorithm is called using WEKA API. We have used the genetic search for attribute selection and once the most relevant features are selected, JRIP algorithm is called to extract rules which later are used for the construction of ontology as summarized in figure 1.

Rules to Ontology Mapper is responsible to construct ontology using the rules extracted in the first component as shown in figure 2. It first creates ontology classes for each dataset (i.e. hypertension, Obstructed labor, Hamorrhage, Septicemia) then the properties are created based on the attributes found in the rules. Furthermore, the range and domain of each property is set based on the information provided by each dataset. Afterwards, for each class, its definition is created that reflects the classification criteria for that class.

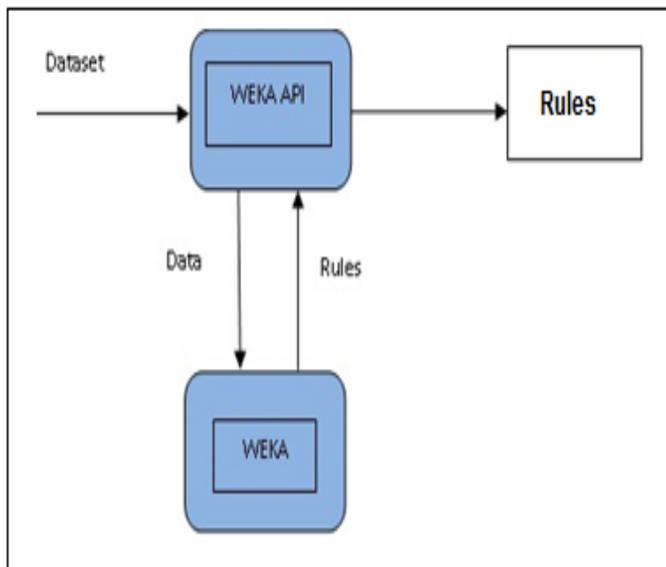


Fig. 1. Rules Extractor

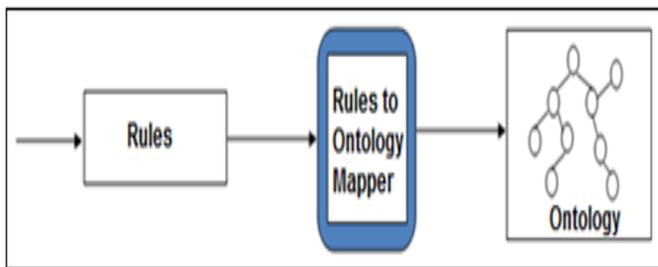


Fig. 2. Rules to Ontology (R2O) Mapper

Figure 3 shows the class definition (containing all attributes and their values) for Obstructed Labor in Protégé.

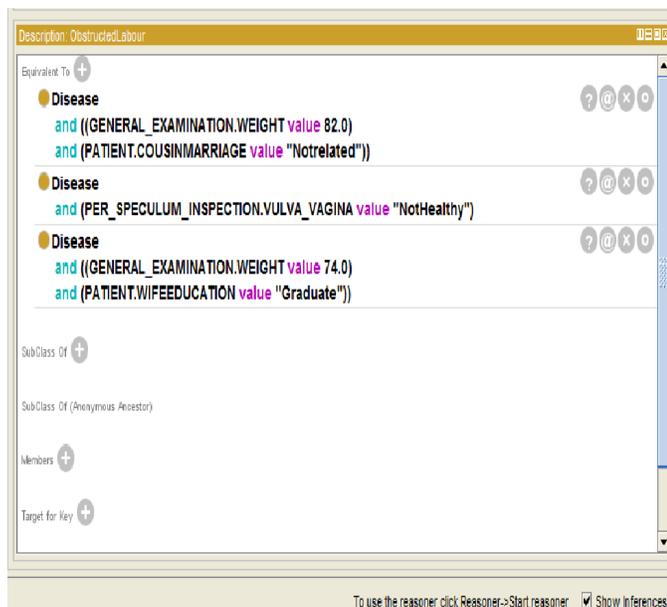


Fig. 3. Obstructed Labor definition in Protege

B. Feedback System

Once a patient is predicted as high risk by our proposed system and recommended to the doctor for treatment, feedback is taken from the doctor whether the patient was correctly identified or not. If the doctor overrules the decision made by the system, the attributes which caused the decision are shown to the doctor. He can add / delete the attribute(s) or tune their parameters / range (i.e. change the values of an attribute to more suitable one). Afterwards, the overruled instance with tuned parameters or different attributes (if any) is added to overruled dataset. Once the number of instances in overruled dataset reaches β (configurable), this dataset is passed as input to Ontology Enhancing Process for updating the ontology. In this way, doctor's knowledge is incorporated in the system and ontology is enhanced accordingly as shown in figure 4.

C. Ontology Enhancing Process

According to Gruber [23], ontology represents knowledge as a hierarchy of concepts and their relation for a specific domain. Furthermore, ontology knowledge representation is based on the concept of conceptualization which can be defined as the objects, the concepts and the relationship that hold among them [23]. New concepts, their properties and relationship can easily be incorporated in ontology knowledge base. Therefore, once the number of overruled instances reaches β , JRIP algorithm is executed on overruled dataset to extract new knowledge, and the same is updated in the ontology either by creating new concepts or by defining new relationship between the old concepts. The complete architecture of our proposed Ontology based Clinical Decision Support System for predicting high risk pregnant woman is shown in figure 4.

IV. EXPERIMENTAL RESULTS

A series of tests have been carried out in order to demonstrate that the proposed system is working properly. The tests have been subdivided in two classes: 1) Automatic construction of ontology, and 2) Ontology enhancement Process. Each of the experiment and its results are described in following:

A. Automatic Construction of Ontology

This experiment is designed to verify the automatic construction of ontology (i.e. how well the rules are mapped into ontology). For this purpose, the ontology and existing rule-base system are tested on the same datasets with the same training and testing ratio. In this experiment, we used four dataset (for each output class i.e. Hypertension, Obstructed Labour, Hamorrhage, Septicemia) where each dataset is divided into two parts (i.e. 70% and 30%) where 70% of data is used for training and 30% is used for testing. The experimentation result showed that ontology and rule based system have same accuracy which means the rules are transformed with 100% accuracy into the ontology as shown in figure 5.

B. Ontology enhancement Process

Enhancing the ontology is the most important aspect and has been deeply validated. In this activity, fifteen doctors from five different hospitals of Pakistan participated, whenever the

patient is identified as high risk by our system, the patient is referred to one of the fifteen doctors. If the doctor overruled the decision, feedback from the doctor is taken and the instance is added to overrule dataset (in this case the β value is 25). When the instances in the overruled dataset reach β , the new

knowledge is incorporated in the ontology. We have observed that after two rounds of ontology update the accuracy of the system increases and the false positives are decreased significantly as shown in figure 6.

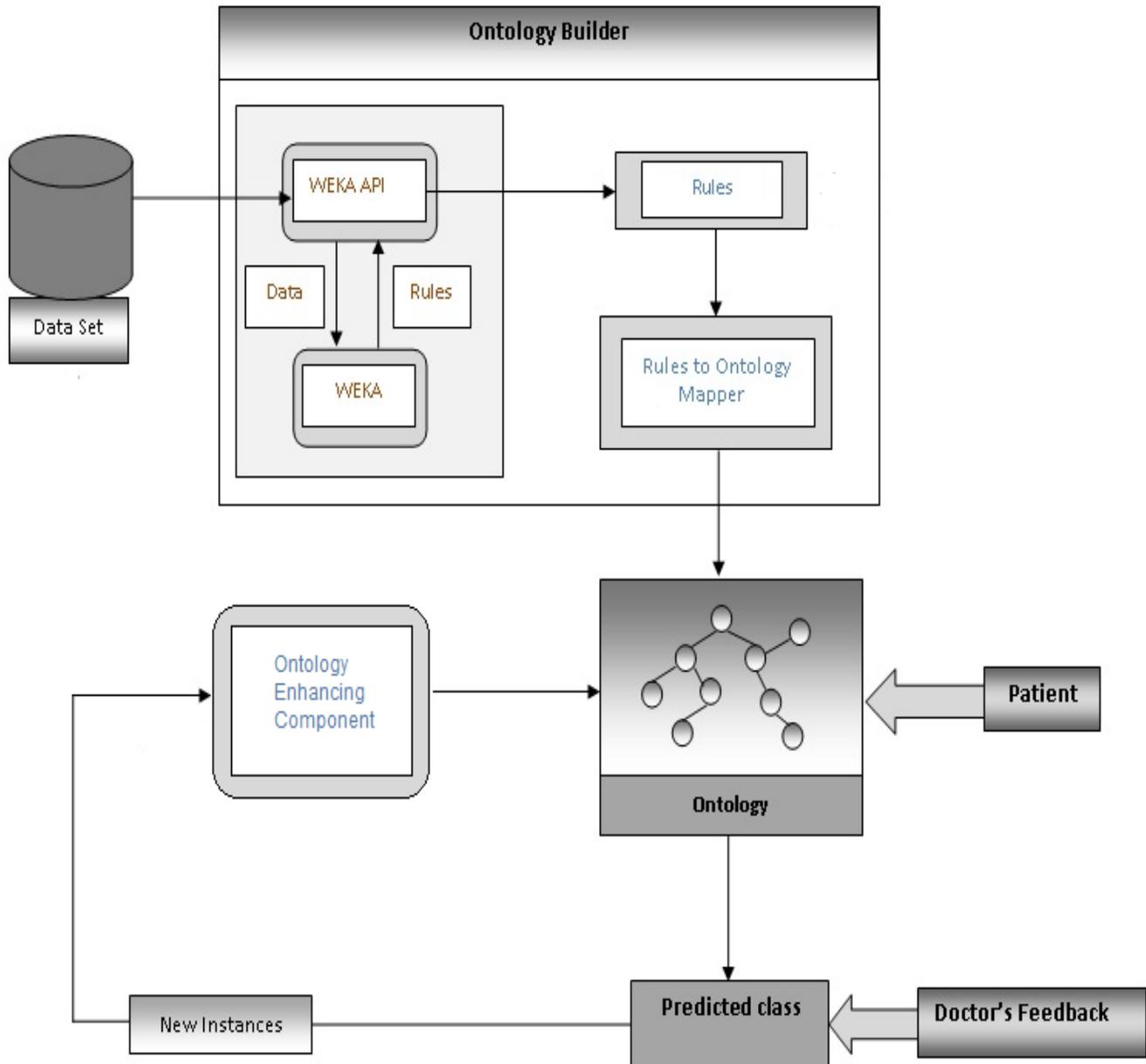


Fig. 4. System architecture of the proposed framework

V. CONCLUSION

According to world health organization (WHO), almost 500,000 women die every year from pregnancy-related complications. Moreover, these complications can be easily treated once diagnosed; however, the major problem is unavailability of proper diagnosis because of shortage of medical doctors. In this paper, we have proposed an ontology-based CDSS for diagnose high-risk pregnant woman and refer them to the qualified medical doctors for treatment. The

proposed framework is tested on a large number of test cases, results are satisfactory and support the implementation of the same. The work can be extended in many directions; one possible direction is including more pregnancy-related disease in the ontology. Second direction is automatic ontology enhancement process, currently based on expert (doctor) feedback the ontology is enhanced, machine learning algorithms should be incorporated in the proposed system to enhance ontology automatically.

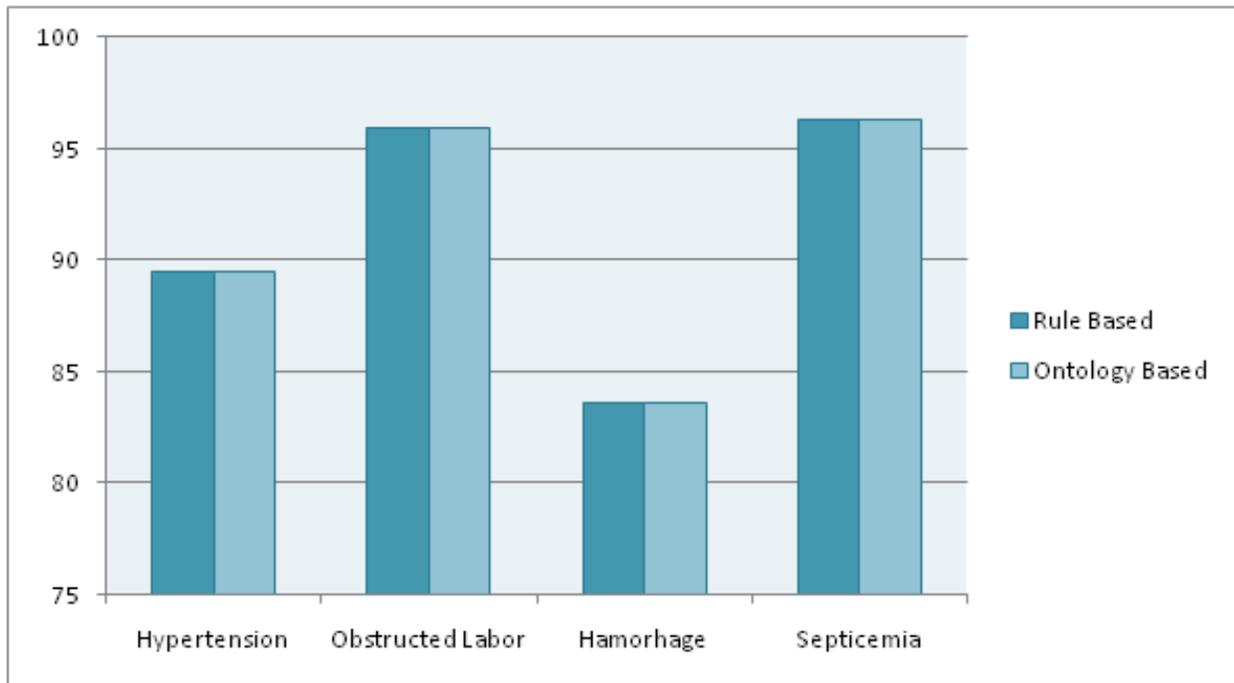


Fig. 5. Accuracy of Rule Based and Ontology Based system

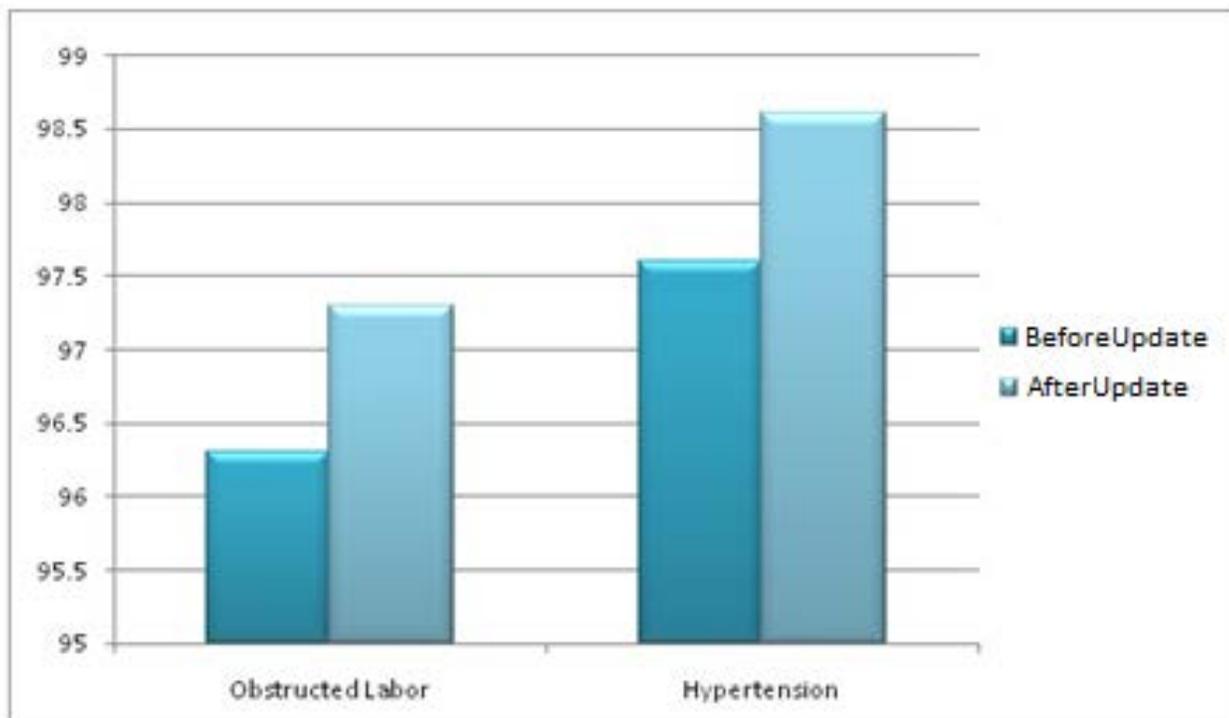


Fig. 6. Classification accuracy comparison before and after updating the ontology

REFERENCES

- [1] American Association of American Medical Colleges. (November, 2011), 2011 State Physician Workforce Data Book.
- [2] Pakistan Medical and Dental Councils, <http://www.pmdc.org.pk/>
- [3] U. Fund, "State of the Worlds Children 2009. Maternal and Newborn Health," New York: UNICEF, vol. 60, 2008.
- [4] Umar Manzoor, Samia Nefti, Yacine Rezgui, "Categorization of malicious behaviors using ontology-based cognitive agents", *Data & Knowledge Engineering*, Volume 85, May 2013, Pages 40–56.
- [5] K.L. Clark, F.G. McCabe, "Ontology schema for an agent belief store", *International Journal of Human-Computer Studies*, Volume 65, Issue 7, July 2007, Pages 640-658.
- [6] Protégé, <http://protege.stanford.edu/>

- [7] WEKA Tool, <http://www.cs.waikato.ac.nz/ml/weka/>
- [8] B. Orgun, J. Vu, "HL7 ontology and mobile agents for interoperability in heterogeneous medical information systems", *Computers in Biology and Medicine*, Volume 36, Issues 7–8, July–August 2006, Pages 817-836.
- [9] Quynh-Nhu Numi Tran, Graham Low "MOBMAS: A methodology for ontology-based multi-agent systems development", *Information and Software Technology*, Volume 50, Issues 7–8, June 2008, Pages 697-722.
- [10] C. Su and C. Yang, "Feature selection for the SVM: An application to hypertension diagnosis," *Expert Systems with Applications*, vol. 34, no. 1, pp. 754–763, 2008.
- [11] Y. Chae, S. Ho, K. Cho, D. Lee, and S. Ji, "Data mining approach to policy analysis in a health insurance domain" *International journal of medical informatics*, vol. 62, no. 2-3, pp. 103–111, 2001.
- [12] A. Tanwani, J. Afridi, M. Shafiq, and M. Farooq, "Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets," in *Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer, 2009, pp. 128–139.
- [13] Umar Manzoor, Samia Nefti "iDetect: Content Based Monitoring of Complex Networks using Mobile Agents", *Applied Soft Computing* 12 (5), 1607-1619.
- [14] Farahidayah Bt. Mahmud, Maryati Mohd Yusof, Shahrul Azman Naoh. "Ontological Based Clinical Decision Support System." *Electrical Engineering and Informatics (ICEEI)*, 2011 International Conference on. Bandung, 17-19 July 2011.
- [15] Hazmy Iman Abas, Maryati Mohd. Yusof, Shahrul Azman Mohd Noah. "The Application of Ontology in a Clinical Decision." *Semantic Technology and Information Retrieval (STAIR)*, 2011 International Conference on. Putrajaya, 28-29 June 2011.
- [16] Eider Sanchez, Carlos Toro, Eduardo Carrasco, Gloria Bueno, Patricia Bonachela, Carlos Parra, Frank Guijarro. "A Knowledge-based Clinical Decision Support System for the diagnosis of Alzheimer Disease." *e-Health Networking Applications and Services (Healthcom)*, 2011 13th IEEE International Conference on. Columbia, MO, 13-15 June 2011. 351 - 357.
- [17] Mehnaz Adnan, Jim Warren, Martin Orr. "Ontology Based Semantic Recommendations for Discharge Summary Medication Information for Patients." *Computer-Based Medical Systems (CBMS)*, 2010 IEEE 23rd International Symposium on. Perth, WA, 12-15 Oct. 2010. 456 - 461.
- [18] Matt-Mouley Bouamrane, Alan Rector, Martin Hurrell. "Development of an Ontology for a Preoperative Risk Assessment." *Computer-Based Medical Systems*, 2009. CBMS 2009. 22nd IEEE International Symposium on. Albuquerque, NM, 2-5 Aug. 2009.
- [19] Abd-Elrahman Elsayed, Samhaa R. El-Beltagy, Mahmoud Rafea, Osman Hegazy. "Applying data mining for ontology building." *The 42nd Annual Conference On Statistics, Computer Science, and Operations Research*. Cairo, 2007.
- [20] Seongwook Youn, Dennis McLeod. "Efficient Spam Email Filtering using Adaptive Ontology." *Information Technology*, 2007. ITNG '07. Fourth International Conference on. Las Vegas, 2-4 April 2007. 249 – 254
- [21] S Nefti, U Manzoor, S Manzoor "Cognitive agent based intelligent warning system to monitor patients suffering from dementia using ambient assisted living", *International Conference on Information Society (i-Society)*, Pages 92-97, 2010.
- [22] F Rea, S Nefti-Meziani, U Manzoor, S Davis "Ontology enhancing process for a situated and curiosity-driven robot", *Robotics and Autonomous Systems* 62 (12), 1837-1847
- [23] Amit Bhagat, Sanjay Sharma, K.R.Pardasani. "Ontological Frequent Patterns Mining by potential use of Neural Network." *International Journal of Computer Applications*, Vol 36, Issue 10, 2011.
- [24] MA Balubaid, U Manzoor, B Zafar, A Qureshi, N Ghani "Ontology Based SMS Controller for Smart Phones", *International Journal of Advanced Computer Science and Applications* 6 (1), 133-139.

Distributed Optimization Model of Wavelet Neuron for Human Iris Verification

Elsayed Radwan^{1,2}

¹Deanship of Scientific Research
Umm Al-Qura University
Makkah , KSA

Mayada Tarek²

²Computer Science Dept
Faculty of Computer and Information
Sciences, Mansoura University,
Egypt

Abdullah Baz^{1,3}

³Computer Engineering Dept
Faculty of Computer and Information
System , Umm Al-Qura University
Makkah , KSA

Abstract—Automatic human iris verification is an active research area with numerous applications in security purposes. Unfortunately, most of feature extraction methods in human iris verification systems are sensitive to noise, scale and rotation. This paper proposes an integrated hybrid model among Discrete Wavelet Transform, Wavelet Neural Network and Genetic Algorithms for optimizing the feature extraction and verification methods. For any iris image, the wavelet features are extracted by Discrete Wavelet Transform without any dependency on scale and pixels' intensity. Besides, Wavelet Neural Network classifier is integrated as a local optimization method to solve the orientation problem and increase the intrinsic features. In solving the down sample process caused by DWT, each human iris should be characterized by a set of parameters of its optimal wavelet analysis function at a determined analysis level. Thus, distributed Genetic Algorithms, meta-heuristic algorithm, is introduced as a global optimization searching technique to discover the optimal parameter values. The details and limitation of this paper will be discussed where a comparative study should appear. Moreover, conclusions and future work are described.

Keywords—Discrete Wavelet Transform (DWT); Wavelet Features; Wavelet Neural Network (WNN); Distributed Genetic Algorithms (GA); Human Iris Verification

I. INTRODUCTION

Since safety communication with others is a fundamental demand, verifying the direct measurements of some human parts, Biometrics, is the unique solution. In the field of human identification, iris verification is regarded as the most reliable and accurate biometric identification system [1]. The texture of the iris is relatively static and stable during the person's lifetime. Thus, iris texture is uniquely identifying individuals. The human iris, the part between the pupil and the sclera, has an extraordinary structure and provides many interlacing minute characteristics. The process of iris recognition depends on two consecutive phases, localization of the iris domain, as depicted in Figure 1 [2], and generation of the feature set of iris images. Hence, a convenient iris classifier should be used. Unfortunately, iris recognition suffers from the scale and rotation invariant problems, a certain fixed resolution, and non-regarding iris features during the stage of feature extraction. Moreover, the time complexity is taken in training by the iris classifier [2, 3, 4, 5].

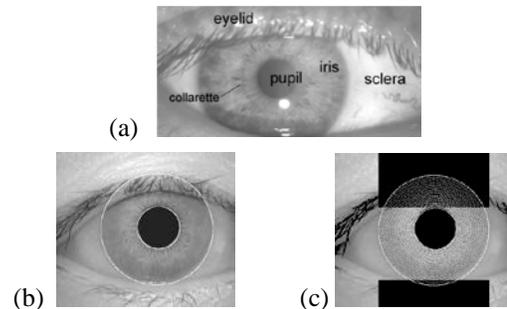


Fig. 1. Example of iris localization: (a) Original image. (b) Iris area localization. (c) Iris native area

In the recent past, some iris verification and recognition techniques have been developed. Based on local feature extraction process, many papers appear. Feng Hao et al. [6] proposed a practical and secure way to integrate the iris biometric into cryptographic applications. They represented the iris key as the repeatable binary string. Yaser Danial Khan et al. [7] extracted the iris feature based on the moments, and classified iris patterns based on k-means. Case-based Reasoning technique is combined in their classification technique. Huaqing Liang et al [8] proposed an iris recognition method based on iris's speckles characters. Mayank et al. [9] create a single high-quality iris image by enhancing the iris image globally. The iris image features are extracted based on 1-D log-polar Gabor transform and local topological using Euler numbers. Ma et al. [10] extracted the features of the iris by key local variations, spatial filter. Unfortunately, these techniques have unpromising results under intraclass variations, different contrast and illumination settings, and miss the geometrical representation of the iris texture. Moreover, these techniques need much time to be processed and classified. On the other hand, spectral methods aim at describing the multi-resolution and directionality of periodic or almost periodic 2-D patterns in an image. Spectral methods refer to the frequency domain where feature are related to statistics of filter response [11]. K. Miyazawa [12] presented an efficient algorithm for iris recognition using phase-based image matching in 2D Discrete Fourier Transforms (DFTs). Unfortunately, DFTs perform poorly in practice, due to its lack of spatial localization. Some papers

enhanced the extracted iris texture features based on Gabor [9, 13] where better spatial localization is provided. Unfortunately, Gabor filter is limited because there is no single filter resolution at which one can localize a spatial structure in iris texture. Because of the wide range of wavelet functions, Wavelet Transforms (WTs) have various resolutions that allow researchers to represent iris texture at the most suitable scale [14, 15, 2, 16]. But WT is still non-supportive to directionality and anisotropy [17]. As the result of the short in WT, each human iris should be characterized by a set of parameters of its optimal wavelet analysis function at a determined analysis level. Moreover, a suitable classifier should be combined to reduce the False Error Rate such as Backpropagation Neural Network (BPNN) and Support Vector Machines (SVM) with Radial Basis Function (RBF)..etc. [18,15].

In short, these analysis and classification methods achieved some accurate results, but these methods still have a lack of characterizing each human iris by a predefined analysis and classification parameters, the interclass similarity problem. Also, recognition performance of iris features still have many gaps to be improved, such as the intraclass variation, as well as the massive number of iris texture parameters [19]. Hence, several accurate iris recognition algorithms with multiscale analysis techniques in addition to a fast classifier are needed as a well-suited representation for iris verification.

Wavelet Transform (WT) is especially suitable for processing an iris image that satisfy these requirements. Since most details could be hardly represented by one function, they could be matched by various versions of the mother wavelet with various translations and dilations [20, 21]. Three problems will face wavelet transform in human iris verification system. First problem is the process of segmenting and normalizing the iris parts from each eye images without any eyelid and eyelash noise. This paper proposes Hough Transform and Daugman's rubber sheet model techniques in segmentation and normalization process [22]. Second, the feature extraction process are associated with the problems of interclass similarity, the down-sample process, and orientation invariant that make loss of some important extracted features from iris image. Thus, choosing the correct wavelet function at a determined level during feature extraction should help in solving this problem. Moreover, this paper proposes a Wavelet Neural Network (WNN) [23] technique as a local optimization method for iris verification to overcome these disadvantages and increase the intrinsic features. The third is the process of selecting the most effective and integrated parameters between DWT and WNN for optimal characterization to each human iris. In DWT the parameters are wavelet analysis function at effective analysis level. Also, WNN parameters are completely determined by wavelet activation function and learning rate value. Thus, choosing the parameters based on the correct wavelet function should affect the feasible domain of the wavelet activation function. Because of the lattice structure of WT Bank [24], a non-specific domain technique, independent from the specified problem, should serve in finding a general and global solution. Thus, a meta-heuristic based technique [25, 26] should be an

effective searching strategy. Because of the slowness of GA and the population diversity problems, this paper introduces a distributed and meta-heuristic searching strategy based on GA [27, 28]. DGA is chosen as a global strategy searching technique to select the optimal integration between DWT and WNN parameters to characterize each human iris. DGA depends on interact among sub-population through the migration process, wherever each sub-population is addressed by a specific analysis level. By this paper, the migrated individuals are selected based on the wavelet entropy value.

This paper proposed an integrated hybrid model among DWT, WNN and distributed GA (new searching strategy based on GA) techniques for optimizing feature extraction and verification method for the human iris verification system. DWT technique analysis iris images to extract wavelet detail coefficients. According to a huge number of coefficients, a statistical model is represented by wavelet energy and entropy values [14]. Because of the problems of interclass similarity and intraclass variation, WNN technique will be used as a suitable classifier and increase the characterization features. DGA try to find the most effective DWT and WNN parameters for optimal characterization to each human iris. A testing stage examines the verification rate to the unseen iris. Moreover, the result will be concluded.

The rest of this paper is organized as; in Section 2, an abbreviation of Discrete Wavelet Transform (DWT), Wavelet Neural Network (WNN), and Genetic Algorithms (GA) are mentioned. Application of human iris verification using a proposed hybrid integrated system is described in Section 3. Section 4 declares the result of the proposed integration system. Moreover, a comparative study determine the verification rate between both strategies for iris verification systems (Searching for suitable integration between DWT and WNN parameters using standard GA and Distributed GA). Finally section 5 concludes the paper and give a recommendation for future work.

II. PRELIMINARIES

A. Wavelet Decomposition Analysis

Wavelets are basis functions that satisfy certain mathematical requirements. They are used to cut up data into different frequency components. Then, a study on the behavior for each component with a resolution matched to its scale [14]. The basic idea of wavelet transform is to represent any arbitrary function as a superposition of wavelets. Any such superposition decomposes the proposed function into different scale levels where each level is further decomposed with a resolution adapted to the level [29]. Thus, Wavelet function is characterized by a varying window size, wide for slow frequencies and narrow for fast ones. Furthermore, wavelet windows are adapted to the transients of each scale, regardless wavelets lack the requirement of stationary. For example, the signal $x(t)$ is characterized by;

$$\left. \begin{aligned} x(t) &= \sum_k s_{j,k} \phi_{j,k} + \sum_k d_{j,k} \psi_{j,k} + \sum_k d_{-1,k} \psi_{j,k} + \dots + \sum_k d_{1,k} \psi_{j,k} \\ s_{j,k} &= \int \phi_{j,k} x(t) dt \\ d_{j,k} &= \int \psi_{j,k} x(t) dt \end{aligned} \right\} \quad (1)$$

Where $\psi_{j,k}(t)$ and $\phi_{j,k}(t)$ are the mother wavelet functions which analogous corresponding to sinusoidal basis function in Fourier Analysis. $s_{j,k}$ and $d_{j,k}$ are wavelet transform coefficient that we call them w_j^k . Also $j = 1, 2, \dots, J$ is the number of multi-resolution levels (or scale) and k is the translation parameter. The Discrete wavelet transform, DWT, is selected to be dyadic scales and positions, i.e. the scales and shifts are based on power of two. Such analysis yielded from DWT is defined as;

$$DWT(j, k) = w_j^k = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} x(t) \psi\left(\frac{t-2^j t}{2^j}\right) dt \quad (2)$$

Where w_j^k is translated as the local remaining error between two successive signal approximations at scales j and $j + 1$.

By implementing DWT on an image, it is actually decomposed into sub-bands and critically into sub-sampled. An efficient way for implementing this scheme is by passing the signal through a series of low pass and high pass filter pairs called as quadrature mirror filter as illustrated in Figure 2 [24]. 1-D level decomposition of DWT arises four sub-bands from separable applications of vertical and horizontal filters where L and H denote the 1-D low pass and high pass filter respectively. Low pass image LL corresponds to the coarse

level coefficients, approximation image. On the other hand, three detail images HL, LH, and HH represent the finest scale as shown in Figure 2a. The LH channel contains image information of low horizontal frequency and high vertical frequency and so on. To obtain the next coarse level of wavelet decomposition, the sub-band LL is further decomposed and 2-D level decomposition is resulted, Figure 2b[29]. Repeatedly, this process is iterated until some final scale is reached, which is considered as one of the main goals in this paper.

By DWT analysis technique, the wavelet coefficients that are gathered from each sub-band (LL, HL, LH, and HH) are very huge to be certified as discrimination features. To overcome this problem, wavelet coefficients can be represented by statistical functions such as mean, median, standard deviation, energy and entropy [30].

Wavelet energy is the measure that keep the main characteristic of the wavelet coefficients and produce the same images with different translation, rotation and scale, having the same wavelet energy values[31, 32]. Wavelet energy values are measured by analyzed iris image to its wavelet sub-image coefficient (LLx, HLx, LHx, HHx) as defined in equations (3) [14] where w_j^k is the wavelet coefficients to sub-band j at k -level.

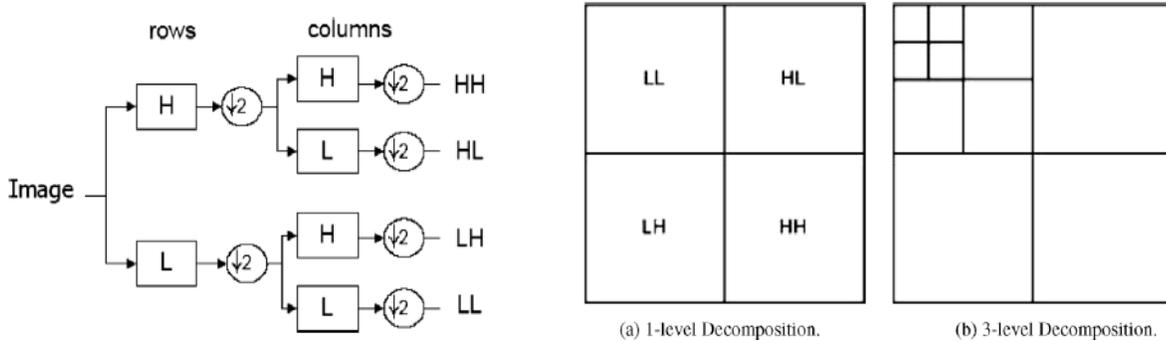


Fig. 2. A one-level wavelet analysis filter bank and Wavelet multi-level frequency decomposition

$$E_w^{(k)} = \frac{1}{J} \sum_{j=0}^J (w_j^k)^2$$

Also, the energy at each resolution level $j = 1..J$ will be the energy of the detail signal

$$E w_j = \sum_k w_j^k \quad (3)$$

wavelet entropy : wavelet entropy is an estimated measure based on the wavelet coefficients to provide quantitative information about the order/complexity of iris image [31]. Its values should be computed after analysis image to its wavelet sub-image coefficient (LLx, HLx, LHx, HHx). There are various wavelet entropy measures. The definition of norm entropy and sure entropy are defined by equations (4) respectively where w_i is the wavelet coefficients to sub-band x at k -level, ϵ is a positive threshold value.

$$\left. \begin{aligned} H_N(w)^{(k)} &= \sum_{i=0}^n |w_i|^p \quad \text{for } (1 \leq p < 2) \\ |w_i| \leq \epsilon &\rightarrow H_S(w)^{(k)} = \sum_{i=0}^n \min(w_i^2, \epsilon^2) \end{aligned} \right\} \quad (4)$$

B. Wavelet Neural Network

When the sigmoid activation function is used in training the neural network(NN), it can recognize any deterministic nonlinear process. But, NN suffer from a series of drawbacks. Random initial weights is generally associated with extended training times and NN may be trapped into local minima. Moreover, there is a shortage between the specific sigmoidal activation function and the admissible neural network architecture [33]. On the opposite, Wavelet neuron solve these problems. WNN is a feedforward neural network that gather both characteristics of neural network and wavelet decomposition. It is a generalization of the Radial based Neural Network(RBNN) by using wavelet as an activation function [34, 33]. RBNN is a bell shaped activation function that scale variable nonlinearity whereas WN does not consider symmetry condition in activation function. The reason for the application of WNN in case of such a problem as classification is that the feature extraction and representation

properties of the wavelet transform are merged into the structure of the ANN to further extend the ability to approximate complicated patterns [35]. Moreover, WN is preserve in high compression ability and updating the function estimate from a new local measure, involves only a small subset of coefficients. The WN depends mainly on the bias that allows the sensitivity of the wavelet activation neuron to be adjusted. The architecture of WNN consists of three-layer structure with an input layer, a wavelet layer, and an output layer as shown in Figure 3.

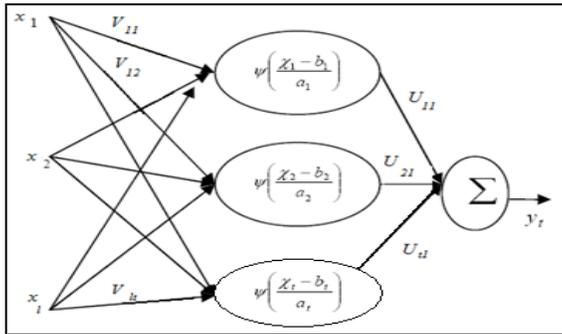


Fig. 3. The structure of the Wavelet Neural Network

In WNN, both the position and dilation of the wavelets as well as the weights are optimized. The basic neuron of a WNN is a multidimensional wavelet in which the dilation and translation coefficients are considered as neuron parameters. The output of WNN is therefore a linear combination of

several multidimensional wavelets [34]. In this WNN model, the hidden neurons have wavelet activation functions ψ and have two parameter a_t, b_t which represent dilation and translation parameter of wavelet function and V is the weight connecting the input layer and hidden layer and U is the weight connecting the hidden layer and output layer.

C. Genetic Algorithms

Genetic Algorithm (GA), introduced by John Holland in 1975, is a computing search technique used in finding a solution in optimization problems. GA applies the principles of evolution found in nature to the problem of finding an optimal solution [36, 25]. GA generates successive populations of alternate solutions that are represented by a chromosome, i.e. a solution to the problem, until acceptable results are obtained. Each chromosome, individual solution, consists of number of binary code called genes. A fitness function assesses the quality of a solution in the evaluation step [28, 37]. The evolution from one generation to the next is performed using three operations: reproduction, crossover and mutation. Chromosomes are selected for reproduction by evaluating the fitness value. The fitter chromosomes have higher priority to be selected into the recombination pool using the roulette wheel or the tournament selection methods. Crossover selects genes from two parent chromosomes using randomly chosen crossover point and creates two new off springs as in Figure 4 (a). Mutation process changes chromosome randomly by altering a single bit as in Figure 4 (b).

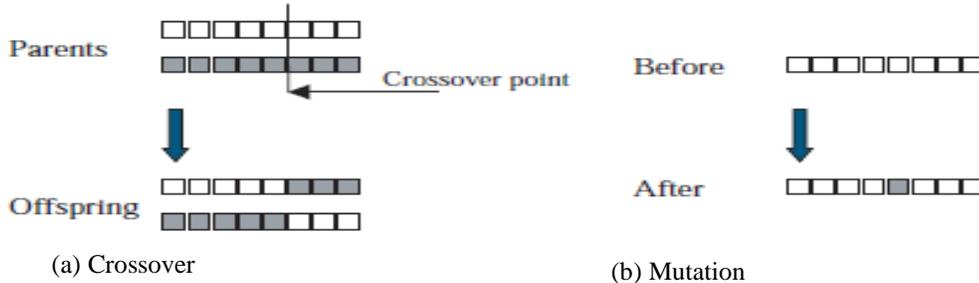


Fig. 4. Genetic algorithm operations

The evolutionary process operates many generations until termination condition satisfy. The termination condition is either reaching the maximum number of generations or a predefined fitness is achieved. Associated with the characteristics of exploitation and exploration search, GA can deal with large search spaces efficiently, and hence has less chance to get local optimal solution than other algorithms.

III. HUMAN IRIS VERIFICATION SYSTEM BASED ON WNN

Human Iris Verification is considered as the most reliable and accurate biometric identification system [3, 38]. This paper presents an implementation for Human Iris Verification System using an integrated model among DWT in feature extraction phase, WNN in increasing the intrinsic features as well a fast classifier as local optimization method and Distributed GA (DGA) as an evolutionary searching strategy and a global optimization method. The proposed Human Iris verification system depends on several stages as depicted in

Figure 5. First, segmenting and normalizing stage. The segmentation process isolates the iris part from an eye image without any eyelash and eyelid noise [4]. Then, normalizing the iris part process yields the corresponding texture based image [7, 22]. Second, extracting the intrinsic features for any iris texture image by discovering the optimal parameters between DWT and WNN. Since, iris texture is uniquely identifying each person based on its own characteristics; a stochastic searching strategy is needed to choose the optimal integration between DWT and WNN parameters. Unfortunately, conventional GA suffer from the premature convergence and the population diversity problems. In this paper, Distributed GA, the best large searching spaces strategy, is able to search for the optimal solution in adequate time[26, 37]. Finally, WNN verification rate, False Error rate, is measured for each human iris using the optimal parameters values.

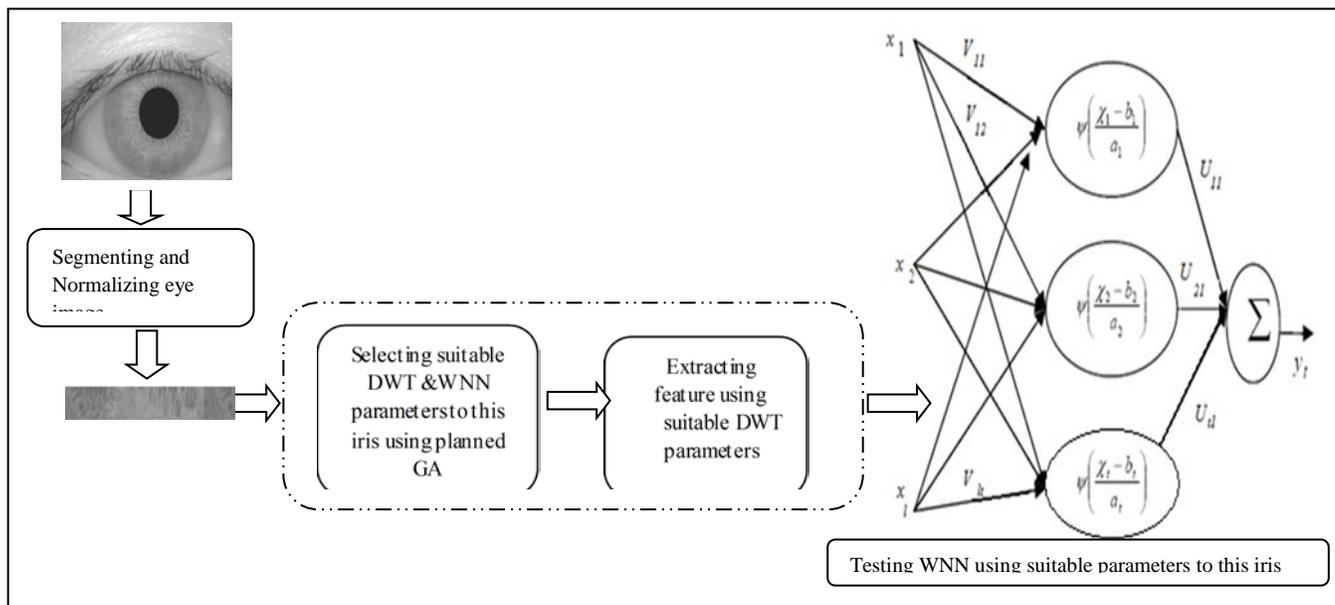


Fig. 5. Proposed Human Iris Verification System

A. Segmenting and Normalizing Stage

Each human eye image is segmented due to localizing the iris area. In this paper, Hough Transform [22] is used to localize the iris area from eye image, as shown in Figure 6. Units



Fig. 6. Example of iris normalization: (a) The iris localized area. (b) The normalized iris area

The Hough transform is a standard computer vision algorithm that can be used to determine the parameters of simple geometric objects, such as lines and circles, present in an image. An automatic segmentation algorithm based on the circular Hough transform can be employed to deduce the radius and center coordinates of the pupil and iris regions [3]. First, an edge map is generated by calculating the first derivatives of intensity values in an eye image and then thresholding the result. From the edge map, votes are cast in Hough space for the parameters of circles passing through each edge point. These parameters are the center coordinates x_c and y_c , and the radius r , which are able to define any circle according to the equation (5).

$$x_c^2 + y_c^2 - r^2 = 0 \tag{5}$$

Then, removing the eyelash from localized iris area is needed to get iris area pure from any noise as shown in Figure 1. In this paper, linear Hough transform [7] remove eyelash from localized iris image by first fitting a line to the upper and lower eyelid. To detect the eyelids, approximating the upper and lower eyelids with parabolic arcs, which are represented by equation (6).

$$-(x - h_j)\sin \theta_j + (y - k_j)\cos \theta_j = a_j \left((x - h_j)\cos \theta_j + (y - k_j)\sin \theta_j \right) \tag{6}$$

where a_j controls the curvature, (h_j, k_j) is the peak of the parabola and θ_j is the angle of rotation relative to the x-axis.

Normalizing the pure iris localized areas is needed to convert these area from different size to the same size as shown in Figure 6. In this paper, Daugman’s rubber sheet model [22] is used to normalize iris for achieving more accurate verification system as in equation (7,8,9).

$$I(x(r, \theta), y(r, \theta)) \rightarrow I(r, \theta) \tag{7}$$

With respect to

$$x(r, \theta) = (1 - r)x_p(\theta) + r x_1(\theta) \tag{8}$$

$$y(r, \theta) = (1 - r)y_p(\theta) + r y_1(\theta) \tag{9}$$

where $I(x, y)$ is the iris region image, (x, y) are the original Cartesian coordinates, (r, θ) are the corresponding normalised polar coordinates, and x_p, y_p and x_1, y_1 are the coordinates of the pupil and iris boundaries along the θ direction.

B. Optimizing the Extracted Features using DGA

Because of the interclass similarity problem among iris textures, increasing the intrinsic feature for each human iris is persistent need. Moreover, as the result of the problem of intraclass variation, discovering the own iris texture parameters is also needed. Thus, a suitable integration between DWT and WNN parameters should serve in characterizing each human iris. Each iris texture should be characterized by its optimal wavelet analysis function at an optimal analysis level. Also, WA parameters with multiresolution analysis should be integrated with an optimal wavelet activation function with an effective learning rate value.

As a result a large feasible space of long string solutions is constructed. Thus a meta-heuristic search strategy is needed. Conventional GA (CGA) is an effective strategy for searching a large space although CGA schema is negatively affected by long defining chromosome. Wherever, schema with long chromosome length are more likely to be disrupted by single point crossover and fall in population diversity problem or premature convergence. Hence, a new searching strategy based on GA, distributed genetic algorithms, is introduced as an effective searching strategy that can deal with large space. The idea of DGA[28, 26] is to divide the large searching space into multiple small searching spaces. These sub-spaces interact together based mainly on the island or fusion models. In island model, Conventional GA is used as an effective searching strategy to search for the effective individuals in each sub-space. Then, a migration for optimal individuals among sub-spaces is run at a predefined number of generation. Since the synthesized wavelet composes much energy into low pass coefficients than the other does, then applying the proposed DWT to many levels should collect more energy in the same number of wavelet coefficients [32]. This paper implement DGA as a global optimization method to run on two consecutive processes. The first process is the division process of the large searching space based on the wavelet analysis level, multiresolution analysis. The individuals in each sub-space with maximum energy are chosen to migrate among the subpopulation at the migration step. This process tries to find the best individuals in each sub-space. Whereas, the second process search for the optimal solution from the best individuals resulting from the last one.

In the first process, Each sub-population is constructed based on the number of DWT analysis level. For each sub-population, the individual chromosome is represented as a combination of WNN parameters and DWT parameters at a predefined analysis level as depicted in Figure 7.

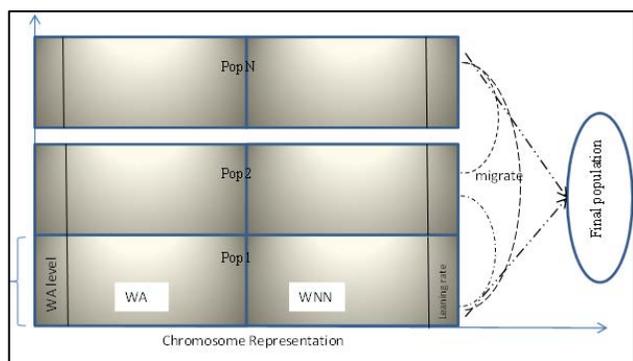


Fig. 7. The DGA level of sub-populations

Since each individual chromosome in the same sub-population have the same analysis level (N constant parameter value in each sub-population), GA schema length should be reduced. So, the searching space will be decreased as an effective way to standard GA in searching space.

a) Chromosome Representation

For each sub-population, GA chromosome is represented as a set of binary bits. These chromosomes should exactly contain six segments. The first segment of an individual chromosome represents the analysis level in DWT. Since different 8 wavelet analysis levels are ranged from N=3 To N=10, three bits are enough for representing this segment. The second segment of an individual chromosome represents the wavelet analysis function in DWT technique at Nth level. According to [3]five bits are enough for representing this segment (ranged from 00000 to 11111) however different 32 wavelet analysis function (db2, db3, db4, db5, db6, db7,db8, db9, db10, db12, db20, bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8,coif1, coif2, coif3, coif4, coif5, sym5) have the ability to analysis iris image and extracting wavelet detail features. The third segment of an individual represents the p-parameter value of the norm entropy , equation 3. In norm entropy, p values should be ranged in [1, 2) [14]. Sensitive p-parameter is considered to be 1/7, though p-parameter is represented by three bits for each individual chromosome (ranged from 000 to 111). Thus, the p-parameter gets one of the values: 1,1.142, 1.285, 1.426, 1.568, 1.71, 1.852 and 1.994. The forth segment of an individual represents ϵ - parameter value of the sure entropy which is mentioned in equation 4. In sure entropy[14], the threshold ϵ should be selected in [1, 8). ϵ -parameter is represented by three bits for each individual chromosome. The ϵ -parameter gets an integer value ranged from 1 to 8 .The fifth segment of an individual represents the wavelet activation function in WNN technique. Three bits are enough for representing this segment since different 8 wavelet activation function (Morlet,RASP1, RASP2, RASP3,POLYWOG1, POLYWOG2, POLYWOG3, POLYWOG4) [33] can be represented as a mathematical function. Finally the last segment of an individual represents the learning rate value in WNN technique. Learning rate value is selected in [0.1, 0.9) to decrease the time complexity for training WNN . Learning rate value is represented by three bits for each individual chromosome . Thus, the learning rate value gets one of the values: 0.2,0.3, 0.4, 0.5, 0.6 ,0.7, 0.8 and 0.9. Thus, each individual chromosome should be 20 binary bits in length. such an example, the encoded chromosome (01000001010011100000) is illustrated in Figure 8;

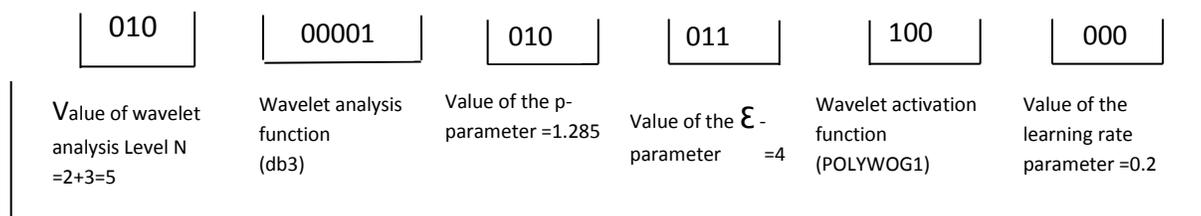


Fig. 8. a random individual chromosome in sub-population of N=5

The initial individual chromosome values in each sub-populations with size ζ_{g_0} of the first generation are initialized randomly, with different binary code. Each chromosome represents combination values among DWT and WNN parameters. The fitness value for each individual in each sub-population is evaluated by: extracting the wavelet energy and entropy values from all trained normalized iris images. These wavelet features is extracted from wavelet detail coefficients according to the corresponding wavelet analysis function and analysis level. These wavelet detail energy and entropy values are taken as input to WNN. The wavelet energy and entropy feature vector which is taken as input to WNN is shown in Figure 9 for $n=1, \dots, N$: n represent the level of wavelet decomposition.

b) Fitness Evaluation

Training the WNN with BP learning algorithm [34] compute the Mean Square Error (MSE) of the WNN at the last iteration as in equation 10. MSE is work as fitness function to each individual in each sub-population. The individual with the least MSE is the best individual in his sub-population. In the next generation, each individual is produced by reproduction, cross-over and mutation process over the individuals in the previous generation [27] using tournament selection strategy. As the result, each individual fitness function is evaluated. This process continues for a fixed number of generations. The best -so-far individual (the individual that has the best fitness overall generations) is designated as the best result for his sub-population.

$$MSE = \frac{1}{Z_{MAX}} \sum_{z=1}^{Z_{MAX}} (Y^z - D^z)^2 \quad (10)$$

Where $z = 1, \dots, Z_{MAX}$, the number of input samples of WNN ; Y^z, D^z represent the actual output of WNN and the expected output respectively for iris pattern z , $D = (1|0)$. $D^z = 1$ if the iris is the correct classified the human iris, and 0 elsewhere.

c) Selection, Crossover, Mutation and Migration Process

Since N , number of resolution levels, sub-population should be constructed, communicated and evolved through

$EHL^{(n)}$	$ELH^{(n)}$	$EHH^{(n)}$	$H_N(HL)^{(n)}$	$H_N(LH)^{(n)}$	$H_N(HH)^{(n)}$	$H_S(HL)^{(n)}$	$H_S(LH)^{(n)}$	$H_S(HH)^{(n)}$
-------------	-------------	-------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------

Fig. 9. WNN input feature vector

Since wavelet representation of an iris image signal is resolved in only one wavelet resolution level, then the relative wavelet energy at any wavelet resolution level should be zero except at the wavelet resolution level that include the representative signal frequency[40,41]. To choose the migrated individual, a comparison among relative wavelet energy values, the wavelet entropy is used to make significate

G_{max} generations, all sub-populations are assumed with an equal size ζ_{g_0} . The first generation is randomly constructed with a constraint on similar individuals to be denied. For each chromosome individual, the WNN is trained with BP learning algorithm [34, 33], where the Mean Square Error (MSE), equation 10, represents each individual fitness value. In subsequent steps, each generation is evolved by constructing the pool mate with λ individuals from n^{th} sub-population $g^n(t)$, using the tournament selection method. A single point crossover method is used with probability p_c to created two new individuals in the next generation. Also, a new individual is created based on the mutation process with varied probability p_m at each generation ϕ [27].

$$p_m = p_{m_0} + \frac{3p_{m_0} * \phi}{G_{max}} \quad (11)$$

G_{max} is the maximum number of generations and p_{m_0} is the initial mutation probability. The best so far individuals are selected as a new members in the next generation. For a fixed number of generations, the best individual so far is the most effective combination parameters between DWT and WNN methods which characterize the texture of human iris.

In the island model, the migration process from each sub-population $g^n(t)$ execute a different evolutionary algorithm corresponding to a single decision variable. All the sub-population interact with themselves through the static hypercube migration process. The best chromosome in each sub-population migrate to another specific sub-population after each generation as depicted in Figure (7).The migration process depending on choosing the minimum number of wavelet coefficients that gain the same energy as the original iris texture. In other words the mother wavelet function that have wavelet energy with maximum values will be taken as migrate individual. Since the number of coefficients produced from the discrete wavelet decomposition is relatively equal or greater than the number of time samples of the original signal [39], then complexity of the signal approximation is chosen to be constant. The complexity of the signal is the ration between number of signal coefficients and the total number of signal samples $M = 2^N$.

values. Thus the wavelet entropy (WE), equation 12, should converge to zero or diverge to very low value. The migrated individual is chosen with WE, as illustrated in Figure 10.

$$WE = - \sum \rho_j \log \rho_j \quad \left. \begin{aligned} \rho_j &= \frac{E_{w_j}}{E_{tot}} \\ E_{tot} &= \sum_j \sum_k w_j^k \end{aligned} \right\} \quad (12)$$

Function Genetic_Algorithms(sub_pop, Fitness_f)

Input : Chromosome Set $(a_1, \dots, a_{z_{g_0}})$, specific scale j, crossover probability p_c mutate probability p_m , the migration

Output: Best match chromosome for specified iris ($Best_{match}$)

1. At a specific scale j, construct z_{g_0} random individual, $g = 1$, tempEntropy=Max_val
2. While ($g \leq G_{max}$) do (Monitor. Enter(obj))
 - a. Foreach (individual a_i in z_{g_0})
 - i. Extract w_j^k for each translation parameter. Compute the wavelet energy E_{w_j} (eq. 3) and wavelet norm and sure entropy with p and ϵ parameter (eq 4). Compute wavelet entropy WE(eq.12). Calculate fitness for each individual , eq. 10, inside the subpopulation in the corresponding client.
 - ii. If ($f(a_i) > f(Best_{match})$)
 - $Best_{match} = a_i$
 - iii. If ($WE_i < tempEntropy$)
 - $tempEntropy = WE_i$
 - $MigratElement = i$
 - b. MigrateList.add($a_{MigratElement}$)
 - c. By tournament selection, construct pool mate with λ individuals, then combine two individuals with probability p_c and mutate individuals with probability p_m , then change p_m (Eq. 11) .
 - d. Replace the worst individuals by the fittest in the current generation
 - e. If ($g \% migrate_{param} \neq 0$)
 - i. Monitor. Pulse(obj)
 - ii. Else
 - Monitor. Wait (obj)
 - replace worst chromosomes by MigrateList from another island.
 - f. $g++$ (Monitor. Exit(obj))
3. Return $Best_{match}$

Fig. 10. DGA sub-space

In the last stage, the WNN input layer represents wavelet energy and entropy values feature vector to WNN. The output layer represents the verified human iris. The middle layer determined the ability to learn the human iris recognition. The result in the output layer is either match (0) or unmatched (1).

IV. SIMULATED RESULTS

This section summarizes the results of using the proposed hybrid integrated system among DWT as feature extraction technique, WNN as classifier technique and Distributed GA as a searching strategy to select the optimal characterization for each human iris texture. This paper uses nine person eye images from CASIA-IrisV3-Interval eyes database[42]. each person has twenty images of his right eye, ten for train stage and another ten for test stage.

In **Segmenting and Normalizing Stage**, For the CASIA database, values of the iris radius range from 90 to 150 pixels, while the pupil radius ranges from 28 to 75 pixels [22]. In order to make the circle detection process more efficient and

accurate, the Hough transform for the iris/sclera boundary was performed first, then the Hough transform for the iris/pupil boundary was performed within the iris region, instead of the whole eye region, since the pupil is always within the iris region.

After this process was complete, six parameters are stored, the radius, and x and y center coordinates for both circles.

In **feature extraction and optimization method**, Distributed GA with two levels searching strategy is used to select the most effective integration between DWT and WNN parameters to each person. The parameters GA, DWT and WNN for each individual in the integrated system are shown in Table1.

TABLE I. GA, DWT AND WNN PARAMETER FOR SELECTING THE MOST EFFECTIVE PARAMETERS TO EACH HUMAN IRIS

GA parameters (First and Second searching Level)	
Sub- Population numbers	8
Population Size	200
Number of generation	100
Number of individual in each tournament selection	5
Reproduction percentage	14 % from population size
Crossover percentage	85.5% from population size
Mutation percentage	0.5% from population size
Number of tries trails	10
DWT analysis parameters	
Number of best analysis level	5 level
WNN architecture and training parameters	
The number of layers	3
The number of neuron on the layers	Input:21 Hidden:42 Output:1
The initial weights and biases	Random values
Learning rule	Back-Propagation
Number of epochs	500

As a result from the integrated system, Table2 shows the most effective integration between DWT and WNN parameters to each human iris from different (32 wavelet analysis function , 8 p-parameter ,8 ϵ -parameter, 8 wavelet activation function and 8 Learning rate values) all at analysis level N=5 as a result from Distributed GA.

To evaluate our proposed system, a comparative study between two strategies for iris verification systems:

1) Searching for suitable integration between DWT and WNN parameters using standard GA (DWT+WNN+GA).

2) Searching for suitable integration between DWT and WNN parameters using planned GA (DWT+WNN+planned GA).

Figure 11 represent the verification rate to each testing human iris data verification rate ranged from 100% to 98% to our proposed system. Since our proposed system has training verification rate 100% to all training data. Figure 11 concluded that our proposed system has the highest verification rate.

TABLE II. THE RESULT OF DISTRIBUTED GA SEARCHING PARAMETERS

Person number	Suitable wavelet analysis function	Suitable p-norm parameter value	Suitable ϵ -sure parameter value	Suitable wavelet activation function	Suitable Learning rate value
P1	bior 5.5	1	1	POLYWOG3	0.2
P2	coif1	1.994	6	Morlet	0.5
P3	db4	1.142	4	RASP2	0.9
P4	coif3	1.71	2	POLYWOG3	0.9
P5	bior 3.9	1	5	RASP3	0.6
P6	db20	1.285	3	POLYWOG3	0.6
P7	sym5	1.994	1	POLYWOG2	0.4
P8	db9	1.42	6	Morlet	0.3
P9	db2	1	8	RASP1	0.2

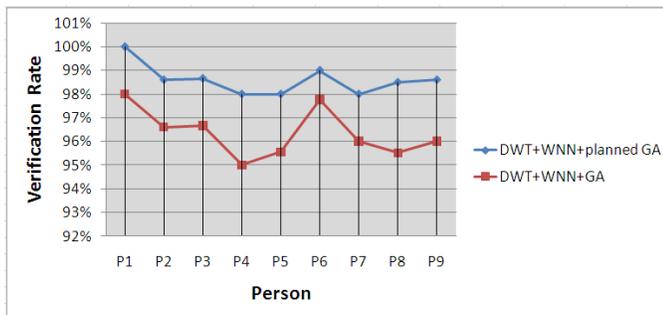


Fig. 11. Comparative Study between two version of the proposed hybrid integration system using conventional GA and distributed GA in Human Iris Verification Systems

Thus the performance of WNN is evaluated based on classification accuracy;

$$\text{Classification Accuracy} = \frac{\{\text{correct clasifed pattern}\}}{\{\text{Total Patterns}\}} = 98.5$$

V. CONCLUSIONS AND FUTURE WORK

Human Iris verification plays an important role in the daily life especially for security purposes. Human iris verification system suffers from several drawbacks, such as the interclass similarity and intraclass variations. Moreover some additional difficulties arose during learning process such as discovering the optimal iris texture and increasing the intrinsic features for each iris image. This paper tries to overcome these problems by optimizing and integrating model of DWT and WNN. The problems appeared by DWT, the down sampling process as well as the optimal wavelet analysis function associated with the most effective analysis level, are solved based on the global optimization method of Distributed GA. Besides WNN, a good and fast classifier, increases the intrinsic feature and works as a local optimization method. Also, WNN overcome the problem of lost information during DWT down sampling process. Because of the large searching space that represents the optimal DWT and WNN parameters, Distributed GA, new searching strategy based on GA, is used to search for an optimal wavelet analysis function at the most effective analysis level. Moreover, Distributed GA search for the optimal WNN activation function and the learning rate value. The benefit of Distributed GA in dealing with large searching space is to avoid the problems of premature convergence and population diversity. By this work, the integration between

DWT and WNN optimal parameters were able to introduce a new verification system that applied on CASIA database. The results demonstrate that this integration achieve good solution. Hence, a comparative study of human iris verification systems is appeared. The conclusion finds that our proposed system has high verification rate.

Unfortunately, Iris recognition system still undesirably increase recall rate so, it is in a need of more and more work. In future work, a new hybrid data fusion classification system should achieve better recognition in the digital iris image. The proposed system as any classification system based on the quality of the set of features characterizes the pattern and the efficiency of the classifier. A data fusion system based on two methods for texture analysis and constructing the set of features characterizes iris texture image is proposed. The first method is a statistical method to analyze the spatial distribution of gray values using co-occurrence matrix. The second method is a filtering method to analyze the frequency content of the iris image using contour-let transform. The data fusion system combines the extracted features, though an augmented features database is constructed.

ACKNOWLEDGMENT

The author deeply express gratitude for Libor Masek for making the iris recognition system code available to us. Also, a great appreciation for Chinese Academy of Sciences who make a public version of the CASIA Iris Database is available. Also, the authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for support.

REFERENCES

- [1] Kevin W. Bowyer *, Karen Hollingsworth, Patrick J. Flynn, Image understanding for iris biometrics: A survey", Computer Vision and Image Understanding 110 (2008) 281–307.
- [2] Omaira Nomir, Elsayed Radwan," Human Identification Using Iris Features", Proceedings of the Sixth IASTED International Conference on Advances in Computer Science and Applications, Sharm El-Sheikh,Egypt, pp 155-158, 2010.
- [3] Thiyam Churjit Meetei, and Shahin Ara Begum, A Comparative Study of Feature Extraction and Classification Methods for Iris Recognition, International Journal of Computer Applications (0975 – 8887), Volume 89 – No.7, pp. 13-20, March 2014.
- [4] R.M. Farouk, R. Kumar, K.A. Riad, " Iris matching using multi-dimensional artificial neural network", IET Computer Vision , Vol. 5, Iss. 3, pp. 178 –184, 2011.
- [5] V. Saishanmuga Raja, and S.P. Rajagopalan, " IRIS Recognition System using Neural Network and Genetic Algorithm", International

- Journal of Computer Applications (0975 – 8887) Volume 68– No.20, April 2013.
- [6] F. Hao, R. Anderson, & J. Daugman, "Combining Crypto with Biometrics Effectively", IEEE Transactions on Computers, p.p 1081-1088 , 2006.
- [7] Yaser Daanial Khan, Sher Afzal Khan, Farooq Ahmad, and Saeed Islam," Iris Recognition Using Image Moments and k-Means Algorithm", Hindawi Publishing Corporation, Scientific World Journal, Volume 2014, Article ID 723595 (<http://dx.doi.org/10.1155/2014/723595>).
- [8] H. Liang, Z. Cai, X. Chen, & K. Shuang, "Iris recognition based on characters of Iris's speckles",: 7th World Congress on Intelligent Control and Automation, p.p 6793-6797 , 2008.
- [9] Mayank Vatsa, Richa Singh, and Afzel Noore," Improving Iris Recognition Performance Using Segmentation, Quality Enhancement, Match Score Fusion, and Indexing",IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 38, NO. 4, pp. 1021-1035, AUGUST 2008.
- [10] Ma L., Tan T., Wang Y. and Zhang D. 2004. Efficient Iris Recognition by Characterizing Key Local Variations. IEEE Trans. Image Processing 13(6):739-750.
- [11] John W. Leis, "Digital Signal Processing using Matlab for Students and Researchers", John Wiley & Sons, Inc., 2011
- [12] K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi, & H. Nakajima, "An Effective Approach for Iris Recognition Using Phase-Based Image Matching", IEEE Transactions on Pattern Analysis and Machine Intelligence, p.p 1741-1756, 2008.
- [13] Zhiping Zhou, Huijun Wu and Qianxing Lv," A New Iris Recognition Method Based on Gabor Wavelet Neural Network",International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008.
- [14] Engin Avci , Abdulkadir Sengur, Davut Hanbay, " An optimum feature extraction method for texture classification", Expert Systems with Applications: An International Journal, Published by Elsevier Ltd, Vol.36 , No. 3,p.p. 6036-6043,2009.
- [15] Mahmoud Elgamal, and Nasser Al-Biqami, An Efficient Feature Extraction Method for Iris Recognition Based on Wavelet Transformation, International Journal of Computer and Information Technology (ISSN: 2279 – 0764), Volume 02– Issue 03, pp. 521-527, May 2013.
- [16] Sandipan P. Narote , Abhilasha S. Narote, Laxman M. Waghmare," Iris Based Recognition System Using Wavelet Transform", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.11, p.p 101-104, 2009.
- [17] Mayada Tarek, Taher Hamza, El-sayed Radwan," Off-line Handwritten Signature Recognition Using Wavelet Neural Network", International Journal of Computer Science and Information Security, Vol. 8, No. 6, p.p. 13-21 , 2010.
- [18] Omaima N. Ahmad AL-Allaf, Shahlla A. AbdAlKader, Abdelfatah Aref Tamimi, Pattern Recognition Neural Network for Improving the Performance of Iris Recognition System,(ISSN 2229-5518) International Journal of Scientific & Engineering Research, Volume 4, Issue 6, pp. 661-667, June-2013.
- [19] Zhenan Sun and Tieniu Tan, Ordinal Measures for Iris Recognition, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 12, pp. 2211-2226, DECEMBER 2009
- [20] Sing-Tze Bow,"Pattern recognition and image pre-processing", Marcel Dekker,Inc, chapter 15,2002.
- [21] Mrs. Minakshi R. Rajput, Iris feature extraction and recognition based on different transforms, International Journal of Engineering Research and Development, Volume 9, Issue 2 (November 2013), PP. 30-35.
- [22] Tania Johar, Pooja Kaushik, "Iris Segmentation and Normalization using Daugman's Rubber Sheet Model," International Journal of Scientific and Technical Advancements, Volume 1, Issue 1, pp. 11-14, 2015.
- [23] Zhang Q. and Benveniste A,"Wavelet Networks", IEEE Trans. On Neural Networks ,Vol.3, p.p. 889-898, 1992.
- [24] Jan Stolarek, " On properties of a lattice structure for a wavelet filter bank Implementation: Part I", Journal of Applied Computer Science", vol. 19, no. 1, pp. 85-116, 2011
- [25] Christian Blum, Jakob Puchinger, Gunther R. Raidl, Andrea Roli," Hybrid metaheuristics in combinatorial optimization: A survey", Applied Soft Computing 11 (2011) 4135–4151.
- [26] Enrique Alba, Gabriel Luque and Sergio Nesmachnow, Parallel metaheuristics: recent advances and new trends", International Transactions in Operational Research, 20 (2013) 1–4.
- [27] Gautam Garai and B. B. Chaudhuri, " A Distributed Heirarchical Genetic Algorithms for Efficient Optimization and Pattern Matching", Pattern Recognition , vol. 40, pp. 212-228, 2007.
- [28] Jan Roupec and Pavel Popela, The Nested Genetic Algorithms for Distributed Optimization Problems", Proceedings of the World Congress on Engineering and Computer Science 2011 Vol I, WCECS 2011, October 19-21, 2011, San Francisco, USA.
- [29] Reem Abd El-Salam El-Deeb, Elsayed Radwan, Taher Hamza, "Hybrid Model of Texture Classification using 2D Discrete Wavelet Transform and Probablistic Neural Network", International Journal of Computer Science and Information Security, vol. 8 no. 5, pp. 148-154, 2010
- [30] A. Wahi, E. Thirumurugan "Recognition of Objects by Supervised Neural Network using Wavelet Features", First International Conference on Emerging Trends in Engineering and Technology, p.p. 56-61,2008.
- [31] Chi-Man Pun and Moon-Chuen Lee ,"Log-Polar Wavelet Energy Signatures for Rotation and Scale Invariant Texture Classification", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 25, No. 5, p.p. 590 – 603 , 2003 .
- [32] JAN STOLAREK, " Improving energy compaction of a wavelet transform using genetic algorithm and fast neural network", Archives of Control Sciences, Volume 20(LVI), 2010, No. 4, pages 417–433.
- [33] Antonios K. Alexandridis and Achilleas D. Zaprakis, " Wavelet neural networks: A practical guide", Neural Networks 42 (2013) 1–27.
- [34] S.Sitharama Lyengar,E.C.Cho,Vir V.Phoh ,"Foundations of Wavelet Networks and Application", chapman&Hall/CRC Press LLC , chapter 4, 2002.
- [35] Xian-Bin Wen, Hua Zhang, and Fa-Yu Wang," A Wavelet Neural Network for SAR Image Segmentation", Sensors , Vol.9, No.9, p.p. 7509-7515,2009.
- [36] Randy L. Haupt, Sue Ellen Haupt, "Practical Genetic Algorithms", Willey- InterScience, Second Edition, chapter 2, 2004.
- [37] Zhihua Cai, Chengyu Hu, Zhuo Kang and Yong Liu, " Advances in Computation and Intelligence", Proceedings of 5th International Symposium, ISICA 2010, Wuhan, China, October 22-24, 2010.
- [38] G.Y. Chen, T.D. Bui, A. Krzyzak," Contour-based handwritten numeral recognition using multi-wavelets and neural networks", Pattern Recognition ,Vol.36 ,p.p. 1597 – 1604, 2003.
- [39] Oltean, G.; Ivanciu, L.-N.; Kirei, B., "Signal approximation using GA guided wavelet decomposition," in Signals, Circuits and Systems (ISSCS), 2015 International Symposium on Lasi , pp.1-4, 9-10 July 2015. doi: 10.1109/ISSCS.2015.7203996.
- [40] Osvaldo A. Rosso, Susana Blanco, Juliana Yordanova Vasil Kolev," Wavelet Entropy: a New Tool For Analysis of Short Duration Brain Electrical Signal", Journal of NeuroScience Methods, vol. 105, pp. 65-75, 2001.
- [41] Yatindra Kumar, Mohan Lal Dewal and Radhey Shyam Anad," Relative Wavelet Energy and Wavelet Entropy based Epileptic Brain Signal Classification", Biomedical Engineering Letters, pp. 147-157, 2012.
- [42] Chinese Academy of Sciences – Institute of Automation. Database of 756 Grayscale Eye Images. <http://biometrics.idealtest.org/dbDetailForUser.do?id=3> Version 3.0, last seen Nov. 2015.

Composable Modeling Method for Generic Test Platform for Cbtc System Based on the Port Object

WAN Yongbing

Shanghai Rail Transit Technical
Research Center
Shanghai, China

WANG Daqing

Shanghai Rail Transit Technical
Research Center
Shanghai, China

MEI Meng

School of Electronic & Information
Engineering
Tongji University
Shanghai, China

Abstract—The Communications-based train control(CBTC) system has gradually become the first choice for signal systems of urban mass transit. However, how to guarantee its safety has become a research hotspot in safety fields. The generic test system with high efficiency has become the main means to verify the function and performance of CBTC system. This paper discusses a composable modeling method for the generic test platform for CBTC system based on the port object. This method defines the port object(PO) model as the basic component for composable modeling, verifies its port behavior and generates its compositional properties. Based on the port description and the test environment description, it builds port sets and environment port cluster, respectively. Then it analyzes and extracts possible crosscutting concerns, and finally generates a variable PO component library. It takes the modeling of block port objects in line simulation of generic test platform for CBTC systems as an example to verify the feasibility of the method.

Keywords—composable modeling; test platform; CBTC; port object; line simulation

I. INTRODUCTION

By allowing trains to operate safely at closer headways, CBTC system can permit more effective utilization of rail transit infrastructure. It has become the preferred standard of urban rail traffic signal system. As a kind of safety critical systems (SCS), however, CBTC system, with the highest safety requirements, is directly responsible for the train operation[1]. Once the system fails, it will lead to a great or even a catastrophic loss of lives, property, and environment.

As one of the important means to improve the safety and assure the quality of the system, testing plays an important guiding role in the process of researching and developing the CBTC system. However, the key issue of the implementation of system testing is how to build a simulation and test environment in accordance with the real operational scenario of CBTC system under test(SUT)[2,3]. Hence, the more attention paid on establishing the test platform for CBTC system, the higher the requirement. The test platform for CBTC system is developing to the direction of network, intelligence and generalization. The test platform is a virtual environment which is used to verify the correctness and reliability of system design. It generally includes the input, processing, validation, and output of signal data, which can not only meet the needs of SUT for functional verification but also some non-functional verification, such as performance test,

pressure test and safety test[4].

II. TEST PLATFORM FOR CBTC SYSTEM

A perfect test platform for CBTC system is the combination of simulation technology and testing technology. The simulation activity builds up the desired external scenario and simulates the external environment for SUT. The key to the following testing process is whether the design of a simulation environment is successful or not. And the testing process is a continuation of the simulation activity and its sequential execution[5,6]. A full process of building the test platform for CBTC system is shown in Figure 1.

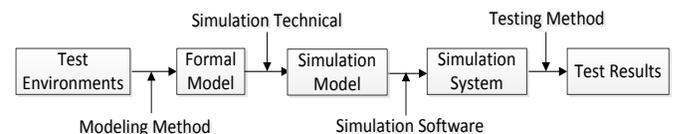


Fig. 1. The process of building the test platform for CBTC system

So far, a widely used and generic test platform for CBTC system hasn't been built. Besides the challenge of developing CBTC system, there are also some problems of building the test platform for CBTC system. One is the architecture of CBTC system differs between the vendors, requiring the test platform makes quick adjustment to adapt to the test of different vendors; another is some of the CBTC interfaces differ between the vendors, requiring the test platform has variable interfaces and is easy to switch and roam seamlessly from interfaces to interfaces.

By aiming at establishing a generic test platform for CBTC systems developed by multiple vendors, it focuses on how to build a generic test model for CBTC system, and how to design effective test platform architecture for CBTC system in this paper. Although [7], [8] and [9] have discussed the port-based approach to integrated modeling and simulation of SCS, respectively, they consider the SUT more without the property to safety, real-time and was not suitable to build the test system for SCS. This paper further the evolution towards a seamless integration of simulation for SCS test platform with the idea of a PO.

III. COMPOSABLE MODELING BASED ON THE PORT OBJECT

Unlike traditional closed-loop system, the need for designing and modeling generic test platform for CBTC system

comes from the external environment[10,11]. Furthermore, the system is designed and modeled merely based on external interface documents and its modeling method should both consider port evolvable and module replaceable, renewable and reusable, which pose a great challenge to the modeling method for the test platform. The concept of Port Object is given below, and a composable modeling method based on Port Object is proposed.

A. Port Object

The Port Object, PO is a kind of novel software abstract that is configurable and replaceable, and the basic unit that generates the system components. As is shown in figure 2, it incorporates the concept of objects and the port automata model used for concurrent processing, and meet the needs of the communication mechanisms in safety-critical areas[12].

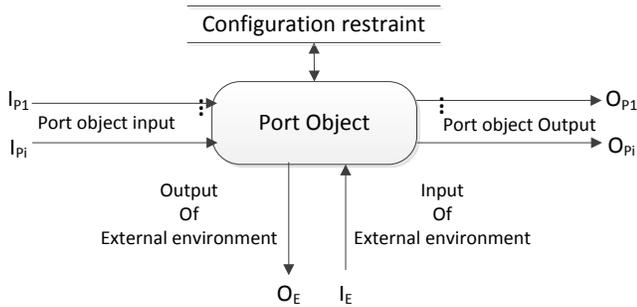


Fig. 2. Port Object

Definition 3.1 (Port Object) is an eight-tuple: $P = \langle S, S_0, I_p, O_p, I_E, O_E, \tau, R \rangle$, where S is a set of states; S_0 is an initial state and contains at most one state; I_p is a set of inputs from other PO; O_p is a set of outputs to other PO; I_E is the input event from external environment, which can be null, namely $I_E = \emptyset$; O_E is the output event to external environment, which can be null, namely $O_E = \emptyset$; $\tau \subseteq S \times I_p \times O_p \times I_E \times O_E$ is the transaction relationship; R is the configuration constraints of PO.

To better understand the PO, and further illustrate its properties, it defines the following relations. Defining that the Port Object P ; two states $s_1, s_2 \in S$; $A_p = I_p \cup O_p$, where A_p is the set of events between POs; $A_E = I_E \cup O_E$, where A_E is the set of events between POs and the CBTC system; $A = I_p \cup O_p \cup I_E \cup O_E$, where A is the set of all events; if $a \in A_p$, $b \in A$, then two finite action sequences $\alpha = a_1 a_2 a_3 \dots a_n \in (A_p)^n$ and $\beta = b_1 b_2 b_3 \dots b_n \in (A)^n$. Under these conditions, it has:

- 1) $s_1 \xrightarrow{a} s_2$, if and only if $(s_1, a, s_2) \in \tau$;

- 2) $s_1 \xrightarrow{\tau} s_2$, if and only if there is an action b which makes $s_1 \xrightarrow{b} s_2$;

- 3) $s_1 \xrightarrow{\alpha} s_2$, if and only if $s_1 \xrightarrow{a_1} p \xrightarrow{a_2} p \dots \xrightarrow{a_n} s_2$, especially $s_1 \xrightarrow{\varepsilon} s_1$;

- 4) $s_1 \xrightarrow{\alpha} p^*$, if and only if there is a state $s_2' \in S$ which makes $s_1 \xrightarrow{\alpha} s_2'$;

- 5) $s_1 \xrightarrow{*} s_2$, if and only if there is a finite action sequence $\gamma \in A$ which makes $s_1 \xrightarrow{\gamma} s_2$;

- 6) $s_1 \xrightarrow{b} s_2$, if and only if $s_1 \xrightarrow{\varepsilon} p \xrightarrow{b} p \xrightarrow{\varepsilon} s_2$;

- 7) $s_1 \xrightarrow{\varepsilon} s_2$, if and only if $s_1 (\xrightarrow{\tau} p)^* s_2$;

- 8) $s_2 \xrightarrow{\beta} s_2$, if and only if $s_1 \xrightarrow{b_1} p \xrightarrow{b_2} p \dots \xrightarrow{b_n} s_2$.

Where $*$ is the reflexive transitive closure, and the union operation means the combination of two relations[13].

B. Ports' Behavior Description of PO Model

The ports' behavior elements of PO model is described as below: $M := \{m\}$, where m is the input identification, which means messages; $R := \{r\}$, where r is the output identification, which means responses; $C := \{c\}$, where c is the Boolean expression.

The internal port behavior of the PO has four different situations:

- 1) *Null output*: $f : M \rightarrow \varepsilon$ which denoted by $m_1 \mapsto \varepsilon$;

- 2) *Only one output*: $f : M \rightarrow R$ which denoted by $m_2 \mapsto r_1$;

;

- 3) *Sequential output*: $f : M \rightarrow R^*$, $R^* = \{r_{k_0}, r_{k_1}, \dots, r_{k_n} \mid r_{k_i} \in R \cup \{\varepsilon\}$ which denoted by $m_3 \mapsto r_2 r_3 r_4$, where $k_i \in \{0, 1, 2, 3, \dots\}$;

- 4) *Branch output*: $f : C \circ M \rightarrow C \circ R^*$ or $f : C' \circ M \rightarrow C' \circ R^*$, where $C' := C \cup \{\varepsilon\}$

To ensure security and real-time communication for CBTC test system, it usually needs timeliness and order calibration. Here, it assumes that the PO model in simulation port as a receiver, and then it receives secure data sent by CBTC system periodically. After receiving the secure data, it will carry out a real-time inspection of the time sequence. If it isn't synchronized, then returns the timeliness calibration request and waits to receive the response from timeliness calibration. If synchronized and passed validation, then it shall proceed with receiving secure data[14]. The internal port behavior of this PO

model is shown in Figure 3. The branch output expression is derived as : $c_1m_4 \mapsto c_1r_5$, $c_2m_4 \mapsto c_2r_6$.

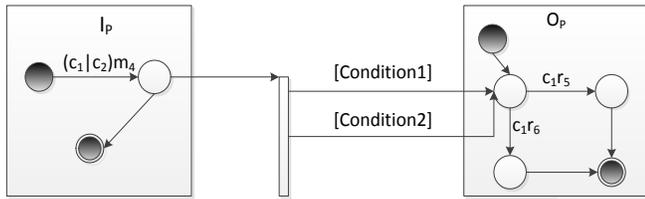


Fig. 3. The internal port behavior of PO mode

C. Configuration Constraint of PO Model

The PO configuration constraint is one of the important parts of the PO model which describes the configuration instruction, external environment interfaces, internal interfaces between objects and properties of PO[15].

- Configuration Content

The structure of PO configuration constraints is shown as Figure 4.

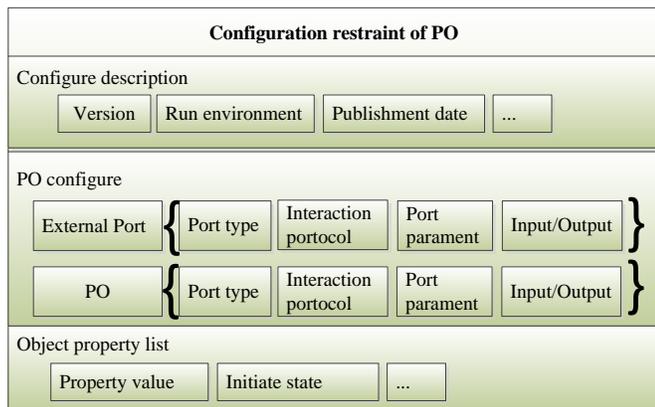


Fig. 4. The structure of PO configuration constraints

1) Configuration Description

The Description of the current configuration constraints, describing the current configuration, such as version, release date and operating environment for PO, mainly facilitates the management of PO, avoids confusion version, as well as the combined behavior or debugging failure due to operating environment errors. Also, configuration description is the main information when components added into the component library.

2) Interface Configuration

Interface configuration comprises the interface configuration with external environment and between POs. The external interface refers to the interfaces with the CBTC system and testers, which includes the interface type, interaction protocols, interface parameters and Input/Output(I/O) description, etc. Depending on its interface mode, the described interface properties shall be different with different modes.

3) Object property table

Object property table describes the relation between interface properties or between properties and values. There

can also be a description of the objects' initial states in the table.

- Configuration Description Based on XML

By nesting and referencing the hierarchical relations between the specific elements, the extensive markup language (XML) uses elements and properties to describe data. The XML configuration description is given below based on the structural characteristics of the configuration constraints in PO model[16]. The XML configuration template is described in Figure 5.

```
<?xml version="1.0" encoding="GB2312"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xsd:element name="Configuration">
    <xsd:sequence>
      <xsd:complexType>
        <xsd:element>
          <xsd:complexType name="Description">
            <xsd:sequence>
              <xsd:element ref="Version"/>
              <xsd:element name="Runtime_Environment" type="xsd:string"/>
              <xsd:element name="Release_Date" type="xsd:string" />
            </xsd:sequence>
          </xsd:complexType>
        </xsd:element>
      </xsd:complexType>
    </xsd:sequence>
  </xsd:element>
  <xsd:complexType name="Ports">
    <xsd:sequence>
      <xsd:complexType name="Environment_Ports">
        <xsd:sequence>
          <xsd:element name="Port_Name" type="xsd:string"/>
          <xsd:element name="Port_Type" type="xsd:string"/>
          <xsd:element name="Port_Protocol" type="xsd:string"/>
          <xsd:element name="Port_Output" type="xsd:string"/>
          <xsd:element name="Port_Iutput" type="xsd:string"/>
          <xsd:element name="Port_Property" type="xsd:string"/>
          .....
        </xsd:sequence>
      </xsd:complexType>
    </xsd:sequence>
  </xsd:complexType>
</xsd:sequence>
</xsd:element>
</xs:schema>
```

Fig. 5. The XML configuration template

D. Composability of PO Model

Definition 3.3 (Composability of PO) The two PO models M and N , if the following conditions are met:

$$A_{P_M} \cap A_{P_N} = \phi, I_{P_M} \cap I_{P_N} = \phi, O_{P_M} \cap O_{P_N} = \phi$$

Then M and N are called composable, and are defines as: $shared(M, N) = (I_{P_M} \cap O_{P_N}) \cup (O_{P_M} \cap I_{P_N})$. To give compatibility and replaceability of the PO model, the related definition is given below.

Definition 3.4 (Product of PO) If PO models M and N are composable, and therefore their product $M \otimes N$ is defined as:

$$S_{M \otimes N} = S_M \times S_N ; S_{0_{M \otimes N}} = S_{0_M} \times S_{0_N} ;$$

$$\begin{aligned}
 I_{P_{M \otimes N}} &= I_{P_M} \times I_{P_N} \setminus \text{shared}(M, N); \\
 O_{P_{M \otimes N}} &= O_{P_M} \times O_{P_N} \setminus \text{shared}(M, N); \\
 I_{E_{M \otimes N}} &= I_{E_M} \times I_{E_N} \setminus \text{shared}(M, N); \\
 O_{E_{M \otimes N}} &= O_{E_M} \times O_{E_N} \setminus \text{shared}(M, N); \\
 \tau_{M \otimes N} &= \{((v, u), a, (v', u')) \mid (v, a, v') \in \tau_M \wedge a \notin \\
 &\quad \text{shared}(M, N) \wedge u \in S_N\} \cup \{((v, u), a, (v, u')) \mid \\
 &\quad (u, a, u') \in \tau_N \wedge a \notin \text{shared}(M, N) \wedge v \in S_M\} \cup \\
 &\quad (v, a, v') \in \tau_M \wedge (u, a, u') \in \tau_N \wedge a \in \text{shared}(M, N)\}
 \end{aligned}$$

Definition 3.5 (Illegal State) If PO models M and N are composable, and therefore the illegal state set $Illegal(M, N) \subseteq (S_M \times S_N)$ in $M \otimes N$ is defined as:

$$\begin{aligned}
 Illegal(M, N) &= \\
 &\left\{ (v, u) \in S_M \times S_N \mid \exists a \in \text{shared}(M, N) \begin{cases} a \in O_{P_M}(v) \wedge a \notin I_{P_N}(u) \\ \vee \\ a \in O_{P_N}(u) \wedge a \notin I_{P_M}(v) \end{cases} \right\}
 \end{aligned}$$

Definition 3.6 (Environment of PO) The environment E of the PO model M shall meet the following conditions: 1) E and M is composable; 2) E is non-null; 3) $I_{P_E} = O_{P_M}$; 4) $Illegal(M, E) = \phi$.

Definition 3.7 (Legal Environment of PO) If PO models M and N are composable, E is the environment of $M \otimes N$, the state of $Illegal(M, N) \times S_E$ is unreachable in $(P \otimes Q) \otimes E$, and then E is called a legal environment of (M, N) ,

Definition 3.8 (Compatibility of PO) If PO models M and N are non-null and composable, additionally, there is a legal environment (M, N) , and then M and N are called compatible.

Definition 3.9 (Replaceability of PO) N is the PO model, E_N is the environment of N in system S , I is the bind action set of N and environment E_N , C Replaced(N, S) is the replace action of N in system S , and therefore C Replaced(N, S) = $(P_N / I) // (P_{E_M} \downarrow \bar{I})$.

IV. COMPOSITE-ORIENTED MODELING PROCESS OF PO MODEL

A. Acquisition of PO model

There are three main sources of demands for the generic test platform for CBTC system: interface document, test environment description and test requirements. The interface document describes all the external interfaces of tested CBTC

system, including the interface type, interface parameters and interaction protocols. The test environment description includes all the required external system, infrastructure and environmental constraints. The test requirement mainly describes the needed tests for CBTC system, including the functional test and performance test. The key to composable modeling a generic test platform for CBTC system is doing requirements analysis based on the interface document, test environment description and test requirements, acquiring the PO model and then generating the component library.

The PO model mainly focuses on the port, so does acquiring the PO model. The acquisition process of PO model is described in Figure 6.

Step1: Based on the interface document, the test system is divided into different ports, and generate port sets.

Step2: Based on the test environment description, the environmental resources is it analyzed, then it is added to those covered by the interfaces to the appropriate port set, and the environment port cluster is generated.

Step3: Based on the test requirements, the man-machine interface is extracted and it is added to the appropriate environment port cluster or as an independent environment port.

Step4: After the environment port cluster is analyzed, and the possible crosscutting concerns is extracted, the problem of requirement distraction and requirement entanglement will be resolved.

Step5: Based on the crosscutting concerns, the environment port cluster is divided into different PO.

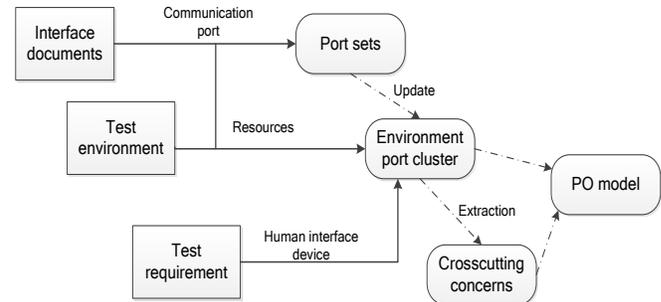


Fig. 6. Process of acquiring the PO model

B. Generate Port Cluster

Description of the test environment comprises all the resource sets needed by the test platform. These resource sets are the state-description objects in PO model. Resources' dependency on the interface comprises.

- 1) *Call*, resources call interface to communicate in operation.
- 2) *Interrupt*, resources interrupt interface communication in operation due to state changes.
- 3) *Modify*, resources (interfaces) modify resources (interfaces) parameters or states in initial start when it is needed.
- 4) *Update*, resources update their states due to interaction information.

5) *Mutually exclusion*, this relationship is less likely to appear, which means the interface will not be activated if resources participated in operation at the very beginning.

Define the resource set as *ResourceSet*, which comprises all the resources of the test system. Define the port set as *PortSet*, which comprise all the communication interfaces of the test system. Define the port cluster as *PortCluster*. *Resources'* dependency on the interface denotes as *r*. The way of generating port clusters is given below, as shown in Figure 7.

```
InPut: ResourceSet,PortSet;
OutPut: PortClusters;
PortCluster= $\phi$ ;
RS  $\in$  ResourceSet;
Do
{
  PortSet' = PortSet;
  Do
  { RS'  $\in$  ResourceSet' ;
    PC  $\in$  PortCluster;
    If (RS'  $\neq$  RS  $\vee$  PCrRS=true)
      PortCluster= PortCluster+{ RS' };
    Endif
    PortSet' = PortSet' - {RS};
    Until {PortSet' =  $\phi$  };
    ResourceSet=ResourceSet-{RS};
    Until{ ResourceSet= $\phi$  }
  }
}
```

Fig. 7. The way of generating port clusters

C. Extract Crosscutting Concern

In the built environment port clusters, some resources will appear in different clusters, which are both related to one PO and another, in other words, exist in the intersection of different environment port clusters.

The crosscutting concerns of a resource in two environment port clusters can be written as $CC=\{Re|Re \in PortCluster1 \wedge Re \in PortCluster2\}$, where CC is the crosscutting concern, *Re* is the resource and *PortCluster* is the environment port clusters.

In an environment port cluster, if there is a resource which has dependency on several ports, then this crosscutting concern can be expressed as: $CC = \{Re | RerPortCluster1 \wedge RerPortCluster2 \wedge \dots \wedge RerPortCluster\}$.

Here an algorithm used seeking all the possible crosscutting concerns is introduced, where CS is the cluster set of all the possible crosscutting concerns, RS is the set of all resources, RRS is the resource relationship set. *ExistR(Re, Po, r)* is a boolean function, if there is a relationship *r* between the resource *Re* and the port cluster *Po*, then its value is true. If not, its value is false. The way of extracting crosscutting concerns is described in Figure 8.

```
CS= $\phi$ ;
RRS= $\phi$ ;
For each RS  $\in$  ResourceSet and Po  $\in$  PortClusters
  If Exist(Re,Po,r)=True
    Then RRS={ (Re,Po,r) }  $\vee$  RRS;
  Endif
For each two (Re1,Po1,r1),(Re2,Po2,r2)  $\in$  RRS
  If r1=r2 & re1=re2
    Then CS={ Re1 }  $\vee$  CS
  Endif
```

Fig. 8. The way of extracting crosscutting concerns

D. Generate PO Model

Once the environment port cluster is formed and the crosscutting concerns which causes distraction and entanglement are extracted, the PO model can be generated based on environment port cluster and crosscutting concerns. Define the port object model set as POMS and CS is the cluster set of all the possible crosscutting concerns. The way of generating the PO model is described in Figure 9.

```
POMS= $\phi$ ;
For each PortCluster
  If PortCluster  $\notin$  CS
    Then PortCluster  $\rightarrow$  PO
  Else
    PortClust=PortCluster-{CC}
    AddObjectPort(CC,PortCluster);
    PortCluster  $\rightarrow$  PO
  If CC  $\notin$  POMS
    Then
      AddObjectPort(CC,PortCluster);
      CC  $\rightarrow$  PO
    Endif
    POMS=POMS+{PO}
  Endif
```

Fig. 9. The way of generating the PO mode

V. INSTANCE ANALYSIS

An example of the PO in the line simulation of test platform for CBTC system is presented, and a detailed modeling process is introduced. Some of the descriptions of the three documents are tabulated in Table 1 for further illustration.

Table 1 only tabulates some of the details about the feature points and appropriate PO models can be built based on these details. The modeling process is shown in Figure 10.

1) *First of all, the port set which obtained through the port description only describes the ports' <type, constraint>, such as 24V, two relay states (0 and 1) and relay connection method.*

TABLE I. PORT DESCRIPTION, ENVIRONMENT DESCRIPTION AND TEST REQUIREMENTS

Number	Document	Content description
1	Port description	Wayside signal system connected to the wayside equipment by 24V relay, 24V relay.
2	Wayside description	Wayside equipment including: section, switch, Signaling, platform screen door, platform emergency stop button, depot/park and drivers protect button, etc. Section includes two states, free and occupation.
3	Test requirement	When the train operation to a track section, CBTC wayside signal equipment should be able to collect this information. When a section is occupation because of a failure, CBTC wayside signal equipment should be able to collect the fault information.

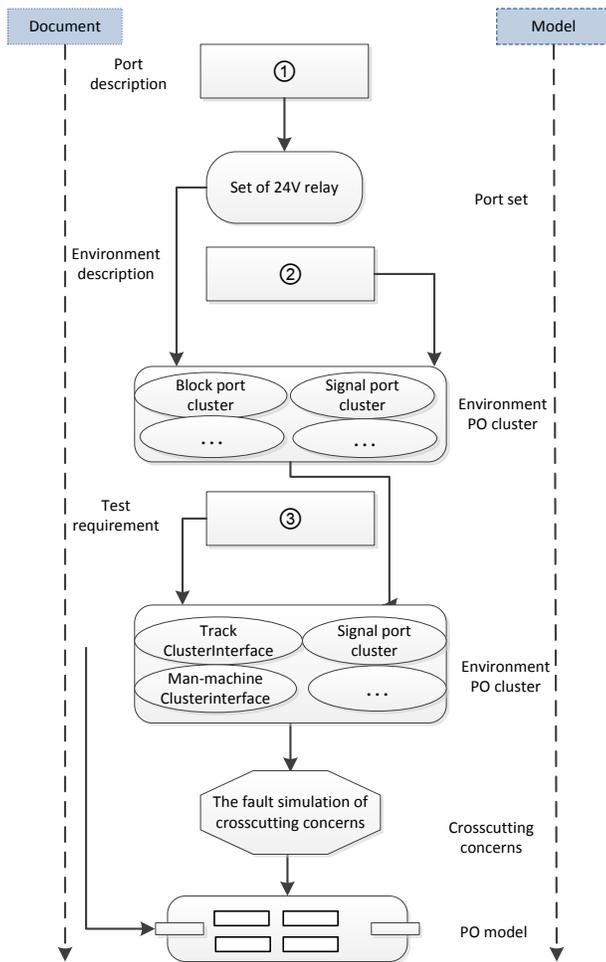


Fig. 10. The modeling process of PO of line simulation

2) After environment description is added, a connection object $\langle \text{port type, port constraint, resource description, dependency} \rangle$ of the 24V relay is generated, where the port type and the port constraints integrate the description of the port set. Resources description includes the section number, the section length and the starting point. Dependency is described as the following. Assume that one section connected with one relay (ignore reacquisition). When the section is occupied and the drive relay is energized, the output is "Occupy", or the section is vacant and the drive relay de-energized, output is "Idle".

3) After test requirements are added, the man-machine interface is added to the environment port cluster. The sector port cluster has two external inputs: the man-machine interface and the train motion simulation PO. Moreover, both of them are POs. The output of the sector port cluster is the environment port of 24V relay.

4) With the further analysis on test requirements, the fault simulation will be appeared in many port clusters. Thus, they are extracted and are named as crosscutting concerns. All of them are POs, their inputs are man-machine port clusters and outputs are other environment port clusters.

5) It is much easier to convert the section port cluster to the section port. Define that the PO's state $S = \{0, 1\}$ where "0" is "Occupied" and "1" is "Idle" and its initial state is 1. Each of the section ports $I_P = \{I_{P1}, I_{P2}\}$ has two POs $I_{P1} = \{\text{Train_Position}\}$, $I_{P2} = \{\text{Fault_Occupy}, \text{Fault_Idle}\}$ and one environment port $O_E = \{\text{Occupy}, \text{Idle}\}$.

The configuration constraints can be easily generated based on the port constraints, as shown in Figure 11.

```
<Configuration>
  <Description>
    <Version>V1.0</Version>
    <Runtime_Environment>WondowsXP,VS2010,
      SQL Server</Runtime_Environment>
    <Release_Date>2013-09-01</Release_Date>
  </Description>
  <Ports>
    <Environment_Ports>
      <Port_Name>XL_LS_Port</Port_Name>
      <Port_Output>Occupy/Idle</Port_Output>
      <Port_Property>24V</Port_Property>
      <Port_Count>128</Port_Count>
    </Environment_Ports>
    <Object_Ports>
      <Port_Name>Fault_Port</Port_Name>
      <Port_Input>Fault_Occupy/Fault_Idle</Port_Input>
      <Port_Output>Fault_Occupy/Fault_Idle</Port_Output>
      <Port_Protocol>Shared Memory</Port_Protocol>
    </Object_Ports>
    <Object_Ports>
      <Port_Name>Train_Port</Port_Name>
      <Port_Input>Train_Position</Port_Input>
      <Port_Output>Occupy/Idle</Port_Output>
      <Port_Protocol>Shared Memory</Port_Protocol>
    </Object_Ports>
  </Ports>
  <Property_List>
  <Track>
    <Name>T2115</Name>
    <ID>0x1C000009</ID>
    <StartE>50</DeviceWidth>
    <EndE>360</EndE>
  </Track>
  ...
  <Track>
    <Name>T2116</Name>
    <ID>0x1C000010</ID>
    <StartE>360</DeviceWidth>
    <EndE>558</EndE>
  </Track>
  </Property_List>
```

Fig. 11. The configuration constraints

VI. CONCLUSION

The test and verification of CBTC system has become one of the important means of ensuring system security. While building the generic test platform for CBTC system is the prerequisite for test and verification. Nowadays, CBTC system having a more complex architecture, more diversified interfaces and more often upgrading, which poses a greater challenge to the modeling and simulation technology. In this paper, some preliminary study on how to build and realize the generic test platform for CBTC system will be done, a composable modeling method for CBTC simulation and test system based on the port object will be suggested, and integral

composite-oriented modeling process as well as verify the feasibility of the method will be illustrated

In this paper, a meaningful technology method are put forward for resolving the problems in the process of development for test platform of CBTC system, but there are still many works to be further researches, such as another important aspects with the combination of system modeling, namely the assembly based on component technology. In view of the PO model based assembly technology, a best method of assembly needs further discussion and analysis.

REFERENCES

- [1] Z. Yujun, X. Zhongwei, M. Meng. "Port-based Composable Modeling and Simulation for Safety Critical System Testbed". Lecture Notes in Computer Science, Vol.7529, pp.51-58, 2012.
- [2] A. Speck, E. Pulvermuller, M. Jerger, et al. Component Composition Validation. International Journal of Applied Mathematics and Computer Science, Vol.12, pp.581-589, 2002.
- [3] G.V. Bochmann, S. Haar, C. Jard, et al. Testing Systems Specified as Partial Order Input/Output Automata. Testing of Software and Communicating Systems. Springer Berlin Heidelberg, Vol.5, pp.169-183, 2008.
- [4] L. Bin, W. Xin, F. Hernan, M. Antonello. A Low-Cost Real-time Hardware-in-the-loop Testing Approach of Power Electronics Controls. IEEE Transactions on Industrial Electronics, Vol.57, pp.919-931, 2007.
- [5] L. Chen, W.P. Wang, Y.F. Zhu. Research on SEB Composable Modeling methodology for system-of-system Combat Simulation. Journal of System Simulation, Vol.19, pp.644-656, 2007.
- [6] G.Y. Wang, Q.W. Hu, W. Liu. Study on Composable Modeling and Simulating technology for equipment battlefield damage. Binggong Xuebao/Acta Armamentarii, Vol.10, pp.1266-1275, 2012.
- [7] C.B. Peter. Port-based Modeling of Mechatronic Systems. Mathematics and Computers in Simulation, Vol.66, pp.99-127, 2004.
- [8] Y.L. Lei, Q. Li, F. Yang, W.P. Wang. A Composable modeling framework for weapon system effectiveness simulation. System Engineering Theory and Practice, Vol.11, pp.2954-2966, 2013.
- [9] C. Yuan, N. Ru, T.H. Xu, T.Tang. Wireless Test Platform of Communication based Train Control(CBTC) System in Urban Mass Transit. Proc. of the 2007 IEEE International Conference on Vehicular Electronics and Safety, pp.39-14, 2008.
- [10] D.I. August, S. Malik, L.S Peh, P. Willmann. Achieving Structural and Composable modeling of Complex Systems. International Journal of Parallel Programming, Vol.2, pp.81-101, 2005.
- [11] C.J.J. Paredis, R. Sinha, P.K. Khosla. Composable Models for Simulation-based design. Engineering with Computers, Vol.2, pp.112-128, 2001.
- [12] C. Berger, M. Chaudron, R. Heldal, O. Landsiedel. Model-based, Composable simulation for the development of autonomous miniature vehicles. Simulation Series, Vol.4, pp.118-125, 2013.
- [13] Z. Zhu, Y.L. Lei, Z. Ning, Y.F. Zhu. Composable modeling frameworks for networked air and missile defense systems. Journal of National University of Defense Technology, Vol.5, pp.186-192, 2014.
- [14] X.X. Chen, D. Wang, H. Huang. Design of Simulation Testing Platform for CBTC System. Railway Computer Application, Vol.8, pp.50-56, 2011.
- [15] N.N. Chen, J.Xu, X.Z. Yin. Desing and Implementation of Eurobalise Simulation Test Platform for Urban Transit CBTC System. Railway Computer Application, Vol.12, pp.59-61, 2013.
- [16] R. Srinon, S. Ramakrishnan. Distributed Simulation Modeling for Manufacturing Systems Design Using XML. Proc. 18th International Conference on System Engineering, pp.395-400, 2005.

JPI UML Software Modeling

Aspect-Oriented Modeling for Modular Software

Cristian Vidal Silva

Escuela de Ingeniería Informática, Facultad de Ingeniería y
Administración, Universidad Bernardo O'Higgins
Santiago, Chile

Leopoldo López

Instituto de Investigación y Desarrollo Educacional, IIIDE
Universidad de Talca
Talca, Chile

Rodolfo Schmal

Escuela de Ingeniería Informática Empresarial
Facultad de Economía y Negocias
Universidad de Talca
Talca, Chile

Rodolfo Villarreal

Escuela de Ingeniería Informática, Facultad de Ingeniería
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Miguel Bustamante

Escuela de Ingeniería Comercial
Facultad de Economía y Negocios, Universidad de Talca,
Talca, Chile

Víctor Rea Sanchez

Facultad de Ciencias de la Ingeniería
Universidad Estatal de Milagro
Milagro, Ecuador

Abstract—Aspect-Oriented Programming AOP extends object-oriented programming OOP with aspects to modularize crosscutting behavior on classes by means of aspects to advise base code in the occurrence of join points according to pointcut rules definition. However, join points introduce dependencies between aspects and base code, a great issue to achieve an effective independent development of software modules. Join Point Interfaces JPI represent join points using interfaces between classes and aspect, thus these modules do not depend of each other. Nevertheless, since like AOP, JPI is a programming methodology; thus, for a complete aspect-oriented software development process, it is necessary to define JPI requirements and JPI modeling phases.

Towards previous goal, this article proposes JPI UML class and sequence diagrams for modeling JPI software solutions. A purpose of these diagrams is to facilitate understanding the structure and behavior of JPI programs. As an application example, this article applies the JPI UML diagrams proposal on a case study and analyzes the associated JPI code to prove their hegemony.

Keywords—JPI; UML; AOP; JPI UML Class Diagram; JPI UML Sequence Diagram

I. INTRODUCTION

Aspect-Oriented Programming, AOP [4] [5] [6] [8] is an extension of Object-Oriented Programming OOP that introduces aspects, i.e., modules that advise classes' behavior or add structural members to base classes. Aspects are intended to isolate and modularize crosscutting concerns in classes and methods of software components.

Even though AOP isolates crosscutting concerns, it also introduces implicit dependencies between advised classes and aspects. First, aspects define pointcut PC rules, which alter advised classes' behavior; base classes are completely

oblivious about changes to their behavior and structure during program execution. Second, changes in the signature of advised methods of target classes can produce ineffective or spurious aspects, i.e., occurrences of *the fragile pointcut problem* [1] [3]. Furthermore, [2] [3] [9] observe that dependencies between classes and aspects compromise independent development of base modules and aspects code. In classic AOP, developers of both, base code and aspects, need some knowledge about of all software modules, i.e., base classes and aspects that might advise them, rules and associated advice code.

For isolating crosscutting concerns and getting modular AOP programs without the mentioned implicit dependencies, [1] proposed the concept of Join Point Interface JPI as new AOP programming methodology. Like classic AOP [4] [5] [6], aspects in JPI isolate crosscutting functionalities; but, unlike classic AOP, JPI aspects do not provide PC rules. Instead, aspects in JPI implement defined join point interfaces. In addition, in JPI, advised classes define like PC rules for the join point interfaces exhibition.

Looking for a complete JPI software development process, this article proposes deploying two types of UML diagrams: class diagrams and sequence diagrams to model JPI programs, and presents a running example of a JPI program. Thus, the main goal of this article is to present diagrams to understand the structure and behavior of JPI programs and apply them to a case study to analyze their hegemony with the associated JPI code for a complete JPI software development process. Clearly, this is basic for the goal of reaching a model-driven JPI development methodology in the future.

This paper is organized as follows: Section II describes traditional UML class diagrams along with proposed extension to support JPI, JPI UML class diagrams. Section II

also describes the running JPI program example, and applies JPI UML class diagrams on it; Section III presents traditional UML sequence diagrams and their extension JPI UML sequence diagrams. Like Section II, Section III applies JPI sequence diagrams on the running example; Section IV presents, for the running example, a consistency analysis of JPI code and JPI UML diagrams; Section V describes related work; and, Section VI presents the conclusions and future work.

II. UML CLASS DIAGRAMS

A. Classic UML Class Diagrams

For object-oriented software modeling, UML class diagrams model the resources used to build and operate the system. Class diagrams model each resource in terms of its structure and relationships to other resources [7].

As an example, taking in account a *Shopping Session System SSS* that preserves a record of costumers, items in the stock, and transactions. The SSS also maintains information about each shopping session that a costumer initiates; a shopping session may include any number of transactions. Figure 1 shows an UML class diagram for the described structure of SSS in which classes include described attributes and methods.

In general, new requirements for SSS will demand changes in the entire system. For example, let us consider the following new system requirements:

- 1) *Frequent customer should receive a discount,*
- 2) *To log all transactions.*

For these requirements, a classic solution consists of adding new attributes and methods to either *ShoppingSession* or *Transaction* class. Hence, either the *buying(..)* method of class *ShoppingSession* or the constructor method of class *Transaction* would invoke new required methods; mentioned methods would include non-natural attributes and behavior not needed for their core purpose. Thus, these extensions represent clear examples of crosscutting concerns.

B. JPI UML Class Diagrams

This article follows ideas of [12] to propose and apply on the *ShoppingSession* system JPI class-based diagram to model JPI systems. The stereotype <<jpi>> labels join point interfaces which may not contain attributes or methods. In addition, a class linked to a JPI *exhibits* that join point interface and possibly defines a *pointcut* PC rule for that exhibition, i.e., a rule that defines a design policy through aspects and thus precludes any design violation at the join point events. In our proposal, aspects are represented as normal classes that define attributes and methods, and stereotyped by <<aspect>>. Since aspects implement join

point interfaces, they directly link to a join point interface class and define a kind of join point (before, around, and after) for the join point interface implementation.

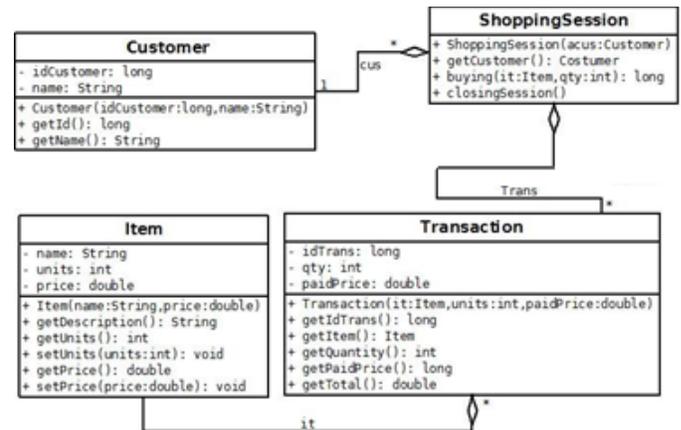


Fig. 1. UML class diagram of the system Shopping Session

Figure 2 shows a JPI UML class diagram for the *JPI SSS* version. Note that there is a join point interface *JPIPreBuying* to link the *ShoppingSession* class and *PreBuying* aspect. In these associations, *JPIPreBuying* defines a method exhibited by class *ShoppingSession* and implemented by aspect *PreBuying*; class *ShoppingSession* defines a PC rule for the *buying(..)* method execution. Furthermore, Figure 2 presents a join point interface *JPIDiscount* to link the *ShoppingSession* class and *Discount* aspect, as well as, a join point interface *JPILogging* to link the *ShoppingSession* class and *Logger* aspect. For the first mentioned association, *ShoppingSession* exhibits the method *JPIDiscount(price, ss)*, a method defined by *JPIDiscount* and implemented by the *Discount* aspect, and defines a PC rule for the *BuyTransaction* class invocation. In this case, *price* is an argument of the constructor whereas *ss* corresponds to the *ShoppingSession* instance that invokes for the execution of *BuyTransaction* class constructor. It is necessary to remark, each link from a class to a join point interface is stereotyped by the name <<exhibits>> to indicate the associated join point interface method and its arguments along with a PC rule to define the join points occurrence. Similarly, the *implements* signature labels links from aspects to join point interfaces. Thus, since JPI UML class diagrams only applies stereotypes for associations and JPI elements; therefore, usual UML tools seems able for JPI UML class modeling.

Next section presents details about a proposal for the behavior modeling of a JPI system by JPI sequence diagrams, and presents example models for scenarios of the *ShoppingSession* system, as well.

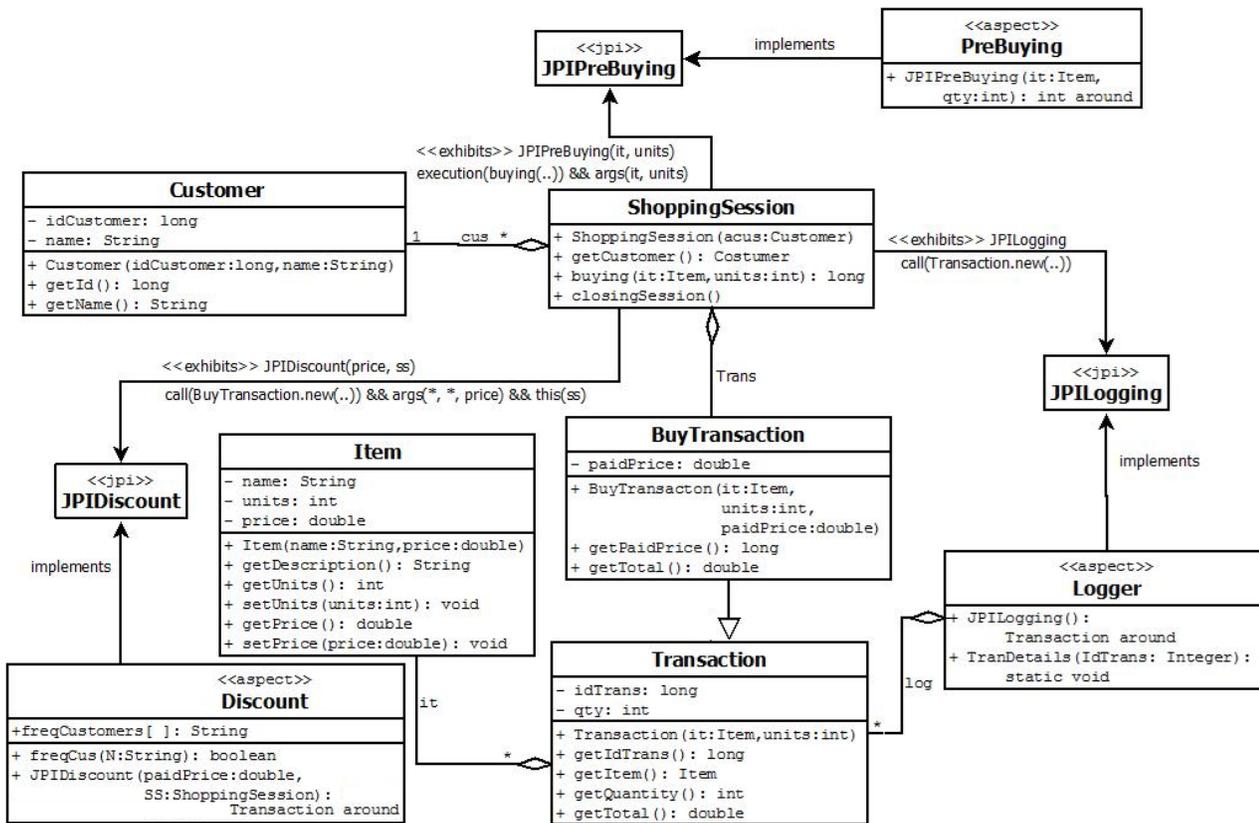


Fig. 2. JPI UML class diagram of updated version of system ShoppingSession

III. JPI UML SEQUENCE DIAGRAMS

UML sequence diagrams model execution scenarios of object-oriented programs [7]. Hence, to model the interaction between participants in the execution of JPI systems, this article proposes the so-called JPI sequence diagram which considers aspects as a sort of participants stereotyped by <<aspect>> in diagrams, and interactions between participant objects and aspects at join points are denoted by *advices*. Our modeling hypothesis is that by means of JPI sequence diagrams, the associated behavior of JPI programs for model-scenarios is deductible.

According to JPI notation, *join point interfaces* act as a bridge to let in the UML class instance catches up the result of the defined aspect's method [1] [2] [3]. Communication between aspects and classes instances is synchronous. Thus, when an instance of an aspect advises an object, i.e., it implements a *join point interface* for that class instance, the advised object, in order to continue its actions, waits for a message from the aspect to proceed.

AOP languages like AspectJ as well as JPI only permit *around advices* to explicitly proceed. Therefore, in AspectJ and JPI, *before* and *after advices* implicitly proceed associated to the advised classes' methods execution, i.e., *before* or *after* advices must execute and then the advised classes can continue their execution. Like for the *around advices* behavior, this proposal considers messages for an explicit activation from aspects to class instances to proceed. Given

these ideas, rules to model JPI program-behavior execution scenarios by means of JPI sequence diagrams correctly are:

- I. Object and aspects in any execution scenario are participants.
- II. As usual, objects participants communicate by synchronous and asynchronous messages represented by \rightarrow and \Rightarrow , respectively. A participant that sends a synchronous message waits for a return message, represented by \dashrightarrow , from the target object to continue its actions.
- III. A participant can create or delete an existent object. For objects creation, a box that represents an object participant is linked to the creation-sent message. An arrow like a return message represents a creation-sent message. Destruction messages, synchronous or asynchronous, imply that the affected object will definitely end its activities and a cross at the bottom of its lifeline after its destruction represents this situation.
- IV. When a participant receives a message, an activation gray line is created until it finishes its associated actions and returns.
- V. When a class B exhibits a join point interface with a *pointcut* PC, and a participant *a'* of class A sends a message *M* to a participant *b'* of class B asking for a method involved in the PC rule, there will be a synchronization point in *b'* lifeline, if PC rule holds. A JPI message denotes the JPI method name and the values of its arguments. These values are usually

conformant with the advised method signature, i.e., matching the number and types of parameters. For example, message 3.0 of Figure 3 shows a JPI message `<<around>> JPIPreBuying(it, qty = units)`, i.e., *it* and *qty* are arguments of the *JPIPreBuying* method call, and *it* takes the value of *it* from the source participant, in this case from *sp1*, and *qty* takes the value of *units* from that source as well. In addition, advice `<<around>>` stereotypes the JPI message.

To preserve the UML sequence diagrams semantic, three important rules are proposed and applied here:

- 1st, a JPI message from a participant object *O* to a participant aspect *A* always appears after the invocation message of the advised method.
- 2nd, for any synchronized message *M* from a participant *X1* to a participant *X2*, then *M* requires a return message from *X2* to *X1*.
- 3rd, for *around* advices, *proceed* calls are nested non-return calls.

In addition, for a JPI message from an object to an aspect, the message signature must be conformant to the JPI interface, which includes details for the advice execution by the aspect, i.e., kind of advice, parameters of the method in the JPI interface along with their values. Before performing any action, advised object waits for a *proceed* message from the aspect.

Proceed messages associated to *before* and *after* advices are like return messages in OOP-languages, whereas *around* advices cause that *proceed* messages behave like nested calls in an imperative language.

In general, *proceed* messages are more like the “pony express” in the Old West that delivers an important information (e.g. “*paidPrice with discount*” in the 1st *proceed* message of Figure 4) to the encamped (waiting) cavalry commander (the method) just before conducting the “correct” attack (method execution) to the enemy.

Following preceding mentioned rules, since a JPI message is a synchronized message, for the previous described sequence of Figure 3, the first aspect-participant sends *proceed* message 3.1 to the participant object *sp1*, which then can perform its actions. A *proceed* message indicates the preserved and updated values of arguments important for the advised method to execute. For example, message 3.1 of Figure 3 shows a *proceed* message, `proceed(it = new Item(“null product”, 0, 0), units = 0)`, for the participant *sp1*, i.e., a new item instantiates the argument *it* of advised method whereas *units* has the value of 0.

With these rules, it is possible to model behavior of JPI programs for particular scenarios. Since UML sequence diagrams allow modeling global scenarios and algorithmic behavior by means of combined fragments, thus this modeling proposal for JPI programs behavior would permit to understand JPI programs participants and their interactions for the reviewed scenarios, i.e., what a JPI program does, to obtain a semantics understanding about model JPI programs.

Figure 3 shows a JPI UML sequence diagram for the scenario in which a frequent customer wants to buy a product not sold by the ShoppingSession system: action 1 shows a *TestDriver* object that obtains *sp1*, an instance of *ShoppingSession*, for a frequent Customer *c1* = {2, ‘Cristian’} who wants to buy 15 units of the item *b1*, a not in stock item, action 2.0 represented by the message `buying(it = b1, units = 15)` from *TestDriver* to *sp1*. Next, action 3.0 represents the `<<around>>` advice `JPIPreBuying(it, qty = units)` activation for the *PreBuying* aspect, meanwhile the action 3.1 represents a *proceed* message from *PreBuying* aspect that changes values of arguments *it* and *units* of the *sp1*’s advised method, i.e., `it = new item(“null product”, 0, 0), units = 0`. Action 4.0 takes into account the `<<around>>` advice in message `JPIDiscount(paidPrice=it.getPrice(), ss=this)` from *sp1* to an instance of aspect *Discount*. By action 4.1, aspect *Discount* sends a message to one of its methods `freqCustomer(ss.getCustomer());` and action 4.2 represents a *proceed* message from *Discount* aspect to *sp1*. The result of this *proceed* message is to update *paidPrice* to `0.9*paidPrice` whereas the price of stocked item *ss* remains invariant. Action 5.0 follows an `<<around>>` advice without arguments from *sp1* to the aspect *Logger*. Action 5.1 is a *proceed* message from the aspect *Logger* to *sp1*. Action 6.0 is a message that creates a new *Transaction* instance *t1* with arguments *it*, *qty*, and *paidPrice*; and action 6.1, `return t1`, is a message from the created *Transaction* object to *sp1* that returns itself. After these actions, since *proceed* messages return in a LIFO order like nested procedure calls, and previous `<<around>>` messages have not returned yet, action 5.2 returns *t1* from *sp1* to the *Logger* aspect, and action 5.3, after transaction *t1* gets in the log, returns *t1* from *Logger* aspect to *sp1*. Likewise, action 4.2 returns *t1* from *sp1* to aspect *Discount*, and action 4.3 returns *t1* from aspect *Discount* to *sp1*. Since the latter is a “null item”, the stock of item *it* does not change (decremented by 0). Likewise, action 3.2 is a return message, in this case, the message returns `idTrans=t1.getIdTrans()`, from *sp1* to *PreBuying* aspect, and message 3.3 returns from *PreBuying* aspect to *sp1* again. Finally, message 2.1 returns *idTrans* from *sp1* to *TestDriver* instance which sends message 7.0 to method `TranDetails(idTrans)` of the *Logger* aspect. Message 7.1 gives back the execution control to instance of *TestDriver* and the execution scenario finishes.

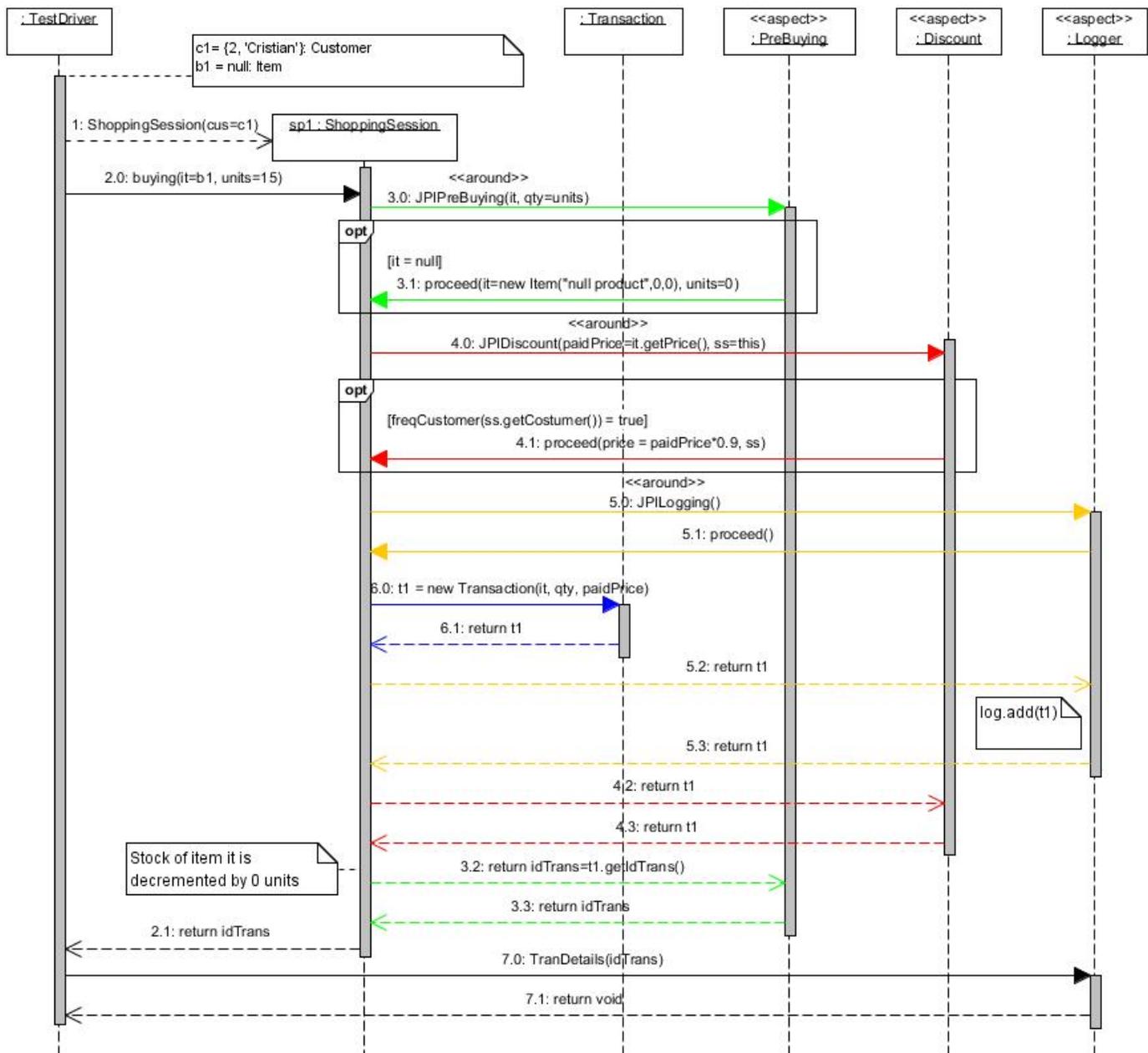


Fig. 3. JPI UML sequence diagram for a frequent customer buying a product not sold by the ShoppingSession system

Figures 4, 5, and 6 describe other execution scenarios for the ShoppingSession system, which semantic is similar to the Figure 3 described semantic.

Next section presents part of the *ShoppingSession* system code, to review and verify its functioning according to the described modeled scenarios.

IV. ANALYZING JPI CODE

Figure 7 presents the code of class *ShoppingSession* that exhibits the join point interfaces *JPIPreBuying*, *JPIDiscount*, and *JPILogger*: execution of *method buying(..)* exhibits *JPIPreBuying* whereas the call of constructor of class *Transaction* exhibits *JPIDiscount* and *JPILogger*. Figures 8, 9, and 10 show the aspects *PreBuying*, *Discount*, and *Logger*

to implement mentioned join point interfaces. In addition, Figure 11 presents the code for the join point interfaces definition. Clearly, this code solution structurally represents associations and components of JPI UML class diagram of Figure 2.

In the functioning logic analysis of the code class *ShoppingSession* and its exhibited aspects, i.e., aspect *PreBuying*, *Discount*, and *Logger*, a clear hegemony exists to the functioning logic of the sequence diagrams of Figures 3, 4, 5, and 6. When a frequent customer buys units of an item *it*, Figure 3, 4, and 5 show the behavior of class *ShoppingSession* and its exhibited aspects when the item *it* represents a *null item*, item *it* is an item in stock, and item *it* is an item without enough units in stock, respectively. Figure 3

shows a sequence diagram that includes a UML 2.0 *opt* combined fragment, between class *ShoppingSession* and aspect *PreBuying*, with a constraint for item *it*, when *it* is a *null item*. For this scenario, aspect *PreBuying* instantiates item *it* to an item named “*null item*”, and proceeds with the new value of *it* and *units = 0*, i.e., buying 0 units of item *it*. Figure 4 shows a sequence diagram that includes an *opt* combined fragment between class *ShoppingSession* and aspect *PreBuying*, with a constraint for item *it* that is fulfilled, i.e., item *it* != null and *it.getUnits()* >= *qty*, *qty* represents the *units*

argument in *PreBuying* aspect. For the latter scenario, aspect *PreBuying* proceeds with *it* and *qty* actual values without changes. Figure 5 shows a sequence diagram that includes an *opt* combined fragment, between class *ShoppingSession* and aspect *PreBuying*, with a constraint for item *it* that is not fulfilled in that scenario, though there are enough units in stock, i.e., item *it* is not a *null item*, but *it.getUnits()* < *qty*. For this scenario, Figure 5 shows that aspect *PreBuying* proceeds with item *it* and *units = 0*.

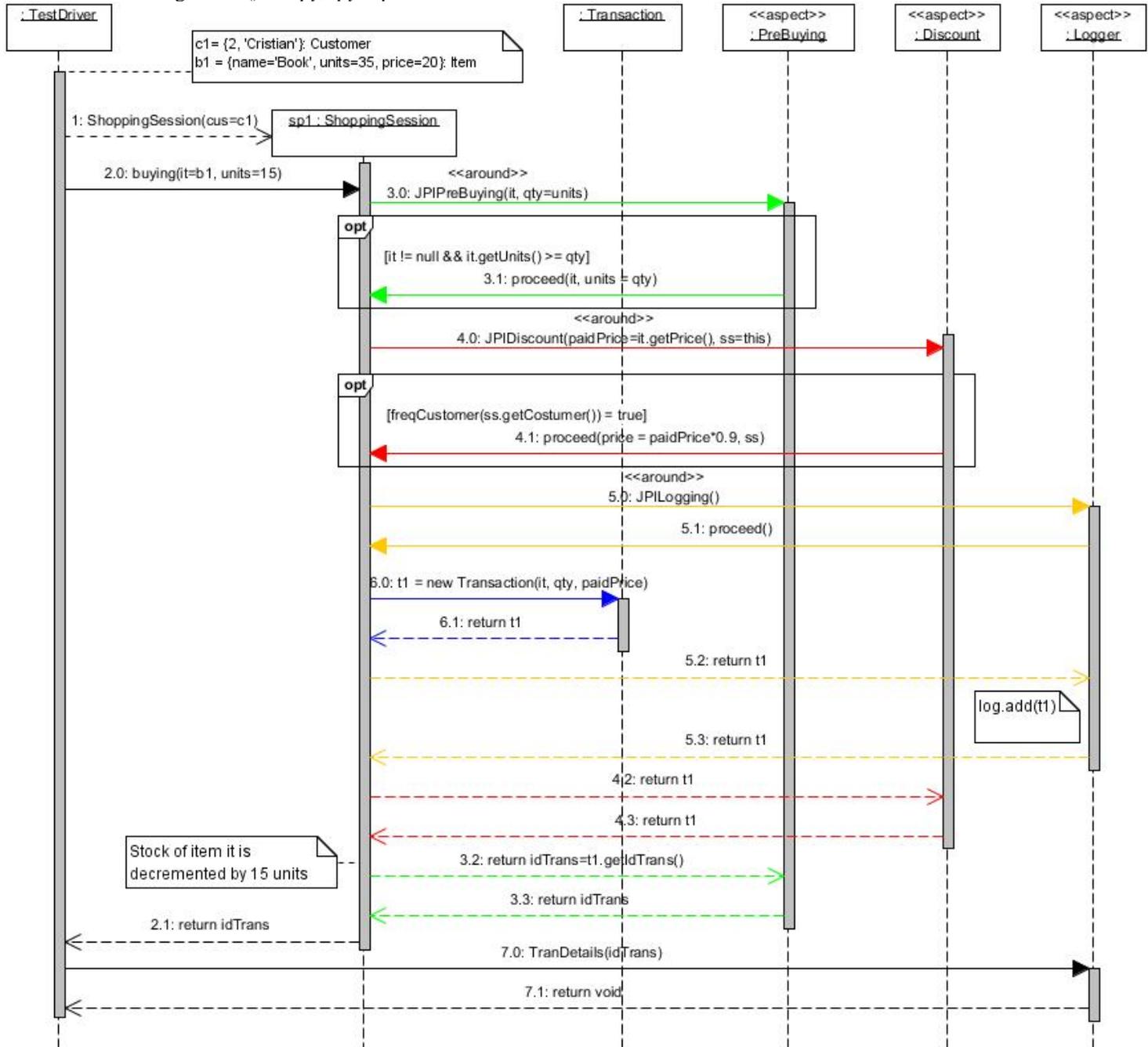


Fig. 4. JPI UML sequence diagram for a frequent customer buying a product in stock at the ShoppingSession system

When a non-frequent customer buys *qty* units of an item *it* in stock, such as Figure 6 shows, aspect PreBuying proceeds without changing *it* and *qty* values.

Concerning aspect Discount that receives *paidPrice*, i.e., the price for the new item, and *ss*, the equivalent instance of ShoppingSession, such as Figures 3, 4, and 5 show, for a frequent customer, the constraint of the second *opt* combined fragment of these figures is fulfilled, thus aspect Discount proceeds with $price = paidPrice * 0.9$, and preserves the *ss* value. Thus, there is always a discount for transactions performed by a frequent customer. However, for a non-

frequent customer, Figure 6 shows that a discount does not apply on the paid price.

Regarding aspect *Logger*, for the mentioned scenarios, this aspect logs final values for each transaction, i.e., log of values of transactions after being updated by aspects *PreBuying* and *Discount*. Since each of these behaviors is modeled by JPI UML sequence diagrams, and they are consistent to the code of classes and aspects, the behavior of aspects *PreBuying*, *Discount*, and *Logger* is consistent with the functioning logic of *ShoppingSession* system execution scenarios expressed in the JPI UML sequence diagrams of Figures 3, 4, 5, and 6.

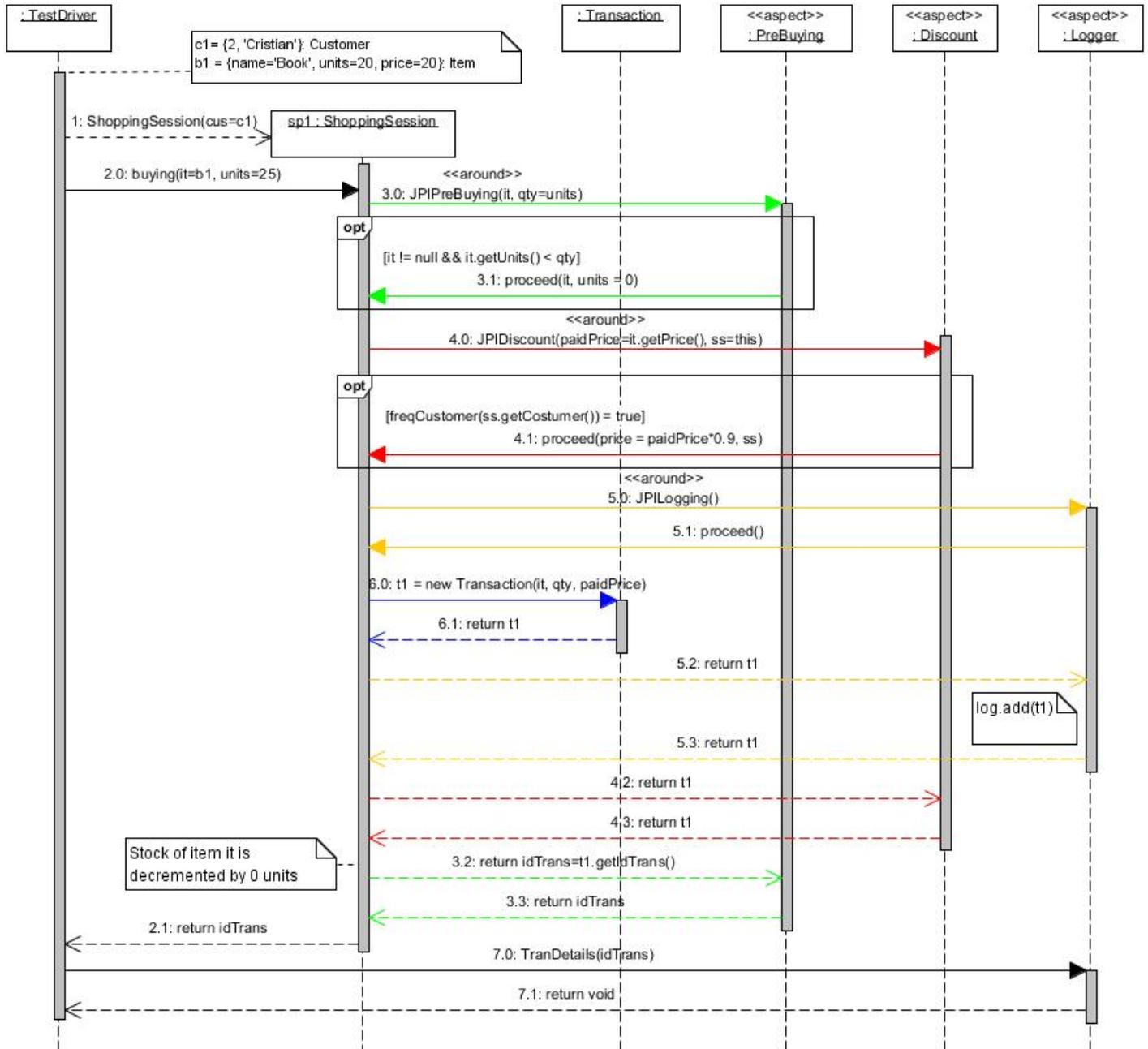


Fig. 5. JPI UML sequence diagram for a frequent customer buying a product without enough stock at the ShoppingSession system

V. RELATED WORK

As was mentioned, AOP represents a software development paradigm to modularize crosscut behavior [5]. For modeling traditional AOP solutions, [10] presents an UML use case application and extension of the formal language AspectZ for the aspect-oriented software requirements specification and analysis. Likewise, [11] describes an aspect-oriented UML class diagram-based and the OOAspectZ formal language for the software structure and requirements specification. In general, [14] surveys UML-based aspect-oriented design approach. Thus, mentioned research does not involve JPI ideas.

For JPI UML-based modeling, [13] presents an AspectZ extension, JPIAspectZ, for the formal modeling of JPI software requirements. Thus, given the JPI benefits for the

modular software production, this research is of a high value looking for a complete JPI software development process.

VI. CONCLUSIONS

JPI is a novel aspect-oriented programming methodology that permits solving classical issues of traditional aspect-oriented programming, i.e., implicit dependencies among classes and aspects. Nevertheless, as traditional aspect-oriented programming, elements such as JPI UML or JPI formal languages like JPI AspectZ [10] [11] [13] do not exist to perform a complete software development process inspired by JPI methodological practices. To partially solve these issues, this article has proposed and applied as well, JPI UML diagrams, i.e., JPI UML class diagrams and JPI UML sequence diagrams, respectively, for modeling structure and behavior of software applications developed using JPI to ease aspect-based programming.

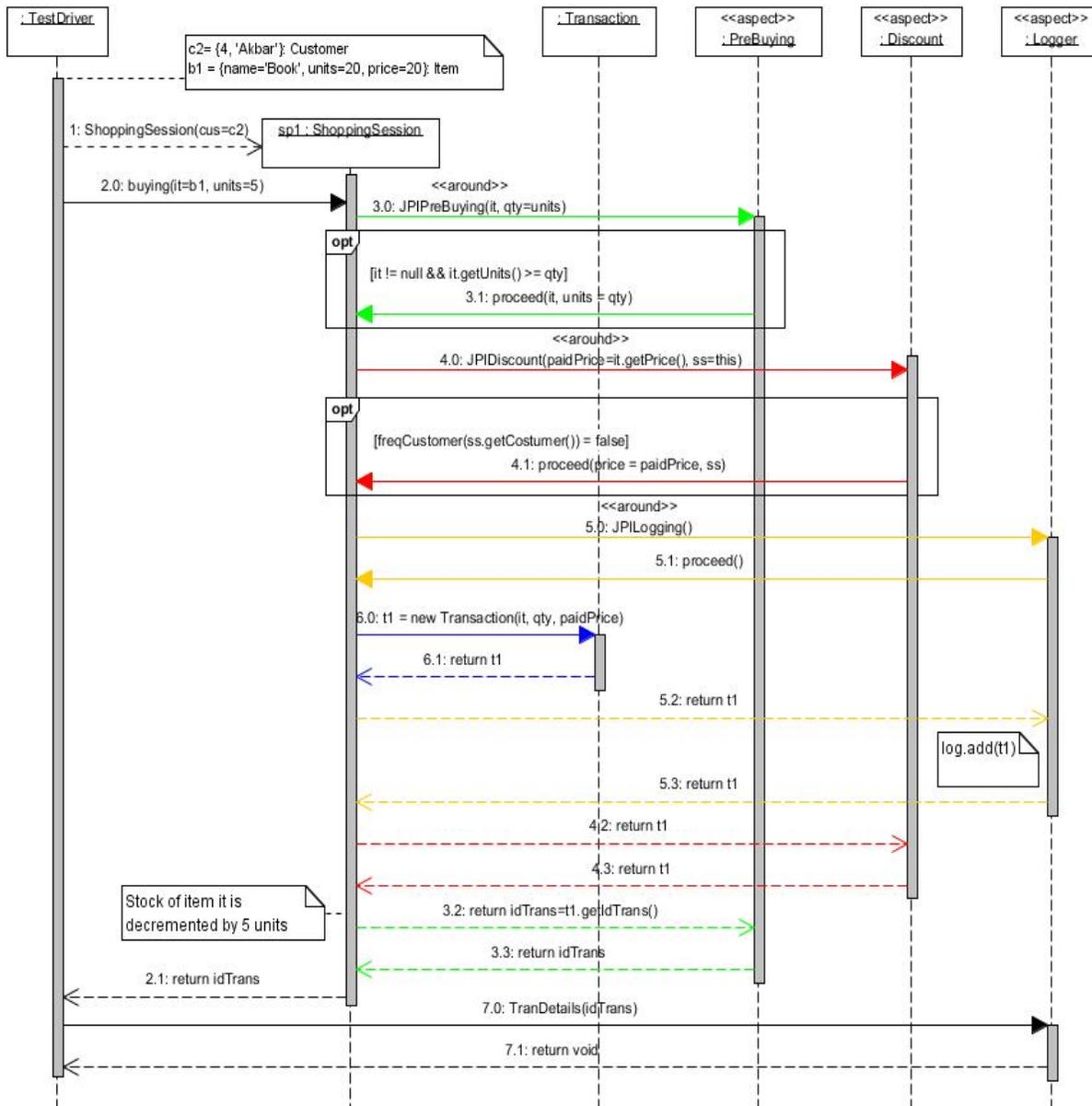


Fig. 6. JPI UML sequence diagram for a non-frequent customer buying a product in stock at the ShoppingSession system

JPI UML class diagrams allow capturing main modules of JPI programs, i.e., classes and join point interfaces, and associations between them. The presented JPI UML proposal clearly established associations among classes and join point interfaces, such as direction, stereotypes for different kind of advices, and *pointcut* rules. By mean of JPI class diagrams, one can know and understand existing relationships among classes and join point interfaces.

JPI UML sequence diagrams capture the functioning logic of modeled execution scenarios of a JPI program, and by means of these diagrams, we hypothesize that the functioning of a program can be deduced. Our proposal used *opt* combined fragments to zoom conditions and behavior for the functioning logic of aspects. After applying JPI UML sequence diagrams and analyzing the code of the modeled program for the *ShoppingSession* example, this article has shown consistency between models and the program code derived by means of our methodological approach. Clearly, using JPI UML diagrams, there is a functioning logic hegemony between modeled execution scenarios and code of main class and the exhibited aspects. This issue permits continuing researching to look for a full JPI software development process.

REFERENCES

- [1] E. Bodden, "Closure Joinpoints: Block Joinpoints without Surprises," *Proceedings of the Tenth International Conference on Aspect-Oriented Software Development, AOSD '11*. ACM, New York, NY, USA, pp. 117-128, March 2011.
- [2] E. Bodden, E. Tanter, M. Inostroza, "A Brief Tour of Join Point Interfaces," *Proceedings of the 12th Annual International Conference Companion on Aspect-Oriented Software Development, AOSD '13 Companion*. ACM, New York, NY, USA, pp. 19-22, March 2013.
- [3] E. Bodden, E. Tanter, M. Inostroza, "Join point interfaces for safe and flexible decoupling of aspects," *ACM Transactions on Software Engineering and Methodology*, ACM, New York, NY, USA, pp. 1-41, February 2014
- [4] B. Griswold, E. Hilsdale, J. Hugunin, W. Isberg, G. Kiczales, M. Kersten, "Aspect-Oriented Programming with AspectJ™," *AspectJ.org, Xerox PARC, 2001*. Tutorial slides online at <http://www.cse.msu.edu/sens/Software/aspectj/aspectj1.0.4/doc/tutorial.pdf> [Accessed: 25-Sep-2015]
- [5] G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. Lopes, J. M. Loingtier, J. Irwin, "Aspect oriented programming," *Proceeding of the European Conference on Object-Oriented Programming (ECOOP)*, Springer-Verlag LNCS 124, Finland, June 1997.
- [6] G. Kiczales, M. Mezini, "Aspect-Oriented Programming and Modular Reasoning," *Proceedings of the 27th International Conference on Software Engineering, ICSE '05*. ACM, New York, NY, USA, pp. 49-58, May 2005.
- [7] T. Pender, "UML Bible," *John Wiley & Sons, Inc.*, New York, NY, USA, 1 Edition, 2003.
- [8] L. Ramnivas, "AspectJ in Action: Practical Aspect-Oriented Programming," *Manning Publications Co.* Greenwich, CT, USA, 2003.
- [9] F. Steimann, "The Paradoxical Success of Aspect-oriented Programming," *Proceedings of the 21st Annual ACM SIGPLAN Conference on Object-oriented Programming Systems, Languages, and Applications, OOPSLA '06*, Portland, Oregon, USA, October 2006 .
- [10] C. Vidal, R. Saens, C. Del Rio, R. Villarroel, "Aspect-Oriented Modeling: Applying Aspect-Oriented UML Use Cases and Extending AspectZ," *Computing and Informatics Journal*, Bratislava, Slovak , pp. 573-593, 2013.
- [11] C. Vidal, R. Saens, C. Del Rio, R. Villarroel, "OOAspectZ and Aspect-Oriented UML Class Diagrams," *Ingeniería e Investigación Journal*, Medellín, Colombia, pp. 66-71, 2013.
- [12] C. Vidal, R. Villarroel, "JPI UML: JPI class and sequence diagrams for aspect-oriented JPI applications," *Proceedings of XXXIII International Conference of the Chilean Computer Society*, Talca, Chile, November 2014.
- [13] C. Vidal, R. Villarroel, C. Pereira, "JPIAspectZ: A formal specification language for aspect-oriented JPI applications," *Proceedings of XXXIII International Conference of the Chilean Computer Society*, Talca, Chile, November 2014.
- [14] M. Wimmer, A. Schauerhuber, G. Kappel, W. Retschitzegger, W. Schwinger, E. Kapsammer, "A survey ofn UML-based aspect-oriented design modeling," *Journal ACM Computing Surveys CSUR*, New York, NY, USA. vol.43, issue 4, pp. 1-28, 2011.

```
package classes;
import java.util.*; import joinpointinterfaces.*;

public class ShoppingSession {
    private HashMap<Integer, Transaction> ShoppingSessionTrans;
    private Customer cus;

    exhibits Integer JPIPreBuying(Item it, int qty): execution(Integer buying(..)) && args(it, qty);
    exhibits BuyTransaction JPIDiscount(double price, ShoppingSession ss): call(BuyTransaction.new(..)) &&
        args(*, *, price) && this(ss);
    exhibits BuyTransaction JPILoggingBuy(): call(BuyTransaction.new(..));

    public ShoppingSession(Customer acus){...}
    public Customer getCustomer(){ return cus;}
    public void closingSession(){ ...}

    public Integer buying(Item it, int units){
        BuyTransaction buyTrans; Integer key;
        buyTrans = new BuyTransaction(it, units, it.getPrice());
        key = buyTrans.getIdTrans();
        ShoppingSessionTrans.put(key, buyTrans);
        it.setUnits(it.getUnits()-units);

        return key;
    }
    ...
}
```

Fig. 7. Class ShoppingSession code of the ShoppingSession system

```
package aspects;
import classes.*; import joinpointinterfaces.*;
public aspect PreBuying{
    Integer around JPISPreBuying(Item it, int qty){
        if (it != null){
            if (qty <= it.getUnits())
                return proceed(it, qty);
            }
        else
            it = new Item("null product", 0, 0);
        return proceed(it, 0);
    }
}
```

Fig. 8. Aspect PreBuying of the ShoppingSession system

```
package aspects;
import classes.*; import joinpointinterfaces.*;
public aspect Discount {
    /*To establish aspects precedence*/
    declare precedence: Discount, Logger;
    final String freqCustomers[] = {"Laurie", "Cristian"};
    boolean frequentCostumer(String N){
        for(int i=0;i<freqCustomers.length; i++){
            if (freqCustomers[i].equals(N))
                return true;
            }
        return false;
    }
    BuyTransaction around JPIDiscount(double paidPrice, ShoppingSession ss){
        double factor = 1;
        if (frequentCostumer(ss.getCustomer().getName()))
            factor = 0.9; return proceed(paidPrice*factor, ss);
    }
}
```

Fig. 9. Aspect Discount of the ShoppingSession system

```
package aspects;
import java.util.*; import classes.*; import joinpointinterfaces.*;
public aspect Logger {
    private static HashMap<Integer, Transaction> log = new HashMap<Integer, Transaction>();
    BuyTransaction around JPILoggingBuy(){
        BuyTransaction BT = proceed();
        if (BT.getQuantity()==0)
            //Non-Successful Logging
        else
            //Successful Logging

        log.put(BT.getIdTrans(), BT);

        return BT;
    }
    public void ListLogger(){ ... //List of Transactions}
    public static void TranDetails(Integer idTrans){ ... // Details of Transaction idTrans}
}
```

Fig. 10. Aspect Logger of the ShoppingSession system

```
package joinpointinterfaces;
import classes.*;

jpi BuyTransaction JPIDiscount(double paidPrice, ShoppingSession SS);
jpi BuyTransaction JPILoggingBuy();
jpi Integer JPISPreBuying(Item it, int qty);
```

Fig. 11. JPI instances of the ShoppingSession system

Association Rule Hiding Techniques for Privacy Preserving Data Mining: A Study

Gayathiri P

Research Scholar,
Department of Computer Science,
Bharathiar University, Coimbatore-641 046
TamilNadu, India

Dr. B Poorna

Principal,
SSS Jain College for Women,
T.Nagar, Chennai,
TamilNadu, India

Abstract—Association rule mining is an efficient data mining technique that recognizes the frequent items and associative rule based on a market basket data analysis for large set of transactional databases. The probability of most frequent data item occurrence of the transactional data items are calculated to present the associative rule that represents the habits of buying products of the customers in demand. Identifying associative rules of a transactional database in data mining may expose the confidentiality and privacy of an organization and individual. Privacy Preserving Data Mining (PPDM) is a solution for privacy threats in data mining. This issue is solved using Association Rule Hiding (ARH) techniques in Privacy Preserving Data Mining (PPDM). This research work on Association Rule Hiding technique in data mining performs the generation of sensitive association rules by the way of hiding based on the transactional data items. The property of hiding rules not the data makes the sensitive rule hiding process is a minimal side effects and higher data utility technique.

Keywords—Association rule mining; transactional data; privacy preservation; Association Rule Hiding (ARH); Privacy Preserving Data Mining (PPDM)

I. INTRODUCTION

Data belongs to a person or an organization may have different sensitive levels. These data are made available only for authorized persons. So ensuring the protection of sensitive data by access restriction is not a complete method. This may affect the utility of the data mining result and with help of the knowledge the user may re-identify sensitive data items from non-sensitive data is known as Inference Problem. The privacy preserving data mining is to provide a solution for protecting sensitive information by developing a data mining techniques which could be applied on databases without affecting the accuracy of data mining result and without violating the privacy of individuals is the motivation for this research.

Data mining is the method of determining patterns in large data sets with artificial intelligence, machine learning, statistics and database systems. The aim of data mining process is to extract information from a huge volume of data set to have logical structural representation of the data item in the transactional database. It is utilized to mine significant and useful information or knowledge from large database. Protected or private information extracted by data mining methods leads to the risk of threats to privacy. Association

rule mining is a technique in data mining to recognize the regularities created in large volume of data. The method is cooperated by allowing third party to recognize and disclose hidden private information for an individual or organization.

Privacy-preserving data mining with association rule denotes the area of data mining that looks to preserve sensitive information from unnecessary or unlawful disclosure. Privacy information comprises personal or confidential information in business like social security numbers, home address, credit card numbers, credit ratings, purchasing behavior, medical records and best-selling services. The privacy preservation data mining requires guarantee for hiding of sensitive information in efficient manner. The association rule hiding technique protects the sensitive data indirectly under the scanner. Also it fails to hide data items which are not sensitive. It affects the privacy of rules and the utility of the data mining results.

This paper is organized as follows: Section 2 discusses survey with existing techniques of Association Rule Hiding (ARH) for Privacy Preserving Data Mining (PPDM), Section 3 shows the Association Rule Hiding (ARH) for Privacy Preserving Data Mining (PPDM), Section 4 identifies the possible comparison between them, Section 5 discusses about the limitations of the existing techniques and Section 6 concludes the paper, key areas of research is given for improving the selection of sensitive rules for enhancing the business transactions. It also preserves the association rules for maintaining the privacy in database.

II. LITERATURE SURVEY

Privacy-Preserving Data Mining of Association Rules from Outsourced Transaction Databases technique [3] is developed with an encryption scheme. Encryption/Decryption (E/D) model was used to change the client data before it is shipped to the server. But, the mined results are not intended for sharing and remain exposed. Attack able to identify the intricacies of the rule preservation and data item property supports are not true supports. To Secure Association Rules, Secure Multi-party Computation (SMC) algorithm [4] is introduced to hide the association rules in a horizontally distributed database. The combination of private item subsets is calculated using SMC algorithm. Though, secure protocol is not relying on the commutative encryption and transfer. The SMC algorithm fails to secure the transaction items.

A perturbation-based PPDM with Multilevel Trust (MLT-PPDM) [5] is developed to preserve the privacy of data and association rules at different levels. This method preserved multiple perturbed copies, data miner perform resistance to diversity attacks and reconstruct the original data more accurately. MLT-PPDM permits the data owners for designs perturbed copies of data for different trust levels. However, the data set does not re-anonymize after it is updated with insertions and deletions. Efficiently Hiding Sensitive Item set with Transaction Deletion Based on Genetic Algorithms [2] is planned to enhance the chosen transactions deleted, so minimizing the side effects in Privacy-preserving data mining (PPDM) technique. But, predefined item set and a missing item set are non-sensitive item set that affect the rules being disclosed.

A Hiding Sensitive Association Rules [1] with Limited Side Effects are designed for PPDM. Heuristic method is involved for raising the hidden sensitive rules quantity in hiding sensitive association rule. The side effects minimized are not taken for correlation among rules that wipes out the creativity of the association rule. Privacy preserving data mining attains data mining goals without showing the privacy information of the individuals to the public users. A novel Hiding-Missing-Artificial Utility (HMAU) algorithm [6] is designed for hiding the sensitive itemsets during the transaction deletion process. Privacy preserving data mining (PPDM) is presented to hide the sensitive information. HMAU algorithm reduces the side effects through transaction depletion and the transaction with minimal HMAU value removed from the database. But, the noise addition and data modification are the significant problems to hide the sensitive information in PPDM.

III. ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Privacy Preserving Data Mining (PPDM) is used to extract relevant knowledge from large amount of data and protects the sensitive information from the data miners simultaneously. Privacy preserving data mining is a hot spot in data mining. Privacy Preserving Data Mining (PPDM) solves the issues of designing accurate models about combined data without access to exact information in individual data record. Association Rule Hiding is a PPDM technique use with Association Rule Mining method in transactional database.

An itemset is a set of products and transaction maintains simultaneously for a given set of items. The support of an itemset I in a transaction database is the percentage of transactions having I in the whole database. An itemset is frequent when the support is higher than a minimum support threshold (MST).

For two itemsets X and Y where $X \cap Y = \emptyset$. The confidence of an association rule $X \rightarrow Y$ is the probability that number of times Y occurs given that X occurs is equal to $Sup_{X \cup Y}$ divided by Sup_X . When $X \rightarrow Y$ holds in the database if $X \cup Y$ is frequent and its confidence is higher than a minimum confidence threshold (MCT). This rule is called the strong association rule. Association rule mining is used to discover all strong rules in the database.

A. Association Rule Mining using Selection Technique for Sensitive Rules

Privacy-preserving data mining (PPDM) is designed to minimize the privacy threats. Privacy threats are decreased by sensitive information hiding process from databases. These types of data having the confidential information result in the privacy threats when the data gets misused. Heuristic methods are used to choose the suitable data for sanitization to hide the sensitive information. In hiding the sensitive information process, side effects of missing cost and artificial cost are created. The beneficial method is used to choose the hidden sensitive information based on the NP-hard problem in sanitization process.

An extensive work on privacy-preserving data mining (PPDM) is carried out in various contexts. A general characteristic of frameworks is the patterns mined from the data that are planned to distribute with parties other than the data owner. The significant difference between the work and issues is: both the fundamental data and the mined results are not planned for sharing and stays private to the data owner. A conservative frequency-based attack model [3] is used where the server recognizes the correct set of items in the owner's data. Furthermore, it recognizes the exact support for all items in the original data.

Patterns with read/write Support Encrypted TDB B*

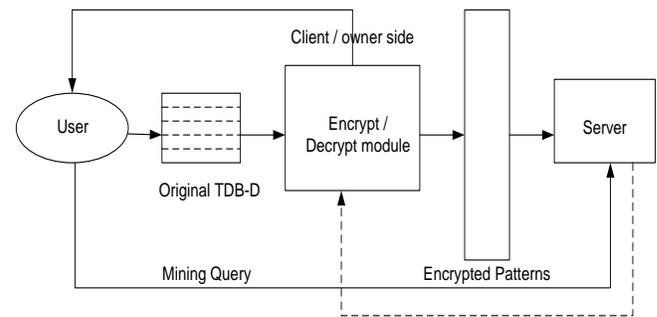


Fig. 1. Architecture of mining-as-service paradigm

The client/owner encrypts the data using encrypt/decrypt (E/D) module which is considered as a black box from its viewpoint. Encrypt/decrypt (E/D) module is used for converting the input data into an encrypted database. On the other hand, the server performs data mining operations and transmitted the patterns in the encrypted form to the owner of the data. The encryption scheme contains property where the revisited supports are not true supports. The E/D module regains the true individuality of the returned patterns and the true supports. An encryption scheme is named as RobFrugal. It is used to change the client data before it send to the server.

An alternative protocol is designed in [4] using simplicity, efficiency and privacy. Particularly, protocol fails in depending on commutative encryption and oblivious transfer. The solution is not completely secured. It gives large information to a small number of feasible combinations not same as protocol that discloses information.

B. Rule hiding for Privacy Preservation

The association rule hiding technique is to remove the sensitive rules from the transactional database during association rule mining. ARH technique protects sensitive data items by concealing the sensitive rules from miners and discloses all the non-sensitive rules to the miners. Data perturbation is used by Privacy Preserving Data Mining (PPDM) approach takes single-level trust on data miners. The technique establishes the ambiguity regarding individual values than the data released to the third parties for data mining purposes. In single trust level assumption, a data owner creates disturbed copy of its data with an amount of uncertainty. This assumption is restricted in many functions where a data owner trusts the data miners at various levels. An innovative dimension of Multi-Level Trust (MLT) [5] contains new demands for perturbation based PPDM. In contradiction to the single-level trust situation where only one perturbed copy is released and several perturbed copies of the similar data is presented for the data miners at various trusted levels.

The additional trust in data miner resulted in the less perturbed copy access. It also contains the access to the perturbed copies exist at lower trust levels. Additionally, data miners access multiple perturbed copies in forms. With diversity maintained across perturbed copies, the data miner on the other hand produced an exact reconstruction of the original data than permitted by the data owner. It is known as the diversity attack. It comprises the colluding attack situation where adversaries join their copies to increase an attack. It also incorporates the situation where an adversary uses public information to execute the attack by themselves. Preventing diversity attacks is the significant issue in solving the MLT-PPDM problem.

A compact prelarge GA-based (cpGA2DT) algorithm is designed in [2] to perform hiding operation of the sensitive itemsets while deleting transaction. The designed algorithm solves the issues of the evolutionary process by implementing both the compact GA-based (cGA) mechanism and the prelarge concept. A fitness function that was flexible in nature was structured using three adjustable weights to identify suitable transactions deleted to securitize the sensitive itemsets with minimal side effects of hiding failure, missing cost and artificial cost. A GA algorithm minimizes the memory needs by not taking the crossover and mutation operations but mimic the performances of traditional GAs.

C. Association Rule Hiding Techniques with Minimal Side Effects

The common technique of PPDM is to sanitize the database for hiding the information that is sensitive. A novel hiding-missing-artificial utility (HMAU) algorithm is designed in [6] to hide sensitive itemsets during transaction deletion. The transaction through the higher ratio of sensitive to non-sensitive one is chosen to delete.

In order to hide sensitive itemsets, three side effects were considered known as hiding failures, missing itemsets and artificial itemsets. Data sanitization is used to hide the sensitive knowledge from reveal in PPDM. To reduce the side effects, minimal distortion of the databases is required.

The transactions with any of the sensitive itemset are designed to locate the minimal HMAU values between transactions. The transaction with minimal HMAU value is directly taken away from the database. The process gets iterated till all sensitive itemsets are hidden. To avoid exposing hidden sensitive itemsets, the minimum count is modernized in the deletion process.

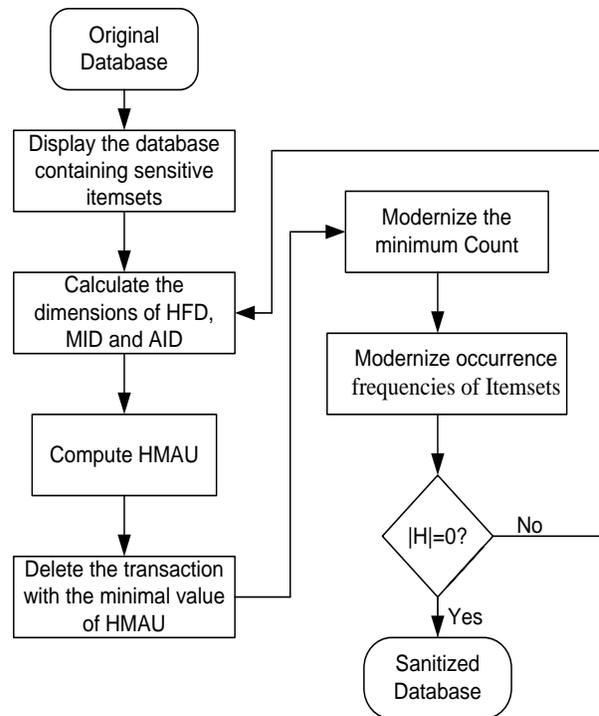


Fig. 2. HMAU Algorithm

A new heuristic method is designed in [1] that changes few transactions in the transaction database to reduce the supports or confidences of sensitive rules without any higher side effects. Connection between rules is not apparent to attain the goal. Heuristic methods are used for incrementing the hidden sensitive rules and minimize the number of modified entries. Rejected side effects are removed in the rule hiding process. The complete sensitive rules are hidden without unauthentic rules that are falsely created.

IV. COMPARISON OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION & SUGGESTIONS

In order to evaluate the privacy preservation using association rule hiding, number of data is taken to execute the experiment. Various parameters are used to calculate the privacy preserving in association rule hiding of the data mining techniques.

A. Privacy Preserving Level

Privacy preserving level is described as the level at which the data is privately transacted to the corresponding user without showing to the public users. It also increases the information delivery to the private users. It is measured in terms of percentage (%).

TABLE I. TABULATION FOR PRIVACY PRESERVING LEVEL OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Number of Data (Number)	Privacy Preserving Level (%)	
	PPMAR-OTD Techniques	SMC Algorithm
10	61	69
20	64	71
30	68	74
40	71	78
50	74	82
60	78	85
70	80	88

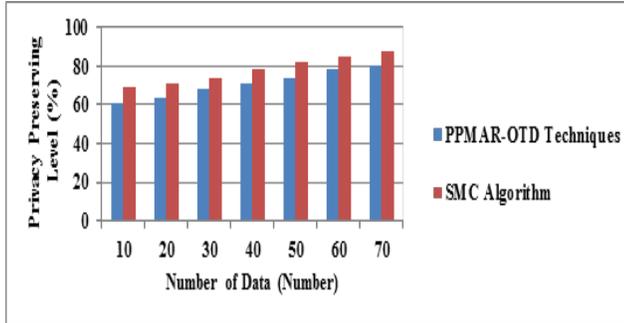


Fig. 3. Privacy Preserving Level of Association Rule Hiding Techniques for Privacy Preservation

Fig. 1 describes the privacy preserving level of association rule hiding techniques for privacy preservation. The privacy preserving level comparison takes place on existing Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases (PPMAR-OTD) technique and Secure Multi-party Computation (SMC) algorithm. The experiment shows that SMC Algorithm has 9.37% higher privacy preserving level than PPMAR-OTD technique.

B. Data Utility Rate

Data utility rate is defined as the amount of data utilized for privacy preserving using association rule hiding techniques. It is measured in terms of percentage (%).

$$\text{Data Utility Rate} = \frac{\text{Amount of data utilized for preserving privacy}}{\text{Total number of data}}$$

TABLE II. TABULATION FOR DATA UTILITY RATE OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Number of Data (Number)	Data Utility Rate (%)	
	PPMAR-OTD Techniques	SMC Algorithm
10	68	51
20	72	54
30	75	56
40	78	58
50	80	61
60	81	64
70	84	68

The data utility rate comparison takes place on existing Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases (PPMAR-OTD) technique and Secure Multi-party Computation (SMC) algorithm.

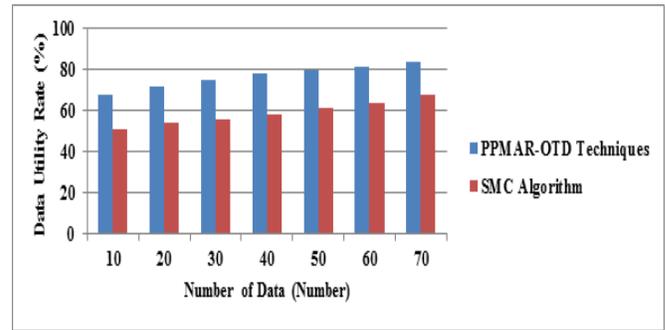


Fig. 4. Data Utility Rate of Association Rule Hiding Techniques for Privacy Preservation

Fig. 2 explains the data utility rate of association rule hiding techniques for privacy preservation. The experiment shows that PPMAR-OTD technique has 23.53% higher data utility rate than SMC Algorithm.

C. Efficiency (in terms of Side Effects)

Efficiency is defined as the number of data hidden without any side effects to the total number of data given. It is measured in terms of percentage.

$$\text{Efficiency (\%)} = \frac{\text{Data hidden without any side effects}}{\text{Total number of data given}}$$

TABLE III. TABULATION FOR EFFICIENCY OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Number of Data (Number)	Efficiency (%)	
	MLT-PPDM	cpGA2DT Algorithm
10	75	65
20	78	68
30	81	71
40	83	74
50	85	77
60	87	81
70	89	83



Fig. 5. Efficiency of Association Rule Hiding Technique for Privacy Preservation

Fig. 3 demonstrates the efficiency of association rule hiding techniques for privacy preservation. The efficiency comparison takes place on existing compact prelarge GA-based (cpGA2DT) Algorithm and Multi-Level Trust Privacy Preserving Data Mining (MLT-PPDM). The experiment shows that MLT-PPDM is 10.34% higher efficient than cpGA2DT Algorithm.

D. Execution Time

Execution time is defined as the time taken to hide the data with minimum side effects. Execution time is measured in terms of milliseconds (ms).

Fig. 4 describes the execution time of association rule hiding techniques for privacy preservation. The execution time comparison takes place on existing compact prelarge GA-based (cpGA2DT) Algorithm and Multi-Level Trust Privacy Preserving Data Mining (MLT-PPDM). The experiment shows that cpGA2DT Algorithm consumes 33.86% lesser time for execution than MLT-PPDM.

TABLE IV. TABULATION FOR EXECUTION TIME OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Number of Data (Number)	Execution Time (ms)	
	MLT-PPDM	cpGA2DT Algorithm
10	21	15
20	25	18
30	28	20
40	31	23
50	34	26
60	37	29
70	40	32



Fig. 6. Execution Time of Association Rule Hiding Technique for Privacy Preservation

E. Memory Requirement

Memory requirement is defined as the amount of memory space required for hiding the data using the association rule hiding techniques. It is measured in terms of mega bytes (MB).

TABLE V. TABULATION FOR MEMORY REQUIREMENT OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Number of Data (Number)	Memory Requirement (MB)	
	HMAU Algorithm	Heuristic Method
10	12	17
20	15	19
30	18	21
40	21	24
50	24	28
60	27	32
70	30	35

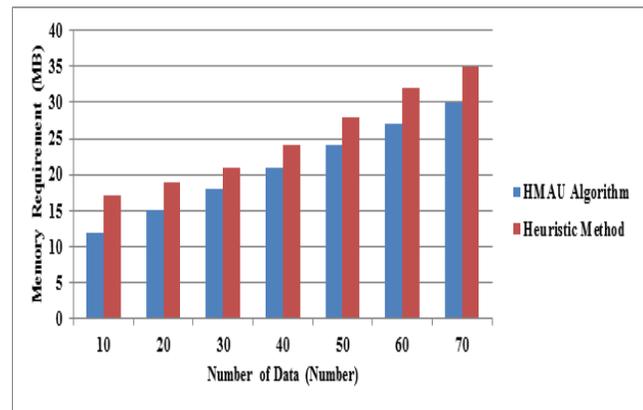


Fig. 7. Memory Requirement of Association Rule Hiding Technique for Privacy Preservation

Fig. 5 illustrates the memory requirement of association rule hiding techniques for privacy preservation. The memory requirement comparison takes place on existing Heuristic Method and Hiding-Missing-Artificial Utility (HMAU) algorithm. The experiment shows that HMAU Algorithm takes 21.59% lesser memory space than Heuristic Method.

F. Hiding Failure Rate (in terms of Side Effects)

Hiding failure rate is defined as the ratio of number of sensitive itemsets before sanitization to the number of sensitive itemsets after sanitization. It is measured in terms of percentage (%).

Fig. 6 shows the hiding failure rate of association rule hiding techniques for privacy preservation. The hiding failure rate comparison takes place on existing Heuristic Method and Hiding-Missing-Artificial Utility (HMAU) algorithm.

TABLE VI. TABULATION FOR HIDING FAILURE RATE OF ASSOCIATION RULE HIDING TECHNIQUES FOR PRIVACY PRESERVATION

Number of Data(Number)	Hiding Failure Rate (%)	
	HMAU Algorithm	Heuristic Method
10	25	18
20	28	21
30	31	24
40	35	27
50	37	30
60	39	33
70	41	36

The experiment shows that Heuristic Method has 26.63% lesser hiding failure rate than HMAU Algorithm.

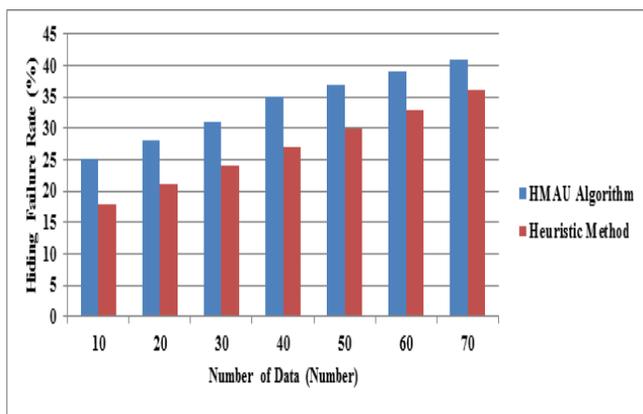


Fig. 8. Hiding Failure Rate of Association Rule Hiding Technique for Privacy Preservation

V. DISCUSSION ON LIMITATION OF PRIVACY PRESERVATION IN TRANSACTIONAL DATABASE

An encryption scheme designed with encryption/decryption module was used to transform client data before it send to the server. An attack model was designed to expose secret information and knowledge for privacy preserving mining. Usually, association rule mining task is performed in a shared privacy-preserving framework. In privacy preserving mining, the mined result is not aimed at sharing and stays as private. Attack model failed to know the details for encryption algorithms. Encryption scheme preserve the support item count values, attacker module can only work based on false support item counts.

Heuristic methods are used to improve the level of hidden sensitive rules quantity. The modified database is used to hide sensitive rules with limited side effects. Efficient mechanisms are needed to increase the speed of the rule hiding process for large databases. The association rules generated from the modified database have item sets does not appear in original transaction database. The side effect minimization fails to retain the correlation among rules on the modified transactional database. Secure Multi-party Computation (SMC) algorithm computes the union of private subsets.

Secure mining of association rules is located in distributed databases in horizontal manner. The leakage information delivers, the protocol of item sets exposed, were insecure. Secure protocol is not based on the commutative encryption and transfer.

MLT-PPDM introduces the flexibility dimension that permits the data owners to make perturbed copies of data for various trust levels. In MLT-PPDM, data miners have an ability to approach several perturbed copies. Multiple perturbed copies and data miners achieves diversity attacks to modernize the original data more correctly. The department fails to have more power in reconstructing the original data with many copies. The data set does not re-anonymized after it is modified with insertions and deletions. Less perturbed copies are not used by data miners at lower trust levels.

A. Related Works

Direct and Indirect Discrimination Prevention method [9] was designed to evaluate the discriminatory frequent item sets between original and modified transactional database during data mining process. Discrimination-free data models are produced from transformed data set without damaging data quality and mining based on single measure. However, discrimination fails to include any measure to remove redundant information. To provide with a minimum extension to the original database, Border based approach with hiding algorithm [11] was designed to present the sensitive knowledge hiding. Border approach provides globally optimal solution for sensitive frequent item set hiding. However, the border approach fails to change the original data set properly, lead to information leakage and redundant frequent item sets. The regenerated frequent patterns were not present in the initial data set.

Locality-Sensitive Hashing (LSH) based Blocking Approach with a Homomorphic Matching Technique [10] is designed for recognizing the candidate record pairs. The matching of pairs is designed using a basic protocol performing simple distance computations. Matching Technique is used for Privacy-Preserving Record Linkage. Though, it fails to create exact results because of the used anonymization format. Because of improper encoding, the initial distances fails to preserve. In order to obtain higher computational overhead, Local NN-search and Global data reorganization technique [8] is implemented for Sensitive Transactional Data. But, anonymization of personal data is not enough in various applications and the approach is not suitable because of the high dimensionality of the data.

An improved Gaussian Function based Perturbation Technique [7] was designed for preserving privacy of association rules and private data of individuals in an outsourced business transaction database. Gaussian Function based perturbation technique [7] preserved the privacy of association rules generated from the dataset and the sensitive frequent item sets. However, it is highly complex for distributed high volume dataset in cloud environment. A group incremental feature selection algorithm [12] was developed to locate the new feature subset in a short interval of time, when multiple objects are added to a decision table. Incremental feature selection algorithm is derived from

information entropy and it manages an effectual as well as well-organized mechanism. Though, the time complexity does not include the computational time of entropies.

B. Future Direction

The future direction of the privacy preservation using association rule hiding techniques needs to handle the confidentiality of sensitive rules in terms of better data utility and optimal side effects on the modified transactional databases. As each user may have different concern over privacy, user-oriented privacy preserving techniques can be developed. Parallel algorithms could be developed to prevent revealing of sensitive association between items and to improve the performance of the algorithm for large and dynamic datasets. Most of the proposed research works are concentrating on side effects and numbers of sensitive rules are hidden from sanitized database. Those are not clearly stated about number of rules are hidden in each iteration, number of levels in multi-level sensitive rule hiding, number of scan needed for the database, computational efficiency in terms of memory and CPU time. In future, these objectives are also being considered and new techniques are to be proposed for hiding the sensitive association rules in privacy preserving data mining.

VI. CONCLUSION

Based on the obtained nature of the survey, existing privacy preservation techniques in data mining using association rule hiding techniques has less privacy preserving level and also involves higher amount of side effects. At the same time, the utility of the data is also very low. As well, it takes higher execution time and so the efficiency gets decreased. The survey shows that while sending the data to the destination, the public user access the data and so the privacy is not maintained when it reaches to the destination. These types of issues decrease the effectiveness of the existing systems. The wide range of experiments on existing techniques calculates the relative performance of several privacy preserving techniques and its limitations. For this reason the new privacy preservation technique using association rules hiding techniques are planned to design. Finally from the result, the research work can be carried out in

privacy preservation using association rule hiding techniques to attain minimal side effects with higher data utility.

REFERENCES

- [1] Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society, "Hiding Sensitive Association Rules with Limited Side Effects", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007.
- [2] Chun-Wei Lin, Binbin Zhang, Kuo-Tung Yang and Tzung-Pei Hong, "Efficiently Hiding Sensitive Itemsets with Transaction Deletion Based on Genetic Algorithms", Hindawi Publishing Corporation, the Scientific World Journal, Volume 2014, Article ID 398269 September 2014.
- [3] FoscaGiannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases", IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 2013.
- [4] TamirTassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014.
- [5] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "Enabling Multilevel Trust in Privacy Preserving Data Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012.
- [6] Chun-Wei Lin, Tzung-Pei Hong and Hung-Chuan Hsu, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", Hindawi Publishing Corporation, the Scientific World Journal, Volume 2014, Article ID 235837 April 2014.
- [7] VineetRichhariya., and PrateekChourey., "A Robust Technique for Privacy Preservation of Outsourced Transaction Database" International Journal of Research in Engineering & Technology (IJRET), Vol. 2, Issue 6, Jun 2014, 51-58.
- [8] Gabriel Ghinita, Member, IEEE, PanosKalnis, and Yufei Tao, "Anonymous Publication of Sensitive Transactional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 2, FEBRUARY 2011.
- [9] Sara Hajian and Josep Domingo-Ferrer, Fellow, IEEE, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013.
- [10] DimitriosKarapiperis and Vassilios S. Verykios, Member, IEEE, "An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 4, APRIL 2015.
- [11] ArisGkoulalas-Divanis, Member, IEEE, and Vassilios S. Verykios, Member, IEEE, "Exact Knowledge Hiding through Database Extension", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009.

Improving Video Streams Summarization Using Synthetic Noisy Video Data

Nada Jasim Al-Musawi
Computer Science, University of Babylon
Baghdad, Iraq

Saad Talib Hasson
Computer Science, University of Babylon
Baghdad, Iraq

Abstract—For monitoring public domains, surveillance camera systems are used. Reviewing and processing any subsequences from large amount of raw video streams is time and space consuming. Many efficient approaches of video summarization were proposed to reduce the amount of irrelevant information. Most of these approaches do not take into consideration the illumination or lighting changes that cause noise in video sequences. In this work, video summarization algorithm for video streams has been proposed using Histogram of Oriented Gradient and Correlation coefficients techniques. This algorithm has been applied on the proposed multi-model dataset which is created by combining the original data and the dynamic synthetic data. This dynamic data is proposed using Random Number Generator function. Experiments on this dataset showed the effectiveness of the proposed algorithm compared with traditional dataset.

Keywords—Video summarization; Histogram of Oriented Gradient (HOG); Correlation coefficients (R); key frames; illumination changes; noise; Random Numbers Generator function

I. INTRODUCTION

Public places may contain many of stationary cameras (such as banks, transport, airports, etc.) for security requirements. This data has rich information; it should be analyzed in order to get the useful information. Processing and storing this huge data is very difficult. It is very important to summarize this data in order to facilitate many tasks such as data mining tasks. Summarization task is a basic key in data mining. Summarization techniques find a compact description of dataset transforming it to a smaller and suitable form for stream data analysis with the maximum information content [1], [2].

Video summary can take two forms: a static summary [3], [4], which is a set of selected key-frames, or a dynamic video which is a short video constructed by concatenating short video segments [5]. Several efficient video summarization approaches have been proposed for surveillance video stream such as [6]-[9]. For real-time, generally video summarization approaches utilized motion object detection and extraction as essential process to extract motion information from video sequence [10]-[12]. Many of summarization approaches generated a video synopsis or summary for a single video stream such as [13]-[15]. Summarization approaches for single camera do not give generalization to multiple cameras and they do not take into account the relationship between the different cameras. Thus, some recent approaches have been proposed to handle the problem of multiple stationary cameras to produce a video summary for related scenes.

Xu et al. (2015) in [16] developed a new video summarization framework using clustering techniques that produces a video summary for multiple videos observing the same scene by computing a shared activity among all scenes. Gygli et al. (2015) in [17] proposed a new dynamic video skimming in which a supervised approach was used in order to learn the useful global information of a summary. The result of the proposed method is an optimal video summary that maintain the diversity of the original video. Kuanar et al. (2015) proposed a video summarization approach using a graph theoretic method. The steps can be summarized: Shot boundary detection is achieved depending on Bag of Visual Words and the global feature such as color, texture and shape to remove the redundant frames. The video summary is constructed using Gaussian entropy algorithm [18]. Sigari et al. (2015) proposed a fast video summarization using an on-demand feature extraction and a fuzzy inference system. Based on an on-demand feature extraction, the input video is partitioning to highlight and analyze video content. Each highlight is assigned a score using a Fuzzy Inference System. The score value indicates the importance of the events occurred in the highlight [19]. Bian et al., (2015) proposed a video event summarization method which comprised three stages: (1) noise removal. (2) Discovering sub events from multiple data types. (3) Generating visualized summary from the microblog streams of multiple data types [38]. Fu et al. proposed a summarization technique using the spatio-temporal shot graph, then the shot graph is divided and clusters of event-centered shots with similar contents are constructed. The video summary is produced by solving a multi-objective optimization problem by shot importance evaluated using a Gaussian entropy fusion technique [39]. Zheng et al. (2015) proposed a novel surveillance videos summarization. The motion feature is extracted using graphics processing units GPU to reduce running time. Then, the result of this step is smoothed to reduce noise, and finally, the video summary is created by selecting frames with local maxima of motion information [40].

Most of these approaches for video summarization do not take into account temporal noise which occurs in scenes under illumination changes or light changes. Noise is a common problem in digital cameras due to some of errors may occur in one of two sensor cameras, or ambiguity in some of the sensor data that is exposed to noise. The video summarization that based on motion detection with noisy video produces wrong motion object vectors. For real world applications, the feature extraction in computer vision and image processing should be robust to brightness or illumination changes or to frame

distortion such as noise or blur. The illumination or brightness changes of some points between consecutive frames in video frames sequence often occur due to variations in parameters of different video cameras, or moving of objects from one part to another part of the scene can be changed with different illuminations [20], [21]. Presence of these issues will cause an inaccurate processing of the video stream. Most of the video summarization methods for a single static camera or multi-camera video do not take into account to illumination changes or to the existing noise signals which are occurred in some of video frames. They depended on the assumption that noise or illumination values are static along video frames.

In this research, multi-sensor video summarization algorithm has been proposed based on Histogram of Oriented Gradient (HOG) procedure which is used as feature extraction and robust similarity or dissimilarity measure which is Correlation Coefficients approach. Unlike some video summarization approaches in the literature, the proposed algorithm framework is not operated directly on the raw pixels. The algorithm uses the feature vector for each frame in video sequence in order to improve the motion detection accuracy under illumination variance and shadowing. Thus, (HOG) is selected for this purpose.

The availability of real or representative data is an important issue for evaluating data analysis algorithms. Because of some of real data are lack or difficult to obtain, synthetic data becomes alternative data. In many research areas such as data mining, image processing, computer vision, sensor networks, and artificial intelligence developed different synthetic data generation schemes for different applications [22], [23]. Corruption of data may come from noise or blur which sometimes comes from different atmospheric conditions. Many approaches used synthetic multi-temporal data generated by Gaussian noise in order to test and evaluate their proposed approaches such as [24]-[26]. These techniques produced the traditional Gaussian noise which is identically distributed noise. Meaning that, the noise values at all pixel locations in all sequenced data are generated from the probability density function with the static mean and the standard deviation values. In this work, the developed synthetic data generation has been proposed. The proposed synthetic data generation method generates sequenced frames using Random Number Generator (RNG) function in order to simulate the original video sequence containing variant noisy frames.

The paper rest is organized as follows: in Section II, background theories that are related to this work are presented. The proposed methodology is provided in Section III. The experimental results and performance assessment are provided in Section IV and V. Finally, Section VI draws conclusions of this work.

II. THEORY

A. Histogram of Oriented Gradient

Histogram of Oriented Gradient HOG is features extraction technique and the most successful used to extract low-level features for object detection and recognition. HOG, can be found in [27], and was originally designed for human

detection [28], [29]. It has low and computational time [30] and robust against shadow and illumination changes [21], [31]. The HOG algorithm for an image can be implemented by four main steps [33]: 1) gradient computation, 2) orientation binning 3) descriptor blocks, and 4) block normalization.

B. Correlation Coefficients

The *correlation coefficient* which is a measure of the calculating the score of relationship similarity between two variables x and y can be defined as [32]:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}} \quad (1)$$

Where x_i and y_i are gray level values of i -th pixel in the first and second frames respectively and \bar{x} and \bar{y} are the means of gray level values of x and y . The values of R always fall in $[-1, +1]$. When R is near to 0 meaning that the relationship between the two variables is little and when R is near to 1 the relationship is greater.

C. Noise Modeling

Real environments are often exposed to unexpected situations which are considered as noise. Gaussian noise is the most natural type of noise which is normally distributed. In MATLAB, noisy signal corrupted by Gaussian noise can be obtained by using the following:

$$g(x, y) = f(x, y) + \sigma * randn(\text{size}(f)) + \mu,$$

or

$$(x, y) = f(x, y) + \text{sqr}t(\sigma^2) * randn(\text{size}(f)) + \mu \quad (2)$$

Where $g(x, y)$ is the signal with additive Gaussian noise; $f(x, y)$ is the original signal; σ is the standard deviation; σ^2 is the variance; and μ is the mean. $randn()$ is MALAB function for generating random numbers with a Gaussian normal distribution. The probability density function for a Gaussian distribution with mean μ and variance σ^2 can be defined as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

D. Performance Evaluation Metrics

The performance evaluation of the proposed video summarization algorithm has been achieved using three types of metrics: (1) Data compression ratio (DCR) which is the ratio between number of frames in the original video and number of frames in the summary video [34], (2) Space savings ($S_{savings} = 1 - (\frac{1}{DCR})$) [35], and (3) Condensed Ratio (CR) which is the ratio between number of frames in the summary video and number of frames in the original video [36].

III. METHODOLOGY

A. The Proposed Method for Synthetic Data Generation

The standard deviation or the variance is the power of Gaussian noise signal. The classical Gaussian noise generator has the same approximated linear power for all given video frames. In dynamic environment, noise features are usually changed over time, it is necessary to use a simulation technique in order to adjust noise features. In this paper, the developed algorithm for generating synthetic video sequence

uses MATLAB random number generator function, the output of this function is added in (2) instead of the global σ values among frames. The new proposed equation can be defined as:

$$g(x, y) = f(x, y) + randn(d) * randn(size(f)) + \mu \quad (4)$$

Where d ($d=1$) for a single generated random number, $g(x, y)$ is noisy frame signal, $f(x, y)$ is the original frame.

Fig. 1 illustrates an example the visual comparison between the classical synthetic noisy data and the proposed synthetic noisy data.

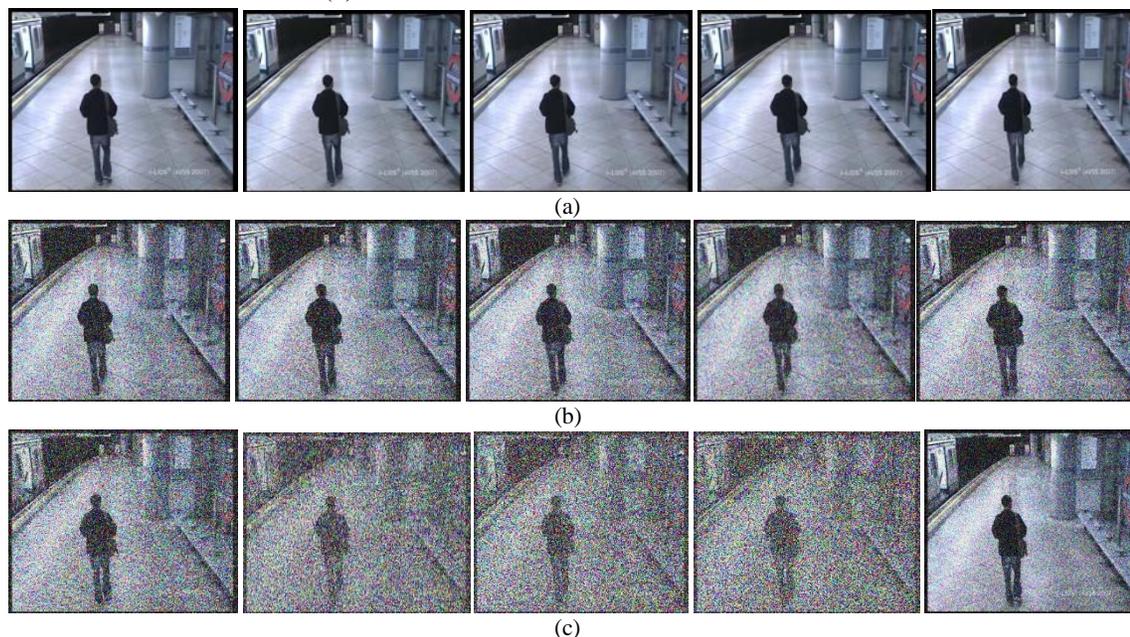


Fig. 1. Comparison between the classical synthetic noisy data and the proposed synthetic noisy data (a) The original video sequence, (b) The traditional noisy video sequence, (c) The proposed noisy video sequence

At each iteration, $randn(1)$ generates a new single random number which is considered as the value of standard deviation. The output of the algorithm is noisy video containing frames with non-linear or dynamic noise components. Mean Square Error (MSE) is used as an error measure between the original video signals and the noisy video signals. Table 1 and Fig. 2 show the comparison between Mean MSE's of the classical synthetic noisy data and MSE's of the proposed synthetic noisy data for ten sequenced frames.

TABLE I. COMPARISON BETWEEN MSE'S OF THE CLASSICAL SYNTHETIC NOISY DATA AND MSE'S OF THE PROPOSED SYNTHETIC NOISY DATA

MSE for traditional noisy sequenced frames	MSE for developed noisy sequenced frames
8	21.86
7.94	23.31
7.92	10.61
7.87	18.83
7.78	21.77
7.76	17.91
7.73	14.3
7.88	6.29
7.78	23.42
7.76	22.9

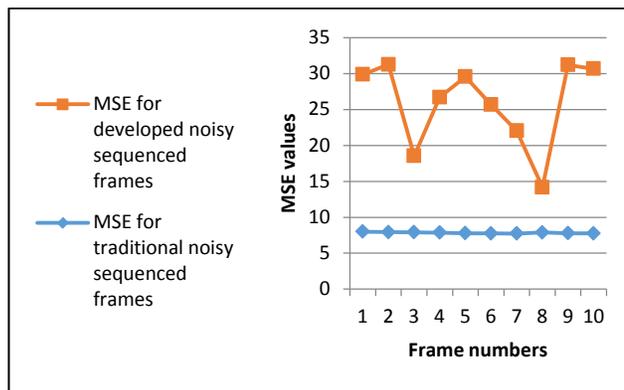


Fig. 2. Graph of comparison between MSE's of the classical synthetic noisy data and MSE's of the proposed synthetic noisy data

From Fig. 2, notice that the MSE's of the proposed synthetic video have nonlinear model, while MSE's of the proposed synthetic noisy video have approximately linear model. Multi-model dataset is constructed by using the original video dataset and the corresponding synthetic corrupted video dataset. In this research, this multi-model dataset is used for testing and evaluating the proposed summarization approach.

B. The Proposed Multi-Model Video Summarization

Multi-model video summarization using Histogram Oriented Gradient (HOG) algorithm for features extraction and Correlation Coefficient as a measure that quantifies the dependency (independency) between two video sequences. This algorithm has been applied on the proposed multi-model video dataset (for two videos). The videos have equal length (n frames). To increase efficiency of the proposed algorithm, HOG is used, which is a faster process and has low small features space. HOG procedure is computed for the current frame f_i ($i = 1..n$) from vid_1 (original video or reference video) and the current frame g_i ($i = 1..n$) from vid_2 (the corresponding noisy video of the original video). The results are two feature vectors. Correlation Coefficients (R) is computed between these feature vectors. Algorithm (1) illustrates the proposed HOG-Correlation Coefficients algorithm.

Algorithm (1) : HOG-Correlation based Summarization

Input: Vid_1 is the original video sequence

Vid_2 is the noisy video sequence

n is the number of video frames.

Output: video summary

Steps:

1. Given two video streams, Vid_1 and Vid_2 .
2. For $i = 1$ to n
3. Read the current frame (f_i) from the video stream (Vid_1).
4. Read the current frame (g_i) from the video stream (Vid_2).
5. Compute $H_1 = HOG(f_i)$ and $H_2 = HOG(g_i)$
6. Compute correlation coefficient $R = \text{correlation}(H_1 \text{ and } H_2)$
7. If $R \geq 0.9$ then
8. Store the current original frame into summary file.
9. else go to step 2 (disregard the current frame)
10. end

To detect redundancy (noisy) frames, we put a constraint on the correlation coefficients that are computed from the step 3. If $R \geq 0.9$ then the current frame from the original video sequence is stored into a summary file, otherwise go to step 2 (the current frames is dropped), and then continuously detect redundant frames for all frames.

IV. EXPERIMENTAL RESULTS

The original samples video datasets for metro and road scenes were selected from [37]. Using these videos, the synthetic data has been generated. The original and the synthetic video were used to construct the multi-model dataset for testing and evaluating the proposed summarization method.

The results analysis of the comparison between the classical synthetic video generation algorithm and the proposed synthetic video generation algorithm can be illustrated as follows:

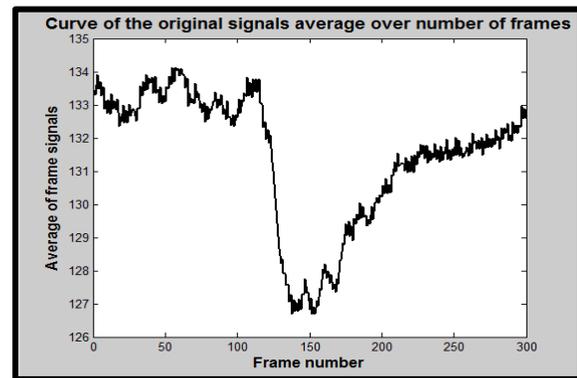
1) The random noise is in normal distribution case $N(0,1)$, $\text{mean}(\mu) = 0$ and $\text{variance}(\sigma^2) = 1$, the result contains high noise values (Mean Square Error (MSE) = $6.4784e+03$).

2) The random noise is in general distribution case $N(u, \sigma^2)$, the result contains very high noise values.

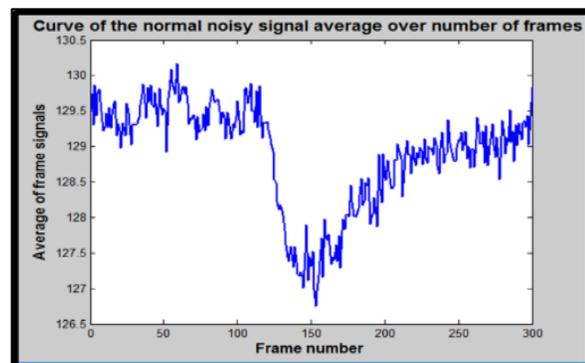
3) μ and σ^2 have values in the range $[0, 1]$, the result will contain lower noise values. The noise values close to zero when σ^2 is near or equal to zero and the result will become similar to the original signal ($\sigma^2 = 0.1$, $MSE = 923.0849$ and $\sigma^2 = 0.1$, $MSE = 321.7665$).

4) Applying the proposed algorithm on the input video sequence is achieved, the result contains high noise components, but these noise components are variant among all frames. Randomly, some of the frames have high values and others have low values depending on the random values of the variance

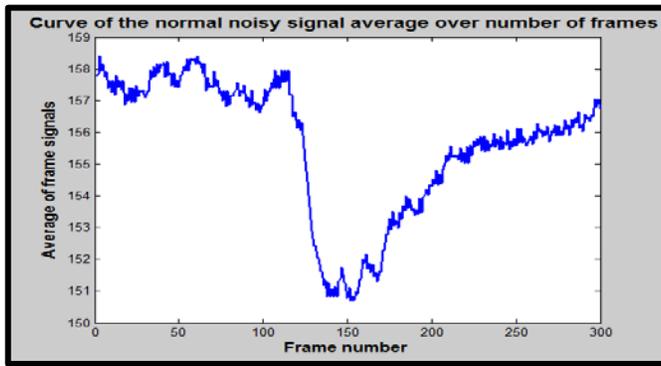
Fig. 3 and Fig. 4 show the curves of comparison between the original video (metro and road) signals and the corresponding synthetic noisy video with different noise levels. The x-axis represents the number of frames and the y-axis represent the average signals of each frame in the sequence.



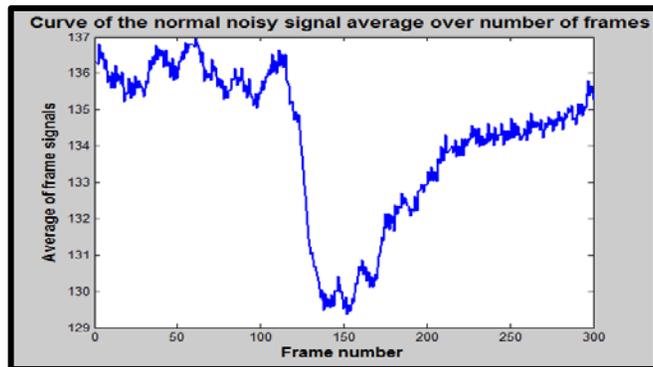
(a)



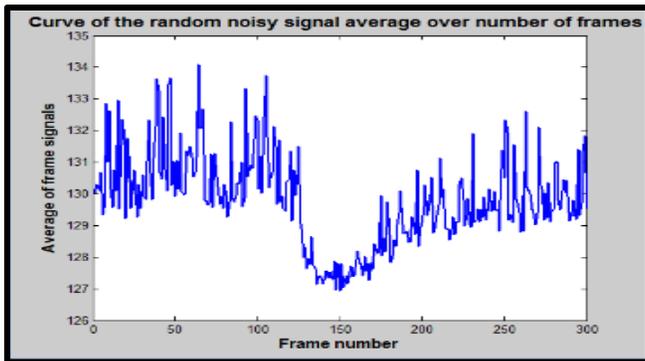
(b)



(c)

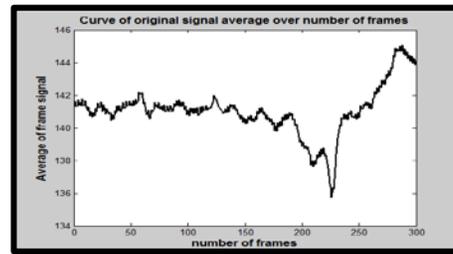


(d)

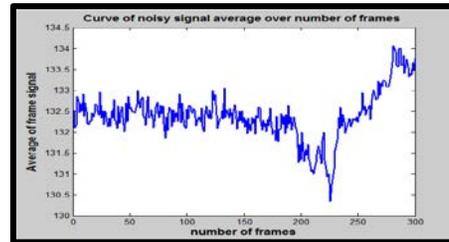


(e)

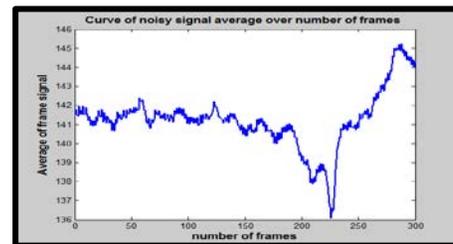
Fig. 3. Sequenced 300 frames from the metro video corrupted by Gaussian noise generated with $\mu = 0$ and different standard deviation values : (a) the original signals. (b) the noisy signals with $\sigma = 1$. (c) the noisy signals with $\sigma = 0.1$. (d) the noisy signals with $\sigma = 0.01$. (e) the noisy signals with $\sigma = randn(1)$



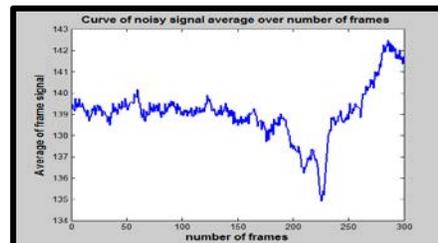
(a)



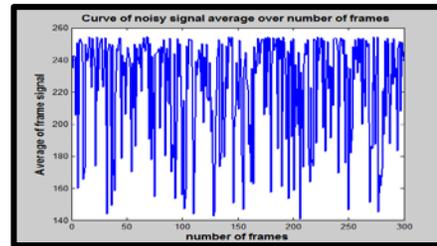
(b)



(c)



(d)



(e)

Fig. 4. Sequenced 300 frames from the road video corrupted by Gaussian noise generated with $\mu = 0$ and different standard deviation values : (a) the original signals. (b) the noisy signals with $\sigma = 1$. (c) the noisy signals with $\sigma = 0.1$. (d) the noisy signals with $\sigma = 0.01$. (e) the noisy signals with $\sigma = randn(1)$

In Fig. 3, and Fig. 4, when the noise values are high, the summarization of the video streams is not occurred because the correlation coefficients are smaller than the threshold ($R < 0.9$). When the noise values are small, all correlation coefficients will be greater than the threshold ($R > 0.9$). In this case, the summarization of the video stream is also not occurred and the output video has redundant frames like the redundant frames in the original video frames.

In order to generate a video sequence containing variant noisy pixels. The proposed algorithm for generating noisy video with dynamic and variant noise values has been applied to solve the above problem. The result of this algorithm is a noisy video which is used with the original (reference) video to form a multi-model dataset. The HOG-correlation based algorithm has been applied on this multi-model dataset. The correlation coefficients (R) have different values. the values that are greater than threshold ($R > 0.9$) are saved, otherwise are removed. The output is a video summary containing important information without redundant and noisy frames.

V. PERFORMANCE ASSESSMENT

The HOG-correlation algorithm for video stream sumrization have been tested on the video sequence (metro scene). The size of short sample of the original metro video is equal to 46.9 MB of 3001 frames. The Table 2 shows the information details before and after applying the HOG-correlation algorithm on the multi-model dataset (the original and the synthetic videos). As shown in Table2, V1 is the video output of HOG-correlation algorithm between the original metro video and the video generated using the traditional noise video generation algorithm, the size of V1 in the memory space is equal 44.9 MB because the original data which is 3D video is converted to 2D video. But the number of frames of V1 is still the same numbers of the original one.

V2 is the output of HOG-correlation algorithm between the original metro video and video generated using the proposed algorithm of dynamic noise generation with using random numbers. The size of V2 is equal to 1.09 MB and the number of frames is only 69 after applying the HOG-correlation based algorithm.

TABLE II. THE INFORMATION DETAILS (VIDEO SIZE AND NUMBER OF FRAMES) BEFORE AND AFTER APPLYING THE PROPOSED CORRELATION BASED METHOD ON THE ORIGINAL METRO VIDEO AND THE SYNTITHICS VIDEOS DATASETS

Metro video sequence	Video size befor	Video size after	Number of frames befor	Number of frames after
V1	101 MB	44.9 MB	3001	3001
V2	200 MB	1.09 MB	3001	69

From the experimental results, the HOG-correlation algorithm gives better results when it is applied on the original video and the video generated by the developed method of dynamic noise generation based on generated random numbers. As a result, it has ability to reduce the size of input video sequence and extract important motion information.

Depending on above information, the performance of HOG-correlation algorithm has been tested. Table3 ilustraites the performance evaluation of the proposed algorithm using

three metrics: Data compression ratio (DCR), Space savings (S_{saving}), and Condensed Ratio (CR).

TABLE III. THE RESULTS OF THE PERFORMANCE EVALUATION OF CORRELATION BASED ALGORITHM FOR METRO VIDEO SEQUENCE

Meto video sequence	DCR	S_{saving}	CR
V1	2.2494	55.54%	0%
V2	183.4862	99.45%	97.70%

As demonstrated in Table3 for the metro video sequence, DCR which is the ratio between the incompact size and compact size gives very good results for V2 against the value of DCR for V1. The S_{saving} , the compact in size relative to the incompact size, for V2 gives better results against the value of S_{saving} for V1. The CR, the ratio between the number of output frames and the number of input frames, also gives very good results (97.70%) for V2 against the value of CR for V1.

From the above performance evaluation results, the HOG-correlation algorithm works better on the noisy video that is ctreated using random numbers.

VI. CONCLUSION AND FUTURE WORK

In real time applications such as surveillance applications, illumination changes or shadowing for motion objects may occur in surveillance video stream. Many video summarization methods that are trying to construct video summary depended on the assumption that noise or illumination values are static along video frames. This leading to the existing video summarization algorithms will stumble in understanding of scene under observation. In this paper, the synthetic noisy video generation algorithm has been developed for testing the proposed video summarization algorithm based on Histogram Oriented Gradient and Correlation Coefficient for multi-model video dataset. The experimental results on the proposed dataset showed good results compared with the classical dataset. For more efficient time and space computation, online video summarization using Histogram of Oriented Gradient and Correlation coefficients techniques will be generated as a future work.

REFERENCES

- [1] Chandola, Varun, and Vipin Kumar. "Summarization-compressing data into an informative representation." Springer, Knowledge and Information Systems 12.3 (2007): 355-378.
- [2] Chandola, Varun, and Vipin Kumar. "Summarization-compressing data into an informative representation." *Knowledge and Information Systems* 12.3 (2007): 355-378.
- [3] Cayllahua-Cahuina, E. J. Y., G. Cámara-Chávez, and D. Menotti. "A static video summarization approach with automatic shot detection using color histograms." Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICCV), 2012.
- [4] Mahmoud, Karim M., Mohamed A. Ismail, and Nagia M. Ghanem. "Vscan: an enhanced video summarization using density-based spatial clustering." *Image Analysis and Processing-ICIAP 2013*. Springer Berlin Heidelberg, 2013. 733-742.
- [5] Li, Y., Meriardo, B., Rouvier, M., & Linares, G. "Static and dynamic video summaries." Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.

- [6] Li, Xiang-Wei, et al. Li, X. W., Zhang, M. X., Zhao, S. P., & Zhu, Y. L. "A Novel Dynamic Video Summarization Approach Based on Rough Sets in Compressed Domain." (2009).
- [7] Raikwar, Suresh Chandra, Charul Bhatnagar, and Anand Singh Jalal. "A Novel Framework for Efficient Extraction of Meaningful Key Frames from Surveillance Video." *International Journal of System Dynamics Applications (IJSDA)* 4.2 (2015): 56-73.
- [8] Zheng, R., Yao, C., Jin, H., Zhu, L., Zhang, Q., & Deng, W. "Parallel key frame extraction for surveillance video service in a smart city." *PLoS one* 10.8 (2015): e0135694.
- [9] Mei, Shaohui, et al. "Video summarization via minimum sparse reconstruction." *Pattern Recognition* 48.2 (2015): 522-533.
- [10] Damjanovic, U., Fernandez, V., Izquierdo, E., & Martinez, J. M "Event detection and clustering for surveillance video summarization." *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*. IEEE, 2008.
- [11] Ji, Zhong, Su, Y., Qian, R., & Ma, J. "Surveillance video summarization based on moving object detection and trajectory extraction." *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*. Vol. 2. IEEE, 2010.
- [12] Dammak, Majdi, and Adel M. Alimi. "Viewers Affective Feedback for Video Summarization." *Journal of information processing systems* 11.1 (2015): 76-94.
- [13] Peleg, Shmuel, Pritch, Y., Rav-Acha, A., & Gutman, A. "Method and system for video indexing and video synopsis." U.S. Patent No. 8,818,038. 26 Aug. 2014.
- [14] Potapov, Danila, Douze, M., Harchaoui, Z., & Schmid, C. "Category-specific video summarization." *Computer Vision-ECCV 2014*. Springer International Publishing, 2014. 540-555.
- [15] Pranjal, Pratyush, Lakhdive, R., Vasave, A., Varma, A., & Barve, A. "An Effective Method of Video Segmentation and Summarization for Surveillance." *International Journal of Computer Applications* 95.2 (2014).
- [16] Xu, Xun, Timothy Hospedales, and Shaogang Gong. "Discovery of Shared Semantic Spaces for Multi-Scene Video Query and Summarization." arXiv preprint arXiv:1507.07458 (2015).
- [17] Gygli, Michael, and Helmut Grabner, Luc Van Gool. "Video Summarization by Learning Submodular Mixtures of Objectives." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [18] Kuanar, Sanjay, Kunal Ranga, and Ananda Chowdhury. "Multi-View Video Summarization using Bipartite Matching Constrained Optimum-Path Forest Clustering." (2015).
- [19] Sigari, Mohamad-Hoseyn, Hamid Soltanian-Zadeh, and Hamid-Reza Pourreza. "Fast Highlight Detection and Scoring for Broadcast Soccer Video Summarization using On-Demand Feature Extraction and Fuzzy Inference." *International Journal of Computer Graphics* 6.1 (2015).
- [20] Tang, Feng, Lim, S. H., Chang, N. L., & Tao, H. "A novel feature descriptor invariant to complex brightness changes." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.
- [21] Rashwan, Hatem A., Mohamed, M. A., García, M. A., Mertsching, B., & Puig, D. "Illumination robust optical flow model based on histogram of oriented gradients." *Pattern recognition*. Springer Berlin Heidelberg, 2013. 354-363.
- [22] Gabrijel, Ivan, and Andrej Dobnikar. "On-line Inference of Finite Automata in Noisy Environments." Springer Vienna, 2005.
- [23] Pei, Yaling, and Osmar Zaiane. "A synthetic data generator for clustering and outlier analysis." Department of Computing science, University of Alberta, edmonton, AB, Canada (2006).
- [24] Rezatofighi, Seyed Hamid, et al. "A framework for generating realistic synthetic sequences of total internal reflection fluorescence microscopy images." *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013.
- [25] Su, Xin, et al. "Two-step multitemporal nonlocal means for synthetic aperture radar images." *Geoscience and Remote Sensing, IEEE Transactions on* 52.10 (2014): 6181-6196.
- [26] Rozantsev, Artem, Vincent Lepetit, and Pascal Fua. "On rendering synthetic images for training an object detector." *Computer Vision and Image Understanding* (2015).
- [27] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [28] Fleyeh, Hasan, and Janina Roch. "Benchmark Evaluation of HOG Descriptors as Features for Classification of Traffic Signs." (2013).
- [29] Ma, Yingdong, Xiankai Chen, and George Chen. "Pedestrian detection and tracking using HOG and oriented-LBP features." *Network and Parallel Computing*. Springer Berlin Heidelberg, 2011. 176-184.
- [30] Li, Changyan, Lijun Guo, and Yichen Hu. "A new method combining HOG and Kalman filter for video-based human detection and tracking." *Image and Signal Processing (CISP), 2010 3rd International Congress on*. Vol. 1. IEEE, 2010.
- [31] Pedersoli, M., González, J., Chakraborty, B., & Villanueva, J. J. "Enhancing real-time human detection based on histograms of oriented gradients." *Computer Recognition Systems 2*. Springer Berlin Heidelberg, 2007. 739-746.
- [32] Roy, Anuradha. "Estimating correlation coefficient between two variables with repeated observations using mixed effects model." *Biometrical journal* 48.2 (2006): 286-301.
- [33] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.
- [34] Davydov, Alexander Y. "Signal and Image Processing with Sinlets." *arXiv preprint arXiv:1206.0692* (2012).
- [35] Dutch, Mike. "Understanding data deduplication ratios." *SNIA Data Management Forum*. 2008.
- [36] Zhu, J., Feng, S., Yi, D., Liao, S., Lei, Z., & Li, S. Z. "High Performance Video Condensation System". *IEEE Transactions On Circuits And Systems For Video Technology*, Vol. 25, No. 7, 2015.
- [37] 2007 IEEE International Conference on Advanced Video and Signal based Surveillance, London (United Kingdom), 2007.
- [38] Bian, Jingwen, et al. "Multimedia Summarization for Social Events in Microblog Stream." *Multimedia, IEEE Transactions on* 17.2 (2015): 216-228.
- [39] Fu, Yanwei, et al. "Multi-view video summarization." *Multimedia, IEEE Transactions on* 12.7 (2010): 717-729.
- [40] Zheng, Ran, et al. "Parallel key frame extraction for surveillance video service in a smart city." *PLoS one* 10.8 (2015): e0135694.

A New Algorithm for Post-Processing Covering Arrays

Carlos Lara-Alvarez

CONACYT Research Fellow.
HCI-Lab, Center for Research in Mathematics (CIMAT).
Av. Universidad 222, 98068 Zacatecas, Mexico

Himer Avila-George

CONACYT Research Fellow - HARAMARA TIC-LAB,
CICESE Unidad Tepic (CICESE-UT³),
Andador 10 #109, 63173 Tepic, Nayarit, Mexico

Abstract—Software testing is a critical component of modern software development. For this reason, it has been one of the most active research topics for several years, resulting in many different algorithms, methodologies and tools. Combinatorial testing is one of the most important testing strategies. The test generation problem for combinatorial testing can be modeled as constructing a matrix which has certain properties, typically this matrix is a covering array. The construction of covering arrays with the fewest rows remains a challenging problem. This paper proposes a post-processing technique that repeatedly adjusts the covering array in an attempt to reduce its number of rows. In the experiment, 85 covering arrays, created by a state-of-the-art algorithm, were subject to the reduction process. The results report a reduction in the size of 28 covering arrays (~33%).

Keywords—Software testing; Combinatorial testing; Covering arrays; Post-Processing

I. INTRODUCTION

The ever increasing complexity, ubiquity, and dynamism of modern software systems demands new approaches to quality assurance. Extensive testing is required to assure that software works correctly, however, in many practical applications the number of configurable parameters may be large, and testing all possible configurations is not possible due to limited testing resources. Combinatorial testing enables the tester to execute a small set of test cases on the system, while achieving very high fault coverage. The pairwise test is one of main approaches in black-box testing. Several studies have demonstrated the effectiveness of *pairwise* testing [1], [2]. By examining fault reports for several systems [3] shown that ~100% of faults can be discovered with *4-wise* to *6-wise* interactions.

The first step to apply combinatorial testing is to construct a parametrized model of the *System Under Test* (SUT). The tester should first identify the input parameters related to the test goal, i.e. parameters affecting the system behavior; they may include but not limited to the following: (a) parameters of method calls; (b) parameters in system settings; and (c) a selection of replaceable system components installed in a test environment, such as hardware devices, system libraries and applications [4].

The key idea of combinatorial testing is that most of the SUT faults can be detected by combinations of a small number of factors. In combinatorial testing, a covering array (CA) is usually used as test suite, which covers parameter combinations involving t factors.

Covering Arrays (CA) are one of the most popular methods for representing pseudo-exhaustive test suites, they are small in comparison with an exhaustive approach but guarantee a level of interaction coverage among the parameters involved. They focus on having minimum cardinality (i.e. minimize the number of test cases), and maximum coverage (i.e. they guarantee to cover all combinations of certain size between the input parameters). To address this problem it has been proposed several methods (e.g., algebraic, exact, greedy and metaheuristic); however, usually they produce quasi-optimal covering arrays that contain combinations of symbols which are covered more than once (redundant). Redundancy opens the possibility for designing post-processing algorithms that eliminate the redundant information in the existing covering arrays with the aim of improve them.

This paper presents a new algorithm called *Post-Processing Covering Arrays* (PPCA) for eliminating redundant tests; it receives a covering array as input, then it tries to reduce the number of tests (rows).

The remainder of this paper is organized in four more sections. Section II, presents a brief overview of the principal techniques and tools for constructing covering arrays; Section III presents the new algorithm for post-processing covering arrays by deleting unnecessary tests. Section IV, shows the complete results for post-processing a benchmark composed by 85 covering arrays. Final remarks are presented in the section V.

II. RELATED WORK

There are several methods for constructing covering arrays; according to the strategy for generating covering arrays, they can be classified into algebraic, exact, greedy and metaheuristic approaches. Additionally, there are some useful operations that can be applied to a covering array previously constructed.

Algebraic approaches use formulas or operations with mathematical objects such as cyclic vectors [5], permutation vectors (Zero-sum method [6]), groups [7], cover starter [8] or covering arrays with small values of t , k , v (doubling [9] and v -plication [10]) Algebraic constructions often provide a better bound in less computational time, but impose serious restrictions on the system configurations to which they can be applied. For example, many approaches for constructing covering arrays require that the domain size be a prime number

or a power of a prime number; this significantly limits the applicability of algebraic approaches for testing.

Greedy approaches are more flexible than algebraic constructions. These methods can generate any covering array using as input t , k , and v . The majority of commercial and open source test data generating tools use greedy approaches for covering arrays construction (TVG [11], ACTS [12], Jenny [13] and *T tuples* tool [14]). The problem with these approaches are the quality of results –greedy methods rarely obtain optimal covering arrays–.

The *exact approaches* are exhaustive methods for the construction of optimal covering arrays. Despite of the fact that some approaches have techniques for accelerating the search process, in general they require an exponential time for completing the task, making them only practical for constructing small covering arrays. Some examples of this type of construction were reported in [15], [16], [17].

Metaheuristic approaches do not guarantee the construction of the optimal covering array but in practice they give good results in a reasonable amount of time. Among the most used metaheuristics are simulated annealing [18], tabu search [19], [20] and genetic algorithms [21].

III. METHODOLOGY

This section presents an algorithm for post-processing covering arrays; it starts with some basic definitions that introduce the problem, and then the proposed algorithm is described.

A. Definitions and Preliminaries

Definition 1: Let N , t , k , and v be positive integers where $t \leq k$. A covering array $CA(N; t, k, v)$ is a matrix of size $N \times k$ and strength t where each column has entries from alphabet Σ of size v . In every $N \times t$ subarray, all possible v^t t -tuples of symbols occurs at least once. Then N is the number of rows, t is the strength of the coverage of interactions, k is the number of factors (also called the degree), and v is the number of symbols for each factor (also called the order).

Definition 2: A t -way interaction is the assignment of specific values to each factor from set of t factors. The array is ‘covering’ in the sense that every t -way interaction is represented by at least one experimental run. In any covering array, the number of $N \times t$ subarrays is $M = \binom{k}{t}$, and the number of t -way interactions to be covered is $\binom{k}{t} v^t$.

Definition 3: The covering array number $CAN(t, k, v)$ is the smallest N for which a $CA(N; t, k, v)$ exists. The CAN is defined according to

$$CAN(t, k, v) = \min\{N : \exists CA(N; t, k, v)\};$$

evidently $CAN(t, k, v) \geq v^t$.

When a covering array is used as test suite:

- Each column represents a parameter of the software under testing (SUT).
- The symbols in the column specify the values for such parameter.
- Each row represents a test case to be performed.

(a)	(b)
$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 2 & 1 & 0 \\ * & * & * & * \end{pmatrix}$

Fig. 1: Detection of a redundant row (a) a covering array $CA(10; 2, 4, 3)$; (b) the last row (made by asterisks) is not required in $CA(9; 2, 4, 3)$.

- The fundamental problem is to determine $CAN(t, k, v)$.

When a covering array is constructed (see section II), it can contain t -way interactions which are covered more than once (in the definition of a covering array, the indication at *least once* means that a combination of symbols can be covered more than once). This fact opens the possibility that some symbols in certain positions are redundant and can be changed for any value without affecting the coverage of a CA, these symbols are referred to as *redundant*. To illustrate the existence of redundant rows, consider the example provided in Fig. 1. If the last row is deleted from the $CA(10; 2, 4, 3)$ shown in Fig. 1(a) then the matrix shown in Fig. 1(b) is obtained which is still a covering array because all 2-combinations of symbols are present. Hence, the last row is *redundant* and can be deleted from the original matrix; then $CA(9; 2, 4, 3)$ is better than the original one.

B. Proposed approach

Let \mathcal{R} be the set of possible realizations (t -tuples of Σ), and $\mathcal{I} = (\mathcal{I}_j)_{j=1}^M$ be the vector of interactions (t -tuples of columns). The i -th row test r_i can be represented by a vector S_i of the form

$$S_i = (s_{ij})_{j=1}^M, \quad (1)$$

where the t -way interaction $s_{ij} = (\mathcal{I}_j, v_{ij})$ associates the interaction \mathcal{I}_j to its realization $v_{ij} \in \mathcal{R}$ in the i -th test.

In the example shown in Fig. 2, the set of possible realizations are $\mathcal{R} = \{00, 01, 10, 11\}$ and the interactions are $(\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_2) = (c_0c_1, c_0c_2, c_1c_2)$ and the row test r_0 is represented by $S_0 = ((c_0c_1, 10), (c_0c_2, 10), (c_1c_2, 00))$, the row test r_1 by $S_1 = ((c_0c_1, 10), (c_0c_2, 11), (c_1c_2, 01))$, and so on.

Elements of S_i can be used for building an index \mathcal{M} that maps t -way interactions to lists of row tests that cover each interaction. That is,

$$\mathcal{M} = ((e_o, \mathcal{L}(e_o))_{o=1}^N \quad (2)$$

where $e_o \in \mathcal{I} \times \mathcal{R}$, and $\mathcal{L}(e_o)$ is the list of rows that test e_o .

For obtaining the reduced covering array, the lists of rows in \mathcal{M} are iteratively modified by removing elements. Given a

map \mathcal{M} , the vector of cardinality $S_i^\#$ of row test r_i is defined as

$$S_i^\# = (f(s_{ij}))_{j=1}^M, \quad (3)$$

where

$$f(s_{ij}) = \begin{cases} \#\mathcal{L}(s_{ij}) & \text{if } s_{ij} \in \text{keys}(\mathcal{M}) \\ N + 1 & \text{otherwise,} \end{cases} \quad (4)$$

and $\#\mathcal{L}(s_{ij})$ is the size of the list $\mathcal{L}(s_{ij})$.

For deciding which rows are included in the reduced covering array, the vector $S_i^\#$ is sorted in ascending order

$$S_i^s = \text{sort}(S_i^\#). \quad (5)$$

This array is an indicator of how a row test is required; when the first element of S_i^s is one it means that the row is strictly required. If all elements are set to $N + 1$, then the i -th row test is unnecessary. Hence, the order of elements in S_i^s is important; then, for obtaining a reduced covering array vectors S_i^s of all the rows are compared; the first row in the lexicographic order –i.e., first unequal elements determine the order– is selected as the best row test in each iteration; the selected row test is included in the reduced covering array

C. Algorithm

Procedure $\text{PPCA}(C)$ shown in the algorithm 1 illustrates the proposed approach. The input C is a covering array of size $N \times k$ and the algorithm produces a reduced covering array C' . The set L is used to store the row tests of C that are included in C' , initially L is set to empty. At line 3, the algorithm creates a map \mathcal{M} by analyzing each row test as stated in (2). After that, the algorithm iterates the following steps while the map \mathcal{M} has entries, i.e. $\text{keys}(\mathcal{M}) \neq \emptyset$: (a) Select the index i_m such that $S_{i_m}^\#$ is the smaller according to the lexicographic order (step 5), (b) Remove entries of \mathcal{M} that include i_m (step 6), and (c) Add i_m to the set L (step 7). Finally, C' is obtained by selecting rows L from C (step 9). Hence, the number of rows of the resulting covering array, $N' \leq N$, is equal to the size of L .

Algorithm 1 A Post-Processing covering array algorithm (PPCA).

Require: A covering array, C , of size $N \times k$
Ensure: A reduced covering array, C' , of size $N' \times k$ with $N' \leq N$

- 1: **procedure** $\text{PPCA}(C)$
- 2: $L \leftarrow \emptyset$
- 3: $\mathcal{M} \leftarrow \text{CREATEMAP}(C)$
- 4: **while** $\text{keys}(\mathcal{M}) \neq \emptyset$ **do**
- 5: $i_m \leftarrow \underset{i \in \{1 \dots N\}}{\text{argmin}} S_i^s$
- 6: $\text{keys}(\mathcal{M}) \leftarrow \text{keys}(\mathcal{M}) \setminus S_{i_m}$
- 7: $L \leftarrow L \cup \{i_m\}$
- 8: **end while**
- 9: $C' \leftarrow \text{Select rows } L \text{ from } C$
- 10: **return** C'
- 11: **end procedure**

D. Example

The toy example shown in Fig. 2 is used for clarifying algorithm 1, the matrix C of size 9×3 illustrates the test cases; but some of them are redundant. For obtaining the reduced covering array C' the PPCA algorithm proceeds as is illustrated in Figure 3 and described in the following:

INITIALIZATION:

After creating the initial map \mathcal{M} from C , and obtaining $S_i^s | i = 0, \dots, 8$; the row $i = 2$ is selected for the first iteration because it is the smaller according to the lexicographic order; i.e. $S_2^s[1, 2, 2]$ is selected because row 2 is strictly required for covering $(c_0c_2, 01)$.

ITERATION 1:

After inserting row 2 into the reduced covering array, and updating the vectors S_i^s for $i = \{1, 6\}$; the row $i = 4$ is selected for the next iteration.

ITERATIONS 2,3:

Row s 3 and 0 were included in C' . Note that if one of the rows 1, 6 or 8 were selected in iteration 3 (by a function other than the proposed), the covering array C' must include one or more additional rows for the complete covering. But, by using (5) the optimal solution can be found because row 0 completes the covering array.

ITERATION 4:

The algorithm finishes because $\text{keys}(\mathcal{M}) = \emptyset$ and the resulting selection $L = \{2, 4, 3, 0\}$ are the row tests included in the reduced covering array.

It is easy to show that all combinations of $t = 2$ are included in the matrix C' that only includes rows $\{2, 4, 3, 0\}$ of C .

IV. RESULTS AND DISCUSSION

This section presents an experimental design and results derived from the methodology described in the previous section. An experiment consisting of 85 covering arrays was designed, each covering array was built using a tool called IPOG (one

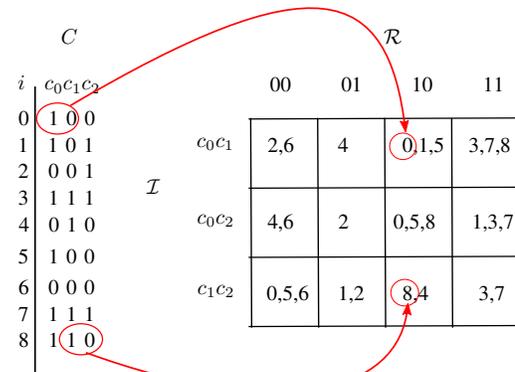


Fig. 2: Left: A covering array with $k = 3$, $t = 2$, and $\Sigma = \{0, 1\}$. Right: an illustration of the t -way interactions of row tests used for generating the map \mathcal{M} that relates realizations \mathcal{R} to interactions \mathcal{I} .

INITIALIZATION

\mathcal{M}		
e_O	$\mathcal{L}(e_O)$	$\#\mathcal{L}(e_O)$
$(\mathcal{I}_0, 00)$	[2,6]	2
$(\mathcal{I}_0, 01)$	[4]	1
$(\mathcal{I}_0, 10)$	[0,1,5]	3
$(\mathcal{I}_0, 11)$	[3,7,8]	3
$(\mathcal{I}_1, 00)$	[4,6]	2
$(\mathcal{I}_1, 01)$	[2]	1
$(\mathcal{I}_1, 10)$	[0,5,8]	3
$(\mathcal{I}_1, 11)$	[1,3,7]	3
$(\mathcal{I}_2, 00)$	[0,5,6]	3
$(\mathcal{I}_2, 01)$	[1,2]	2
$(\mathcal{I}_2, 10)$	[8,4]	2
$(\mathcal{I}_2, 11)$	[3,7]	2

ITERATION 1

\mathcal{M}		
e_O	$\mathcal{L}(e_O)$	$\#\mathcal{L}(e_O)$
$(\mathcal{I}_0, 01)$	[4]	1
$(\mathcal{I}_0, 10)$	[0,1,5]	3
$(\mathcal{I}_0, 11)$	[3,7,8]	3
$(\mathcal{I}_1, 00)$	[4,6]	2
$(\mathcal{I}_1, 10)$	[0,5,8]	3
$(\mathcal{I}_1, 11)$	[1,3,7]	3
$(\mathcal{I}_2, 00)$	[0,5,6]	3
$(\mathcal{I}_2, 10)$	[8,4]	2
$(\mathcal{I}_2, 11)$	[3,7]	2

ITERATION 2

\mathcal{M}		
e_O	$\mathcal{L}(e_O)$	$\#\mathcal{L}(e_O)$
$(\mathcal{I}_0, 10)$	[0,1,5]	3
$(\mathcal{I}_0, 11)$	[3,7,8]	3
$(\mathcal{I}_1, 10)$	[0,5,8]	3
$(\mathcal{I}_1, 11)$	[1,3,7]	3
$(\mathcal{I}_2, 00)$	[0,5,6]	3
$(\mathcal{I}_2, 11)$	[3,7]	2

ITERATION 3

\mathcal{M}		
e_O	$\mathcal{L}(e_O)$	$\#\mathcal{L}(e_O)$
$(\mathcal{I}_0, 10)$	[0,1,5]	3
$(\mathcal{I}_1, 10)$	[0,5,8]	3
$(\mathcal{I}_2, 00)$	[0,5,6]	3

ITERATION 4

\mathcal{M}

Fig. 3: PPCA for reducing the covering array instance shown in Fig. 2, the reduced covering array C' is obtained by selecting the rows $\{2, 4, 3, 0\}$ from C .

i	S_i^s
0	[3,3,3]
1	[2,3,3]
2	[1,2,2] \Leftarrow
3	[2,3,3]
4	[1,2,2]
5	[3,3,3]
6	[2,2,3]
7	[2,3,3]
8	[2,3,3]

$L = \emptyset$

i	S_i^s
0	[3,3,3]
1	[3,3,9]
3	[2,3,3] \Leftarrow
4	[1,2,2] \Leftarrow
5	[3,3,3]
6	[2,3,9]
7	[2,3,3]
8	[2,3,3]

$L = \{2\}$

i	S_i^s
0	[3,3,3]
1	[3,3,9]
3	[2,3,3] \Leftarrow
5	[3,3,3]
6	[3,9,9]
7	[2,3,3]
8	[3,3,9]

$L = \{2, 4\}$

i	S_i^s
0	[3,3,3] \Leftarrow
1	[3,9,9]
5	[3,3,3]
6	[3,9,9]
8	[3,9,9]

$L = \{2, 4, 3\}$

$L = \{2, 4, 3, 0\}$

TABLE I: Results of post-processing binary covering arrays, with $2 \leq t \leq 6$ and $k \leq 50$. The number in each entry is the value $N - N'$ for the instance with values k, t .

k	t					total
	2	3	4	5	6	
2	0	-	-	-	-	
3	1	0	-	-	-	
4	1	0	0	-	-	
5	1	1	1	0	-	
6	1	1	0	9	0	
7	0	0	3	3	4	
8	0	0	1	3	4	
9	1	0	0	2	3	
10	1	1	0	2	3	
11	0	1	0	0	1	
12	0	0	0	0	0	
13	0	0	0	0	0	
14	1	0	1	0	0	
15	0	0	0	0	0	
16	0	0	0	0	0	
17	1	0	0	0	0	
18	0	0	0	0	1	
19	0	0	0	0	0	
20	1	0	0	0	0	
# tested instances	19	18	17	16	15	85
# reduced instances	9	4	4	5	6	28

of the most popular tools in the state-of-the-art of covering arrays construction).

The results derived from our experiment are shown in table I. In this analysis, binary covering arrays are grouped by the number of their columns and their strength. Every group of t contains the different values of the alphabet for each covering array. Every cell of the this table shows the number of rows reduced in the corresponding binary covering array. As seen in the last row, the results reported a reduction in the size of 28 covering arrays (~33%).

Section II summarizes the techniques for constructing covering arrays, they can be grouped into: algebraic, greedy, exact and metaheuristics techniques. The best known solutions for CA with $t = 2, 3, \dots, 6$ are publicly available [8]. By analyzing that results, one can see that metaheuristics techniques produce better bounds but they are computationally expensive. For this reason, these techniques have concentrated on the construction of CA with $k < 100$. Algebraic and greedy techniques are better suited for large covering arrays, i.e. $v > 3$, $k > 100$ and $t > 3$; therefore, PPCA algorithm can be used for post-processing solutions constructed by these heuristics.

V. CONCLUDING REMARKS AND FUTURE WORK

This paper presents a post-processing strategy, called PPCA, for reducing the size of a covering array. The post-processing reduces the number of rows of a covering array through iteratively including the best row in the reduced covering array –the row that is most important for guaranteeing covering–. In some cases, the reduced covering array could be optimized but here we are interested just in reducing the size of a previously constructed CA, not in building a new one.

A dataset of 85 covering arrays constructed by the state-of-the-art algorithm IPOG was used to test the PPCA algorithm. The results show a reduction in ~33% of the instances.

In conclusion, PPCA has already proved being effective for reducing a wide variety of covering arrays.

We are designing a parallel version of the PPCA algorithm, in order to address problems with high strength, many factors or rows.

ACKNOWLEDGMENT

The following projects have funded the research reported in this paper: 148784 - Fondo Mixto CONACyT y Gobierno del Estado de Nayarit; 2143 - Cátedras CONACyT (H. Avila), 3163 - Cátedras CONACyT (C. Lara).

REFERENCES

- [1] R. Kuhn, Y. Lei, and R. Kacker, "Practical combinatorial testing: Beyond pairwise," *IT Professional*, vol. 10, no. 3, pp. 19–23, 2008.
- [2] B. Pérez Lamancha, M. Polo, and M. Piattini, "PROW: A pairwise algorithm with constraints, order and weight," *Journal of Systems and Software*, vol. 99, pp. 1 – 19, 2015.
- [3] D. R. Kuhn, D. R. Wallace, and A. M. Gallo, "Software fault interactions and implications for software testing," *IEEE Transactions on Software Engineering*, vol. 30, no. 6, pp. 418–421, 2004.
- [4] Z. Zhang, J. Yan, Y. Zhao, and J. Zhang, "Generating combinatorial test suite using combinatorial optimization," *Journal of Systems and Software*, vol. 98, pp. 191 – 207, 2014.
- [5] C. J. Colbourn, "Covering arrays from cyclotomy," *Designs, Codes and Cryptography*, vol. 55, no. 2-3, pp. 201–219, 2010.
- [6] G. Sherwood, "On the construction of orthogonal arrays and covering arrays using permutation groups," 2015, available online at <http://testcover.com/pub/background/cover.htm>. Accessed October 20, 2015.
- [7] K. Meagher and B. Stevens, "Group construction of covering arrays," *Journal of Combinatorial Designs*, vol. 13, no. 1, pp. 70–77, 2005.
- [8] C. J. Colbourn, "Covering array tables," 2015, available online at <http://www.public.asu.edu/~ccolbou/src/tabby/3-3-ca.html>. Accessed on October 11, 2015.
- [9] M. Chateauneuf and D. L. Kreher, "On the state of strength-three covering arrays," *Journal of Combinatorial Designs*, vol. 10, no. 4, pp. 217–238, 2002.
- [10] N. J. A. Sloane, "Covering arrays and intersecting codes," *Journal of Combinatorial Designs*, vol. 1, no. 1, pp. 51–63, 1993.
- [11] P. J. Schroeder, E. Kim, J. Arshem, and P. Bolaki, "Combining behavior and data modeling in automated test case generation," in *Proceedings of the Third International Conference on Quality Software*, 2003, pp. 247–254.
- [12] M. Forbes, J. Lawrence, Y. Lei, R. N. Kacker, and D. R. Kuhn, "Refining the in-parameter-order strategy for constructing covering arrays," *Journal of Research of the National Institute of Standards and Technology*, vol. 113, no. 5, pp. 287–297, 2008.
- [13] B. Jenkins, "Jenny: a pairwise testing tool," 2015, available online at <http://burtleburtle.net/bob/math/jenny.html>. Accessed on October 11, 2015.
- [14] A. Calvagna and A. Gargantini, "T-wise combinatorial interaction test suites construction based on coverage inheritance," *Software Testing, Verification and Reliability*, vol. 22, pp. 507–526, 2012.
- [15] J. Martinez-Pena and J. Torres-Jimenez, "A branch and bound algorithm for ternary covering arrays construction using trinomial coefficients," *Research in Computing Science*, vol. 49, pp. 61–71, 2010.
- [16] M. Banbara, H. Matsunaka, N. Tamura, and K. Inoue, "Generating combinatorial test cases by efficient SAT encodings suitable for CDCL SAT solvers," in *Proceedings of the 17th international conference on Logic for programming, artificial intelligence, and reasoning*, 2010, pp. 112–126.
- [17] C. Ansótegui, I. Izquierdo, F. Manyá, and J. Torres-Jimenez, "A Max-SAT-Based approach to constructing optimal covering arrays," in *Artificial Intelligence Research and Development*, 2013, pp. 51–59.
- [18] M. B. Cohen, C. J. Colbourn, and A. C. H. Ling, "Constructing strength three covering arrays with augmented annealing," *Discrete Mathematics*, vol. 308, no. 13, pp. 2709 – 2722, 2008.
- [19] K. J. Nurmela, "Upper bounds for covering arrays by tabu search," *Discrete Applied Mathematics*, vol. 138, pp. 143–152, 2004.
- [20] R. A. Walker II and C. J. Colbourn, "Tabu search for covering arrays using permutation vectors," *Journal of Statistical Planning and Inference*, vol. 139, no. 1, pp. 69–80, 2009.
- [21] T. Shiba, T. Tsuchiya, and T. Kikuno, "Using artificial life techniques to generate test cases for combinatorial testing," in *Proc. of the 28th Annual Intl. Computer Software and Applications Conf.* IEEE Computer Society, 2004, pp. 72–77.

Database Preservation: The DBPreserve Approach

Arif Ur Rahman
and Muhammad Muzammal
Department of Computer Science
Bahria University, Islamabad
Pakistan

Gabriel David
and Cristina Ribeiro
Departamento de Engenharia Informática
Faculdade de Engenharia, Universidade do Porto
Rua Roberto Frias, Porto, Portugal

Abstract—In many institutions relational databases are used as a tool for managing information related to day to day activities. Institutions may be required to keep the information stored in relational databases accessible because of many reasons including legal requirements and institutional policies. However, the evolution in technology and change in users with the passage of time put the information stored in relational databases in danger. In the long term the information may become inaccessible when the operating system, database management system or the application software is not available any more or the contextual information not stored in the database may be lost thus affecting the authenticity and understandability of the information.

This paper presents an approach for preserving relational databases for the long-term. The proposal involves migrating a relational database to a dimensional model which is simple to understand and easy to write queries against. Practical transformation rules are developed by carrying out multiple case studies. One of the case studies is presented as a running example in the paper. Systematic implementation of the rules ensures no loss of information in the process except for the unwanted details. The database preserved using the approach is converted to an open format but may be reloaded to a database management system in the long-term.

Keywords—Database Preservation, Transformation Rules

I. INTRODUCTION

There are many advantages of working digitally. Digitally stored data is easily accessible, manageable and helps in providing faster and better services to users than their paper based counterparts. However, with the benefits also come some trade-offs. For example the constant change in software and hardware technologies affect the data. This change may turn data unreadable as the old hardware, operating system and application software used for creating, storing and managing may not be supported by the latest technology. However, data may be relevant for the purpose of evidence of activities, institutional memory, scientific significance or historical significance and therefore preserving them may be required. Sometimes keeping organizational data is not a choice but it is mandated by law or by organizational policy. An example is the duration for which a university should keep the grades obtained by students in the courses they studied. This information is no longer required in day-to-day matters after a student completes his degree but it may be required as an evidence in the long-term.

Keeping in view the significance of data, digital preservation may be required for ensuring the constant availability

of data over time. Selecting a preservation strategy, tools and formats for preserving data is a complicated task. Typically, decisions depend on the aims for given settings and institutional needs [5]. However, study shows that the selection of open formats is better for preservation than proprietary formats whenever possible [18]. Data stored in open formats may be easily accessible in the long term. New software may be developed for accessing data stored in open formats in case the existing software becomes obsolete. Moreover, widely used formats should be used for preservation as they raise the prospect that they will continue to be used for a long time [28]. In addition to this, formats for which a variety of writing and rendering tools are available should be given preference [20]. Metadata inherent to data and about the whole environment around them should be collected. It includes all the technical metadata like the software used for creation and preservation of data and non-technical metadata like the people involved in the creation of data and the reasons behind the creation. Another good practice may be to minimize the dependence of digital objects on users and software from the operational environment and preservation environment.

Different types of digital objects are created and managed by organizations for their operation including text documents, images, graphics, audio, videos, databases, websites and software. Different preservation approaches are required to preserve different types of digital objects as their nature and structure are different [4], [25]. For example emulation may be used as a preservation strategy if it is required to provide access to obsolete software but if the requirement is to provide access to data then migration may be used. The focus of this paper is on relational database preservation.

Software engineering good practices enforce the three layers approach to information systems design. The interface layer is accessible to users and get access to the data layer through the business rules layer. The data layer is implemented in a relational database. Relational databases are complex digital objects with a well-defined data structure. They are based on the formal foundations of relational modeling proposed by [9]. They organize collections of data items into formally-described tables. They are designed using the rules of normalization which is a step-by-step reversible process of replacing a given collection of relations by successive collections in which the relations have a progressively simpler and more regular structure [6], [16]. Normalization involves splitting large data tables into smaller and smaller tables, until reaching a point in which all functional dependencies among columns are

dependencies on the primary key of the corresponding table. Thus, preserving the uniqueness of the primary key ensures the uniqueness of the representation of the fact subject to the dependency. This helps databases to be consistent and efficient in capturing facts. However, because of normalization the model of a database may become too fragmented and difficult to understand.

The interface and business rules layers are implemented in code (triggers, functions, stored procedures and application). It is a fact that preserving code is an even harder issue than preserving data. Code is normally platform-dependent hence the requirement is to preserve the whole engine required to run it. Otherwise, there is a danger of losing the derived information which the code is able to generate. The preservation of code may refer to the generic software preservation problem which has been addressed in several ways e.g. emulation and technology preservation, but none of them makes it platform-independent [11], [19], [23]. The interface part dealing with presentation and interaction aspects is less relevant from an information preservation perspective. Part of the business rules deal with access control, compliance with organizational policy of the new transactions, and other operational aspects which are not very relevant as well. However, part of the business rules contain functions able to compute complex derived data which may be absent in the actual database. An example of this is a ranking function for scholarship granting.

Data stored in database systems are vulnerable to loss because it may become inaccessible and unreadable when the software needed to interpret them or the hardware on which that software runs becomes obsolete or are lost. Data may also become difficult to understand if the contextual information needed to interpret them is not known or is lost. Furthermore, if the structure of the information is too complex it may become a hard and time consuming task to query the information stored in a database. It is also possible to have wrong or missing information in operational systems which may lead to wrong results when queries will be made in the long term.

Different database systems require different solutions for solving the issues of complex structure, information embedded in code and missing or wrong information. For example a database system may have reached the end of its active life and is not used in day to day activities (retired databases). A database may retire in situations like an organization develops a new human resource management system using the latest technologies and the old one is not used anymore. Similarly, databases may retire if the activity for which they were developed comes to an end. Another kind of databases may be databases which are operational but a part of them is not used any more in day to day activities. For example once the fiscal year of an organization finishes, the data for that year may not be required in day to day activities.

The proposed solution works in cases where a retired database needs to be preserved or a snapshot of a database is taken and needs to be preserved. It follows the general framework of the Open Archival Information System reference model (OAIS) [7]. The producer sends a submission information package (SIP) containing the database and supporting information which may be helpful in the database preservation procedure. A digital preservation team preserves the database and produces the archival information package (AIP). The AIP

is produced following the model migration approach which proposes to migrate a relational database to a dimensional model in the preservation procedure. Moreover, practical transformation rules are proposed which help in carrying out the model migration procedure. The AIP is stored for the long-term which may be accessed in the future using a simple to use and platform-independent tool. The proposed approach is a step further on the existing approaches for database preservation.

II. RELATED WORK

Significant research has already been conducted for preserving relational databases for the long-term. The conclusions discard approaches like building technology museums for preserving specimens of machines, system software and applications, in all their main versions, so that the backups of every significant system could be used whenever required. Emulation is another approach which suggests simulating the old hardware or software in new machines [10], [13]. However it is not a permanent solution as technology changes very fast and writing new emulators is required whenever a change in technology occurs. More promising research suggests the conversion of a database into an open and neutral format with a significant amount of semantics associated, hence making it independent of the details of the actual DBMS. Such an approach is the software-independent archival of relational database approach presented in the sequel.

A. SIARD

The Swiss Federal Archives (SFA) proposed the Software-Independent Archival of Relational Databases (SIARD) for preserving relational databases for the future [12]. A software package known as SIARD Suite was developed which facilitates the conversion of a relational database to an open format known as the SIARD format [26]. It can convert a database originally in Oracle, Microsoft SQL Server, Microsoft Access and MySQL to the format. A database archived using the SIARD Suite needs to be loaded back to a DBMS to be able to query it. The SIARD format is based on open standards such as XML, SQL:1999, ZIP and UNICODE. It is able to record metadata at different levels including database, schema, table, column, trigger, routine, user, role and privilege. Most of the metadata is automatically retrieved from the data dictionary. However parts of metadata like the description of database and database objects, life span, owner and the people performing the preservation procedure need to be manually entered.

The DBPreserve Suite (presented in the sequel) is used to extend the metadata in a SIARD archive.

B. DBPreserve Suite

The DBPreserve Suite is a tool developed in Java for generating dimensional modeling metadata [2]. The metadata includes the names of stars and the relevant fact tables and dimensions. Moreover, it includes the metadata about the levels and hierarchies based on the levels in dimensions. The tool automatically extracts most of the metadata from the data dictionary of the DBMS. However, some information including the descriptions of stars, dimensions and dimension levels has to be manually added. The tool adds a new file with the dimensional modeling metadata in a SIARD Archive.

III. MODEL MIGRATION APPROACH

A preservation procedure must be initiated when a database is approaching the end of its active life. The end of a database active life may be determined by the end of the reason for its creation but very often the case is of a replacement of the corresponding information system by a new one. Even in the latter case, though migration of the previous data may occur, it may be just partial. So, in all these cases a global appraisal of the database must be performed and consider the regulations in force, the probability of data loss, the data value in terms of evidence for the recorded facts, of scientific relevance and of organizational memory. This archival analysis must be accompanied by a technical and economic analysis on the feasibility of the preservation procedure.

Figure 1 presents the model migration approach for database preservation which involves the migration of a database to a dimensional model in the preservation procedure. The two issues to be considered in the migration procedure are the complexity of a relational model and the embedding in code of important knowledge from the application domain. The requirement of normalization does not remain in force once the active life of a database finishes and it becomes closed or when a snapshot is taken and a frozen database is to be preserved. The data will no longer be updated but it may be queried for purposes like verifying evidence of activities. This change of usage and a change in consumers bring a change of requirements. The need is now to preserve a database in a form that is easier for consumers to understand and write queries against. It means that it should be simple and it should give quick access. This can be achieved by migrating the database from a relational model to a dimensional model as suggested by Figure 1. The how-to issue will be dealt latter.

A. Dimensional Modeling

Dimensional modeling is a logical design technique that seeks to present data in a standard framework which is intuitive, allows for high-performance access and is resilient to change [1], [3], [27]. It is compatible with relational modeling and there is a straight forward mapping between them [8], [21]. It organizes data in tables of two natures, namely dimensions and fact tables. Dimensions store the detailed data about objects or entities. Furthermore, they can have levels and hierarchies which enable users to view data at various levels of detail. Fact tables store references to the set of relevant dimensions involved in a business process as well as the values representing real world facts [14], [15], [17]. A representation where a fact table is surrounded by dimensions involved in a business process is known as a star. Dimensions can be shared by different stars. A bus-matrix is constructed in the process of developing a dimensional modeling for a source relational model. It identifies business processes and the dimensions involved in them. Business processes are listed as matrix rows and dimensions are listed as matrix columns. The cells of the matrix are marked to represent which process includes which dimensions. Data transformations may be needed in the process of developing a dimensional model from a relational model as changes may occur in the structure, representation or content of data [22].

B. Migration Procedure

Figure 1 presents the database preservation procedure using the model migration approach. It can be seen that extraction, transformation and loading (ETL) is needed to migrate a database to its dimensional model. ETL is the process of taking out data from one or more operational systems (extract), modify them to fit into the dimensional model formatting needs (transform) and finally insert them into the dimensional model (load) [24]. In the migration process the relevant functions in the database are executed and the derived information are explicitly stored in new tables or in new columns in existing tables. This removes the dependency of a database on the underlying DBMS or computation environment and improves accessibility. The step of migrating the database to XML format is included to make it completely platform-independent. In this manner the main preservation concern of preserving the database content and the information implicit in business rules is accomplished.

The next section presents the transformation rules for carrying out the model migration procedure.

IV. TRANSFORMATION RULES

To make the process of model migration efficient and traceable a set of transformation rules are proposed. These rules have three main purposes. The first is to design each component of the dimensional model from the systematic analysis of the relational model. The second is to perform a form of extraction, transformation and loading process on the data, with the care of keeping track of the original records. A final verification of the result against the original is important. The third purpose is to add enough contextual, content, technical, and provenance metadata to ensure understandability and authenticity of the database. The overall goal of these rules is to make sure that as much as possible of the information in a relational database is transferred to a dimensional model without any changes or loss. Thus, the rules help in carrying out the process in a way which ensures that the integrity of the database is not lost. Moreover, the rules ensure that it is still possible to query the information for the same purposes as it was possible with the source systems. Each step taken following the rules is documented. This helps to ensure the authenticity of the preserved database. The rules were obtained from carrying out case studies. One of them is used as a running example for presenting the transformation rules. The case study involved a database system which was used to manage all the information required by the institution about teachers as well other staff members. Oracle was used as the DBMS and an application developed in Visual Basic was used for interaction with the database.

A few years ago the old database system was replaced by a new state-of-the-art information system. The old system became frozen as the data stored in it was not updated anymore. However, the data is important and it may be required in the future as an evidence of past activities and institutional memory. Therefore, the database needs to be preserved for the long-term. A part of the model of the database is presented in Figure 2. It includes different tables including a table for the basic information of the personnel, their contracts and the contract categories. It also includes a table for the salary

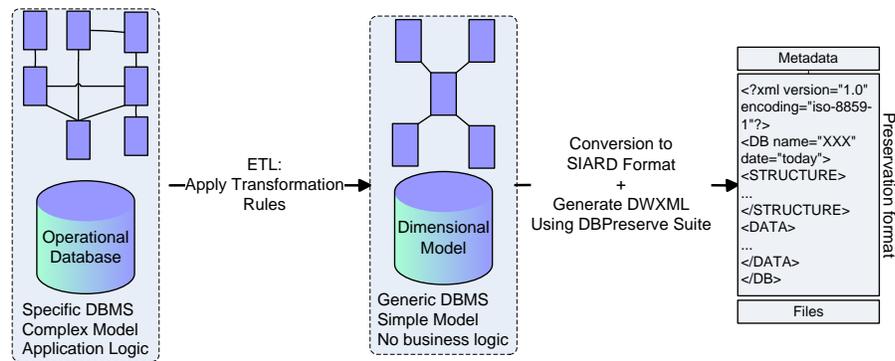


Fig. 1: Model Migration Approach for Database Preservation

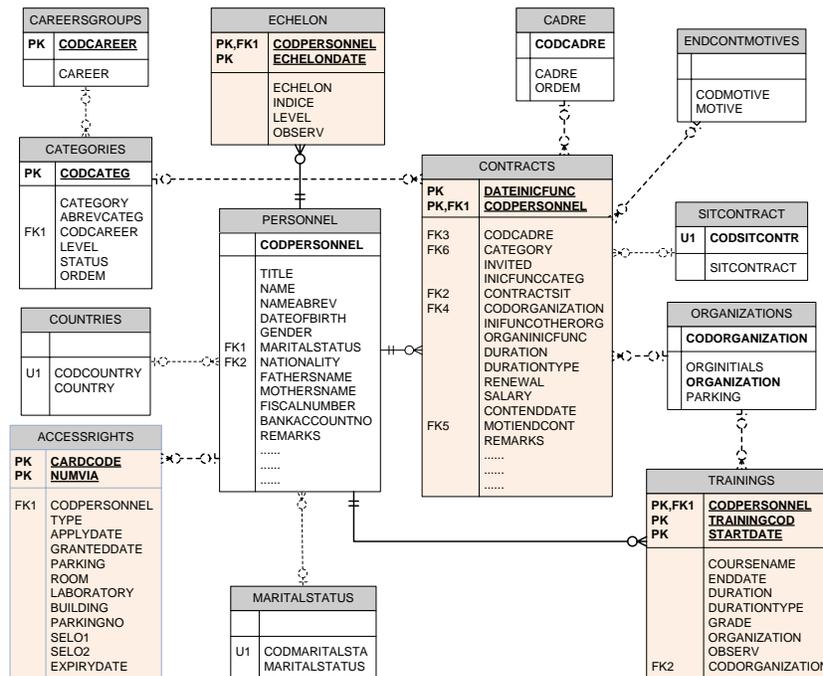


Fig. 2: A Part of the Model of the Human Resources Department Database

echelon and the motives of the end of contracts. The rules are grouped based on the phases in which they are implemented.

The rules are grouped based on the phases in which they are implemented.

1. Preparation and Cleanup

In this phase the source database tables are analyzed which includes writing a description of each one of them and identifying empty and temporary tables. The following rules are implanted in this phase.

- i. Describe tables and columns: *A description of each table is prepared with name, meaning, source, number of rows, number of columns, date of creation, last update, number of LOB columns and constraints. In addition to this, information about each table column is collected including name, meaning, source, data type, number of distinct values, minimum and maximum value and number of nulls.*

An initial analysis of the case study database was done. It was

found that there were 129 tables in the database, collectively having 999 columns. Following the transformation rules the descriptions of the tables were written. Statistics about each table were recorded including the number of rows, number of columns and number of distinct values in a column and maximum and minimum value in a column. The description of Contracts database table is shown at table level and at column level in Table I and Table II respectively. These descriptions are stored separately in two database tables in the staging area. Parts of the descriptive information were automatically retrieved from the data dictionary. However some information needed to be manually entered including the source and meaning.

- ii. Remove empty tables and columns: *Empty tables may exist in a database because of various reasons e.g. an anticipated but unimplemented functionality. Such tables may be ignored after a thorough analysis. The analysis may include a check on the meaning of its columns and its relationships with other tables. Similarly, there may be columns in tables which are*

TABLE I: Table Description at Table Level

Table Name	CONTRACTS
Meaning	This table contains data about the contracts of employees. Information like the start and end date, salary, type, duration, renewable or not about the contract are included in the table.
Source	Data is added to the table each time a new employee is hired or his contract is renewed. There is a form for adding a new employee in the application software.
Number of Rows	6220
Number of Columns	36
Date of Creation	
Date of Last Update	
Number of LOB Columns	0
Comments	This table contains data about all the contracts of an employee. The <code>personnel</code> table also contains data about the current contract of an employee which seems repetition and may be ignored.

TABLE II: Table Description at Column Level

Column Name	Meaning	Source	Data Type	Number of Nulls	Distinct Values	Minimum Value	Maximum Value
codpessoal	Reference to personnel table	New employee addition form	char	0	2224	AA	ZMAV
organization	The organization where the employee works	char	1256	10	0032	6
codcadre	Cadre Code	char	4	13	1	9
category	Contract Category	char	1657	260	0	930
invited	If the employee is invited from another institution	char	5666	9	0	V
inicfunccateg	Start date in a category	date	5924	175	1969.08.22	2008.12.31
contractsit	char	6023	12	A	XXX
organinicfunc	date	5673	27	0	99016
duration	number	268	40	0	51
durationtype	char	3219	2	A	M
renewal	char	4247	5		S
motiendcont	char	629	20	0	22
remarks	char	4036	1998		ú

empty for all the rows. Such columns may also be ignored after an analysis.

The example database tables were analyzed. It was found that there were 26 empty tables and 67 columns with all nulls in the non-empty tables. After this step the empty tables and empty columns were ignored.

iii. Remove temporary tables: *Sometimes snapshots of tables are taken on a specific date and kept in a database. For example in the process of adding a new functionality in a database system, a snapshot of the existing tables involved may be taken as a backup in case of any problems. If documentation on the database exists, namely its data model, a check against it may help on the identification of temporary tables. Tables having similar names and having the same columns are candidates to be analyzed. If such a copy exists, it may be discarded and only the table which has the latest information is kept. The table which has the latest information may have a greater number of rows. Moreover, database logs may also be queried to see which copy of the table was updated last. The analysis may also include comparison of results of queries on both copies of a table.*

Temporary tables of two types were found in the example database. These tables were either storing data about a middle step in a process like recording data about all the applicants for a position before selection or they were snapshots of tables taken on a specific date. For example the Personnel table contains data about all the employees in the higher education institution. However, the Docdee table had data about teachers working in a specific department. This table should actually be a View as it presented a part of the Personnel table. Similarly, the Contests, Conteststea and Contestsntea tables were storing

data about the contests for a specific position having 1, 16 and 1 row respectively. No reference to these tables was found in other tables. The low number of rows in the tables shows that they were added because of a newly added functionality in the system. However, the functionality was not fully implemented and it was not used. After a confirmation from the users of the system the temporary tables and the empty tables were ignored.

2. Identify Keys

In this phase the primary keys and foreign keys identified. The rules presented below are followed for their identification.

i. Identify primary keys: *It may be required to merge and split tables in the migration process. Typically, it is required to know the unique identifier in a table to carry out such operations. Moreover, if a table is a candidate to become a single dimension, it requires a primary key for linking with a fact table. Though it may not be always required, it is helpful to identify primary keys of tables before proceeding to any operations on the tables. There are different scenarios which may happen.*

(a) *The primary key is implemented through database constraints and it can be identified just by looking at the constraints of a table.*

(b) *The primary key is not implemented through database constraints and there is a single candidate key. In this case there is no other choice, therefore the candidate key is declared as the primary key.*

(c) *There is no primary key implemented through database constraints and there are multiple candidate keys. In this situation, actual foreign keys in the other tables referencing candidate keys in the current table are searched for. The*

primary key is chosen among the candidate keys involved in the external references. Otherwise, any candidate key may be considered as the primary key.

(d) There is no primary key and no candidate key in a table. In this situation a primary key column may be added and filled with numeric values.

The primary keys of the tables in the example database were identified. It was found that the primary keys conceptually existed but they were physically not implemented. For example there was a column in the Personnel table with the initials of the names of the employees. The number of distinct values in the column was equal to the number of rows in the table so it was an obvious candidate key. Furthermore, it turned out that this column was used in many other tables as a reference to this table. Therefore, this column was recorded as the primary key of the table. In addition to this, a check was performed on the Personnel table to see if it contains repeated records for persons with the same name. It was found that there were ten people with the same name. They were considered different records as the primary key values for these rows were different.

ii. Identify foreign keys: *All the tables are analyzed and any foreign keys are recorded. In some cases foreign keys are declared through database constraints. To identify the other, the fields of all the tables are analyzed and their meaning and contents studied to find out if they are references to other tables. If a field references any other table, it should be checked if it contains any orphan child records (OCR). There are three ways to resolve the situation if OCRs are found.*

(a) The OCR may be unwanted data e.g. inserted as a result of poor data validation applied prior to record insertion. In such situations the records may be ignored and deleted. However, any deletion is double checked in order not to damage the integrity of the database.

(b) The second way is to insert a parent record for each set of OCRs who needs the same parent record in the parent table. This way the problem is solved but it may create problems as the insertion of a new record may mislead users in the future. Users may consider the new entry a part of the data. For example the creation of an extra department may mislead users in the future to consider it a functional department. Therefore, when using this option it is ensured that it is not confusing for future users.

(c) The third way is to insert a single parent record for all OCRs meaning precisely unknown parent record. However, the issue with this choice is that the keys in the column need to be changed to the newly inserted record. This way the actual value inserted in the life of the database is lost.

In the next step the foreign keys in the tables were identified. It was found that relationships among the tables existed conceptually but in most cases they were not physically implemented. This had resulted in many data quality problems like orphan child records and illegal values in many tables. The foreign keys for the Contracts table are presented in Table III. For example in Contracts table the Codorganization column is a reference to the Organization table. It included values like xtCodand for 32 and 44 rows respectively. These are illegal values and are not in the domain. After an analysis of the application it was concluded that the application software had no validation in place before inserting a new record. Therefore, the system accepted whatever values were entered by users. To resolve the situation, records were inserted in the Organization table with the values to avoid losing the records in the loading

process. A similar situation arose in the Department column of the Personnel table. The Department column was a reference to the Department table. In the reference there were some records with a value iCC The value was correct but because of case difference OCRs were detected. The actual value in the parent table was ICC The value was updated to the same as in the parent table. However in the same table there were some employees with no department. Therefore a record was inserted in the parent table with as 'Unknown Department' with a code as D All these actions taken for removing the anomalies from the data were documented and kept for future reference. The PL/SQL code for these actions was also recorded in textual format. Table IV presents the preservation actions log. This log is also included in the AIP.

TABLE III: Foreign Keys in the Personnel table

Column	Referenced Table
codcadre	cadre
category	category
regime	regime
coddegree	degree
codorganization	organization
codstatus	status
codprogram	program
contracttype	contracttype
department	department
section	section
codteachertype	teachertype

3. **Normalization** In this phase the source database tables are analyzed to see if they are properly normalized or not. The rules presented below are followed in this phase.

i. Expand abbreviations: *Codes may have been used in the data to represent the status of something e.g. A - Approved, N - Not approved, C - Cancelled. Such codes may be replaced with meaningful words making them clearer and easily distinguished among them. Alternatively, the encoding may be recorded as metadata in the corresponding column of the preserved database. Information on the encoding may come from the application interface or the code generating it, from the database documentation or from domain experts.*

In some tables abbreviations were used, for example in the Contracts table 'Y' for year, 'M' for month and 'D' for day were used in the Durationtype column. The abbreviations were replaced with the full word.

ii. Normalize tables: *Each table is checked against the rules of relational modeling. Sometimes there are tables in a database that are not properly normalized. If a table is not properly normalized it may need to be split into multiple tables which may then become sources for different dimensions in the resulting dimensional model. In the normalization procedure, the extraction of columns into a new table goes along with the definition of its primary key and leaves a corresponding foreign key in the original table.*

In the example database, there were tables which needed to be merged. For example the Personnel table was storing basic data about employees as well as some data about the current contracts. It was needed to verify if the information about the current contract of an employee is also stored in the Contracts table. Queries were made on both tables and the results were compared. It came out that Contracts table stored information about all the contracts including any past contracts and the current one. Another table Personnelinfo was storing

TABLE IV: Preservation Actions Log

SNo	Action Description	Rule	Code
1	Assigning foreign key to personnelinfo (Column: maritalstatus) referencing maritalstatus table.	2	alter table personnelinfo add constraint fk_perso0_data_maritstat foreign key (maritalstatus) references maritalstatus (codmaritalstatus);
2	Some persons are without department in the personnel table (department is null), so I inserted a row in department table	2	insert into department (sigla, nomedep) values ('UD', 'Unknown Department');
3	Rows in the personnel table with department = null, updated with department = UD (190 rows)	2	update personnel set department='UD' where department is null;

detailed data about personnel like names of parents, date of birth, marital status, address, country of origin, bank account information and so on. A merge was performed but in the Personnelinfo table there were records for which there was no matching record in the Personnel table. In order not to lose information, the records which did not fulfil the joining condition were manually inserted in the resulting merged table. Furthermore, the columns which were storing information about the current contract of an employee in the Personnel table were ignored as this information was recorded in the Contracts table.

4. Define Stars

In this phase the tables are clustered, the dimensions are identified and then the bus matrix is defined. The rules presented below are followed in this phase.

i. Cluster tables: *All the tables are clustered considering their foreign keys and under the broader context of the relevant processes in the organization. The organizational processes about which the system stores data are found out. A discussion with the users of the system may be helpful. Moreover, looking at the application program and understanding the way it works is also important.*

For each organizational process there is a set of tables in the model which stores related data. For all the processes the set of participating tables are listed. In addition to this, in each set of tables there is normally a central table which is identified. This table normally has larger number of rows than the tables it references. Usually, it has no or a small number of incoming references but it references other tables. The central tables have the real world facts recorded in the organizational processes. They may be candidates to be loaded into fact tables and become the centers of stars. In the end all the tables must be included in one or more stars, otherwise new stars are defined to encompass the remaining tables. To help in the clustering task, the following guidelines are provided, to be applied to the set of all tables after the normalization steps.

- *An isolated table is a cluster and it is its central table.*
- *When a table has no incoming references but it references at least one table, it becomes the central table of a cluster and the tables it references are recursively included in the cluster.*
- *When a table, though having incoming references, is referencing at least two other tables and has meaning, in the context of an organizational process, as a fact recording table at some level of granularity, it becomes the central table of a cluster and the tables it references are recursively included in the cluster.*

The union of clusters must contain all the tables. A table may belong to more than one cluster. Sub-clusters may be contained in larger clusters.

The following organizational processes were identified in the

part of the model of human resources department database presented in Figure 2.

- **Contracts:** This process is about the contract of an employee. An employee is hired by an organization, the contract is according to a cadre and it may be for a certain period with an explicit termination motive.
- **Echelon:** This process manages data about the salary echelon of an employee. An employee is assigned an echelon which is further divided into different levels.
- **Access Rights:** This process manages the access rights of an employee. For example access to different rooms, buildings and car parking of the institution.
- **Trainings:** This process manages the trainings which an employee takes. The trainings may take place in a specific organization.

Now keeping in view the organizational processes, the tables Contracts, Echelon, Trainings and AccessRights seem to be the central tables of clusters. Starting from the Contracts table, it can be seen that it references tables, namely Personnel, Categories, Cadre, Endcontmotives, Sitcontrat and Organizations. It can be seen that the Personnel table references tables, namely Countries and MaritalStatus. Furthermore, the Categories table references the CareerGroups table. The tables referenced by the Personnel table and the Categories table are included in the Contracts cluster.

The Echelon and AccessRights tables references the Personnel table. Therefore, they become the central tables of two different clusters. Both reference the Personnel table. Therefore the Personnel table and the tables referenced by it are included in the clusters. The Trainings table references two tables, namely Personnel and Organizations. Therefore, the Trainings table becomes the center of another cluster. The central table references the Organizations table, the Personnel table. Therefore, the Organizations table, the Personnel table and the tables referenced by it are included in the cluster. The Table V shows the result of this step.

Though the tables Personnel and Categories reference other tables, they are not considered central tables of a cluster as they are not fact recording tables about an organizational process. Furthermore, it can be seen that all the tables in the model belong at least to one cluster. Therefore, no table is left in the process.

ii. Identify dimensions: *Each cluster of tables is analyzed to identify dimensions in the future stars. Moreover, levels and hierarchies in each dimension are identified. A single dimension may be a target for multiple source tables. The identification is best guided by the knowledge of the organizational processes and of their main entities.*

The following analysis should be applied to all the clusters, starting with the smaller ones. In each cluster analyze each of the foreign keys out of the base table. If the referenced table is

TABLE V: Table Clustering

Clusters	Categories	CareerGroups	Countries	AccessRights	Personnel	MaritalStatus	Echelon	Contracts	Cadre	Trainings	Organizations	SitContract	EndContMotives
Contracts	X	X	X		X	X		X	X		X	X	X
Echelon			X		X	X	X						
Access Rights			X	X	X	X							
Trainings			X		X	X					X		

not part of a dimension or a sub cluster, build a dimension starting with that table as the lowest level and the tables it references, on the next level, recursively. The hierarchy is defined by the sequence of references. If the referenced table is the base table of a sub-cluster, build a dimension with just the non-foreign key attributes. The foreign key attributes may be added to the base table of the main cluster if considered relevant to understand it.

In the example consider the Contracts cluster. The central table i.e. Contracts, has several keys. Following FK1 the Personnel table becomes the lowest level in a dimension. Following FK1 of the Personnel table, the maritalStatus table becomes the next level in the dimension. Similarly, following the FK2, the Country table becomes another level in the dimension. Following the FK2, FK3, FK4 and FK5 of the central table, the SitContract, Cadre, Organizations and EndContMotives tables respectively become single level dimensions. Following FK6, the Categories table become another dimension. The FK1 of the Categories table lead to the creation of another level in the dimension by adding the CareerGroups table. Now considering the AccessRights cluster, the AccessRights table is identified as the central table of the cluster. At first look it may seem that the columns, namely Room, Laboratory and Building may become a separate dimension which levels including Rooms and Buildings. However, an analysis of the data shows that these columns have only two distinct values i.e. '0' and '1'. Therefore, they are kept in the central table without creating a dimension. The FK1 of the central table is a reference to the Personnel table. The Personnel dimension is already identified in the previous cluster. Therefore, the same dimension is included in this star. Similarly the Echelon cluster has the central table Echelon which references the Personnel table using its FK1. The Personnel dimension is already identified so the same dimension can be shared by this star. The Trainings table is the central table of the Trainings cluster. The central table seems to have columns for the course an employee takes, namely Codcourse and Coursename. These columns are separated and a separate dimension is created, namely Courses. The central table references the Organizations table using FK2. Though a column for the name of the organization exists in the table, it is better to use the Organization dimension.

iii. Define the bus matrix: *Decide whether all the clusters will be converted into stars. For each cluster, the central table becomes the fact table in the corresponding star and the dimensions identified in the previous step are also included in the star. Once all the dimensions and fact tables are identified a bus-matrix is constructed. The bus-matrix makes it clear which dimensions are part of which star. In some cases, the connection between a certain fact table and a dimension may*

require a bridge table of some sort. The source tables in the original model for the bridge tables are also identified.

It is necessary to check whether all the columns excluding the ones which are deliberately left-out because of a known reason, belong to at least one map in the end of this mapping process.

The bus matrix for the example database is shown in Table VI. It can be seen that four stars are built. The Personnel dimension is shared by all the stars where as the Organizations dimension is shared by two stars, namely Contracts FT and Trainings FT.

TABLE VI: Bus Matrix

Dim	Personnel	SitContract	EndContMotives	Organizations	Categories	Cadre	Courses
Contracts_FT	X	X	X	X	X	X	
AccessRights_FT	X						
Echelon_FT	X						
Trainings_FT	X			X			X

5. Implement Dimensions

In this phase the dimensions are implemented. In some cases it may be needed to use degenerate dimensions. The rules presented below are followed for implementing dimensions.

i. **Implement dimensions:** *The migrated model should be made of simple stars, to be easy to query. One technique to achieve this is to de-normalize the dimensions, including in each one of them simple or multiple hierarchies. This corresponds to merging tables in the original model, but keeping the primary key for each level as a level key. There may be situations in which some columns from one table may need to be moved to another table.*

While merging tables, situations may arise where some records in one table mismatch records in other tables. Such situations are carefully verified not to lose data and the records may be manually inserted in the resulting merged table if required.

The identified dimensions were implemented. This required joining and splitting tables. The Personnel dimension required joining three tables, namely Personnel, Countries and MaritalStatus. These tables were joined and level keys were added. Similarly, for the creation Categories dimension the Categories and CareerGroups were joined. The creation of Courses dimension required splitting the Trainings table. The columns related to the course taken, namely CodCourse and CourseName were added to the Courses dimension. The other dimensions did not require splitting or merging tables and they became single level dimensions. The resulting model

included four stars, namely Contracts, Echelon, Trainings and AccessRights. The Contracts star is shown in Figure 3.

ii. Include degenerate dimensions: *There may be columns storing descriptive information e.g. 'remarks'. Such a column may be added to a dimension if the values are solely dependent on its key. However, it may be added as a degenerate dimension value in the fact table if it depends on a combination of dimensions.*

In the Contracts star fact table there column remarks could become a dimension. However, the number of distinct values in the column was almost equal to the number of rows in the fact table; it was decided to keep it in the fact table.

6. Implement Fact Tables

In this phase the fact tables are implemented. They may require the use of a star schema or a snowflake schema. The rules presented below are followed in this phase.

i. Use snowflake schema: There may be situations where a set of tables in the operational system may need to be joined for constructing a dimension and one of them is a lookup table with more records than actually used by the lower level in the hierarchy. In such situations a snowflake schema is constructed which helps in not losing the higher level rows.

ii. Handle nulls: *Nulls are abnormal values in a column in the sense that, in some situations, they lead to unexpected results and they require specific operations. When they appear in a column that is a foreign key, the corresponding line is stripped from inner joins with the referenced table. One solution is to use outer joins. Another one is to add a line for the unknown or applicable value in the referenced table and update the null values in the foreign key column to become the corresponding key.*

There were some records in central tables of the clusters with a null in the foreign key columns. For example the SitContract column in the Contracts table (Figure 2) had some lines with nulls. Therefore, a row was inserted in the SitContract table with CODSITCONTR = 0 and SITCONTRACT = 'Unknown' and the records in the Contracts table were pointed to this record.

iii. Convert the records: *Before loading data to the target dimensional model the stars may be compared among them and it may be checked if two stars are candidates to be combined. They may be candidates to be combined if they share the same dimensions, record information at the same level of granularity and the timing and nature of events is the same. Data may be loaded to the target dimensional model once the check is performed.*

In the example database there were no similar stars. Therefore, the dimensions were loaded from the source tables.

7. Code

In this phase the application software or code in other forms including functions, stored procedures and triggers are dealt with. The rules presented below are followed in this phase.

i. Analyse application forms: *The application program may have forms for adding new data, displaying the data already in the system and generating reports. The forms are analyzed and the functionality is recorded. This may be used to verify that the resulting dimensional model is able to answer the queries used in the forms for retrieving information from the database. Furthermore, screen shots of the forms are taken and kept for future reference.*

Associating different menu items of the application with the

database tables may be helpful in understanding the functionality of the form in the future.

Screen shots of the application software were analyzed. The forms in the application software for inserting new data did not have validation in place. This was one of the reasons of orphan child records in some tables.

ii. Make short descriptions of algorithms (code) *If there are functions, stored procedures or code in any form to derive information from the data stored in a database, they are executed and the results are explicitly stored. In addition to this, a description of each piece of code explaining it and the information it produces is written and kept in the preserved database.*

In the application software there was no such code which generated derived data not explicitly stored in the tables. However, there were triggers to show notifications about the last date of a contract on the contract end date. Description of the triggers were written and kept for future reference.

8. Metadata and Verification

Metadata is recorded in the migration procedure. In this phase the results are verified using some of the metadata collected in different phases. The rules presented below are followed for collection of metadata and verification of the resulting model and the data.

i. Record metadata: *There is a set of metadata elements that should be recorded along with the migration procedure, at several stages. These elements include reference, contextual, technical, and provenance metadata at several levels like global, organizational process level, star, dimension and column levels. Even the mappings which are the basis for the ETL process from the source database to the target dimensional model are kept for future reference. The mappings record the origin of data in the dimensional model and document the whole process. They may be used to facilitate any verification procedure.*

The recorded metadata was initially kept in the staging area in database tables. It was moved to the dimensional model schema once the migration process completed.

ii. Verify the result: *In the end of the process a complete verification of the number of rows in each table against the original database must be performed, in order guarantee that no relevant information has been lost. Notice that in rule 8 a completeness check is performed at the level of tables, in rule 9 at the level of columns, and in rule 11 at the level of records, so no data is lost in the migration procedure, except for the assumed irrelevant or empty tables and columns.*

In the end of the migration procedure, it was verified that all the data from the source system was transferred to the dimensional model. For this purpose the table clustering done following rule 8 was checked once again. Moreover, it was verified that the identified dimensions contained all the columns of all the tables. For this purpose the bus matrix constructed following rule 10 was reviewed. Furthermore, it was also verified that the stars contained the relevant dimensions.

The model migrated database becomes free from the operational environment. It has a simple model which is easy to understand. Moreover, the metadata gathered in the migration procedure help in assessing the authenticity of the database. After the model migration procedure the Archival Information package is prepared and kept for the long-term.

The complete number of the transformation rules is high but

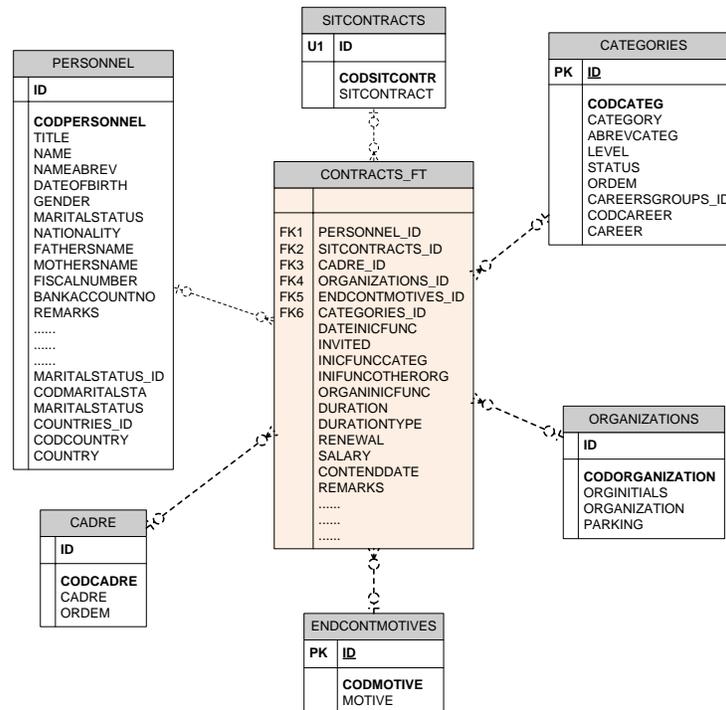


Fig. 3: Contracts Star

in many processes not all of them will be actually requiring effort. The implementation of some of the rules is immediate if the original database to be preserved is well designed and with no junk. For example, the identification of keys (Group II) will be very quick if they are implemented through database constraints and normalization (Group III) may not be required. The best opportunity to preserve a database is during its deactivation but before the system is dismantled. If the system is still able to run, several steps of the procedure become easier. If the DBMS is no longer available, or the platform to run the application, or the know-how on its use, it may become hard to restore the backups and establish the environment for the migration to take place.

V. ARCHIVAL INFORMATION PACKAGE

The AIP contains all the information considered relevant for future use. It contains the following items.

- The model migrated database is converted to SIARD format for making it completely platform-independent and then it is included in the AIP. It includes the data values with a simple model as well as some contextual metadata. SIARD Suite allows for the addition of table descriptions, column descriptions, the name and contact details of the contact person, the owner of the database and so on.
- The dimensional modeling metadata is generated using the DBPreserve Suite. The DBPreserve Suite automatically adds the metadata to the SIARD Archive.
- The preservation log containing the record of all the steps taken in the database preservation procedure. It

includes the steps taken according to the transformation rules. The log is stored in a text file.

- Screen shots of the application forms used to add new data to the database are included in the AIP. A text file explaining the application forms is included in the AIP.
- The schema of the original database and the model migrated database is also included in the AIP.

The structure of the AIP is based on the structure of the SIARD archive. It is shown in Figure 4.

VI. DATABASE DISSEMINATION

A database preserved using the proposed approach can be accessed in two ways. Firstly, the database is loaded to a DBMS using the SIARD Suite. This will require users to have the knowledge of a query language e.g. SQL. Furthermore, it will also require a DBMS installed on a machine. The other choice is the use of the DBPreserve Archive Browser (DAB). The DAB is a browser developed using Java. The use of DAB may avoid the need to load back a preserved database to a DBMS in many use cases. A user needs to select a table from one combo box thus making it the current focus (base table). The names of all the tables referenced by the base table are displayed in another combo box. Users may then open one of the tables to browse its contents. The browser provides different browsing functionalities including the following.

- **Table Filtering:** A user may filter a table by searching a string in the entire table. Moreover, the search may be made limited to a single column. The filtering is

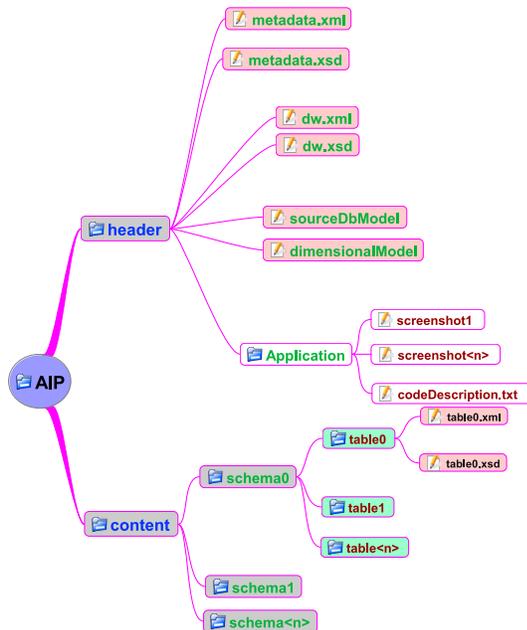


Fig. 4: Archival Information Package

done regardless of the case (upper case and lower case) of the string.

- **Following Keys:** The browser also allows navigating from one table to another table using keys. If a user clicks on a value in a foreign key column, the corresponding row in the parent table is displayed. Similarly, clicking a value in the primary key column leads to all the corresponding values in the selected base table
- **Table Joins:** The browser also allows joining the base table with the tables it references (if any).

VII. EVALUATION

Evaluation of digital preservation procedures and the result is a complicated task. The evaluation of a database preservation procedure is even more difficult as there are no studies available identifying the parameters for the evaluation. It is hard to anticipate the tools available in the future and technical expertise of the people who will be using the preserved database. Moreover, the usage of a preserved database may be different from the anticipated usage during the preservation procedure. However, some points related to any type of digital objects and procedures can be verified including the use of open formats, addition of metadata and independence from the operational environment.

The proposed approach uses the SIARD format as the archival format for a database. SIARD format is an open format used by many libraries and archives around the world for archiving databases. It allows adding some technical and non-technical metadata. However, the format does not allow to add dimensional modeling metadata. Therefore, the DWXML was developed which is also openly available. The DWXML is

added to the archive using the DBPreserve Suite. Moreover, all the steps taken for the migration of a database are documented and kept for future reference. This is helpful in controlling the loss of information and ensuring the authenticity of the database.

The structure of the database is simplified and can be easily understood by users in the long term. The simplicity of the structure also eases the development of new software for dissemination of a database preserved using the proposed approach. The DAB is an example which may be helpful in many use cases for querying a database without the need of a DBMS.

The database is not dependent on the operational environment anymore. The application software used in the active life of a database is not required in the long term to query the database. In conclusion the simplification of database structure, usage of open formats, the collection of metadata, minimizing the dependence on tools from the operational environment and preservation environment is a step further in the current state-of-the-art solutions for database preservation.

VIII. CONCLUSION

Database preservation is a complex process that needs to be carried out in a step by step and traceable manner. The loss of a single piece of information in the process may make the authenticity of a preserved database questionable. The proposed transformation rules for database preservation are implemented in a systematic manner which guarantee that no loss of information occurs in the procedure except for the irrelevant and unwanted details. The resulting preserved database has a different model from the original one. However, the model of the preserved database is simple to understand and easy to write queries against for people who did not develop it or used it in its active life.

The implementation of the transformation rules shares some intuitions and techniques with traditional data warehouse systems. However the ultimate goal of preserving a database is very different from the usual goal of building a decision-support system. This has a main consequence in the nature of the fact tables, which often lack clear measures or the measures included are just secondary elements.

The model migrated database is converted to the SIARD format which makes it completely platform independent. The proposal includes the use of a platform-independent and easy to use tool for dissemination of a preserved database. The tool does not require users to have knowledge of a query language thus avoiding the need of an expert for disseminating a preserved database. The use of the tool also avoids the need of installing and configuring a DBMS for loading a preserved database to it.

REFERENCES

- [1] R. Agrawal, A. Gupta, and S. Sarawagi, "Modeling multidimensional databases," in *Proceedings of the Thirteenth International Conference on Data Engineering*, ser. ICDE '97. Washington, DC, USA: IEEE Computer Society, 1997, pp. 232–243. [Online]. Available: <http://portal.acm.org/citation.cfm?id=645482.653299>
- [2] C. Aldeias, G. David, and C. Ribeiro, "DWXML - A Preservation Format for Data Warehouses," in *XML: Aplicações e Tecnologias Associadas*, Vila do Conde, Portugal, June 2011.

- [3] C. Ballard, D. M. Farrell, A. Gupta, C. Mazuela, and S. Vohnik, *Dimensional Modeling: In a Business Intelligence Environment*. Vervante, 2006.
- [4] C. Becker and A. Rauber, "Decision criteria in digital preservation: What to measure and how," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 6, pp. 1009–1028, June 2011. [Online]. Available: <http://dx.doi.org/10.1002/asi.21527>
- [5] D. Burda and F. Teuteberg, "Sustaining accessibility of information through digital preservation: A literature review," *J. Inf. Sci.*, vol. 39, no. 4, pp. 442–458, Aug. 2013. [Online]. Available: <http://dx.doi.org/10.1177/0165551513480107>
- [6] E. F. Codd, "Normalized data base structure: a brief tutorial," in *Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control*, ser. SIGFIDET '71. New York, NY, USA: ACM, 1971, pp. 1–17. [Online]. Available: <http://doi.acm.org/10.1145/1734714.1734716>
- [7] *Reference Model for an Open Archival Information System (OAIS)*, Consultative Committee for Space Data Systems (CCSDS) Std. ISO 14721:2003, Rev. Magenta Book, Recommended Practice, Issue 2, June 2012. [Online]. Available: <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [8] Daniel L. Moody and Mark A. R. Kortink, "From ER models to dimensional models part II: Advanced design issues," *Journal of Business Intelligence*, 2003.
- [9] Edgar Frank Codd, "A relational model of data for large shared data banks," *Communications of the ACM - Special 25th Anniversary Issue*, vol. 13, no. 6, pp. 377–387, June 1970. [Online]. Available: <http://doi.acm.org/10.1145/362384.362685>
- [10] S. Granger, "Emulation as a digital preservation strategy," Tech. Rep., 2000.
- [11] M. Guttenbrunner and A. Rauber, "A measurement framework for evaluating emulators for digital preservation," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 14:1–14:28, May 2012. [Online]. Available: <http://doi.acm.org/10.1145/2180868.2180876>
- [12] S. Heuscher, S. Järman, P. Keller-Marxer, and F. Möhle, "Providing authentic long-term archival access to complex relational data," in *Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. European Space Agency, 2004.
- [13] J. V. D. Hoeven and H. V. Wijngaarden, "Modular emulation as a long-term preservation strategy for digital objects," in *5th International Web Archiving Workshop*, 2005.
- [14] C. Imhoff, N. Galemno, and J. G. Geiger, *Mastering Data Warehouse Design: Relational and Dimensional Techniques*, R. Elliott, Ed. Joe Wikert, 2003.
- [15] W. H. Inmon, *Building the data warehouse*, 4th ed. Wellesley, MA, USA: QED Information Sciences, Inc., 2005.
- [16] W. Kent, "A simple guide to five normal forms in relational database theory," *Communications of the ACM*, vol. 26, no. 2, pp. 120–125, February 1983. [Online]. Available: <http://doi.acm.org/10.1145/358024.358054>
- [17] R. Kimball, L. Reeves, W. Thornthwaite, M. Ross, and W. Thornwaite, *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses*. New York, NY, USA: John Wiley & Sons, Inc., 1998.
- [18] R. Latham, "Digital preservation formats," Tasmanian Archive and Heritage Office, Tech. Rep., 2012.
- [19] B. Matthews, B. McIlwrath, D. Giaretta, and E. Conway, "The significant properties of software - a study," Joint Information Systems Committee (JISC), Tech. Rep., December 2008.
- [20] E. P. McLellan, "Selecting formats for digital preservation: Lessons learned during the archivematica project," *Information Standards Quarterly*, vol. 22, no. 2, pp. 30–33, 2010.
- [21] D. L. Moody and M. A. R. Kortink, "From ER models to dimensional models: Bridging the gap between OLTP and OLAP design, part I," *Journal of Business Intelligence*, 2003.
- [22] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3–13, 2000.
- [23] J. Rothenberg, "Avoiding technological quicksand: Finding a viable technical foundation for digital preservation," Council on Library and Information Resources, Washington DC, Tech. Rep., January 1999.
- [24] A. Simitsis, P. Vassiliadis, and T. Sellis, "State-space optimization of ETL workflows," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 17, no. 10, pp. 1404 – 1419, OCTOBER 2005.
- [25] S. Strodl, C. Becker, R. Neumayer, and A. Rauber, "How to choose a digital preservation strategy: evaluating a preservation planning procedure," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ser. JCDL '07. New York, NY, USA: ACM, 2007, pp. 29–38. [Online]. Available: <http://doi.acm.org/10.1145/1255175.1255181>
- [26] H. Thomas, *SIARD Suite Manual*, Swiss Federal Archives, SFA, Archivstrasse 24, 3003 Bern, Switzerland, May 2009.
- [27] R. Torlone, "Multidimensional databases," M. Rafanelli, Ed. Hershey, PA, USA: IGI Publishing, 2003, ch. Conceptual multidimensional models, pp. 69–90. [Online]. Available: <http://dl.acm.org/citation.cfm?id=887433.887438>
- [28] C. Webb, *Guidelines for the Preservation of Digital Heritage*, Information Society Division, United Nations Educational, Scientific and Cultural Organization Std., March 2003.

Detection of Denial of Service Attack in Wireless Network using Dominance based Rough Set

N. Syed Siraj Ahmed

School of Computing science and Engineering
VIT University
Vellore, India 632014

D. P. Acharjya

School of Computing science and Engineering
VIT University
Vellore, India 632014

Abstract—Denial-of-service (DoS) attack is aim to block the services of victim system either temporarily or permanently by sending huge amount of garbage traffic data in various types of protocols such as transmission control protocol, user datagram protocol, internet connecting message protocol, and hypertext transfer protocol using single or multiple attacker nodes. Maintenance of uninterrupted service system is technically difficult as well as economically costly. With the invention of new vulnerabilities to system new techniques for determining these vulnerabilities have been implemented. In general, probabilistic packet marking (PPM) and deterministic packet marking (DPM) is used to identify DoS attacks. Later, intelligent decision prototype was proposed. The main advantage is that it can be used with both PPM and DPM. But it is observed that, data available in the wireless network information system contains uncertainties. Therefore, an effort has been made to detect DoS attack using dominance based rough set. The accuracy of the proposed model obtained over the KDD cup dataset is 99.76 and it is higher than the accuracy achieved by resilient back propagation (RBP) model.

Keywords—Denial of service; Rough set; Lower and upper approximation; Dominance relation; Data analysis

I. INTRODUCTION

Denial-of-service attack is one of the most threatening security issues in wireless networks. Over the past few years, it is observed that while surfing websites on the internet a computer in the network host may have been the target of denial-of-service attacks using various protocols such as TCP, UDP, ICMP, and HTTP. Among which TCP flooding is the most prevalent [1]. This results in disruption of services at high cost. The main objective of denial-of-service attack is to consume a large amount of resources, thus preventing legitimate users from receiving service with some minimum performance. TCP flooding [1] exploits TCPs three-way handshake procedure, and specifically its limitation in maintaining half-open connections. Denial of service attack is a technique to make a host or network resource block to its intended users. The attack temporarily or permanently interrupts or suspends services of a computer in the network host connected to the Internet. A permanent denial-of-service attack damages a system so badly that it requires replacement or reinstallation of hardware such as routers, printers, or other network hardware. Hence in general, detection is required before the spread of this attack. Detection of such an attack is often a part of information security [2, 3]. Therefore, it is essential to secure wireless networks from such an attack.

A distributed denial of service (DDoS) attack is a simultaneous network attack on a victim from a large number of compromised hosts, which may be distributed widely among different, independent networks [4]. By exploiting asymmetry between network wide resources, and local capacities of a victim a DDoS attack can build up an intended congestion very quickly. The Internet routing infrastructure, which is stateless and based mainly on destination addresses, appears extremely vulnerable to such coordinated attacks. It is a type of cyber attacks in which the victim will be overloaded and will not able to perform any normal functions. Many researchers have presented their work in various directions. Gavrilis and Dermatas uses radial basis function neural network and statistical features to achieve accurate classification of abnormal activity under DDoS attack without interfering normal traffic [5]. The advantage of this method is that it can block the traffic selectively based on the attack. Wang et al. introduced a queuing model for the evaluation of the denial of service attacks in computer networks. The network is characterized by a two-dimensional embedded Markov chain model. It helps in developing a memory-efficient algorithm for finding the stationary probability distribution which can be used to find other interesting performance metrics such as connection loss probability and buffer occupancy percentages of half-open connections [6]. Gelenbe and Lukes proposed a model to defense denial of service attack using cognitive packet network infrastructure. The technique uses smart packets to select paths based on quality of service [7].

Mell introduces resistant intrusion detection system architecture to counter denial of service attack. The components of intrusion detection system architecture are invisible to the attacker and also this architecture relocates intrusion detection system components from attacked hosts. This is achieved by using mobile agent technology [8]. Hamdi uses outbound and inbound demilitarized zone to detect denial of service attack. The major advantage is that it also identifies synchronize-flooding attack [9]. Later, Chen et al., applied targeted filtering method to identify a distributed denial of service attack. The advantage is that it can be deployed at a local firewall. But, it takes extra time to detect the attack [10]. Rajkumar and Selvakumar proposed a model using Resilient back propagation (RBP) algorithm as the base classifier for the detection of denial of service attack [11]. From the literature survey, it is understood that much research is carried out for the detection of denial of service attack and distributed denial of service attack.

Denial-of-service attacks commonly block the services of legitimate user in a wireless network either temporarily or permanently by supplying either short term or long term harmful artificial traffic. Additionally, it is observed that the information system pertaining to denial-of-service attack in wireless network contains uncertainties and the attributes involved in the information system have some specific order. To deal with such uncertainties, criteria, and specific order the concept of dominance based rough set can be used. This motivation helps us to think of an alternative approach using dominance based rough set.

In this paper, we propose an alternative method using dominance based rough set for the detection of denial of service attack. The rest of the paper is organized as follows: we discuss basic concepts of dominance based rough set in section 2. Section 3 discusses dominance principle. A case study is presented in section 4 to analyze and track denial of service attack using dominance based rough set. Finally, the paper is concluded with a conclusion.

II. FOUNDATION OF INFORMATION SYSTEM

An information system provides an expedient to describe a finite set of objects called the universe with a finite set of attributes thereby representing all the available information and knowledge. Formally, it is defined as a four tuple $T = (U, A, V, f)$ where $U = \{x_1, x_2, \dots, x_n\}$ is a non-empty finite set of objects called the universe, $A = \{a_1, a_2, \dots, a_n\}$ is a nonempty finite set of attributes. The component V is defined as $V = \cup_{a \in A} V_a$, where V_a is the set of attribute values that an attribute a may take. The component $f : (U \times A) \rightarrow V$ is an information function. The information system is said to be a decision system if $A = C \cup \{d\}$, $C \neq \phi$, $\{d\} \neq \phi$ and $C \cap \{d\} = \phi$ where C is a set of conditional attributes and d is the decision [12].

Let $B \subseteq A$. Two objects x_i and x_j are said to be B -indiscernible if $f(x_i, a) = f(x_j, a)$ for all $a \in B$. Mathematically, we denote it as $IND(B)$ is defined as below and we write $x_i I_B x_j$.

$$IND(B) = \{(x_i, x_j) \in U^2 : f(x_i, a) = f(x_j, a) \forall a \in B\}$$

Object x_j dominates object x_i on criteria a if $V_a^{x_j} \leq V_a^{x_i}$, where $V_a^{x_j}$ is the attribute value of object x_j on criteria a . Let $Q \subseteq C$ be a criteria set. Let us define a dominance relation $dm(Q)$ on U as

$$dm(Q) = \{(x_i, x_j) \in U^2 : V_a^{x_j} \leq V_a^{x_i} \forall a \in Q\} \quad (1)$$

If $(x_i, x_j) \in dm(Q)$, then we write $x_j D_Q x_i$. Let $P \subseteq C$ is a criteria set. Let us define $D_P^+(x_i)$, P -dominating x_i as below.

$$D_P^+(x_i) = \{x_j \in U : x_j D_P x_i\} \quad (2)$$

Similarly, we define a set $D_P^-(x_i)$, P -dominated by x_i as below.

$$D_P^-(x_i) = \{x_j \in U : x_i D_P x_j\} \quad (3)$$

Two objects x_i and x_j are said to be inconsistent, if their criteria do not satisfy dominance principle with ordered decision class [13].

III. DOMINANCE BASED ROUGH SET

Rough set of Pawlak is a mathematical tool used in data analysis in particular to analyze uncertainties [14]. But it fails to analyze data containing preference order and may lead to loss of information. To overcome the limitations the concept of dominance based rough set is introduced [15, 16, 17]. In dominance based rough set, given a set of objects, there is a criterion at least among condition attributes. Additionally, attributes like color, country may not be of preference order. Therefore, criteria attributes are divided into ordered decision classes based on decision attribute. Also criteria in condition attributes are correlated semantically with ordered decision attribute by means of dominance relation.

Formally, dominance based rough set (DRS) is based on the concept of dominance principle to extract knowledge from the information system. Here, the classification is carried out based on decision class (d). Therefore, the decision (d) divides the universe U into a finite number of classes, CL , such as

$$CL = \{CL_i : i \in T\}; T = \{1, 2, 3, \dots, m\}$$

Additionally, these classes are ordered. It means that, if $r, s \in T$ and $r > s$, then the objects of class CL_r are preferred then the objects of class CL_s . The upward and downward unions of every element CL_i of CL is given as CL_i^{\geq} and CL_i^{\leq} respectively. Mathematically, it is defined as

$$CL_i^{\geq} = \cup_{j \geq i} CL_j; CL_i^{\leq} = \cup_{j \leq i} CL_j$$

Let $Q \subseteq C$, objects certainly belong to CL_i^{\geq} and CL_i^{\leq} are in their lower approximations $\underline{Q}(CL_i^{\geq})$ and $\underline{Q}(CL_i^{\leq})$ respectively. The lower approximations are defined as below.

$$\underline{Q}(CL_i^{\geq}) = \{x \in U : D_Q^+ \subseteq CL_i^{\geq}\} \quad (4)$$

$$\underline{Q}(CL_i^{\leq}) = \{x \in U : D_Q^- \subseteq CL_i^{\leq}\} \quad (5)$$

Similarly, objects possibly belong to CL_i^{\geq} and CL_i^{\leq} are in their upper approximations $\overline{Q}(CL_i^{\geq})$ and $\overline{Q}(CL_i^{\leq})$ respectively. It is defined as below.

$$\overline{Q}(CL_i^{\geq}) = \bigcup_{x \in CL_i^{\geq}} D_Q^+(x) \quad (6)$$

$$\overline{Q}(CL_i^{\leq}) = \bigcup_{x \in CL_i^{\leq}} D_Q^-(x) \quad (7)$$

The boundary region of CL_i^{\geq} and CL_i^{\leq} , which contains ambiguous elements are defined as below

$$BN_Q(CL_i^{\geq}) = \overline{Q}(CL_i^{\geq}) - \underline{Q}(CL_i^{\geq})$$

$$BN_Q(CL_i^{\leq}) = \overline{Q}(CL_i^{\leq}) - \underline{Q}(CL_i^{\leq})$$

A. Dominance Relation Based Rule Formation

For a given information system, the dominance principle is capable of deducing more generalized description of objects. This can be done by means of upward and downward union of rough approximation. This is a fundamental concept in a knowledge discovery.

Let $Q \subseteq C$ be a conditional attributes. Based on the rough approximation, the Q -lower and Q -upper approximations are computed on criterion attribute to extract the knowledge. The rules generated from criterion attribute using upward and downward union of Q -lower, Q -upper approximations are of the form "If Condition then Decision".

In real life situation, the data collected may be uncertain, vague and imprecise which may leads to inconsistency. The inconsistency data are identified in rough set by means of indiscernible relation. Likewise the inconsistency presents in the collected data are identified in dominance based rough set on employing dominance relation. The two objects are said to be inconsistent when the criteria attributes do not satisfy dominance principle with decision attribute. Further such inconsistency exists in logic must be removed try as it leads to error decision. The simplest way to remove such inconsistency is to omit the inconsistent objects. The five kinds of determinate rules associated with dominance based rough set are defined as follows [13].

- 1) For all criteria $a_i \in Q \subseteq C$; if $f(x, a_1) \geq V_{a_1}^x$ and $f(x, a_2) \geq V_{a_2}^x$ and $\dots f(x, a_i) \geq V_{a_i}^x$, then $x \in Cl_t^{\geq}$ where $t \in \{2, 3, \dots, n\}$. Rules generated in such way called as certain D_{\geq} decision rules. These rules are obtained from $\underline{Q}(Cl_t^{\geq})$.
- 2) For all criteria $a_i \in Q \subseteq C$; if $f(x, a_1) \geq V_{a_1}^x$ and $f(x, a_2) \geq V_{a_2}^x$ and $\dots f(x, a_i) \geq V_{a_i}^x$, then $x \in Cl_t^{\geq}$ where $t \in \{2, 3, \dots, n\}$. Rules generated in such way called as possible D_{\geq} decision rules. These rules are obtained from $\overline{Q}(Cl_t^{\geq})$.
- 3) For all criteria $a_i \in Q \subseteq C$; if $f(x, a_1) \leq V_{a_1}^x$ and $f(x, a_2) \leq V_{a_2}^x$ and $\dots f(x, a_i) \leq V_{a_i}^x$, then $x \in Cl_t^{\leq}$ where $t \in \{1, 2, \dots, (n-1)\}$. Rules generated in such way called as certain D_{\leq} decision rules. These rules are obtained from $\underline{Q}(Cl_t^{\leq})$.
- 4) For all criteria $a_i \in Q \subseteq C$; if $f(x, a_1) \leq V_{a_1}^x$ and $f(x, a_2) \leq V_{a_2}^x$ and $\dots f(x, a_i) \leq V_{a_i}^x$, then $x \in Cl_t^{\leq}$ where $t \in \{1, 2, \dots, (n-1)\}$. Rules generated in such way called as possible D_{\leq} decision rules. These rules are obtained from $\overline{Q}(Cl_t^{\leq})$.
- 5) Let $O_1 = \{a_1, a_2, \dots, a_k\} \subseteq C$; $O_2 = \{a_{k+1}, a_{k+2}, \dots, a_i\} \subseteq C$; $Q = (O_1 \cup O_2)$; O_1 and O_2 are not necessarily disjoint. If $f(x, a_1) \geq V_{a_1}^x$ and $f(x, a_2) \geq V_{a_2}^x, \dots$, and $f(x, a_k) \geq V_{a_k}^x$ and $f(x, a_{k+1}) \leq V_{a_{k+1}}^x$ and $f(x, a_{k+2}) \leq V_{a_{k+2}}^x, \dots$ and $f(x, a_i) \leq V_{a_i}^x$, then $x \in Cl_u \cup Cl_{u+1} \cup \dots \cup Cl_v$, where $r \leq u \leq v \leq t$ and $r, u, v, t \in T$. Rules generated in such way called as approximate $D_{\geq \leq}$ decision rules. These rules are obtained from $\overline{Q}(Cl_r^{\leq}) \cap \underline{Q}(Cl_t^{\geq})$.

The rules 1 and 3 represent certain knowledge whereas rules 2 and 4 represent possible knowledge that can be ex-

tracted from the information system. The rules 5 represent ambiguous knowledge. If $y \in \underline{Q}(Cl_t^{\geq})$ such that $f(y, a_1) = V_{a_1}^y$, $f(y, a_2) = V_{a_2}^y, \dots, f(y, a_i) = V_{a_i}^y$, then y is called as basis of the rule. An object which matches both condition and decision parts of a rule supports the decision rule. An object which meets only condition part of a rule is covered by a decision rule. Decision rules either certain or approximate is said to be complete if it satisfies following conditions.

- 1) Each $x \in \underline{Q}(Cl_t^{\geq})$ must support at least one certain D_{\geq} decision rule whose consequent is $x \in Cl_r^{\geq}$ where $r, t \in \{2, 3, \dots, n\}$ and $r \geq t$.
- 2) Each $x \in \overline{Q}(Cl_t^{\leq})$ must support at least one certain D_{\leq} decision rule whose consequent is $x \in Cl_r^{\leq}$ where $r, t \in \{1, 2, \dots, (n-1)\}$ and $r \leq t$.
- 3) Each $x \in (\overline{Q}(Cl_r^{\leq}) \cap \underline{Q}(Cl_t^{\geq}))$ must support at least one approximate $D_{\geq \leq}$ decision rule whose consequent is $x \in Cl_u \cup Cl_{u+1} \cup \dots \cup Cl_v$ where $r \leq u \leq v \leq t$ and $r, u, v, t \in T$.

It means that, the set of rules must cover all objects of the information system. Additionally, it assigns consistent objects to their original classes and inconsistent objects to clusters of classes pertaining to this inconsistency.

IV. PROPOSED RESEARCH DESIGN

A common type of attack used to block the service of the wireless network in recent years is denial of service attack. Therefore, recognizing such an attack is of great challenge. To this end, in this section, we purpose our research design for detecting dos attack. The following Figure 1 depicts an abstract view of the model. The initial step of any model development is problem identification that includes basic knowledge of the problem undertaken. The data collected initially preprocessed. The main objective is to transform the raw input data into an appropriate format for subsequent analysis. The various steps involved are merging of data from data repositories, data cleaning which removes noise and duplicate observations and then selecting relevant observations as per the requirement of the problem undertaken. The selection of observations is done in order to analyze only one decision denial-of-service. The processed data is partitioned into two categories such as training data of 55% and testing data of 45%. The training data is analyzed using dominance based rough set to identify the decision class that effects the decision. We apply DOMLEM algorithm to obtain the rules. algorithm:

A. DOMLEM Algorithm

In rough set theory several algorithms are proposed for induction of decision rules [18, 19, 20]. Some of these algorithms also generate minimum number of rules. Generally, we use heuristic approach to deduce rules because of NP-hard nature [18]. In this paper we use DOMLEM algorithm as proposed by Greco et al [13] for the detection of denial-of-service attack. The algorithm is repeatedly applied for all lower or upper approximations of the upward (downward) unions of decision classes. Considering preference order of decision classes and of getting minimum rules, the algorithm is applied repeatedly starting from the strongest union of classes. Therefore, decision rules of the lower approximations of upward unions of classes

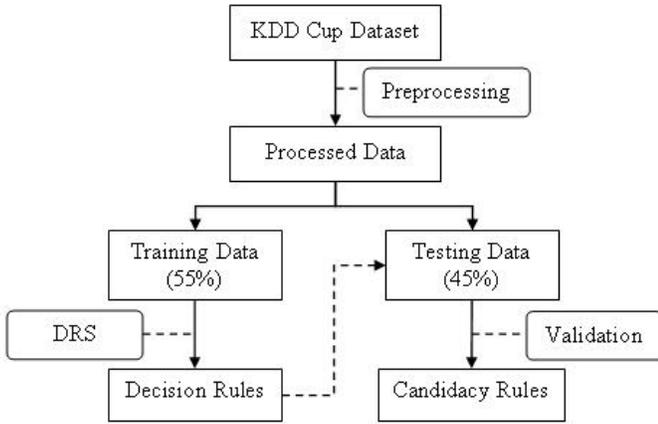


Fig. 1: Abstract View of Research Design

should be taken into consideration in decreasing order. The following notations are used in the DOMLEM algorithm.

C : Denotes set of conditional attributes

Q : Denotes set of criteria, $a_i \in Q \subseteq C$

E : Denotes conjunction of elementary conditions $e = \{f(x, a_i) \geq V_{a_i}^x\}$

$[E]$: Denotes set of objects in E ; $[E] = \{x : f(x, a_i) \geq V_{a_i}^x\}$

FM_k : Denotes the first measure

SM_k : Denotes the second measure

Algorithm 1: (DOMLEM)

Input: Lower approximation of upward union; $\underline{Q}(Cl_i^{\geq}), i = m, (m-1), \dots, 2$

Output: Set of D_{\geq} decision rules

Begin

$$D_{\geq} = \phi$$

for each $\underline{Q}(Cl_i^{\geq})$, do

$$\mathbf{E} = \text{Find_Rules}(\underline{Q}(Cl_i^{\geq}))$$

for each rule $r \in \mathbf{E}$, do

if r is a minimal rule, then $D_{\geq} = D_{\geq} \cup \{r\}$

End

Function Find_Rules

Begin

$$G = \underline{Q}(Cl_i^{\geq})$$

$$\mathbf{E} = \phi$$

while $G \neq \phi$, do

$$E = \phi$$

$$S = G$$

while $E = \phi$ or not ($[E] \subseteq \underline{Q}(Cl_i^{\geq})$), do

$$best = \phi$$

for each criteria $a_i \in Q$ do

$$Cond = \{f(x, a_i) \geq V_{a_i}^x : \exists x \in S, f(x, a_i) = V_{a_i}^x\}$$

for each $e_k \in Cond$, do

$$FM_k = |[e_k] \cap G| / |[e_k]|$$

$$SM_k = |[e_k] \cap G|$$

find e_k for which FM_k and SM_k is maximum

$$best = best \cup \{e_k\}$$

end for

end for

$$E = E \cup \{best\}$$

$$S = S \cap [best]$$

end while

for each $e_k \in E$, do

if $[E - \{e_k\}] \subseteq \underline{Q}(Cl_i^{\geq})$, then $E = E - \{e_k\}$

$$\mathbf{E} = \mathbf{E} \cup \{E\}$$

$$G = \underline{Q}(Cl_i^{\geq}) - \cup_{e \in E} [E]$$

end while

End

B. An Illustration of DOMLEM Algorithm

This section explains how the above concepts can be applied in analyzing denial-of-service attack in a wireless network. To analyze the above concepts, we have considered the dataset discussed by various authors in their papers [15, 21, 22, 23]. We present the dataset in the following Table 1. The various attributes considered are packets received or sent per seconds (Mbps), number of attacker nodes, types of protocol, service block period, and damage. We denote these attributes as a_1, a_2, a_3, a_4 , and a_5 respectively. The attribute a_3 may take values TCP, UDP, or ICMP. Similarly, different values the attribute a_4 may take are zero (Zo), short (So), long (Lo), or permanent (Pt). Finally, the different values that the attribute a_5 may take are hardware fail (HF), software fail (SF), system hang (SH), system reset (SR), time waste (TW), or no damage (ND). The decision attribute (d) describes category of denial of service attack such as permanent denial of service attack (PDA), distributed denial of service attack (DDA), simple denial of service attack (SDA), and no attack (NA). Consider the attributes $Q = \{a_1, a_2, a_4\}$ as criteria among all conditional attributes a_1, a_2, a_3, a_4, a_5 .

The above table contains 13 objects of denial-of-service attack in a wireless network and its various conditional attribute values, where U denotes node number. For analysis purpose, the dataset is divided into two training dataset of 7 objects (55%) and testing dataset of 6 objects (45%). We employ dominance based rough set data analysis on training dataset to obtain candidacy classes. The testing dataset is used to detect over fitting of the decision classes based on the predefined threshold value 70%. The decision divides the training dataset of universe into finite number of classes, CL , as below.

$$CL = \{Cl_1, Cl_2, Cl_3, Cl_4\}$$

where $Cl_1 = \{x_1, x_7\}$; $Cl_2 = \{x_2, x_6\}$; $Cl_3 = \{x_3\}$ and $Cl_4 = \{x_4, x_5\}$. It is also observed that the class Cl_4 has more

TABLE I: An information system of denial-of-service attack in a wireless network

U	a_1	a_2	a_3	a_4	a_5	d
x_1	1.3	2	TCP	Zo	ND	NA
x_2	2.67	1	UDP	So	TW	SDA
x_3	2.5	4	ICMP	Lo	SF	DDA
x_4	3.0	5	UDP	Lo	SH	PDA
x_5	2.4	2	TCP	Pt	SF	PDA
x_6	2.6	7	TCP	Lo	SF	SDA
x_7	2.68	1	ICMP	So	ND	NA
x_8	3.1	4	UDP	Lo	SR	DDA
x_9	2.68	1	ICMP	So	TW	SDA
x_{10}	2.5	6	UDP	Pt	HF	PDA
x_{11}	2.7	2	TCP	So	TW	SDA
x_{12}	3.2	3	ICMP	Lo	SH	NA
x_{13}	1.5	0	UDP	Zo	ND	NA

delay than Cl_3 ; Cl_3 has more delay than Cl_2 ; and Cl_2 has more delay than Cl_1 . The downward unions of every element Cl_i , $i = 1, 2, 3$ of CL are given below.

$$\begin{aligned} Cl_1^{\leq} &= \{x_1, x_7\} \\ Cl_2^{\leq} &= \cup_{j \leq 2} Cl_j = Cl_1 \cup Cl_2 = \{x_1, x_2, x_6, x_7\} \\ Cl_3^{\leq} &= \{x_1, x_2, x_3, x_6, x_7\} \end{aligned}$$

Similarly, the upward unions of training dataset element Cl_i , $i = 4, 3, 2$ of CL are given below.

$$\begin{aligned} Cl_4^{\geq} &= \cup_{j \geq 4} Cl_j = Cl_4 = \{x_4, x_5\} \\ Cl_3^{\geq} &= \cup_{j \geq 3} Cl_j = Cl_3 \cup Cl_4 = \{x_3, x_4, x_5\} \\ Cl_2^{\geq} &= \{x_2, x_3, x_4, x_5, x_6\} \end{aligned}$$

Let us consider the downward union $Cl_1^{\leq} = \{x_1, x_7\}$ on considering the criteria $Q = \{a_1, a_2, a_4\} \subseteq C$, the lower and upper approximations are given as $\underline{Q}(Cl_1^{\leq}) = \{x_1\}$ and $\overline{Q}(Cl_1^{\leq}) = \{x_1, x_2, x_7\}$ respectively. Therefore, the boundary objects are $BN_Q(Cl_1^{\leq}) = \{x_2, x_7\}$. It is because the objects x_2 and x_7 violates the dominance principle. This can be seen from the information system presented in Table I. From Table 1, it is clear that object x_7 dominates object x_2 on criteria Q , but the decision corresponding to the object x_7 is finer then the decision corresponding to the object x_2 . Hence, they are inconsistent. Also, it can be shown that objects x_3 and x_6 are also inconsistent. Similarly the lower, upper approximations, and boundary of downward and upward unions of other classes are presented below.

$$\begin{aligned} \underline{Q}(Cl_2^{\leq}) &= \{x_1, x_2, x_7\}, \overline{Q}(Cl_2^{\leq}) = \{x_1, x_2, x_3, x_6, x_7\} \\ BN_Q(Cl_2^{\leq}) &= \{x_3, x_6\} \\ \underline{Q}(Cl_3^{\leq}) &= \{x_1, x_2, x_3, x_6, x_7\}, \\ \overline{Q}(Cl_3^{\leq}) &= \{x_1, x_2, x_3, x_6, x_7\}, BN_Q(Cl_3^{\leq}) = \{\phi\} \\ \underline{Q}(Cl_4^{\geq}) &= \{x_4, x_5\}, \overline{Q}(Cl_4^{\geq}) = \{x_4, x_5\} \\ BN_Q(Cl_4^{\geq}) &= \{\phi\} \\ \underline{Q}(Cl_3^{\geq}) &= \{x_4, x_5\}, \overline{Q}(Cl_3^{\geq}) = \{x_3, x_4, x_5, x_6\} \\ BN_Q(Cl_3^{\geq}) &= \{x_3, x_6\} \\ \underline{Q}(Cl_2^{\geq}) &= \{x_3, x_4, x_5, x_6\} \\ \overline{Q}(Cl_2^{\geq}) &= \{x_2, x_3, x_4, x_5, x_6, x_7\}, BN_Q(Cl_2^{\geq}) = \{x_2, x_7\} \end{aligned}$$

Now, we explain how certain D_{\geq} decision rules are induced for the upward union. Let us consider the class Cl_4^{\geq} and the lower approximation $\underline{Q}(Cl_4^{\geq}) = \{x_4, x_5\}$ for obtaining D_{\geq} decision rules. Employing the DOMLEM algorithm on $\underline{Q}(Cl_4^{\geq})$, we get the elementary conditions as below.

$$\begin{aligned} e_1 &= \{f(x, a_1) \geq 3.0\} = \{x_4\}; 1/1; 1 \\ e_2 &= \{f(x, a_1) \geq 2.4\} = \{x_2, x_3, x_4, x_5, x_6, x_7\}; 2/6; 2 \\ e_3 &= \{f(x, a_2) \geq 2.0\} = \{x_1, x_3, x_4, x_5, x_6\}; 2/5; 2 \\ e_4 &= \{f(x, a_2) \geq 5.0\} = \{x_4, x_6\}; 1/2; 1 \\ e_5 &= \{f(x, a_4) \geq Lo\} = \{x_3, x_4, x_5, x_6\}; 2/4; 2 \\ e_6 &= \{f(x, a_4) \geq Pt\} = \{x_5\}; 1/1; 1 \end{aligned}$$

The elementary conditions e_1, e_6 produce the highest first measure and second measure. But, both elementary conditions covers only one distinct positive example. Further both $[e_1]$, $[e_6]$ are the subsets of $\underline{Q}(Cl_4^{\geq})$. We choose elementary condition e_1 initially which covers the object x_4 and is used to introduce the rule. However, we can also choose the elementary condition e_6 . Further, the object x_4 is removed from G and the remaining object is to be covered is x_5 . Thus, we have 4 elementary conditions as below to cover the object x_5 .

$$\begin{aligned} e_7 &= \{f(x, a_2) \geq 2.0\} = \{x_1, x_3, x_5, x_6\}; 1/4; 1 \\ e_8 &= \{f(x, a_1) \geq 2.4\} = \{x_2, x_3, x_5, x_6, x_7\}; 1/5; 1 \\ e_9 &= \{f(x, a_4) \geq Lo\} = \{x_3, x_5, x_6\}; 1/3; 1 \\ e_{10} &= \{f(x, a_4) \geq Pt\} = \{x_5\}; 1/1; 1 \end{aligned}$$

Next, we can pick the elementary condition e_{10} because of the highest first and second measure which covers the object x_5 . Thus no need to proceed further and the rule can be written as:

$$\begin{aligned} \text{if } f(x, a_1) \geq 3.0, \text{ then } x \in Cl_4^{\geq} \\ \text{if } f(x, a_4) \geq Pt, \text{ then } x \in Cl_4^{\geq} \end{aligned}$$

Similarly, consider $\underline{Q}(Cl_2^{\geq})$ to obtain the rules for the class $x \in Cl_2^{\geq}$. On employing the DOMLEM algorithm we get the following elementary conditions.

$$\begin{aligned} e_1 &= \{f(x, a_1) \geq 2.5\} = \{x_2, x_3, x_4, x_6, x_7\}; 3/5; 3 \\ e_2 &= \{f(x, a_1) \geq 3\} = \{x_4\}; 1/1; 1 \\ e_3 &= \{f(x, a_1) \geq 2.4\} = \{x_2, x_3, x_4, x_5, x_6, x_7\}; 4/6; 4 \\ e_4 &= \{f(x, a_1) \geq 2.6\} = \{x_2, x_6, x_7\}; 1/3; 1 \\ e_5 &= \{f(x, a_2) \geq 4\} = \{x_3, x_4, x_6\}; 3/3; 1 \\ e_6 &= \{f(x, a_2) \geq 5\} = \{x_4, x_6\}; 2/2; 2 \\ e_7 &= \{f(x, a_2) \geq 2\} = \{x_1, x_3, x_4, x_5, x_6\}; 4/5; 4 \\ e_8 &= \{f(x, a_2) \geq 7\} = \{x_6\}; 1/1; 1 \\ e_9 &= \{f(x, a_4) \geq Lo\} = \{x_3, x_4, x_5, x_6\}; 4/4; 4 \\ e_{10} &= \{f(x, a_4) \geq Pt\} = \{x_5\}; 1/1; 1 \end{aligned}$$

The elementary conditions e_2, e_5, e_6, e_8 , and e_9 have the highest first measure but the elementary condition e_9 has the highest second measure and so we choose the elementary condition e_9 . Further $[e_9]$ is subset of $\underline{Q}(Cl_2^{\geq})$ and covers all

positive examples. Thus the process terminates and the rule can be written as:

$$\text{if } f(x, a_4) \geq \text{Lo, then } x \in Cl_2^{\geq}$$

Likewise, we explain how certain D_{\leq} decision rules are induced for the downward union. Let us consider the class Cl_1^{\leq} and the lower approximation $\underline{Q}(Cl_1^{\leq}) = \{x_1\}$ for obtaining D_{\leq} decision rules. The elementary conditions obtained are given below.

$$e_1 = \{f(x, a_1) \leq 1.3\} = \{x_1\}; 1/1; 1$$

$$e_2 = \{f(x, a_4) \leq Zo\} = \{x_1\}; 1/1; 1$$

$$e_3 = \{f(x, a_2) \leq 2\} = \{x_1, x_2, x_7\}; 1/3; 1$$

The elementary conditions e_1 , and e_2 have the highest first measure and covers all the positive examples. Further both $[e_1]$, $[e_2]$ are subsets of $\underline{Q}(Cl_1^{\leq})$. Therefore, the process terminates and the rules can be stated as:

$$\text{if } f(x, a_4) \leq Zo, \text{ then } x \in Cl_1^{\leq}$$

$$\text{if } f(x, a_1) \leq 1.3, \text{ then } x \in Cl_1^{\leq}$$

Similarly, we consider $\underline{Q}(Cl_2^{\leq}) = \{x_1, x_2, x_7\}$ to obtain the rules for the class Cl_2^{\leq} . The elementary conditions obtained are listed below.

$$e_1 = \{f(x, a_1) \leq 1.3\} = \{x_1\}; 1/1; 1$$

$$e_2 = \{f(x, a_1) \leq 2.67\} = \{x_1, x_2, x_3, x_5, x_6\}; 2/5; 2$$

$$e_3 = \{f(x, a_1) \leq 2.68\} = \{x_1, x_2, x_3, x_5, x_6, x_7\}; 3/6; 3$$

$$e_4 = \{f(x, a_2) \leq 2\} = \{x_1, x_2, x_7\}; 3/3; 3$$

$$e_5 = \{f(x, a_2) \leq 1\} = \{x_2, x_7\}; 2/2; 2$$

$$e_6 = \{f(x, a_4) \leq Zo\} = \{x_1\}; 1/1; 1$$

$$e_7 = \{f(x, a_4) \leq So\} = \{x_1, x_2, x_7\}; 3/3; 3$$

The elementary conditions e_1, e_4 , and e_7 have the highest first measure and the condition e_1 covers only one positive example. Alternatively, the conditions e_4 , and e_7 have the highest second measure and covers all the positive examples. Further, both $[e_4]$, and $[e_7]$ are subsets of $\underline{Q}(Cl_2^{\leq})$. Therefore, the process terminates and the rule can be stated as:

$$\text{if } f(x, a_2) \leq 2, \text{ then } x \in Cl_2^{\leq}$$

$$\text{if } f(x, a_4) \leq So, \text{ then } x \in Cl_2^{\leq}$$

Likewise, consider $\underline{Q}(Cl_3^{\leq}) = \{x_1, x_2, x_3, x_6, x_7\}$ to obtain the decision rules for the class Cl_3^{\leq} . The elementary conditions obtained are listed below.

$$e_1 = \{f(x, a_1) \leq 1.3\} = \{x_1\}; 1/1; 1$$

$$e_2 = \{f(x, a_1) \leq 2.67\} = \{x_1, x_2, x_3, x_5, x_6\}; 4/5; 4$$

$$e_3 = \{f(x, a_1) \leq 2.6\} = \{x_1, x_3, x_5, x_6\}; 3/4; 3$$

$$e_4 = \{f(x, a_1) \leq 2.68\} = \{x_1, x_2, x_3, x_5, x_6, x_7\}; 5/6; 5$$

$$e_5 = \{f(x, a_1) \leq 2.5\} = \{x_1, x_3, x_5\}; 2/3; 2$$

$$e_6 = \{f(x, a_2) \leq 1\} = \{x_2, x_7\}; 2/2; 2$$

$$e_7 = \{f(x, a_2) \leq 2\} = \{x_1, x_2, x_7\}; 3/3; 3$$

$$e_8 = \{f(x, a_2) \leq 4\} = \{x_1, x_2, x_3, x_5, x_7\}; 4/5; 4$$

$$e_9 = \{f(x, a_2) \leq 7\} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}; 5/7; 5$$

$$e_{10} = \{f(x, a_4) \leq Zo\} = \{x_1\}; 1/1; 1$$

$$e_{11} = \{f(x, a_4) \leq So\} = \{x_1, x_2, x_7\}; 3/3; 3$$

$$e_{12} = \{f(x, a_4) \leq Lo\} = \{x_1, x_2, x_3, x_4, x_6, x_7\}; 5/6; 5$$

The elementary conditions e_1, e_6, e_7, e_{10} , and e_{11} have highest first measure whereas e_7 and e_{11} have highest second measure. But, both elementary conditions e_7 and e_{11} covers same positive examples. Further both $[e_7]$, and $[e_{11}]$ are the subsets of $\underline{Q}(Cl_3^{\leq})$. Therefore, we can choose either of the elementary conditions e_7 and e_{11} . Let us choose the elementary condition e_7 that covers objects x_1, x_2 , and x_7 . To proceed further, the objects x_1, x_2 , and x_7 are removed from G and the process is repeated. The remaining objects are to be covered are x_3 , and x_6 . Therefore, the above elementary conditions leads to 7 elementary conditions as below.

$$e_{13} = \{f(x, a_1) \leq 2.67\} = \{x_3, x_5, x_6\}; 2/3; 2$$

$$e_{14} = \{f(x, a_1) \leq 2.6\} = \{x_3, x_5, x_6\}; 2/3; 2$$

$$e_{15} = \{f(x, a_1) \leq 2.68\} = \{x_3, x_5, x_6\}; 2/3; 2$$

$$e_{16} = \{f(x, a_1) \leq 2.5\} = \{x_3, x_5\}; 1/2; 1$$

$$e_{17} = \{f(x, a_2) \leq 4\} = \{x_3, x_5\}; 1/2; 1$$

$$e_{18} = \{f(x, a_2) \leq 7\} = \{x_3, x_4, x_5, x_6\}; 2/4; 2$$

$$e_{19} = \{f(x, a_4) \leq Lo\} = \{x_3, x_4, x_6\}; 2/3; 2$$

The elementary conditions e_{13}, e_{14}, e_{15} , and e_{19} have the highest first measure. Also, the second measure of these conditions are same. But, it is not sufficient to create decision rules using any of the conditions because all these conditions cover objects either x_5 or x_4 which is a negative example. Therefore, one has to consider complexes $(e_{13} \cap e_{19})$, $(e_{14} \cap e_{19})$, and $(e_{15} \cap e_{19})$. All the complexes have highest first measure and covers positive examples. Therefore, we get the following decision rules.

$$\text{if } f(x, a_2) \leq 2, \text{ then } x \in Cl_3^{\leq}$$

$$\text{if } f(x, a_4) \leq So, \text{ then } x \in Cl_3^{\leq}$$

$$\text{if } f(x, a_1) \leq 2.67 \text{ and } f(x, a_4) \leq Lo, \text{ then } x \in Cl_3^{\leq}$$

$$\text{if } f(x, a_1) \leq 2.6 \text{ and } f(x, a_4) \leq Lo, \text{ then } x \in Cl_3^{\leq}$$

$$\text{if } f(x, a_1) \leq 2.68 \text{ and } f(x, a_4) \leq Lo, \text{ then } x \in Cl_3^{\leq}$$

Now we explain how approximate D_{\geq} approximate decision rules are induced form $\overline{Q}(Cl_1^{\leq}) \cap \overline{Q}(Cl_2^{\leq}) = \{x_2, x_7\}$. Let $O_1 = \{a_1, a_2\}$ and $O_2 = \{a_1, a_4\}$. The elementary conditions obtained are listed below.

$$e_1 = \{f(x, a_1) \geq 2.67\} = \{x_2, x_4, x_7\}; 2/3; 2$$

$$e_2 = \{f(x, a_1) \geq 2.68\} = \{x_4, x_7\}; 1/2; 1$$

$$e_3 = \{f(x, a_2) \geq 1\} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}; 2/7; 2$$

$$e_4 = \{f(x, a_1) \leq 2.67\} = \{x_1, x_2, x_3, x_5, x_6\}; 1/5; 1$$

$$e_5 = \{f(x, a_1) \leq 2.68\} = \{x_1, x_2, x_3, x_5, x_6, x_7\}; 2/6; 2$$

$$e_6 = \{f(x, a_4) \leq So\} = \{x_1, x_2, x_7\}; 2/3; 2$$

The elementary conditions e_1 , and e_6 produces the highest first measure. But, both elementary conditions e_1 and e_6 covers

the positive and negative example. Further both $[e_1], [e_6]$ are not the subsets of $\overline{Q}(Cl_1^{\leq}) \cap \overline{Q}(Cl_2^{\geq})$. Thus one has to consider complex $(e_1 \cap e_6)$. It is also a subset of $\overline{Q}(Cl_1^{\leq}) \cap \overline{Q}(Cl_1^{\leq})$. Additionally, it produces the highest first and second measure. Therefore, the rule can be stated as below:

if $(f(x, a_1) \geq 2.67$ and $f(x, a_4) \leq \text{So})$ then $x \in Cl_1 \cup Cl_2$

Similarly, on considering $\overline{Q}(Cl_2^{\leq}) \cap \overline{Q}(Cl_3^{\geq}) = \{x_3, x_6\}$ and O_1, O_2 as stated above, the approximate D_{\geq} rules are computed. The elementary conditions obtained are listed below.

$$e_1 = \{f(x, a_1) \geq 2.5\} = \{x_2, x_3, x_4, x_6, x_7\}; 2/5; 2$$

$$e_2 = \{f(x, a_1) \geq 2.6\} = \{x_2, x_4, x_6, x_7\}; 1/4; 1$$

$$e_3 = \{f(x, a_2) \geq 4\} = \{x_3, x_4, x_6\}; 2/3; 2$$

$$e_4 = \{f(x, a_2) \geq 7\} = \{x_6\}; 1/1; 1$$

The elementary condition e_4 produces the highest first measure, covers positive example, and $[e_4]$ is a subsets of $\overline{Q}(Cl_2^{\leq}) \cap \overline{Q}(Cl_3^{\geq})$. Therefore, the elementary condition e_4 is considered to generate rule. Further, the object x_6 is removed and elementary conditions are obtained to include the object x_3 .

$$e_5 = \{f(x, a_1) \geq 2.5\} = \{x_2, x_3, x_4, x_7\}; 1/4; 1$$

$$e_6 = \{f(x, a_2) \geq 4\} = \{x_3, x_4\}; 1/2; 1$$

$$e_7 = \{f(x, a_1) \leq 2.5\} = \{x_1, x_3, x_5\}; 1/3; 1$$

$$e_8 = \{f(x, a_1) \leq 2.6\} = \{x_1, x_2, x_3, x_5\}; 1/4; 1$$

$$e_9 = \{f(x, a_4) \leq \text{Lo}\} = \{x_1, x_2, x_3, x_4, x_7\}; 1/5; 1$$

The elementary conditions e_6 produces the highest first measure, covers both positive and negative example, and is not a subset of $\overline{Q}(Cl_2^{\leq}) \cap \overline{Q}(Cl_3^{\geq})$. Thus we have to consider the complex $(e_6 \cap e_7)$ to cover the positive example x_3 . The rules generated in this way are listed below.

if $(f(x, a_2) \geq 7)$ then $x \in Cl_2 \cup Cl_3$

if $(f(x, a_1) \leq 2.5$ and $f(x, a_2) \geq 4)$ then $x \in Cl_2 \cup Cl_3$

Now, collectively we write the decision rules obtained as below.

- 1) if $f(x, a_1) \geq 3.0$, then $x \in Cl_4^{\geq}$
- 2) if $f(x, a_4) \geq \text{Pt}$, then $x \in Cl_4^{\geq}$
- 3) if $f(x, a_4) \geq \text{Lo}$, then $x \in Cl_2^{\geq}$
- 4) if $f(x, a_4) \leq \text{Zo}$, then $x \in Cl_1^{\leq}$
- 5) if $f(x, a_1) \leq 1.3$, then $x \in Cl_1^{\leq}$
- 6) if $f(x, a_2) \leq 2$, then $x \in Cl_2^{\leq}$
- 7) if $f(x, a_4) \leq \text{So}$, then $x \in Cl_2^{\leq}$
- 8) if $f(x, a_2) \leq 2$, then $x \in Cl_3^{\leq}$
- 9) if $f(x, a_4) \leq \text{So}$, then $x \in Cl_3^{\leq}$
- 10) if $f(x, a_1) \leq 2.67$ and $f(x, a_4) \leq \text{Lo}$, then $x \in Cl_3^{\leq}$
- 11) if $f(x, a_1) \leq 2.6$ and $f(x, a_4) \leq \text{Lo}$, then $x \in Cl_3^{\leq}$
- 12) if $f(x, a_1) \leq 2.68$ and $f(x, a_4) \leq \text{Lo}$, then $x \in Cl_3^{\leq}$
- 13) if $f(x, a_1) \geq 2.67$ and $f(x, a_4) \leq \text{So}$, then $x \in (Cl_1 \cup Cl_2)$
- 14) if $f(x, a_2) \geq 7$, then $x \in (Cl_2 \cup Cl_3)$
- 15) if $f(x, a_1) \leq 2.5$ and $f(x, a_2) \geq 4$, then $x \in (Cl_2 \cup Cl_3)$

Finally, the rules obtained are validated with the testing dataset on computing the accuracy (Acc.) basing on precision (Prec.) and recall (Rec.). The precision, recall, and accuracy are computed using the equations (8), (9), and (10). The notation T_P is used for correct classification of cases to decisions whereas F_P is used for incorrect classification of cases to decisions. The notation T_N is the number of cases which correctly classified as negative whereas F_N is the number of incorrect cases classified as positive. Additionally a rule is also discarded if the accuracy falls less than the predefined threshold value 70%.

$$Prec. = \frac{|T_P|}{|T_P + F_P|} \quad (8)$$

$$Rec. = \frac{|T_P|}{|T_P + F_N|} \quad (9)$$

$$Acc. = \frac{|T_P + T_N|}{|T_P + F_P + T_N + F_N|} \quad (10)$$

The computation of precision, recall, and accuracy for the testing objects is presented in Table II. It is clear that the accuracy of rules 1, 5, 8, 9, 10, 11, 12, 14, and 15 are less than the predefined threshold value and hence discarded.

TABLE II: Rule validation of denial-of-service attacks in a wireless network

Rule	Sup. Obj.	T_P	F_N	F_P	T_N	Prec.	Rec.	Acc.
1	-	0	1	2	3	0	0	50
2	x_{10}	1	0	0	5	1	1	100
3	x_8, x_{10}	2	0	1	3	1	0.5	83.33
4	x_{13}	1	1	0	4	1	0.5	83.33
5	-	0	2	0	4	0	0	66.67
6,7	x_9, x_{11}, x_{13}	3	1	0	2	1	0.75	83.33
8,9	x_9, x_{11}, x_{13}	3	2	0	1	1	0.67	66.67
10	x_{13}	1	4	0	1	1	0.2	33.33
11	x_{13}	1	4	0	1	1	0.2	33.33
12	x_9, x_{13}	2	3	0	1	1	0.4	50
13	x_9, x_{11}, x_{13}	3	1	0	2	1	0.75	83.33
14	-	0	0	3	3	0	0	50
15	-	0	1	3	2	0	0	33.33

V. EMPIRICAL STUDY OF DOS ATTACK

This section describes how the proposed technique is used for a dataset. The dataset is preprocessed so that it may be able to give as an input to our developed system. Collection of data is a critical problem. This can be done by three ways as by using real traffic, by using sanitized traffic, and by using simulated traffic. However difficulties exist in using these approaches. Real traffic approach is very costly while sanitized approach is risky. The creating of simulation is also a difficult task. Further, in order to model various wireless networks, different types of traffic is needed. In order to avoid dealing with these difficulties, Knowledge Discovery Dataset (KDD)-cup dataset is considered for experimental analysis.

The dataset contains 11,160 records in which decisions for 3,260 records are normal whereas for 7,900 records are

various dos attacks such as neptune, udp storm, smurf, ping of death (PoD), back, teardrop, land, mailbomb, process table. Each sample of the dataset represents a connection between two wireless network hosts according to network protocols. It is described by 42 features as depicted in Table III. Out of 42 features, 41 are conditional features and one is decision. The set of 41 features are divided into four subsets such as basic feature set, data flow feature set, host based feature set, and content feature set. The basic feature set, a_1 to a_9 , is used to check the status of the flags, number of source bytes, number of destination bytes, types of protocols used, and duration of the period while information is communicated. The content feature set, a_{10} to a_{22} , is used to check the number of logins failed, number of compromised, number of logged-in, and number of guest login etc. Likewise the data flow feature set, a_{23} to a_{31} , is used to verify the sending and receiving errors during communication between source and destination. Similarly, the host based feature set, a_{32} to a_{41} , is used to get the information of receiving host and sending host errors while communication. From 41 features, 38 features are continuous or discrete (quantitative) and remaining 3 features are qualitative or categorical.

Each sample of decision feature is labeled as either normal or various dos attack. The dataset contains 10 class labels out of which one class is normal and remaining classes are different dos attacks such as neptune, udp storm, smurf, pod, back, teardrop, land, mail bomb, process table respectively. Some dos attacks such as mail bomb, neptune, or smurf abuse a perfectly legitimate feature. The teardrop, pod create malformed packets that confuse the TCP/IP stack of the machine that is trying to reconstruct the packet. The other dos attacks such as back, land takes the advantage of bugs in a particular network daemon.

A. Experimental Analysis

We implement wireless network dos detection system with C programming language and perform experiments in a computer with 2.67 GHz Intel core i3 processor, and 2 GB RAM. Total 11,600 records are divided into two categories such as training dataset of 6,138 (55%) records and testing dataset of 5022 (45%) records. The details of training, testing, total dataset and its various classifications are given in Table IV. Out of 41 conditional features 18 features such as $a_1, a_3, a_4, a_6, a_{13}, a_{14}, a_{16}, a_{17}, a_{19}, a_{20}, a_{21}, a_{22}, a_{33}, a_{34}, a_{35}, a_{37}, a_{40}, a_{41}$ are considered as criterion as suggested by various authors [24, 25]. For better visualization of the dataset, a graphical representation is shown in Figure 2.

Experimental analysis is carried out on each class of training dataset. Initially, we employed DOMLEM algorithm on 1887 records that are falling under the category normal. The total number of rules generated are 23. The rules generated are presented on Table V. These rules are further validated with 1373 records of testing dataset and found that rules 6, 9, 10, 16 and 18 are having accuracy less than the predefined threshold value. Hence, these rules are discarded. A graphical representation is shown in Figure 3. Likewise 740 records of data that are falling under the category of neptune, 767 records of data of udp storm, 762 records of data of smurf, 1042 records of data of pod, 188 records of data of back, 285 records of data of tear-drop, 155 records of data of land, 162 records of data of mail-bomb, and 150 records of data of

TABLE III: Features set of denial-of-service attack

S. No.	Features	Notation	Type
I	Basic Feature		
1	duration	a_1	continuous
2	protocol-type	a_2	symbolic
3	service	a_3	symbolic
4	flag	a_4	symbolic
5	src-bytes	a_5	continuous
6	dst-bytes	a_6	continuous
7	land	a_7	discrete
8	wrong-fragment	a_8	continuous
9	urgent	a_9	continuous
II	Content Feature		
10	hot	a_{10}	discrete
11	num-failed-logins	a_{11}	continuous
12	logged-in	a_{12}	discrete
13	num-compromised	a_{13}	continuous
14	root-shell	a_{14}	discrete
15	su-attempted	a_{15}	discrete
16	num-root	a_{16}	continuous
17	num-file-creations	a_{17}	continuous
18	num-shells	a_{18}	continuous
19	num-access-files	a_{19}	continuous
20	num-outbound-cmds	a_{20}	continuous
21	is-host-login	a_{21}	discrete
22	is-guest-login	a_{22}	discrete
III	Data Flow Feature		
23	count	a_{23}	continuous
24	srv-count	a_{24}	continuous
25	error-rate	a_{25}	continuous
26	srv-error-rate	a_{26}	continuous
27	error-rate	a_{27}	continuous
28	srv-error-rate	a_{28}	continuous
29	same-srv-rate	a_{29}	continuous
30	diff-srv-rate	a_{30}	continuous
31	srv-diff-host-rate	a_{31}	continuous
IV	Host Based Feature		
32	dst-host-count	a_{32}	continuous
33	dst-host-srv-count	a_{33}	continuous
34	dst-host-same-srv-rate	a_{34}	continuous
35	dst-host-diff-srv-rate	a_{35}	continuous
36	dst-host-same-src-port-rate	a_{36}	continuous
37	dst-host-srv-diff-host-rate	a_{37}	continuous
38	dst-host-error-rate	a_{38}	continuous
39	dst-host-srv-error-rate	a_{39}	continuous
40	dst-host-error-rate	a_{40}	continuous
41	dst-host-srv-error-rate	a_{41}	continuous
V	Decision		
42	decision	d	symbolic

process table are passed to DOMLEM algorithm. The total number of rules generated are 146. The category neptune generated 30 rules, category udp storm generated 18 rules, category smurf generated 17 rules, category pod generated 20 rules, category back generated 15 rules, category tear-drop generated 12 rules, category land generated 10 rules, category mail bomb generated 13 rules, and the category process table generated 11 rules. These rules are further validated with the testing dataset as mentioned in Table IV. The number of rules discarded for the categories neptune, udp storm, smurf, pod,

TABLE IV: Training, testing classification of datasets

S. No.	Description	Training Set	Testing Set	Total Set
1	normal	1,887	1,373	3,260
2	neptune	740	1,270	2,010
3	udp-strom	767	478	1,245
4	smurf	762	579	1,341
5	pod	1,042	680	1722
6	back	188	24	212
7	tear-drop	285	29	314
8	land	155	150	305
9	mail-bomb	162	250	412
10	process-table	150	189	339
	Total	6,138 (55%)	5,022 (45%)	11,160

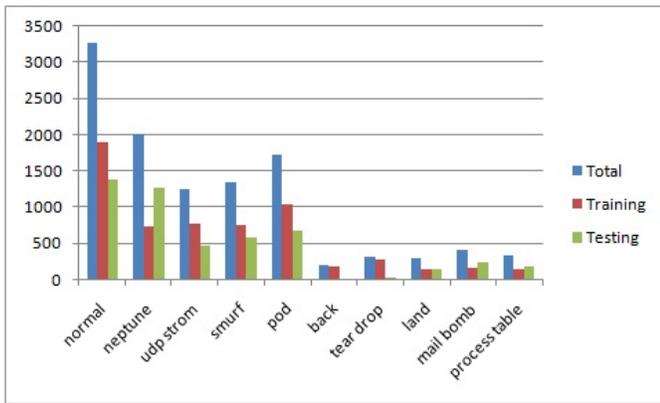


Fig. 2: Characteristics of Dataset

TABLE V: Selected list of normal rules

Rule No.	Description	Acc.
1	If $a_{34} \geq 0.34$ then d=Normal	100
2	If $a_{35} \geq 0.32$ then d=Normal	99
3	If $a_{37} \geq 0.34$ then d=Normal	100
4	If $a_{40} \geq 1$ then d=Normal	99
5	If $a_{19} \leq 0$ then d=Normal	100
6	If $a_{21} \geq 1$ then d=Normal	63
7	If $a_1 \leq 0$ then d=Normal	100
8	If $a_{34} \geq 0.34$ and $a_{19} \leq 0$ then d=Normal	100
9	If $a_{22} \leq 1$ then d=Normal	33.33
10	If $a_{20} \leq 0$ then d=Normal	67.66
11	If $a_{34} \geq 0.34$ and $a_1 \leq 0$ then d=Normal	99
12	If $a_{37} \geq 0.34$ and $a_{19} \leq 0$ then d=Normal	99
13	If $a_{37} \geq 0.34$ and $a_1 \leq 0$ then d=Normal	100
14	If $a_{22} \geq 1$ and $a_{20} \leq 0$ then d=Normal	100
15	If $a_{21} \geq 1$ and $a_1 \leq 0$ then d=Normal	99
16	If $a_{21} \geq 1$ and $a_{20} \leq 0$ then d=Normal	57.67
17	If $a_{34} \geq 0.34$ and $a_{21} \leq 1$ then d=Normal	100
18	If $a_{21} \geq 1$ and $a_{22} \leq 1$ then d=Normal	63.15
19	If $a_{34} \geq 0.34$ and $a_{37} \leq 0.34$ then d=Normal	98
20	If $a_{35} \geq 0.32$ and $a_1 \leq 0$ then d=Normal	100
21	If $a_{35} \geq 0.32$ and $a_{20} \leq 0$ then d=Normal	100
22	If $a_{37} \geq 0.34$ and $a_{22} \leq 1$ then d=Normal	98
23	If $a_{37} \geq 0.34$ and $a_{21} \geq 1$ then d=Normal	100

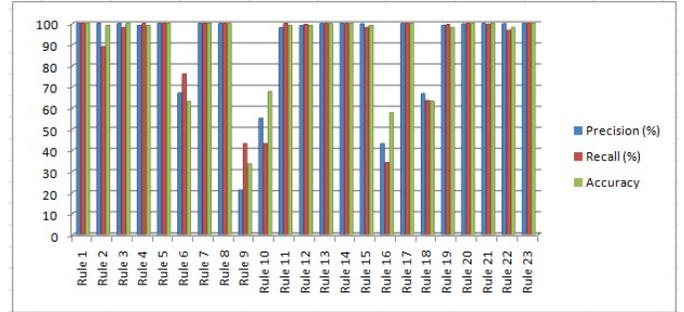


Fig. 3: Graphical view of precision, recall, accuracy

back, tear-drop, land, mail bomb, and process table are 6, 3, 2, 2, 3, 2, 2, 3, and 3 respectively. The final rules selected for various categories naptune, udp storm, smurf, pod, back, tear-drop, land, mail bomb, and process table are presented in Table VI, Table VII, Table VIII, Table IX, Table X, Table XI, Table XII, Table XIII, and Table XIV respectively.

TABLE VI: Selected list of neptune rules

Rule No.	Description	Acc.
1	If $a_{33} \geq 304$ then d=Neptune, Smurf	97.77
2	If $a_{34} \geq 0.36$ then d=Neptune, UDP Storm	99.65
3	If $a_{35} \geq 0.76$ then d=Neptune, Smurf, POD	100
4	If $a_{37} \geq 0.25$ then d=Neptune, Smurf	98.78
5	If $a_{40} \geq 0.24$ then d=Neptune	100
6	If $a_{41} \geq 0.15$ then d=Neptune	99.65
7	If $a_{13} \geq 4$ then d=Neptune, Smurf	99.65
8	If $a_{14} \geq 255$ then d=Neptune, POD	100
9	If $a_{16} \leq 531$ then d=Neptune	100
10	If $a_{17} \geq 8854$ then d=Neptune	99.65
11	If $a_{19} \geq 104$ then d=Neptune	100
12	If $a_{20} \leq 148$ then d=Neptune	100
13	If $a_{13} \geq 1$ then d=Neptune, POD	100
14	If $a_{14} \geq 1$ then d=Neptune, Smurf	100
15	If $a_{34} \leq 0.10$ then d=Neptune	100
16	If $a_{37} \geq 0.61$ then d=Neptune	99.65
17	If $a_1 \geq 31$ and $a_{20} \leq 104$ then d=Neptune	100
18	If $a_6 \geq 2252$ and $a_{22} \geq 1$ then d=Neptune	100
19	If $a_6 \geq 8854$ and $a_{33} \geq 255$ then d=Neptune, Smurf	100
20	If $a_6 \geq 1461$ and $a_{33} \leq 148$ then d=Neptune	100
21	If $a_{17} \geq 304$ and $a_{34} \leq 0.04$ then d=Neptune, POD	100
22	If $a_{16} \geq 245$ and $a_6 \geq 3634$ then d=Neptune, UDP Storm	100
23	If $a_{14} \geq 76$ and $a_{22} \geq 1$ then d=Neptune	100
24	If $a_{34} \geq 0.61$ and $a_{19} \leq 148$ then d=Neptune	100

B. Comparison with different approach

In this section, we compare results of proposed model with five different models such as resilient back propagation (RBP) [11], markov chain model (MCM) [6], radial basis function (RBF) [5], resistant architecture model (RAM) [8], and wavelet transform model (WTM) [9]. Unlike Table XV, the computation is carried out for each case across each technique. The following TABLE XVI presents the comparative analysis of all the techniques mentioned above. The accuracy of the proposed model over the KDD cup dataset is 99.76 whereas

TABLE VII: Selected list of UDP strom rules

Rule No.	Description	Acc.
1	If $a_{33} \geq 42$ then d=UDP Strom, Smurf	100
2	If $a_6 \geq 42$ then d=UDP Strom, POD	99.87
3	If $a_{35} \geq 0.28$ then d=UDP Strom, Neptune	100
4	If $a_{37} \geq 1$ then d=UDP Strom	100
5	If $a_{40} \geq 0$ then d=UDP Strom, Back	100
6	If $a_{41} \leq 0.25$ then d=UDP Strom, Smurf, Back	99.87
7	If $a_{14} \geq 7$ then d=UDP Strom, Back	100
8	If $a_{16} \geq 40$ then d=UDP Strom, Land	100
9	If $a_{17} \leq 40$ then d=UDP Strom, POD	99.87
10	If $a_{20} \geq 1$ then d=UDP Strom, Teardrop	100
11	If $a_{33} \geq 253$ then d=UDP Strom	100
12	If $a_6 \geq 40$ and $a_{14} \leq 40$ then d=UDP Strom	100
13	If $a_{21} \geq 1$ and $a_{33} \leq 255$ then d=UDP Strom	100
14	If $a_6 \geq 40$ and $a_{33} \geq 7$ then d=UDP Strom	100
15	If $a_{33} \geq 77$ and $a_6 \geq 0$ then d=UDP Strom	100

TABLE VIII: Selected list of smurf rules

Rule No.	Description	Acc.
1	If $a_{40} \geq 0.31$ then d=Smurf, UDP Storm	100
2	If $a_{41} \geq 0.14$ then d=Smurf, POD	100
3	If $a_{13} \geq 23$ then d=Smurf	100
4	If $a_{14} \geq 30$ then d=Smurf, Neptune	100
5	If $a_{16} \geq 93$ then d=Smurf, Back	100
6	If $a_{17} \geq 64$ then d=Smurf, Teardrop	100
7	If $a_{19} \geq 185$ then d=Smurf	100
8	If $a_{40} \geq 0.31$ and $a_{41} \geq 0.14$ then d=Smurf	100
9	If $a_{41} \leq 0.14$ and $a_{13} \geq 23$ then d=Smurf	100
10	If $a_{14} \geq 30$ and $a_{16} \geq 93$ then d=Smurf	100
11	If $a_{14} \geq 30$ and $a_{19} \geq 185$ then d=Smurf	100
12	If $a_{17} \geq 64$ and $a_{19} \geq 185$ then d=Smurf	100
13	If $a_{16} \leq 93$ and $a_{13} \geq 23$ then d=Smurf	100
14	If $a_{19} \geq 185$ and $a_{16} \geq 93$ then d=Smurf	100
15	If $a_{41} \leq 0.14$ and $a_{19} \geq 185$ then d=Smurf	100

TABLE IX: Selected list of POD rules

Rule No.	Description	Acc.
1	If $a_{33} \geq 829$ then d=POD, Smurf	100
2	If $a_{34} \geq 0.32$ then d=POD, Back	100
3	If $a_{35} \geq 0.08$ then d=POD, Neptune	100
4	If $a_{37} \geq 0.11$ then d=POD	100
5	If $a_{40} \geq 0.47$ then d=POD, Land	100
6	If $a_{41} \geq 0.03$ then d=POD	100
7	If $a_{33} \geq 829$ and $a_{34} \leq 0.32$ then d=POD	99.45
8	If $a_{33} \geq 829$ and $a_{35} \geq 0.08$ then d=POD	100
9	If $a_{40} \geq 0$ and $a_{34} \leq 0.32$ then d=POD	100
10	If $a_{40} \geq 0$ and $a_{35} \geq 0.08$ then d=POD	100
11	If $a_{37} \geq 0.11$ and $a_{34} \leq 0.32$ then d=POD	100
12	If $a_{37} \geq 0.11$ and $a_{35} \geq 0.08$ then d=POD	100
13	If $a_{33} \leq 829$ and $a_{40} \geq 0$ then d=POD	100
14	If $a_{37} \leq 0.11$ and $a_{34} \leq 0.32$ then d=POD	100
15	If $a_{37} \leq 0.11$ and $a_{35} \geq 0.08$ then d=POD	100
16	If $a_{34} \geq 0.32$ and $a_{41} \geq 0.03$ then d=POD	100
17	If $a_{35} \geq 0.08$ and $a_{34} \geq 0.32$ then d=POD	100
18	If $a_{34} \leq 0.32$ and $a_{35} \geq 0.08$ and $a_{33} \geq 829$ then d=POD	99.45

TABLE X: Selected list of back rules

Rule No.	Description	Acc.
1	If $a_{13} \geq 105$ then d=Back, POD	100
2	If $a_{14} \geq 146$ then d=Back, Land	100
3	If $a_{16} \geq 6$ then d=Back, Process table	100
4	If $a_{17} \geq 20$ then d=Back, Mailbomb	100
5	If $a_{19} \geq 1032$ then d=Back	100
6	If $a_{20} \geq 7$ then d=Back, Land	100
7	If $a_{13} \geq 105$ and $a_{14} \geq 146$ then d=Back	100
8	If $a_{16} \geq 6$ and $a_{13} \geq 105$ then d=Back	100
9	If $a_{17} \geq 20$ and $a_{14} \geq 146$ then d=Back	100
10	If $a_{19} \geq 1032$ and $a_{20} \leq 7$ then d=Back	100
11	If $a_{20} \geq 7$ and $a_{14} \geq 146$ then d=Back	100
12	If $a_{17} \leq 20$ and $a_{13} \geq 105$ then d=Back	100

TABLE XI: Selected list of teardrop rules

Rule No.	Description	Acc.
1	If $a_{40} \geq 0.52$ then d=Teardrop, Back	100
2	If $a_{41} \geq 0.51$ then d=Teardrop, Neptune	100
3	If $a_{33} \geq 20$ then d=Teardrop	100
4	If $a_{37} \geq 0.17$ then d=Teardrop, Land	100
5	If $a_{40} \geq 0.52$ and $a_{33} \geq 20$ then d=Teardrop, Land	100
6	If $a_{40} \geq 0.52$ and $a_{37} \geq 0.17$ then d=Teardrop, POD	100
7	If $a_{41} \geq 0.51$ and $a_{33} \geq 20$ then d=Teardrop	100
8	If $a_{41} \geq 0.51$ and $a_{37} \leq 0.17$ then d=Teardrop	100
9	If $a_6 \leq 520$ and $a_{35} \leq 0.20$ then d=Teardrop	100
10	If $a_6 \geq 520$ and $a_{34} \leq 0.17$ then d=Teardrop	100

TABLE XII: Selected list of land rules

Rule No.	Description	Acc.
1	If $a_{22} \geq 1$ then d=Land, Teardrop	100
2	If $a_{21} \geq 1$ then d=Land, Back	100
3	If $a_{20} \geq 79$ then d=Land, POD	99.97
4	If $a_{16} \geq 18$ then d=Land, Smurf	100
5	If $a_6 \geq 511$ and $a_6 \geq 145$ then d=Land	100
6	If $a_{40} \geq 0.51$ and $a_{41} \leq 0.79$ then d=Land, Mailbomb	100
7	If $a_6 \geq 145$ and $a_{34} \geq 0.18$ then d=Land	99.97
8	If $a_{22} \geq 1$ and $a_{33} \leq 18$ then d=Land	100

TABLE XIII: Selected list of mailbomb rules

Rule No.	Description	Acc.
1	If $a_{16} \geq 1000$ then d=Mailbomb, Land	100
2	If $a_6 \geq 1024$ then d=Mailbomb, Process table	100
3	If $a_{33} \geq 7$ then d=Mailbomb, Back, Land	100
4	If $a_{34} \geq 0.25$ then d=Mailbomb, Smurf	100
5	If $a_{33} \geq 114$ then d=Mailbomb, Neptune	100
6	If $a_{16} \geq 1000$ and $a_{33} \geq 7$ then d=Mailbomb	100
7	If $a_{16} \leq 1000$ and $a_{34} \geq 0.25$ then d=Mailbomb	100
8	If $a_6 \geq 1024$ and $a_{33} \geq 114$ then d=Mailbomb, Process table	100
9	If $a_6 \geq 1024$ and $a_{34} \geq 0.25$ then d=Mailbomb	100
10	If $a_{16} \leq 1000$ and $a_{33} \geq 114$ then d=Mailbomb, Land	100

TABLE XIV: Selected list of process table rules

Rule No.	Description	Acc.
1	If $a_{37} \geq 0.10$ then d=Process table, Mailbomb	100
2	If $a_{33} \geq 224$ then d=Process table, Back	100
3	If $a_{37} \geq 0.10$ and $a_4 \geq 0$ then d=Process table, POD	100
4	If $a_{33} \geq 224$ and $a_4 \geq 0$ then d=Process table, Smurf	100
5	If $a_{22} \geq 0$ and $a_6 \geq 1024$ then d=Process table	100
6	If $a_{13} \geq 224$ and $a_{19} \geq 1024$ then d=Process table, Mailbomb	100
7	If $a_{37} \leq 0.10$ and $a_{35} \geq 0.10$ then d=Process table	100
8	If $a_{33} \leq 224$ and $a_{37} \geq 0.10$ then d=Process table	100

TABLE XV: Precision, recall, accuracy of denial-of-service attack

S. No.	Descr.	T_P	F_N	F_P	T_N	Prec.	Rec.	Acc.
1	normal	1360	3	10	3649	0.99	1	99.74
2	neptune	1,258	2	10	3752	0.99	1	99.76
3	udp strom	465	5	8	4544	0.98	0.99	99.74
4	smurf	556	8	15	4443	0.97	0.99	99.54
5	pod	605	6	9	4342	0.99	0.99	99.70
6	back	21	2	1	4998	0.95	0.91	99.94
7	tear drop	26	1	2	4993	0.92	0.96	99.94
8	land	139	8	3	4872	0.99	0.95	99.78
9	mail bomb	238	7	5	4772	0.98	0.97	99.76
10	process table	175	7	7	4833	0.96	0.96	99.72
	Total	4903	49	70	45198	0.99	0.99	99.76

the accuracy of the RBP model over the same dataset is 99.35. It indicates that the accuracy of the proposed model is 0.41 higher than the RBP model. For better visualization, a graphical representation of the comparative analysis is shown in Figure 4. Figure 5 depicts the number of rules generated, number of rules discarded, and the number of rules finally selected for each class. The total number of rules generated are 169, and 18% number of rules are discarded through validation. This results the number of rules minimized to 82%.

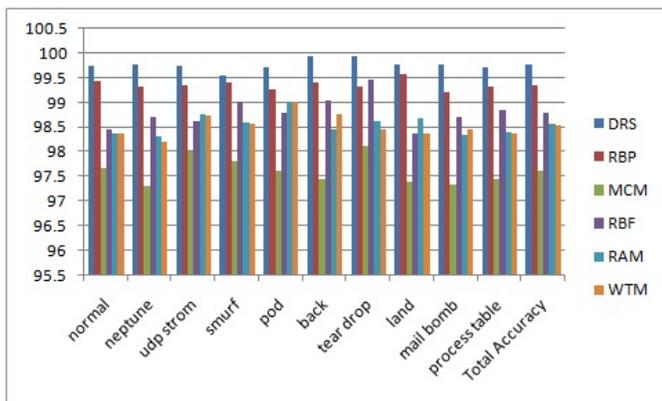


Fig. 4: Graphical Presentation of Comparative Analysis

TABLE XVI: Comparative analysis

S. No.	Descr.	DRS Acc.	RBP Acc.	MCM Acc.	RBF Acc.	RAM Acc.	WTM Acc.
1	normal	99.74	99.42	97.67	98.45	98.37	98.36
2	neptune	99.76	99.32	97.30	98.71	98.31	98.21
3	udp strom	99.74	99.35	98.00	98.63	98.75	98.74
4	smurf	99.54	99.41	97.79	99.01	98.59	98.57
5	pod	99.70	99.27	97.61	98.79	99.01	99.00
6	back	99.94	99.40	97.45	99.04	98.45	98.77
7	tear drop	99.94	99.31	98.11	99.45	98.63	98.45
8	land	99.78	99.56	97.38	98.37	98.67	98.37
9	mail bomb	99.76	99.22	97.32	98.71	98.35	98.44
10	process table	99.72	99.32	97.43	98.84	98.38	98.37
	Total Acc.	99.76	99.35	97.60	98.80	98.55	98.53

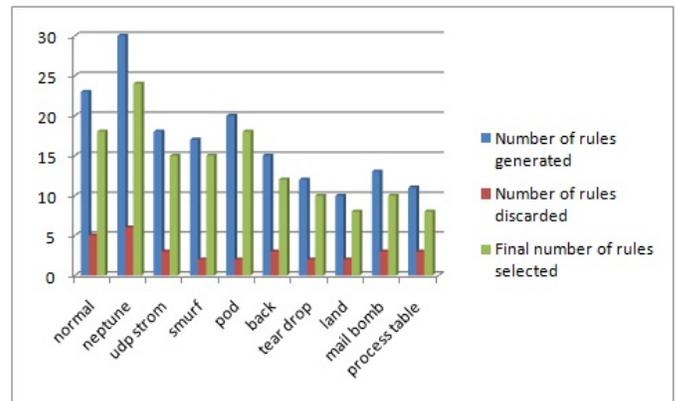


Fig. 5: Graphical view of numbers of rules selected

VI. CONCLUSION

Denial-of-service attack is one of the key security threats in wireless networks. Defending against DoS attack is of prime importance for industries, and internet service providers. To overcome this attack many techniques are proposed by various researchers [5, 6, 8, 9, 11]. In this paper, we propose a model for the detection of denial of service attack in wireless networks using dominance based rough set. The proposed model is analyzed with the help of KDD cup dataset. The total number of rules generated are 169, and 18% number of rules are discarded through validation. This results the number of rules minimized to 82%. Additionally, it is compared with existing techniques and found better accuracy. The accuracy of the proposed model is 99.76 whereas the accuracy of the RBP model is 99.35. This shows that the proposed model is 0.41 higher than the RBP model.

REFERENCES

- [1] X. Chuiyi, Z. Yizhi, B. Yuan, L. Shuoshan and X. Qin, *A distributed intrusion detection system against flooding denial-of-services attacks*, International Conference on Artificial Computing Technology, 2011, pp. 878-881.

- [2] X. Ren, (2009) *Intrusion detection method using protocol classification and rough set based support vector machine*, Computer and Information Science, 2 (2009), pp. 100-108.
- [3] R. C. Chen, K. F. Cheng and C. F. Hsieh, *Using rough set and support vector machine for network intrusion detection*, International Journal of Network Security and Its Applications, 1 (2009), pp. 1-13.
- [4] Y. Wang, *Analysis of a distributed denial of service*, Computer Network, Elsevier, 41 (2009), pp. 1200-1210.
- [5] D. Gavrilis and E. Dermatas, *Real-time detection of distributed denial-of-service attacks using RBF networks and statistical features*, Computer Network, Elsevier, 48 (2005), pp. 235-245.
- [6] Y. Wang, C. Lin, Q. L. Li and Y. Fang, *A queueing analysis for the denial of service (DoS) attacks in computer networks*, Computer Networks, Elsevier, 51 (2007), pp. 3564-3573.
- [7] E. Gelenbe and G. Loukas, *A self-aware approach to denial of service defence*, Computer Networks, Elsevier, 51 (2007), pp. 1299-1314.
- [8] P. Mell, D. Marks and M. McLarnon, *A denial-of-service resistant intrusion detection architecture*, Computer Networks, Elsevier, 34 (2000), pp. 641-658.
- [9] M. Hamdi and N. Boudriga, *Detecting denial-of-service attacks using the wavelet transform*, Computer Networks, Elsevier, 30 (2007), pp. 3203-3213.
- [10] S. Chen, Y. Tang and W. Du, *Stateful DDoS attacks and targeted filtering*, Journal of Network and Computer Applications, Elsevier, 30 (2007), pp. 823-840.
- [11] P. A. R. Kumar and S. Selvakumar, *Distributed denial of service attack detection using an ensemble of neural classifier*, Computer Communication, Elsevier, 30 (2011), pp. 1328-1341.
- [12] Z. Pawlak, *Rough set: theoretical aspects of reasoning about data*, Springer+Business Media, Springer, 1 (1991).
- [13] S. Greco, B. Matarazzo and R. Slowinski, *The use of rough sets and fuzzy sets in MCDM*, In T. Gal, T. Stewart and T. Hanne (eds.) *Advances in Multiple Criteria Decision Making*, chapter 14, Kluwer Academic Publishers, 1999, pp. 14.1-14.59.
- [14] Z. Pawlak, *Rough Sets*, International Journal of Computer and Information Sciences, 11 (1982), pp. 341-356.
- [15] D. B. Parker, *Demonstrating the elements of information security with treats*, Proceeding of the 17th National Computer Security Conference, 1994, pp. 421-430.
- [16] S. Greco, B. Matarazzo and R. Slowinski, J. Stefanowski, *An algorithm for induction of decision rules consistent with the dominance principle*, European Journal of Operational Research, 117 (1999), pp. 63-83.
- [17] J. Blaszczynski, S. Greco, B. Matarazzo, R. Slowinski and M. Szelag, *Dominance-based rough set data analysis framework*, Users Guide, pp. 1-19.
- [18] J. W. Grzymala-Busse, *LEERS - a system for learning from examples based on rough sets*. In R. Slowinski (eds.) *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, 1992, pp. 3-18.
- [19] J. Komorowski, Z. Pawlak, L. Polkowski and A. Skowron, *Rough Sets: tutorial*. In A. Skowron (eds.) *Rough Fuzzy Hybridization. A new trend in decision making*, Springer Verlag, Singapore, 1999, pp. 3-98.
- [20] J. Komorowski, *On rough set based approaches to induction of decision rules*. In L. Polkowski, A. Skowron (eds.) *Rough Sets in Data Mining and Knowledge Discovery*, Physica-Verlag, 1 (1998), pp. 500-529.
- [21] A. Chonka, W. Zhou, J. Singh and Y. Xiang, *Detecting and tracing distributed denial-of-service attacks by intelligent decision prototype*, International Conference on Pervasive Computing and Communications, 1994, pp. 421-430.
- [22] J. Yuan and K. Mills, *Monitoring the macroscopic effect of distributed denial-of-service flooding attack*, IEEE Transactions on Dependable and Secure Computing, 2 (2005), pp. 1-12.
- [23] T. Peng, C. Leckie and K. Ramamohanarao, *Survey of network-based defense mechanisms countering the DoS and DDoS problems*, ACM Computing Surveys, 39 (2007), pp. 370-373.
- [24] Y. Qing, W. Xiaoping, L. Yongqing and H. Gaofeng, *A hybrid model of RST and DST with its applications in intrusion detection*. International Symposium on Intelligent Information Technology and Security Information, 2010, pp. 202-205.
- [25] R. Shanmugavadivu and N. Nagarajan, *Network intrusion detection system using fuzzy logic*. Indian Journal of Computer Science and Engineering, 2 (2011), pp. 101-111.

Enhanced Version of Multi-algorithm Genetically Adaptive for Multiobjective optimization

Wali Khan Mashwani

Department of Mathematics,
Kohat University of Science & Technology,
Khyber Pakhtunkhwa (KPK), Pakistan

Abdellah Salhi

Department of Mathematical Sciences,
University of Essex,
Wivenhoe Park, Colchester, UK

Muhammad Asif jan

Department of Mathematics,
Kohat University of Science & Technology,
Khyber Pakhtunkhwa (KPK), Pakistan

Rashida Adeeb Khanum

Department of Mathematics,
Jinnah College for Women Peshawar,
Khyber Pakhtunkhwa (KPK), Pakistan

Muhammad Sulaiman

Department of Mathematics,
Abdul Wali Khan University, Mardan,
Khyber Pakhtunkhwa (KPK), Pakistan

Abstract—Multi-objective EAs (MOEAs) are well established population-based techniques for solving various search and optimization problems. MOEAs employ different evolutionary operators to evolve populations of solutions for approximating the set of optimal solutions of the problem at hand in a single simulation run. Different evolutionary operators suite different problems. The use of multiple operators with a self-adaptive capability can further improve the performance of existing MOEAs. This paper suggests an enhanced version of a genetically adaptive multi-algorithm for multi-objective (AMALGAM) optimisation which includes differential evolution (DE), particle swarm optimization (PSO), simulated binary crossover (SBX), Pareto archive evolution strategy (PAES) and simplex crossover (SPX) for population evolution during the course of optimization. We examine the performance of this enhanced version of AMALGAM experimentally over two different test suites, the ZDT test problems and the test instances designed recently for the special session on MOEA's competition at the Congress of Evolutionary Computing of 2009 (CEC'09). The suggested algorithm has found better approximate solutions on most test problems in terms of inverted generational distance (IGD) as the metric indicator.

Keywords—Multi-objective optimization, Multi-objective Evolutionary algorithms (MOEAs), Pareto Optimality, Multi-objective Memetic Algorithm (MOMAs).

I. INTRODUCTION

Multi-objective evolutionary optimization is a subject of intense interest in all fields of Science, Engineering, Economics, Logistics and others. Multi-objective optimization problems (MOPs) have more than one conflicting objective function and they have many real-world applications [6], [59]. A general MOP can mathematically be formulated as follows.

$$\begin{aligned} &\text{minimize } F(x) = (f_1(x), \dots, f_m(x))^T \\ &\text{subject to } x \in \Omega \end{aligned} \quad (1)$$

where Ω is the decision variable space, $x = (x_1, x_2, \dots, x_n)^T$ is an individual or solution and $x_i, i = 1, \dots, n$ are their decision variables, $F(x) : \Omega \rightarrow R^m$ consists of m real valued objective functions and R^m is called the objective space. If Ω is a closed and connected region in R^n and all

the objective functions in (1) are continuous for x , we call it a continuous MOP. Furthermore, if $m \geq 3$, then problem (1) is said to be a many objectives problem. In single objective optimization, the main focus is on the decision space while in multi-objective optimization, the focus is mainly on the objective space because objective values are used in checking for optimality [43]. In practical applications of optimization, it is very common that the objective functions of the MOP conflict with one another or are mostly incommensurable. One needs a set of optimal solutions to solve these problems. A solution $u = (u_1, u_2, \dots, u_n) \in \Omega$ is said to be Pareto optimal if there exist no another solution $v = (v_1, v_2, \dots, v_n) \in \Omega$ such that $f_j(u) \leq f_j(v)$ for all $j = 1, \dots, m$ and also $f_j(u) < f_j(v)$ for at least index k . An objective vector is said to be Pareto optimal if their corresponding decision vector is Pareto optimal. All Pareto optimal solutions in the decision space of MOP is called Pareto set (PS) and their corresponding image in their objective space is called Pareto front (PF). The idea Pareto optimality was first proposed by Francis Ysidro Edgeworth in 1881 and then later on generalized by Vilfredo Pareto in 1986 as discussed in [12], [10].

MOEAs are highly effective and powerful stochastic techniques which can find a set optimal solutions in a single simulation run due to their population-based nature, unlike traditional mathematical programming. In the past two decades, and since the inception of vector evaluated GA (VEGA) [48], different types of MOEAs have been suggested, the Pareto dominance based MOEAs [11], [13], [61], [60], [44], [19], [18], [9], [27]), the decomposition based MOEAs [21], [20], [54], [8], [7], [30], [56], [58], [55], [1], [34], [32], [39], [42], [41], [26], [25], [35], [37], [33], [41]), and Indicator Based algorithms [63], [5], [3], [22], [4], [2], [14]. They mainly emphasize three conflicting goals: firstly, the final approximate Pareto front (PF) should be as close as possible to the true PF; secondly, the final set of Pareto optimal solutions should be uniformly distributed and diverse over the true PF of the problem (1); thirdly, the approximated PF should capture the whole spectrum of the true PF. Different fitness assignment procedures, elitism and diversity promoting strategies are found in the current literature of evolutionary computing (EC).

The Pareto dominance concept of MOEAs is very common for solving MOPs [28], [12]. To promote diversity, most of these algorithms use different diversity techniques such as fitness sharing, niching, the kernel approach, the nearest neighbour approach, the histogram technique, crowding or clustering, a relaxed form of dominance and restricted mating [11]. Among them, a fast non-dominated sorting algorithm (NSGA-II) [13], SPEA2 or improving the strength Pareto evolutionary algorithm [60], the Pareto archive evolution strategy (PAES) [27], multi-objective genetic algorithm (MOGA) [18], and niched Pareto genetic algorithm (NPGA) [19] are long-familiar and well known approaches. They have shown good behaviors in several comparative analysis.

Multiobjective evolutionary algorithms (MOEAs) are extremely useful for dealing with MOPs. They evolve their population solutions and provide Pareto optimal solutions in single simulation unlike traditional optimization techniques. In the past two decades, since the inception of vector evaluated genetic algorithm (VEGA) [48], several MOEAs have been suggested and [13], [61], [60], [17], [44], [19], [9], [11], [49], [59], [32], [36], [38], they have successfully tackled various types of MOPs [16], [25], [35], [37], [33]. In general, classical MOEAs can be divided into three main different classes, namely, the Pareto dominance based MOEAs (e.g., [11], [13], [61], [60], [44], [19], [18], [9], [27]), the decomposition based MOEAs (e.g., [21], [20], [54], [8], [7], [30], [56], [58], [55], [1], [34], [32], [39], [42], [41], [26]), and Indicator Based algorithms (e.g., [63], [5], [3], [22], [4], [2], [14]). All these algorithms try to obtain a set of Pareto optimal solutions with three main features, firstly, It should close as much as possible to the true PF. Secondly, the approximated set should require expand uniformly distributive all over the true PF of the problem (1). Thirdly, final Pareto optimal solutions obtained by particular MOEA should require to desirably capture the whole spectrum of the PF of the problems. The existing algorithms implement different fitness assignment procedures to evolve their population in order to achieve aforementioned three goals subject to No free Lunch concept [29]. The Pareto dominance concept based MOEAs thoroughly applied for coping with MOPs [28], [12]. To promote diversity, most of these algorithms are utilizing different diversity techniques such as fitness sharing, niching approach, Kernel approach, nearest neighbour approach, histogram technique, crowding or clustering, relaxed form of dominance and restricted mating [11]. Among them, a fast non-dominated sorting algorithm (NSGA-II) [13], SPEA2 or improving the strength Pareto evolutionary algorithm [60], the Pareto archive evolution strategy (PAES) [27], multi-objective genetic algorithm (MOGA) [18], and niched Pareto genetic algorithm (NPGA) [19] have been chosen in several comparative analysis.

Multi-objective memetic algorithms (MOMAs) form a new and attractive area of research in EC. They are inspired by models of adaptation found in nature. They are known as Baldwinian EAs, Lamarckian EAs, cultural algorithms, or genetic local search and hybrid MOEAs [45]. Hybrid MOEAs have been developed with the aim to overcome the shortcomings of stand-alone MOEAs [34], [32], [39], [42], [40].

A genetically adaptive multi-algorithm for multi-objective (AMALGAM) optimisation is recently developed for solving both multi-objective optimization problems [52] and single

optimization problems [53]. It employs multiple search operators for its population evolution. The search operators used include the particle swarm optimizer (PSO) [15], differential evolution (DE) [47] and NSGA-II [13] and allocates resources dynamically to each search operators based on their individual performances. It does not involve any decomposition as in MOEA/D (multio-bjective evolutionary algorithm based on decomposition) [54].

MOEA/D [54] decomposes the approximated PF of the given MOP into a number of different single objective optimization subproblems (SOPs). It then optimizes all SOPs simultaneously using generic evolutionary algorithm. MOEA/D paradigm have tackled diverse benchmark functions and it has several enhanced versions [31], [34], [32], [39], [36], [42]. In this paper, our main objective is to further improve the algorithmic performance of ALMAGAM by employing by employing multiple search operators including the differential evolution (DE) [46], particle swarm optimization (PSO) [15], simulated binary crossover (SBX) [24], Pareto archive evolution strategy (PAES) [23] and simplex crossover (SPX) [50] with self-adaptive alternative procedures for dealing with both CEC'09 test instances [57] and ZDT test problems [62].

The main objective in this paper to develop an enhanced version of ALMAGAM by employing multiple search operators such as differential evolution (DE) [46], particle swarm optimization (PSO) [15], simulated binary crossover (SBX) [24], Pareto archive evolution strategy (PAES) [23] and simplex crossover (SPX) [50] with self-adaptive procedures for dealing with both CEC'09 test instances [57] and ZDT test problems [62].

The rest of this paper is organized as follows. Section II outlines the framework of the enhanced version of the genetically adaptive multi-algorithm multi-objective (AMALGAM) method. Section III presents experimental results obtained with the enhanced AMALGAM on both CEC'09 [57] and five ZDT test problems [62]. Section IV is devoted to a discussion on experimental results. Section V finally concludes this paper and suggest further areas of research on this topic an related ones.

II. ENHANCED VERSION OF MULTI-ALGORITHM GENETICALLY ADAPTIVE FOR MULTIOBJECTIVE OPTIMIZATION

Algorithm 1 outlines the framework of the an enhanced version of the AMALGAM. In Sept 1, a population P with size N has been generated uniformly and randomly within the search space of the given MOPs. We then evaluate the fitness values of solution of population P . We calculate the crowding distance of each member of population after categorize them into different layers by using fast non-dominating sorting procedure adopted in NSGA-II [13] framework. After this, an Algorithm 1 divide the whole population according to k number of search operators in order to work each search operator on specified number of sub-populations N_1, N_2, N_3 to generate Q offspring population of sizes whose sum is equal to N . We have used five different search operators such as differential evolution (DE) [46], particle swarm optimization (PSO) [15], simulated binary crossover (SBX) [24] and Pareto archive evolution strategy (PAES) [23] and simplex crossover (SPX) [50] in

the evolutionary process of the suggested algorithm. Each individual search operator is getting resources at population level according to their current individual performance based on self-adaptive procedure as explained in subsection II-A.

Algorithm 1 Enhanced Version of Multi-algorithm Genetically Adaptive for Multiobjective optimization

- 1: **Input:**
 - 2: MOP: the multiobjective optimization problem; N : the population size and other main parameters; \mathbf{F}_{eval} : maximum function evaluations;
 - 3: **Output:** $\{x^1, \dots, x^N\}$ and $\{F(x^1), \dots, F(x^N)\}$;
 - 4: Generate an initial population P of size N uniformly and randomly.
 - 5: Calculate the F-function values of each member of the P population.
 - 6: Assign rank to each member of P using fast non-dominating procedure.
 - 7: Assign sub-populations $P = \{P_1, P_2, \dots, P_k\}$ to k operators for creating an offspring population $Q = \{Q_1, Q_2, \dots, Q_k\}$ of size N .
 - 8: Calculate F-function values of Q offspring population.
 - 9: Assign rank to each member of Q using fast non-dominating procedure.
 - 10: Combine the new and old population P and Q , $R = P \cup Q$.
 - 11: Select population P of size N from population R of size $2N$ based on their ranks and crowding distances for next generation.
 - 12: Update N best individuals among C population with high ranks and crowding density.
 - 13: Update $P = \{P_1, P_2, \dots, P_k\}$ (Explanation can be found in subsection II-A) based on the individual performances of each search operator.
-

A. Alternative Adaptive Resources Allocation Scheme

- We calculate the number of solutions that successfully enter to the next generation in the evolutionary process of enhanced version of AMALGAM. A successful solution is rewarded by 1 and unsuccessful by 0. An efficient operator gets more resources in the form of subpopulation to be operate on them as compared to weaker one.
- Let δ_k , $k = 1, 2, \dots, q$ are total number of non-dominated solutions produced by q search operators (i.e, differential evolution (DE) [46], particle swarm optimization (PSO) [15], simulated binary crossover (SBX) [24], Pareto archive evolution strategy (PAES) [23] and simplex crossover (SPX) [50]) that enter successfully to next generation are convert into normalized form to develop probability formula (3)

$$P_k = \frac{\zeta_k}{\sum_{k=1}^q \zeta_k}, \text{ where } \zeta_k = \frac{\delta_k}{\sum_{k=1}^q \delta_k} \quad (2)$$

$$P_k = \alpha P_{k-1} \times N + (1 - \alpha) P_k \times N \quad (3)$$

Where P_k is the current and P_{k-1} is the previous probability of successes of the k search operators. More importantly, the above mentioned dynamic resources allocation did not switch on at every generation of proposed algorithm. It can allocate

resources at every multiple of 5th generation to tackle ZDT test problems [62]. The suggested enhanced AMALGAM allocates resources to each of its embedded search operator at every multiple of 10th generation for dealing with CEC'09 test instances [57].

III. PARAMETERS SETTING AND EXPERIMENTAL RESULTS

We have carried out Experiments using benchmark functions with two and three objectives. The ZDT test [62] were tackled with parameter settings as explained in the subsection III-A while CEC'09 [57] were handled with parameter settings are explained in the subsection III-B, respectively.

A. Parameter Settings for ZDT Problems

- $N = 100$: population size for 2-objective test instances.
- $F = 0.5$: scaling factor of the DE;
- $CR = 0.5$: crossover probability for DE;
- w is the inertia factor which lies in $[0.8, 1.2]$;
- c_1 and c_2 are the two acceleration constant or acceleration coefficients that usually lies between 1 and 4;
- $u_r \in [-1, 1]$ is a continuous uniform random number
- $w = 0.5 + rand/2$: inertia factor which lies in $[0.8, 1.2]$ and $\xi = 1$;
- $c_1 = c_2 = 1.5$: acceleration constant or acceleration coefficients that usually lies between 1 and 4;
- $F_{eval} = 25000$: maximum function evaluations;

B. Parameter Settings for CEC'09 Test Instances

This subsection explains the parameters setting to validate enhanced AMALGAM on CEC'09 test instances [57].

- $N = 600$: population size for 2-objective test instances;
- $N = 1000$: for 3-objective test instances;
- $F = 0.5$: scaling factor of the DE;
- $CR = 1$: crossover probability for DE;
- $F_{eval} = 300,000$: maximum function evaluations;

C. Performance Indicators

Two main goals for dealing with multiobjective evolutionary optimization are very important: 1) convergence towards the Pareto-optimal front, 2) to find uniform and well-distributive set of multiple solutions that cover the whole true PF of the problem at hand [12]. Several performance metrics are found in the specialized literature of evolutionary computing (EC) [51], [13], [12], [64] which are using to judge which algorithm is better than others and in what aspects. Inverted generational distance (IGD) [64], [57], relative hypervolume [51], [12], Gamma Υ and delta Δ [12], [13] which are commonly in several comparative analysis of different algorithms. The aforementioned performance indicators can

use only if the reference set for the test problems are known in advance or available. In this paper, we have used the Inverted Generational Distance (IGD) as performance indicator to judge the quality of final approximated set of Pareto optimal solutions obtained by proposed algorithm in comparison with other MOEAs.

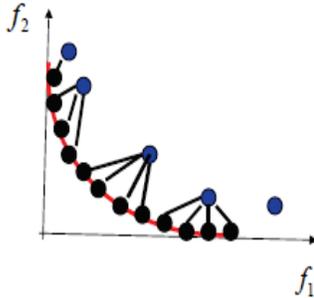


Fig. 1. Inverted Generation Distance (IGD) has been used as a Performance Indicator. The above solutions (blue fill circle) are approximated solution of the algorithm and (black fill circle) are the Pareto solutions uniformly distributive along the Pareto Front (PF).

Let P^* be a set of uniformly distributed points along the PF. Let A be an approximate set to the PF, the average distance from P^* to A is defined as [57]:

$$D(A, P) = \frac{\sum_{v \in P^*} d(v, A)}{|P^*|}$$

where $d(v, A)$ is the minimum Euclidean distance between v and the points in A . If P^* is large enough to represent the PF very well, $D(A, P)$ could measure both the diversity and convergence of A in a sense. The closer the IGD metric values, better is the approximation set. We have used $P^* = 500$ in our carried experiments to tackle 2-objectives test instances and $P^* = 1000$ to solve 3-objectives problems. For two objectives case, the IGD metric is pictorially depicted in the figure 1. It measures both the convergence and the spread of the obtained solutions. Smaller IGD-metric values, better will be approximated set of Pareto optimal of solutions of corresponding MOEAs.

IV. DISCUSSION OF THE EXPERIMENTAL RESULTS OBTAINED ON ZDT AND CEC'09 TEST INSTANCES

The simulation related parameters are as follows.

- Operating system: Windows XP Professional
- Programming language of the algorithms: Matlab
- CPU: Core 2 Quad 2.4 GHz
- RAM: 4 GB DDR2 1066 MHz
- 30 independent runs were performed on each test problem.

A. Discussion of IGD-metric Values

Table I and Table II record the IGD-metric values in terms of minimum (Best), Median, Mean, and standard deviation(std) and maximum (worst) which are found by enhanced AMALGAM and NSGA-II [13], respectively. These statistics have been collected by executing each algorithm 30 times independently with different random seeds on each ZDT test problem [62] and CEC'09 test instance [57]. The average CPU time spent by each algorithm are also provided in the last columns of I and Table II. It is evident from these Tables, that enhanced AMALGAM has found better approximated solution set with reduced the IGD-metric values as compared to NSGA-II [13] for most test problems. In most test problems, global convergence has been got for both test suites of problems. However, the complete Pareto front for some CEC'09 test instances have not been attained by enhanced AMALGAM in multiobjective optimization context. The primary reasons of this weak performance could be reason of the complicated objective functions profile of some CEC'09 test instances [57] which are mostly multi-modal near the global Pareto-optimal frontier and a slight perturbation in their optimization variables causes their solutions to become dominated.

B. Discussion of the Pareto Fronts of ZDT and CEC'09 Test Instances.

Table I and Table II record the IGD-metric values in terms of minimum (Best), Median, Mean, and standard deviation (std) and maximum (worst) which are found by enhanced AMALGAM and NSGA-II [13], respectively. These statistics have been collected by executing each algorithm 30 times independently with different random seeds on each ZDT test problem [62] and CEC'09 test instance [57]. The average CPU time spent by each algorithm are also provided in the last columns of I and Table II. It is evident from statistics gathered in these Tables, that enhanced AMALGAM has found a better approximate solution set with reduced IGD-metric values compared to those of NSGA-II [13] for most test problems. In most test problems, global convergence has been achieved for both test suites of problems. However, the complete Pareto front for some CEC'09 test instances has not been attained by the enhanced AMALGAM algorithm, in the multi-objective optimization context. The primary reasons for this weak performance could be the complicated objective functions profile in some CEC'09 test instances [57]; these are mostly multi-modal near the global Pareto-optimal frontier and a slight perturbation in their optimization variables causes their solutions to become dominated.

Figure 2 and Figure 6 depict the approximated Pareto front (PF) against the real PF of ZDT test problems displayed by enhanced AMALGAM and NSGA-II [13], respectively. These figures indicate that both algorithms have found better approximated PF in their best run among 30 independent runs on each ZDT test problem. We have plotted 30 PFs together in Figures 3 and Figures 7 of the enhanced AMALGAM and NSGA-II [13], respectively. These figures indicate that enhanced AMALGAM has displayed all 30 in better distribution ranges in all 30 independent runs as compared to NSGA-II [13].

Figures 4 display the best approximated PF of the CEC'09 test instances as demonstrated by enhanced AMALGAM

TABLE I. THE IGD-METRIC VALUES OF THE ENHANCED AMALGAM FOR ZDT1-ZDT4 AND ZDT6. AVG-T MEANS AVERAGE CPU TIME IN SECONDS.

ZDT	Best	Median	Mean	St.Dev.	worst	AVG-T
ZDT1	0.004301	0.004521	0.004603	0.000223	0.005065	14.940230
ZDT2	0.004235	0.004794	0.004613	0.000258	0.005643	14.633287
ZDT3	0.005067	0.005498	0.005565	0.000170	0.006134	14.580179
ZDT4	0.004696	0.005175	0.005235	0.000162	0.005575	14.696021
ZDT6	0.003615	0.004037	0.004045	0.000175	0.004691	14.395623

TABLE II. THE IGD-METRIC VALUES OF THE NSGA-II [13] FOR DEALING WITH ZDT1-ZDT4 AND ZDT6. AVG-T MEANS AVERAGE CPU TIME IN SECONDS.

ZDT	Best	Median	Mean	St.Dev.	worst	AVG-T
ZDT1	0.0042193	0.004472	0.004369	0.000139	0.004258	18.01
ZDT2	0.0043213	0.004649	0.004656	0.000182	0.005011	22.85
ZDT3	0.005132	0.00546	0.00912	0.01388	0.0602182042	17.596
ZDT4	0.00482	0.006421	0.00825	0.009649	0.059017370	22.85
ZDT6	0.005606	0.007045	0.007003	0.0005878	0.0080474634	19.90

TABLE III. THE IGD-METRIC VALUES OBTAINED BY ENHANCED AMALGAM OVER CEC'09 TEST INSTANCES. AVG-T MEANS AVERAGE CPU TIME IN SECONDS.

CEC'09	best	mean	median	st. dev.	worst	AvG-T
UF1	0.028431	0.058596	0.057886	0.008465	0.070089	286.307532
UF2	0.011235	0.018219	0.013157	0.001349	0.016859	288.558110
UF3	0.091864	0.134375	0.136495	0.022836	0.198629	297.999586
UF4	0.040348	0.041052	0.041039	0.000337	0.041667	286.159044
UF5	0.165346	0.171338	0.171421	0.002796	0.176307	258.823135
UF6	0.068597	0.079037	0.078632	0.005978	0.088905	308.462841
UF7	0.014935	0.017689	0.017787	0.001267	0.020866	290.402351
UF8	0.103734	0.234131	0.230672	0.026091	0.261546	720.849149
UF9	0.056715	0.067789	0.114643	0.085653	0.325885	700.875474
UF10	0.273393	0.327878	0.326937	0.020029	0.360963	686.032888

TABLE IV. THE IGD-METRIC VALUES GENERATE BY NSGA-II [13] IN 30 INDEPENDENT RUNS FOR CEC'09 TEST INSTANCES. AVG-T MEANS AVERAGE CPU TIME IN SECONDS.

CEC'09	best	mean	median	st. dev.	worst	AvG-T
UF1	0.051996	0.106873	0.096076	0.024862	0.128739	759.27
UF2	0.016012	0.019849	0.020050	0.001407	0.023589	518.07
UF3	0.066353	0.098234	0.097065	0.017958	0.134235	491.95
UF4	0.052199	0.054388	0.054551	0.001274	0.056679	393.60
UF5	1.523087	1.671735	1.676288	0.099452	1.844279	792.28
UF6	0.705834	0.762023	0.762271	0.028052	0.831784	822.79
UF7	0.067270	0.114403	0.112305	0.012055	0.125719	722.11
UF8	0.095436	0.108548	0.120433	0.030475	0.195112	1443.73
UF9	0.088857	0.188603	0.160832	0.047975	0.218993	1270.73
UF10	0.473865	0.744428	0.781509	0.134987	1.043141	1359.30

within 30 times independent execution for dealing with each problem. The Figure 8 is of the NSGA-II [13] produced in its best run among 30 independent runs for CEC'09 test instances [57]. The PFs displayed by enhanced AMALGAM are more promising than NSGA-II [13] in terms of diversity and proximity.

We have also plotted 30 PFs altogether estimated by enhanced AMALGAM in the figure 5 and the figure 9 exhibited by NSGA-II [13] in all 30times independent runs over the problems UF1-UF4 and UF7-UF10. These figures clearly indicate that enhanced AMALGAM has tackled the most CEC'09 test instances more effectively and spend less CPU time dealing with each test problem as compared to NSGA-II [13].

The problems UF5 and UF6 have not been tackled by both algorithms as per demand and genetic drift has been occurred in their respective population due to the presence of highly multi-modality in these problems like UF5-UF6. The search process of both algorithms are get stuck in the local basin of attraction of these problems and due to this both algorithms have not shown further improvement dealing with

these problems.

V. CONCLUSION AND FURTHER WORK

Different operators suite different optimization and search problems. The dynamic use of multiple operators in the EA framework has exhibited good performance on complicated MOPs, [31], [39], [34], [36], [32], [41], [40]. A multi-algorithm genetically adaptive multi-objective (AMALGAM) has recently been developed for solving both single objective [53] and multi-objective optimization problems [52]. In this paper, we have suggested new adaptive resource allocation procedure and developed an enhanced version of AMALGAM to cope with CEC'09, [57], and ZDT test problems [53]. The suggested algorithm has shown promising results on most test problems compared to NSGA-II [13] in terms of proximity and diversity. Furthermore, the suggested algorithm is more efficient which is desirable when solving real-world problems where time can be an issue. In the future, we intend to examine the performance of enhanced AMALGAM over real-world problems such as:

- Tubular permanent magnet linear synchronous motor

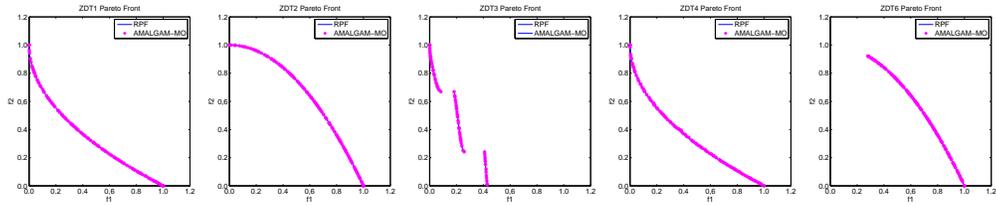


Fig. 2. Plots of the final non-dominated solutions in the objective space displayed by enhanced AMALGAM in its best run among 30 independent runs over ZDT1-ZDT4 and ZDT6 problems.

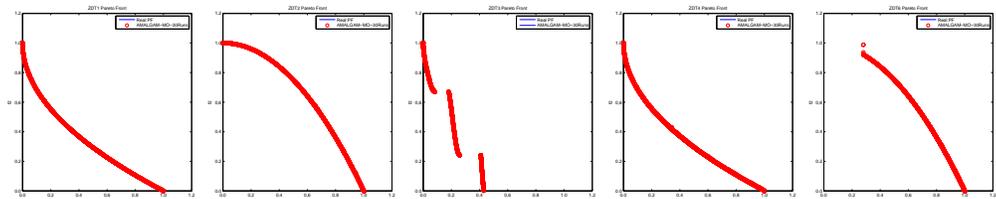


Fig. 3. Plots of the 30 PFs altogether in the objective space displayed by enhanced AMALGAM for ZDT1-ZDT4 and ZD6 problems.

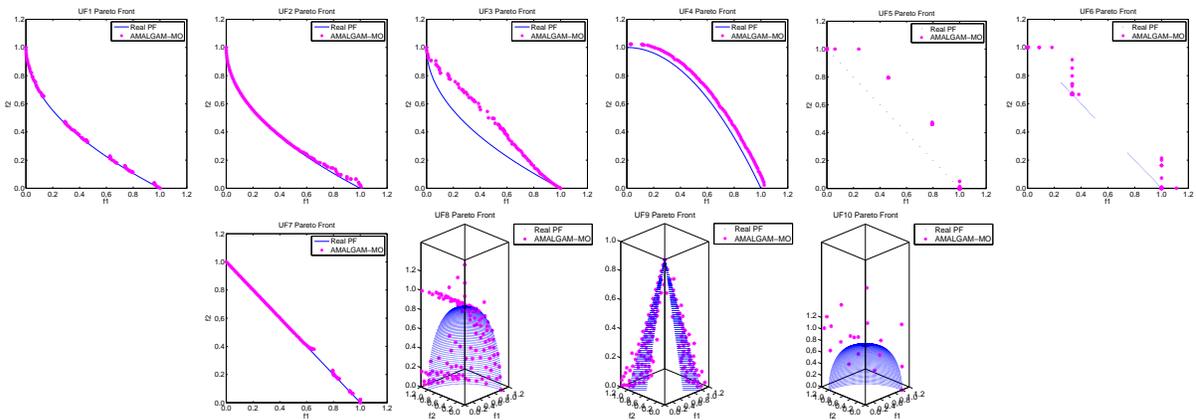


Fig. 4. Plots of the final non-dominated solutions in the objective space displayed by enhanced AMALGAM in its best run among 30 independent runs over UF1-UF10 problems.

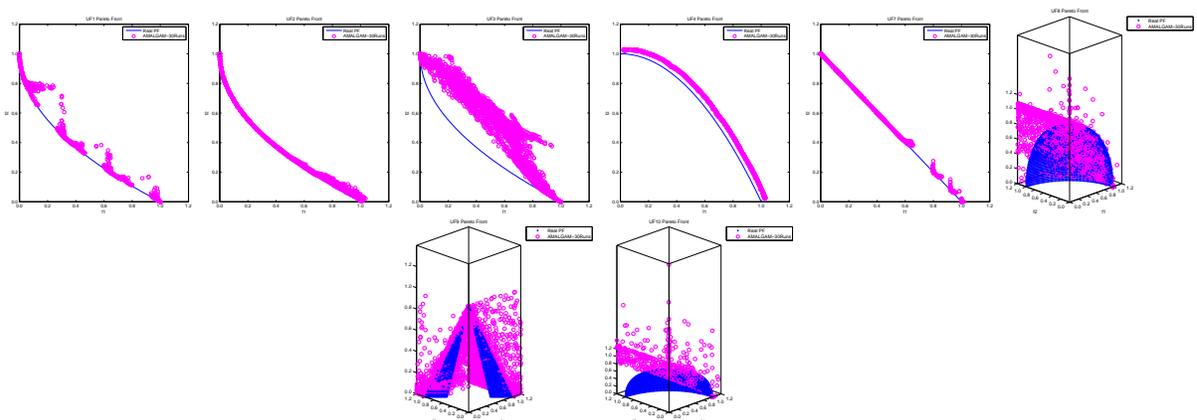


Fig. 5. Plots of the 30 PFs altogether in the objective space displayed by enhanced AMALGAM for UF1-UF4 and UF7-UF10 problems.

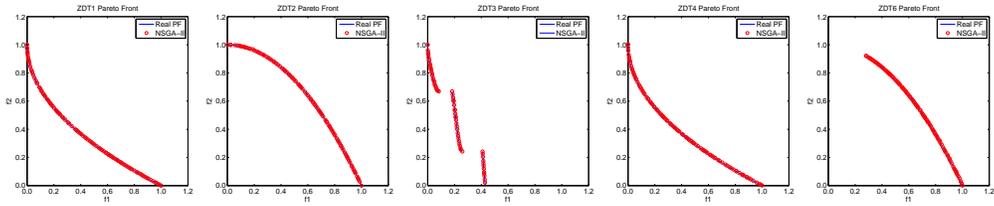


Fig. 6. Plots of the final non-dominated solutions in the objective space displayed by NSGA-II [13] in its best run among 30 independent runs over ZDT1-ZDT4 and ZDT6 problems.

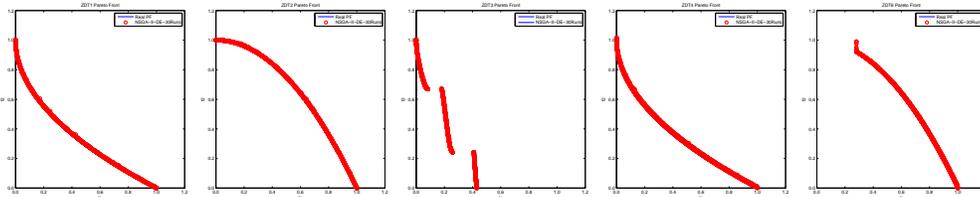


Fig. 7. Plots of the 30 PFs altogether in the objective space produced by NSGA-II [13] for ZDT1-ZD4 and ZDT6 problems.

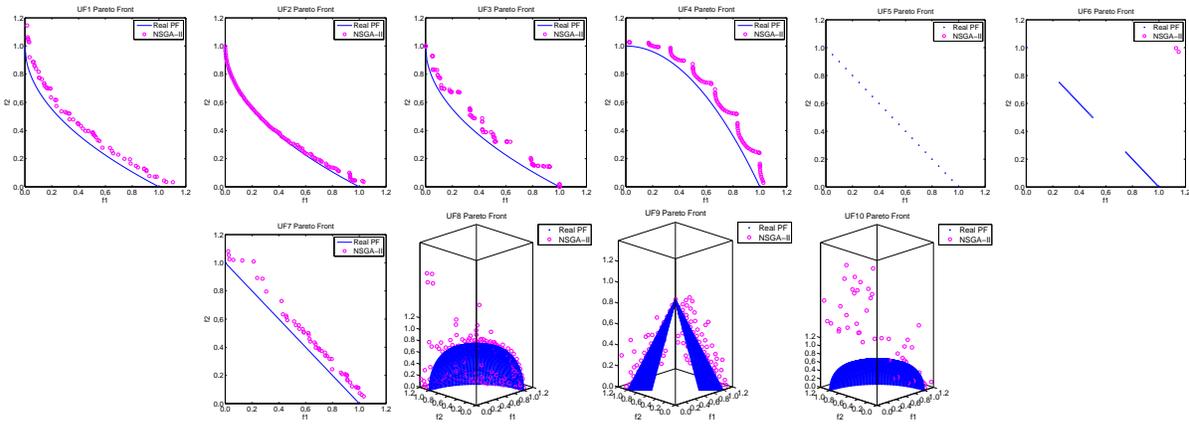


Fig. 8. Plots of the final non-dominated solutions in the objective space display by NSGA-II [13] in its best run among 30 independent runs over UF1-UF10 problems.

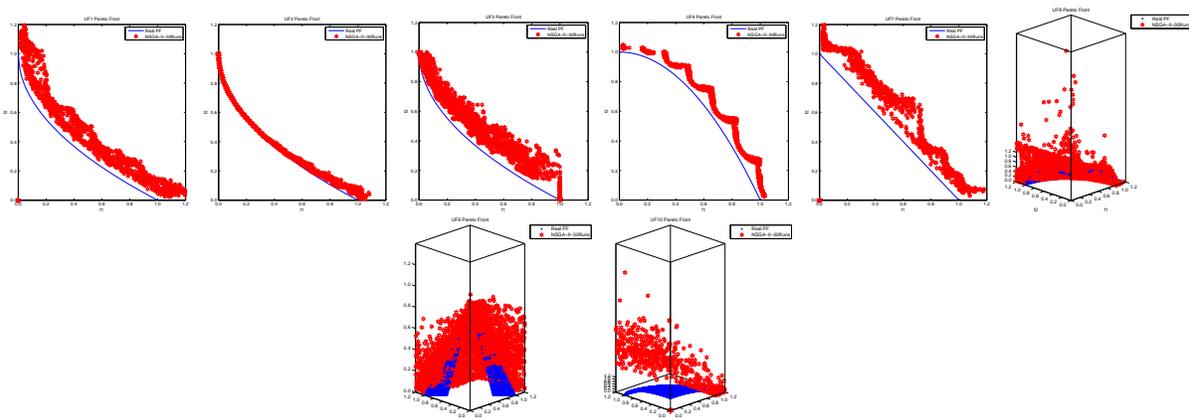


Fig. 9. Plots of the 30 PFs altogether in the objective space display by NSGA-II [13] for UF1-UF4 and UF7-UF10 problems.

(TPMLSM).

- Cancer chemotherapy problems.
- Fuzzy multi-objective reliability redundancy allocation problem.
- Multiobjective Engineering design problems.
- Multi-Objective Capacitated Arc Routing Problem.
- Passive Vehicle Suspension Optimization.
- Multi-objective mobile agent-based Sensor Network Routing.
- Oil and Gas Well Drilling problems.

We will also examine the effect of various enhanced versions of differential evolution [38] with self-adaptive capabilities in our proposed algorithm on the problems mentioned above as well as other complicated test problems.

REFERENCES

- [1] N. Al Moubayed, A. Petrovski, and J. McCall, "A novel smart multi-objective particle swarm optimisation using decomposition," in *Proceedings of the 11th international conference on Parallel problem solving from nature: Part II*, ser. PPSN'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–10.
- [2] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, "Hypervolume-based multiobjective optimization: Theoretical foundations and Practical Implications," *Theoretical Computer Science*, vol. 425, pp. 75–103, 2012.
- [3] J. Bader, "Hypervolume-Based Search for Multiobjective Optimization: Theory and Methods," Ph.D. dissertation, ETH Zurich, Switzerland, 2010.
- [4] J. Bader and E. Zitzler, "HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization," *Evolutionary Computation*, vol. 19, no. 1, pp. 45–76, 2011.
- [5] N. Beume, B. Naujoks, and M. Emmerich, "SMS-EMOA: Multiobjective Selection based on Dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.
- [6] Carlos and R. L. Becerra, "Evolutionary Multi-Objective Optimization in Materials Science and Engineering," *Materials and Manufacturing Processes*, vol. 24, no. 2, pp. 119–129, 2009.
- [7] P. C. Chang, S. H. Chen, Q. Zhang, and J. L. Lin, "MOEA/D for Flowshop Scheduling Problems," in *IEEE Congress on Evolutionary Computation, CEC'08*, June 2008, pp. 1433–1438.
- [8] C.-M. Chen, Y.-P. Chen, and Q. Zhang, "Enhancing MOEA/D with Guided Mutation and Priority Update for Multi-Objective Optimization," in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009*, 2009, pp. 209–216.
- [9] C. A. C. Coello, "A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques," *Knowledge and Information Systems*, vol. 1, pp. 269–308, 1999.
- [10] C. A. C. Coello, G. B. Lamont, and D. A. V. Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [11] C. A. Coello Coello, G. B. Lamont, and D. A. Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, New York, March 2002.
- [12] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, 2nd ed., S. Ross and R. Weber, Eds. John Wiley and Sons Ltd, 2002.
- [13] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," *IEEE Transactions On Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.
- [14] D.H.Phan and J.Suzuki, "R2-IBEA: R2 indicator based evolutionary algorithm for multiobjective optimization," in *2013 IEEE Congress on Evolutionary Computation (CEC)*, June 2013, pp. 1836–1845.
- [15] R. Eberhart and J. Kennedy, "A New Optimizer Using Particle Swarm Theory," in *Proceedings of the Sixth International Symposium on Micro Machine and Human Science, MHS'95*, Oct. 1995, pp. 39–43.
- [16] E. Zitzler and L. Thiele, "Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Evolutionary Algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, 1999.
- [17] C. Fonseca and P. Fleming, "Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization," in *Proceedings of the 5th International Conference on Genetic Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, pp. 416–423.
- [18] —, "An Overview of Evolutionary Algorithm in Multi-Objective Optimization," *Evolutionary Computation*, vol. 3, no. 1, pp. 1–16, 1995.
- [19] J. Horn, N. Nafpliotis, and D. E. Goldberg, "A Niche Pareto Genetic Algorithm for Multiobjective Optimization," in *Proceedings of the First IEEE Conference on Evolutionary Computation, CEC'94*, 1994.
- [20] H. Ishibuchi and T. Murata, "Multi-Objective Genetic Local Search Algorithm and Its Application to Flowshop Scheduling," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, no. 3, pp. 392–403, 1998.
- [21] H. Ishibuchi and T. Murata, "Multi-Objective Genetic Local Search Algorithm," in *Proceedings of the Third IEEE International Conference on Evolutionary Computation*, I. T. Fukuda and T. Furuhashi, Eds., Nagoya, Japan, 1996, pp. 119–124.
- [22] H. Ishibuchi, N. Tsukamoto, Y. Sakane, and Y. Nojima, "Indicator-based Evolutionary Algorithm with Hypervolume Approximation by Achievement Scalarizing Functions," in *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '10. New York, NY, USA: ACM, 2010, pp. 527–534.
- [23] J.D.Knowles and D.W.Corne, "Local Search, Multiobjective Optimization and the Pareto Archived Evolution Strategy," in *Proceedings of the Third Australia-Japan Joint Workshop on Intelligent and Evolutionary Systems*. Ashikaga Institute of Technology, Japan, November 1999, pp. 209–216.
- [24] K. Deb and R. Agrawal, "Simulated Binary Crossover for Continuous Search Space," *Complex System*, vol. 9, pp. 115–148, 1995.
- [25] W. Khan, "Hybrid multiobjective evolutionary algorithm based on decomposition," PhD, Department of Mathematical Sciences, University of Essex, Wivenhoe Park, CO4 3SQ, Colchester, UK, January 2012.
- [26] W. Khan and Q. Zhang, "MOEA/D-DRA with Two Crossover Operators," in *UK Workshop on Computational Intelligence (UKCI)*, September 2010, pp. 1–6.
- [27] J. Knowles and D. Corne, "The Pareto Archived Evolution Strategy: A new Baseline Algorithm for Pareto Multiobjective Optimization," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC'99)*, Piscataway, NJ, JULY 1999, pp. 98–105.
- [28] A. Konak, D. W. Coit, and A. E. Smith, "Multi-Objective Optimization Using Genetic Algorithms: A tutorial," *Reliability Engineering and System Safety*, vol. 91, no. 7, p. 9921007, July 2006.
- [29] M. Koppen, "On the benchmarking of multiobjective optimization algorithm," in *Knowledge-Based Intelligent Information and Engineering Systems, Proceedings Lecture Notes in Artificial Intelligence, 2773*, 2003, pp. 379–385.
- [30] H. Li and Q. Zhang, "Multiobjective Optimization Problems With Complicated Pareto Sets: MOEA/D and NSGA-II," *IEEE Transactions On Evolutionary Computation*, vol. 13, no. 2, pp. 284–302, April 2009.
- [31] W. K. Mashwani, "A Multimethod Search Approach Based on Adaptive Generations Level," in *Seventh International Conference on Natural Computation (ICNC'11)*, Shanghai, China, 26-28 July, 2011, pp. 23–27.
- [32] —, "Hybrid Multiobjective Evolutionary Algorithms: A Survey of the State-of-the-art," *International Journal of Computer Science Issues*, vol. 8, no. 6, pp. 374–392, 2011.
- [33] —, "Integration of NSGA-II and MOEA/D in Multimethod Search Approach: Algorithms," in *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*. ACM, 2011, pp. 75–76.
- [34] —, "MOEA/D with DE and PSO: MOEA/D-DE+PSO," in *The Thirty-first SGAI International Conference on Innovative Techniques*

- and Applications of Artificial Intelligence, Cambridge, UK, December, 2011, pp. 217–221.
- [35] —, “Comparison of Evolutionary Algorithm over Multiobjective Optimization Problems,” in *Proceeding of International Conference on Modeling and Simulation (ICOMS)*, Air University Islamabad, Pakistan, 2013.
- [36] —, “Comprehensive Survey of the Hybrid Evolutionary Algorithms,” *International Journal of Applied Evolutionary Computation (IJAE)*, vol. 4, no. 2, pp. 1–19, July 2013.
- [37] —, “Performance of AMALGAM over CEC’09 Test Instances,” in *Proceeding Third International Conference on Aerospace Science and Engineering (ICASE’13)*, 2013.
- [38] —, “Enhanced versions of Differential Evolution: state-of-the-art survey,” *International Journal Computing Sciences and Mathematics*, vol. 5, no. 2, pp. 107–126, 2014.
- [39] W. K. Mashwani and A. Salhi, “A Decomposition-Based Hybrid Multiobjective Evolutionary Algorithm with Dynamic Resource Allocation,” *Applied Soft Computing*, vol. 12, no. 9, pp. 2765–2780, 2012.
- [40] —, “Multiobjective Evolutionary Algorithm Based on Multimethod with Dynamic Resources Allocation,” *Applied Soft Computing*, vol. 39, pp. 292–309, 2016.
- [41] W. K. Mashwani, A. Salhi, M. A. Jan, R.A.Khanum, and M. Sulaiman, “Impact Analysis of Crossovers in Multiobjective Evolutionary Algorithm,” *Science International Journal*, vol. 27, no. 6, pp. 4943–4956, December 2015.
- [42] W. K. Mashwani and A. Salhi, “Multiobjective Memetic Algorithm Based on Decomposition,” *Applied Soft Computing*, vol. 21, pp. 221–243, 2014.
- [43] K. M. Miettinen, *Nonlinear Multiobjective Optimization*, ser. Kluwer’s International Series. Norwell, MA: Academic Publishers Kluwer, 1999.
- [44] N.Srinivas and K.Deb, “A Multiobjective Optimization using Nondominated Sorting in Genetic Algorithms,” *J Evol Comput*, vol. 2, no. 3, pp. 221–248, 1994.
- [45] P.Moscato, “On evolution, Search, Optimization, Genetic Algorithms and martial arts: Towards memetic algorithms,” California Institute of Technology, Pasadena, California, USA, Tech. Rep. Caltech Concurrent Computation Program 826, 1989.
- [46] R.Storn and K.V.Price, “Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces,” ICSI, Technical Report TR-95-012, 1995.
- [47] —, “Differential Evolution - a Simple and Efficient Heuristic for Global Optimization over Continuous Spaces,” *J.Global Opt*, vol. 11, no. 4, pp. 341–359, December 1997.
- [48] J. D. Schaffer, “Multiple Objective Optimization with Vector Evaluated Genetic Algorithms,” in *Proceedings of the 1st International Conference on Genetic Algorithms (ICGA)*, Pittsburgh, USA, July. Lawrence Erlbaum Associates, 1985, pp. 93–100.
- [49] D. Simon, *Evolutionary Optimization Algorithms: Biologically Inspired and Population-Based Approches to Computer Intelligence*. John Wiley & Sons, 2013.
- [50] S. Tsutsui, M. Yamamura, and T. Higuchi, “Multi-Parent Recombination with Simplex Crossover in Real coded Genetic Algorithms,” in *Proceeding of GECCO-99*, 1999, pp. 657–374.
- [51] D. A. V. Veldhuizen, “Multiobjective Evolutionary Algorithms: Classifications, Analyses, and New Innovations,” Graduate School of Engineering of the Air Force Institute of Technology Air University, Tech. Rep., 1999.
- [52] J. A. Vrugt and B. A. Robinson, “Improved Evolutionary Optimization from Genetically Adaptive Multimethod Search,” *Proceedings of the National Academy of Sciences of the United States of America: PNAS (USA)*, vol. 104, no. 3, pp. 708–701, 16th Jaanuary 2007.
- [53] J. A. Vrugt, B. A. Robinson, and J. M. Hyman, “Self-Adaptive Multimethod Search for Global Optimization in Real-Parameter Spaces,” *IEEE Transsation On Evolutionary Computation*, vol. 13, no. 2, pp. 243–259, April 2009.
- [54] Q. Zhang and H. Li, “MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition,” *IEEE transaction on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, December 2007.
- [55] Q. Zhang, H. Li, D. Maringer, and E. P. K. Tsang, “MOEA/D with NBI-style Tchebycheff Approach for Portfolio Management,” in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2010, Barcelona, Spain, 18-23 July 2010*, 2010, pp. 1–8.
- [56] Q. Zhang, W. Liu, and H. Li, “The Performance of a New Version of MOEA/D on CEC’09 Unconstrained MOP Test Instances,” *IEEE Congress On Evolutionary Computation (IEEE CEC 2009), Trondheim, Norway*, pp. 203–208, May, 18–21 2009.
- [57] Q. Zhang, A. Zhou, S. Zhaoy, P. N. Suganthany, W. Liu, and S. Tiwariz, “Multiobjective Optimization Test Instances for the CEC 2009 Special Session and Competition,” Technical Report CES-487, 2009.
- [58] S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, “Decomposition-Based Multiobjective Evolutionary Algorithm With an Ensemble of Neighborhood Sizes,” *IEEE Trans. Evolutionary Computation*, vol. 16, no. 3, pp. 442–446, 2012.
- [59] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang., “Multiobjective evolutionary algorithms: A survey of the state-of-the-art,” *Swarm and Evolutionary Computation*, vol. 1, pp. 32–49, 16 March 2011, online published.
- [60] E. Zitzler, M. Laumanns, and L. Thiele, “SPEA2: Improving the Strength Pareto Evolutionary Algorithm,” Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Zurich, Switzerland, TIK Report 103, 2001.
- [61] E. Zitzler and L. Thiele, “An Evolutionary Approach for Multiobjective Optimization: The Strength Pareto Approach,” Computer Engineering and Networks Laboratory (TIK), ETH Zurich, TIK Report 43, May 1998.
- [62] E. Zitzler, K. Deb, and L. Thiele, “Comparision of Multiobjective Evolutionary Algorithms: Emperical Results,” *Evolutionary Computation*, vol. 8, no. 2, pp. 173–195, 2000.
- [63] E. Zitzler and S. Knzli, “Indicator-Based Selection in Multiobjective Search,” in *Parallel Problem Solving from Nature - PPSN VIII*, ser. Lecture Notes in Computer Science, X. Yao, E. Burke, J. Lozano, J. Smith, J. Merelo-Guervs, J. Bullinaria, J. Rowe, P. Tino, A. Kabn, and H.-P. Schwefel, Eds. Springer Berlin Heidelberg, 2004, vol. 3242, pp. 832–842.
- [64] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca, “Performance Assessment of Multiobjective Optimizers: An Analysis and Review,” *IEEE Transactions on Evolutionary Computation*, vol. 7, pp. 117–132, 2003.

Extracting Topics from the Holy Quran Using Generative Models

Mohammad Alhawarat
Department of Computer Science,
Prince Sattam Bin Abdulaziz University,
Alkharj, Saudi Arabi

Abstract—The holy Quran is one of the Holy Books of God. It is considered one of the main references for an estimated 1.6 billion of Muslims around the world. The Holy Quran language is Arabic. Specialized as well as non-specialized people in religion need to search and lookup certain information from the Holy Quran. Most research projects concentrate on the translation of the holy Quran in different languages. Nevertheless, few research projects pay attention to original text of the holy Quran in Arabic language. Keyword search is one of the Information Retrieval (IR) methods but will retrieve what is called exact search. Semantic search aims at finding deeper meanings of a text, and it is a hot field of study in Natural Language Processing (NLP). In this paper topic modeling techniques are explored to setup a framework for semantic search in the holy Quran. As the Holy Quran is the word of God, its meanings are unlimited. In this paper the words of chapter Joseph (Peace Be Upon Him (PBUH)) from the Holy Quran is analyzed based on topic modeling techniques as a case study. Latent Dirichlet Allocation (LDA) topic modeling technique has been applied in this paper into two structures (Hizb Quarters and verses) of Joseph chapter as: words, roots and stems. The log-Likelihood has been calculated for the two structures of the chapter. Results show that the best structure to use is verses, which gives the least energy for data. Some of the results of the attained topics are shown. These results suggest that topic modeling techniques failed to capture in an accurate manner the coherent topics of the chapter.

Keywords—Statistical models; Latent Dirichlet Analysis (LDA); Holy Quran; Unsupervised Learning

I. INTRODUCTION

The holy Quran is considered an essential reference for Muslims where they read in a regular basis. They usually need to search it and retrieve relevant information based on more than just simple keyword search techniques.

Dealing with the holy Quran is different from dealing with regular Arabic corpora that is usually extracted from Newspapers and speeches, and hence is the word of human. The holy Quran is the word of God and the meanings of its words are unlimited. The sequence of text is different from human words. For example, one topic could repeat in different places in the holy Quran with different details and sometimes in different contexts. Also, one chapter usually has many topics. While one topic might be started in one verse, another topic may starts immediately in the next verse. Also, one verse may have different topics. Moreover, there are different authentic interpretations for the verses of the holy

Quran; therefore it is very hard for a computer to manage them in the way scholars do especially in situations where meanings are seem opposite to each other. Finally, there is much relevant information that is found in prophet Mohammad (PBUH) sayings (Hadith) that interpret many verses of the holy Quran. For all of these reasons, it sometimes hard to resolve a disambiguation if a word has many synonyms and different senses.

Research in Arabic NLP still young and have many challenges [1]. This is because that Arabic language is different from many other natural languages [2], [3]. Words in Arabic language have many derivations and have also complex Diglossia (modern and colloquial) [4]. Also, Arabic letters appear in different shapes according to their position in the word. Another characteristic of the Arabic language is the diacritic. Some of these diacritical marks are usually not written, but is understood by Arabic readers. Therefore, two exact written words without diacritical marks have totally different meanings. All of these and other characteristics of the Arabic language should be taken in consideration when processing Arabic text.

The holy Quran can be considered as a "Golden Text" to use in Text mining and NLP fields. This might be true for different reasons: it's the word of God, it's limited in terms of text size and it has many translations and many interpretations. These all together encourage building a semantic comprehensive source for the holy Quran that will allow advanced semantic search and knowledge extraction.

Searching in the holy Quran is an essential task for Muslims as well as non-Muslims who study it. Many applications have been built to allow search in the holy Quran. Most of these search engines allow simple search techniques where some of them are mentioned in [5]. However, few research projects are concerned with advanced search in the holy Quran using some NLP techniques such as the papers presented in The holy Quran and new technology workshop that held by King Fahad Complex for printing the holy Quran in Al-Madinah Al-Munawwarah, Saudi Arabia in 2008. The workshop participants discussed different issues related to the holy Quran including searching techniques. Also more papers are presented in another event in Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences that held in Al-Madinah Al-

Munawwarah, Saudi Arabia in 2013. The presented papers are related to a wide range of topics concerning the holy Quran including natural processing issues, security, education and many more.

There are different approaches to model and cluster topics in text documents such as LDA, Latent Semantic Analysis (LSA) and traditional clustering techniques such as K-means. In this research LDA is used for several reasons including accuracy, scalability and comprehension [6], [7], [8].

LDA has been developed to extract topics from text using statistical methods [9], [10], [11], [12], [13]. LDA is one of the techniques that belongs to a large family called probabilistic modeling. The basic intuition behind LDA is that a text document has multiple topics where each topic is defined as a distribution over a set of words. There are many flavors of the LDA model; a thorough review of the LDA topic modeling techniques can be found in [14]. Topic modeling has been applied to many field of study such as Information Retrieval IR, geographical IR, computational linguistics and NLP [15], [16], [17], [18], [19], [20].

This paper aims to build up the first stage in a framework that will allow possible semantic search in the holy Quran. This is done by applying LDA topic modeling to chapter Joseph of the holy Quran as a case study. This chapter has been chosen because it includes relative topics regarding story of the prophet Joseph (PBUH). The LDA topic modeling has been applied to words, roots and stems of that chapter. Next stages might include: studying the topics of the whole holy Quran, linking the text of the holy Quran to both authenticate interpretation of the holy Quran and the related Sayings of the prophet Mohammad (PBUH). These might be achieved using machine learning, text mining as well as NLP techniques.

It should be stated explicitly here that this research is not a religious study; rather it is a statistical study that might result in information that would guide specialized religious people to understand more about the word of God.

The paper is organized as follows: in section II related work is presented, in section III topic modeling is introduced, in section IV the methodology as well as preparation of the Data Set is explained, in section V experimental setups are explained, section VI includes discussion of the results attained in the paper and finally section VII contains conclusion.

II. RELATED WORK

Shoaib et. al. [5] have proposed a simple WordNet for the English translation of the second chapter of the holy Quran (Al-Baqrah). They have created topic-synonym relations between the words in that chapter with different priorities. They have defined different relations that are used in traditional WordNet such as: synonymy, polysemy, hyperonymy, hyponymy, holonymy and meronymy. Then they developed a semantic search algorithm that will fetch all verses that contains the query word and its synonyms with high priority. It is not clear how the authors build their simple WordNet. In similar studies, usually authentic religion references should be used such as interpretation of the holy Quran or meanings of the words of the holy Quran. However, the results show that the developed semantic search outperform simple search algorithms.

Similar work has been carried out to extract verses from the holy Quran using an expert system that use Web Ontology Language (OWL) [21]. Again the work use English translation of the holy Quran and not Arabic language.

Another work explored the structure of a simple domain Quran ontology for birds and animals that are mentioned in the holy Quran [22]. The authors propose a framework for semantic search in the holy Quran using their domain ontology and they have evaluated it using SPARQL query language. This work uses English translation of the holy Quran.

Data mining techniques such as SVM and naive Bayesian classifiers are used cluster chapters of the holy Quran based on Major Phases of Prophet Mohammads (PBUH) Messengership [23]. This work classifies chapters of the holy Quran rather than verses or words of the holy Quran.

LDA topic modeling technique has been used to extract topics from an Arabic corpora composed of Newspapers [24]. The authors have developed a preprocessing lemma-based stemming algorithm and then applied the LDA technique on Arabic processed text.

In [25] author has used clustering techniques in machine learning to extract topics of the holy Quran. The extraction of topics was based on a corpus that is composed of the verses of the holy Quran using nonnegative matrix factorization. The author used Buckwalter code for Arabic letters [3]. Topics are visualized and related verses for each topic are shown for selected topics based on the topic main keywords. One of the shortcoming of his work is that verses are dealt with separately as each as a document. The author claims that he has extracted and identified the underlying topics of the holy Quran. However, this claim is far from reality as no one could identify the underlying topics of the holy Quran even well-known scholars of Quran studies. Also, the it is totally unclear how he has linked the keywords of each topic with the related verses that correspond to topic keywords. Nevertheless, the findings are promising and might help in revealing deeper meanings of the holy Quran by specialized people in Quranic studies.

LDA technique has been compared LDA with K-means clustering technique [8]. The authors have applied both LDA and K-means technique on a set of Arabic documents from OSAC (Open Source Arabic Corpora). The results show that LDA outperforms K-means in most instances.

III. TOPIC MODELING

Topic modeling is a hot field of study in both machine learning and NLP. Topic models are generative models that are based on probability distributions of multiple topics in a document over a set of words. Such models basically depend on term-frequencies in a document. One of these models is LDA. As mentioned previously, LDA is better than other models such as LSA for several reasons[6], [7], [8]. LDA outperforms LSA in many applications including semantic representation [12] and have been used in different fields in the last decade or so including NLP [15], [16], [17]. It is used by researchers to extract important and hot topics; usually from large corpora.

The basic intuition behind LDA is that a set of words of documents are randomly pre-assigned with probability distributions that would represent multiple-topic latent structure on those documents. After that, latent structure of the topics of documents is inferred statistically in a reverse-engineering manner.

Initially, a number of topics T should be specified. Then, a term distribution ϕ over a parameter β is chosen for each topic. After that, ratios θ of topic distribution for document d are specified. Then, a topic z_i is chosen and after that a word is chosen conditioned on that topic over a parameter α . Both ϕ and θ are Dirichlet distributions.

The probability of the i th word in a specific document is given by:

$$P(w_i) = \sum_{j=1}^T P(w_i|z_i = j)P(z_i = j) \quad (1)$$

where z_i represents a latent variable that designates the topic for the drawn i th word. $P(w_i|z_i = j)$ represents the probability of the word w_i under topic j . $P(z_i = j)$ represents the probability of a word from topic j of a document.

Note that $P(w|z)$ can be represented by a multinomial distributions ϕ over a term distribution such that $P(w|z = j) = \phi_w^{(j)}$ and $P(z)$ can be represented by a multinomial distributions θ over a topic distribution over D documents such that $P(z = j) = \theta_j^{(d)}$.

Then an estimation method is used to infer the latent structure of the topics of documents. Different estimation methods can be used in this context including: Variational Expectation-Maximization (VEM) method and Gibbs sampling. For more information about details of these methods please refer to [10], [11], [13], [26].

Besides LDA, Correlated Topic Model (CTM) can be used to extract correlated topics from documents. CTM is an extension of LDA. LDA usually uses Gibbs sampling for model estimation.

IV. DATA SET PREPARATION AND METHODOLOGY

The text of chapter Joseph in the format of CP1256 has been taken from [27] in the shape of two structures: Hizb quarters and verses, all without diacritic. The frequency details of these selected structures are shown in table I. For more information about the text structure of the holy Quran please refer to [27].

TABLE I: The number of documents for the Joseph chapter based on different structures and for words, roots and stems after applying tf-idf measure on DTMs

	No. of Hizb quarters/terms	No. of Verses/terms
Original No. (TF)	6/721	111/721
With TF-IDF (words)	6/299	89/323
With TF-IDF (roots)	6/327	108/163
With TF-IDF (stems)	6/398	103/193

These two structures will be used in the topic modeling process in three shapes: words, roots and stems. Because

the text of the holy Quran is the word of God, there is no margin for errors in the process of extracting both roots and stems. Therefore, the roots and stems have been constructed manually; based on two web sites [28], [29] and verified by the authors according to their experience in Arabic language and as native speakers.

These data sets will be used as the input for the implementation of the LDA to reveal the main topics for the text of the chapter of Joseph (PBUH). Different experimental setups are prepared to compute the topic models for the text of that chapter based on the aforementioned structures.

V. EXPERIMENTS

Both packages `tm` and `topicmodels` of R are used in experiments (a practical guide for `topicmodels` can be found in [30]). First, the `tm` package will be used for text preparation and processing as building the corpus, removing stop words and building the Document Term matrix (DTM). Second, the `topicmodels` package will be used to build and fit LDA model for all structures of the text with the three shapes of word.

The text with two structures has been processed where the stop words are removed. Then, three DTMs have been built for text as: words, roots and stems. The content of the DTM is basically calculated using Term Frequencies (TF) measure. After that, the tf-idf measure has been applied on each DTM to remove frequent terms that appears on most documents, and hence are not recognized as important terms. This has been done by calculating the median and choosing high-frequent terms with frequency more than the calculated median.

TABLE II: The number of topics along with the log-Likelihood for the fitted topic models for the Joseph chapter estimated by Gibbs sampling with 10-fold cross-validation

	Hizb quarters	Verses
	Topics No./Log-Likelihood	Topics No./Log-Likelihood
Word	17/-1258	8/-250
Root	44/-827	5/-169
Stem	27/-860	19/-172

After that, different experimental setups are prepared to find the main topics in the chapter of Joseph (PBUH). These are found first using TF measure and then using different estimation techniques for LDA besides Correlated Topics Model (CTM)-where CTM can use VEM only:

- VEM.
- VEM with fixed α .
- Gibbs

Then, a validation technique that is based on the log-Likelihood of the data set is calculated. This is performed to find the best number of topics for each structure of that chapter. The best number of the topics is calculated using 10-fold cross-validation technique for the two structures with the three term shapes, results of log-likelihood and number of topics are shown in table II. Then the topics are recorded for all cases using the best topic numbers that are calculated according to the aforementioned technique. In some cases different topics

number is chosen because the energy-based topics number is large. The main parameters are set as suggested by [11] where $\alpha = 50/k$ (where k is the number of topics) and $\beta = 0.1$. In many of the experiment setups, the seed parameter of the LDA and CTM models are set to the number of terms according to table I.

Samples of the results of the topics are shown in figures 1 - 13 for the two structures with three shapes of the terms: words, roots and stems. Figures 1 - 9 represent the *Verses* structure where figures 1 - 3 are for words, figures 4 - 6 are for stems and figures 10 - 13 are for roots. Figures 10 - 13 represent the *Hizb Quarters* structure for words, roots and stems.

Topic.2	Topic.4	Topic.8	Topic.11	Topic.13	Topic.15
روح	آيات	الراحمين	الجب	يشعرون	أبا
أباه	ادخلوا	نعزي	أبرئ	أبيهم	الآخرة
أبيننا	أشدده	أرباب	آتيانه	أجمعين	الخاننين
اجعلوا	اقتلوا	أياهم	الذنب	أخنه	الواحد
إخوة	الذنب	الرحيم	السقاية	أستخلصه	آمنوا
أسفى	القصص	السوء	النفس	الساعة	أهلهم
أكله	أمة	القديم	أمرهم	تقتلوا	بالله
الباب	بشمن	تزرعون	أنتكم	خاطئين	بصيرة
الصاغرين	تبتئس	ذكر	بالله	سنبله	بغثة
أنزلناه	تعقلون	رحله	تفقدون	عبادنا	جننا

Fig. 1: Sample of topics for words based on **Verses** where **Gibbs sampling** is used (Topics Number is 17)

Topic.1	Topic.4	Topic.10	Topic.12	Topic.15	Topic.17
وعاء	الجب	الذنب	كيدكن	بالله	تفقدون
تفقدون	غيابت	تأتيهم	تعلمون	يأتيكما	واقبلوا
واقبلوا	أبا	مناعنا	أبيهم	الآخرة	الآخرة
الآخرة	السيارة	يشعرون	أشكو	قليلا	أجر
أجر	بأمرهم	عصبة	الكاذبين	أرباب	تعقلون
تعقلون	تقتلوا	أفأمنوا	أهلها	أكثرهم	أرسله
أرسله	ذهبوا	أكله	بني	القهار	تعلمون
تعلمون	شيخا	الساعة	راودتني	الواحد	جاء
جاء	فاعلين	بغثة	رجعوا	آمنوا	الصادقين
الصادقين	فخذ	تأمننا	شاهد	تأكلون	

Fig. 2: Sample of topics for words based on **Verses** where **CTM** is used (Topics Number is 17)

Topic.2	Topic.3	Topic.5	Topic.10	Topic.12	Topic.17
يوسف	سبع	قالوا	قالوا	أبانا	الناس
أجر	الملك	ربه	أبانا	قالوا	أكثر
اخرج	بقرات	السجن	أراني	كيل	تأويل
أرسلت	خضر	تالله	الذنب	أخانا	ربي
أسفى	سمان	يوسف	الله	العير	وقال
أكبرنه	سنبلات	أبيكم	عصبة	الكيل	مكننا
الآيات	عجاف	أحلام	قال	بجهازهم	يشاء
الحزن	وأخر	اذكرني	أبيننا	جهازهم	الأرض
آيات	وسع	أرضنا	أحب	ردت	ليوسف
أيديهن	وقال	أصغاث	أحمل	قال	آبائي

Fig. 3: Sample of topics for words based on **Verses** where **TF** is used (Topics Number is 17)

Topic.3	Topic.5	Topic.9	Topic.13	Topic.15	Topic.19
كيل	جزاء	سرق	سبع	توكل	آية
جهاز	ظالم	شهد	حزن	سبع	سماء
جهاز	جعل	كاذب	ابيض	قليل	الر
سارق	حلم	ابن	أخاف	أحصن	بدا
أوفى	عرف	إستخلص	أسف	إخوة	رحمن
رحل	انقلب	أمين	تولى	حصد	رحيم
سقاية	جزى	حافظ	ذنب	خلا	كتاب
عير	حفيظ	شاهد	شكا	دأب	معرض
فسد	خزانة	صدق	عين	زرع	إخوة
منزل	رحال	غيب	غافل	سائل	سائل

Fig. 4: Sample of topics for stems based on **Verses** where **VEM** is used (Topics Number is 19)

Topic.3	Topic.7	Topic.9	Topic.10	Topic.14	Topic.18
سجن	كيل	قميص	يوسف	سبع	قال
رأى	بضاعة	ذهب	خاطئ	أكل	صادق
ذكر	وجد	جزاء	سرق	سنبله	عير
نبأ	قال	باب	استغفر	آخر	سأل
خمر	متاع	أهل	قال	رؤيا	جهاز
رأس	أهل	دير	ذنب	أخضر	جهاز
طير	قتل	الله	مكان	أفتى	سارق
صاحب	فاعل	يأس	الله	بقرة	رحل
عصر	غياب	روح	ابيض	سمنين	أقبل
قضى	بعير	استيق	أسف	عجاف	أذن

Fig. 5: Sample of topics for stems based on **Verses** where **VEM with fixed α** is used (Topics Number is 19)

Topic.1	Topic.3	Topic.7	Topic.11	Topic.14	Topic.18
قال	دخل	رأى	يوسف	سبع	كيل
كيد	باب	قال	جاء	وعاء	قال
أمن	قميص	سجن	قال	علم	بضاعة
جزاء	دبر	آخر	رؤيا	أكل	أهل
ضلال	قال	أكل	شيطان	آخر	بعير
مبين	توكل	خبر	فاعل	أخضر	جاء
أحب	حكم	رأس	قتل	أفتى	رجع
رأى	سجن	طير	أحسن	بقرة	فقد
ظالم	متفرق	عصر	أخرج	رؤيا	متاع
وجد	واحد	محسن	أرض	سمين	وجد

Fig. 6: Sample of topics for **stems** based on Verses where **TF** is used (Topics Number is 19)

Topic.1	Topic.3	Topic.5	Topic.8	Topic.11	Topic.14
إله	إله	أبو	سبع	أكل	أني
يأس	قول	قول	أكل	ذهب	قول
صرف	دخل	يوسف	رأي	نبأ	ريب
وثق	جزى	علم	سنبل	قول	علم
أبو	وكل	جهز	فتو	أمر	ملك
حكم	ظلم	سرق	قول	جمع	رسل
خلص	بوب	أخو	آخر	ذنب	أله
روح	جياً	سأل	بقر	رأي	أول
قوم	فرق	عير	خضرت	سجن	تم
نحو	وجد	أذن	سمن	آخر	سمع

Fig. 9: Sample of topics for **roots** based on Verses where TF is used (Topics Number is 15)

Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
سبع	جزى	رحم	بوب	غيب
سرق	كيل	وحي	جمع	رحم
سمو	ذكر	أذن	وكل	خطأ
جهز	كيد	بصر	ظلم	ذنب
كذب	ضلل	حزن	غفر	شهد
أجر	عير	شرك	فعل	غفر
عبد	أبي	قصص	دلو	تم
نزل	بأس	وعى	ذنب	خون
وثق	بدو	سقى	سأل	قدم
أوي	تبع	شعر	شدد	قرأ

Fig. 7: Sample of topics for **roots** based on Verses where **Gibbs sampling** is used (Topics Number is 5)

Topic.1	Topic.2	Topic.6	Topic.9	Topic.12	Topic.15
سأل	حزن	سبع	جزى	ذكر	بوب
خطأ	ولي	ضلل	ظلم	جهز	وكل
غفر	ذنب	شدد	وعى	أذن	دبر
دلو	أسف	قدم	تم	عير	وثق
ذنب	بنت	قلل	شري	رحل	صحب
عرض	بيض	بلغ	بخس	سقى	فرق
أثر	حرض	حصن	ثمن	جعل	قهتر
حرص	خوف	خلو	درهم	عرف	جزى
عير	شكي	طرح	زهد	سرق	سقى
أبي	شمس	قتل	عدد	أمم	ظلم

Fig. 8: Sample of topics for **roots** based on Verses where **VEM** is used (Topics Number is 15)

VI. RESULTS AND DISCUSSION

The number of documents and terms of the chapter of Joseph (PBUH) is shown in table I. Both TF and TF-IDF measures are used and then the number of documents and terms are recorded for words, roots and stems. The results of applying LDA model to the text of the chapter with Gibbs sampling technique is shown in table II. Note that the term with low energy are roots and stems compared with high energy for words.

Many experiment setups have been carried out with different parameter settings apart from the aforementioned setups in section V. Sample of the results are shown in figures 1 - 13. All the results of all the experiments show that most of the resulted topics are a mix of more than one topic. However, very few topics form one coherent topic such as topic number three of figure 3 and topic number three and fourteen of figure 5.

Some topics include a mix of two to may be five topics. In some cases all of the terms of the topic are coherent except one or two words such as topic number 12 of figure 10.

Regarding the shapes of the word; on one hand the roots are considered problematic as there are many shared words between topics such as the topics that appear in figure 12. One of the reasons behind this is that there are some different words in meaning but their root in Arabic language is the same. On the other hand, both words and stems show better results as it appear in most of the figures. For words it is obvious that each word has usually its own semantic in one context. For stems, although there is more than a word with the same stem but they have the same semantic in similar contexts.

The estimation methods that are used in this study show different "percentage of successful" with different shapes of words. For example, TF measure gives better results than TF-IDF measure in certain cases. On another occasion, CTM gives better results. The same is true for VEM, VEM with fixed α and Gibbs sampling.

Also, it is important to mention that all of the numerical results including best number of topics as well as log-Likelihood of the data are based on the seed parameter

for LDA and CTM models. However, many experiments are executed with different values for seed parameter without affecting the quality of the resulted topics.

In other set of experiments, the parameter alpha is set to smaller numbers than that suggested by [11] where $\alpha = 50/k$ (k is the number of topics). When α is set to $1/k$, the results show topics with slightly better quality.

Although the topic modeling techniques used in this study failed to extract coherent topics, still the results are promising as some topics are coherent even that they are very few.

Topic.1	Topic.4	Topic.5	Topic.9	Topic.12	Topic.16
قالوا	قميصه	قالوا	ربك	سبع	قالوا
دخلوا	الذئب	جزاء	واسحاق	السجن	أخيه
تعقلون	والله	سيارة	تعقلون	بتأويله	جزاؤه
ربك	دبر	يشعرون	إبراهيم	ربه	كيل
الله	كيدكن	ربه	الرحيم	الله	أخانا
الرحيم	الجب	أعرض	أبت	أراني	أخاه
أمرهم	غيابت	دبر	يعقوب	الآخر	بجهازهم
أبت	عصبة	والله	آيات	الطير	بضاعتهم
القوم	رأى	هيت	ويتهم	العزير	بعير
ليوسف	وجاءوا	بضاعة	للإنسان	امرات	جهازهم

Fig. 10: Sample of topics for **words** based on **Hizb Quarters** where VEM is used (Topics Number is 17)

Topic.6	Topic.8	Topic.13	Topic.18	Topic.22	Topic.27
قميص	أخاف	أحضر	باب	سنيلة	أكل
ذئب	إسحاق	عزيز	عزيز	روح	آخر
سيارة	بأس	يأس	كيل	طير	إسحاق
صالح	بخس	إبراهيم	وعاء	قرى	بكي
إبراهيم	بعير	أفتى	أتم	معدود	خمر
أراد	ساجد	بقرة	إخوة	إبراهيم	دار
أعرض	سبيل	جهاز	أكل	إتبع	رحل
تفصيل	سماء	خمر	عربي	اتخذ	عجاف
شاهد	سمين	غياب	قرآن	أتم	قصص
غلق	سنيلة	إتبع	إبراهيم	اجتبي	إبراهيم

Fig. 11: Sample of topics for **stems** based on **Hizb Quarters** where Gibbs sampling is used (Topics Number is 26)

Topic.1	Topic.9	Topic.16	Topic.28	Topic.31	Topic.40
أخو	قول	سبع	قول	قول	قول
كيل	أخو	أكل	أخو	كيد	سرق
قول	رحم	فتو	ملك	ملك	عزز
دخل	ملك	آخر	آخر	دخل	أخو
رحم	دخل	سنبل	رحم	بين	رحم
وكل	شياً	كيد	دخل	آخر	يأس
جهاز	آخر	خون	فتو	عزز	دخل
شياً	سرق	نسو	سرق	سمو	بصر
رحل	كيل	ملل	شياً	سمن	شياً
وعي	بوب	عصر	كيل	إبراهيم	روح

Fig. 12: Sample of topics for **roots** based on **Hizb Quarters** where VEM is used (Topics Number is 44)

Topic.1	Topic.4	Topic.6	Topic.8	Topic.12	Topic.14
قول	علم	قول	قول	إله	علم
رأي	إله	نفس	يوسف	أبو	إله
سبع	أله	ربب	أبو	قول	يوسف
سجن	جياً	علم	علم	يوسف	رأي
ربب	أمن	أله	قصص	علم	أول
أكل	أتي	أمن	أكل	أخو	ربب
علم	أمر	أبو	أله	غفر	قول
أتي	خير	نبأ	أمر	أله	أبو
آخر	أهل	رود	بوب	دخل	حكم
فتو	أول	أتي	جياً	سرق	حسن

Fig. 13: Sample of topics for **roots** based on **Hizb Quarters** where TF is used (Topics Number is 15)

VII. CONCLUSION

The topicmodels R package has been used to analyse the underlying topics of the chapter Joseph (PBUH). First the best number of topics for the two structures have been calculated for the three shapes of words and the results are shown in table I. After that, several experiment setups are executed for both of the document structures with three term shapes: word, root and stem. Then, results are recorded and samples of the result are shown in figures 1 - 13. The results are evaluated based on understanding of the meanings and interpretation of the chapter of Joseph (PBUH). The results suggest that verses structure is better than Hizb quarters one in forming more coherent topics. Most of the resulted topics include a mix of more than one topic out of the main topics of the chapter of Joseph (PBUH). However, few of the resulted topics contain one coherent topic.

Semantic search in the holy Quran can be supported by finding accurate coherent topics which helps in finding

contextual terms related to the user search terms. The holy Quran contains hundreds of topics if not thousands. While one verse may contain multiple topics, another set of verses may comprise one topic. Also, one topic may repeat in several contexts and in more than one chapter. If the results are enhanced by combining LDA with another technique then they can be then used together to search for relevant words according to the distribution of topics over words.

The results of this study strongly suggests that while statistical methods succeeded in extracting important topics from text corpora of humans -as many studies show, it failed to achieve the same results with the word of God. This is obvious because the words of God are unlimited in meaning and are one of the attributes/characters of God.

Future work may include exploring more statistical methods and/or combining the methods used in this study with other data mining techniques. Also, if the text of the holy Quran would be linked to one of its authentic interpretations, then topic modeling might find coherent topics because interpretations are the word of human.

REFERENCES

- [1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," vol. 8, no. 4, pp. 14:1–14:22, Dec. 2009. [Online]. Available: <http://doi.acm.org/10.1145/1644879.1644881>
- [2] M. Saad and W. Ashour, "Arabic morphological tools for text mining," in *6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010*, 2010, p. 112117.
- [3] N. Y. Habash, *Introduction to Arabic Natural Language Processing*, G. Hirst, Ed. Morgan and Claypool Publishers, 2010.
- [4] M. DIAB and N. HABASH, "Arabic dialect tutorial," in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL07)*, 2007, pp. 29–34.
- [5] M. Shoaib, M. Nadeem Yasin, U. Hikmat, M. Saeed, and M. Khiyal, "Relational wordnet model for semantic search in holy quran," in *International Conference on Emerging Technologies, 2009. ICET 2009.*, Oct 2009, pp. 29–34.
- [6] I. Biro, "Document classification with latent dirichlet allocation," Ph.D. dissertation, Eötvös Loránd University, 2009.
- [7] P. Crossno, A. Wilson, T. Shead, and D. Dunlavy, "Topicview: Visually comparing topic models of text collections," in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, Nov 2011, pp. 936–943.
- [8] A. Kelaiaia and H. Merouani, "Clustering with probabilistic topic models on arabic texts," in *Modeling Approaches and Algorithms for Advanced Computer Applications*, ser. Studies in Computational Intelligence, A. Amine, A. M. Otmane, and L. Bellatreche, Eds. Springer International Publishing, 2013, vol. 488, pp. 65–74.
- [9] D. Blei and J. Lafferty, "Topic models," in *Text Mining: Theory and Applications*, Srivastava and M. Sahami, Eds. Taylor and Francis, 2006.
- [10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [11] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.
- [12] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, "Topics in semantic representation," *Psychological Review*, vol. 114, p. 2007, 2007.
- [13] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning.*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2006.
- [14] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.
- [15] G. K. Gerber, R. D. Dowell, T. Jaakkola, and D. K. Gifford, "Automated discovery of functional generality of human gene expression programs." *PLoS Computational Biology*, vol. 3, no. 8, 2007.
- [16] J. Boyd-Graber, D. M. Blei, and X. Zhu, "A topic model for word sense disambiguation," in *Empirical Methods in Natural Language Processing*, 2007.
- [17] S. Gerrish and D. M. Blei, "Predicting legislative roll calls from text." in *ICML*, L. Getoor and T. Scheffer, Eds. Omnipress, 2011, pp. 489–496.
- [18] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 178–185.
- [19] Z. Li, C. Wang, X. Xie, X. Wang, and W.-Y. Ma, "Exploring lda-based document model for geographic information retrieval," in *Advances in Multilingual and Multimodal Information Retrieval*, ser. Lecture Notes in Computer Science, C. Peters, V. Jijkoun, T. Mandl, H. Mller, D. Oard, A. Peas, V. Petras, and D. Santos, Eds. Springer Berlin Heidelberg, 2008, vol. 5152, pp. 842–849.
- [20] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Association for Computational Linguistics, 2008, pp. 363–371.
- [21] A. A. Aliyu Rufai Yauri, Rabiah Abdul Kadir and M. A. A. Murad, "Quranic verse extraction base on concepts using owl-dl ontology." vol. 6, no. 23, pp. 4492–4498, 2013.
- [22] M. S. Hikmat Ullah Khan, Syed Muhammad Saqlain and M. Sher, "Ontology-based semantic search in holy quran," vol. 2, no. 6, pp. 562–566, 2013.
- [23] M. Nassourou, "Using machine learning algorithms for categorizing quranic chapters by major phases of prophet mohammads messenger-ship," vol. 2, no. 11, pp. 863–871, 2012.
- [24] A. Brahmi, A. Ech-Cherif, and A. Benyettou, "An arabic lemma-based stemmer for latent topic modeling," *Int. Arab J. Inf. Technol.*, vol. 10, no. 2, pp. 160–168, 2013.
- [25] M. H. Panju, "Statistical extraction and visualization of topics in the qur'an corpus," Master's thesis, University of Waterloo, 2014.
- [26] W. M. Darling, "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 642–647.
- [27] M. Alhawarat, M. Hegazi, and A. Hilal, "Processing the text of the holy quran: a text mining study," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 6, no. 2, pp. 262–267, February 2015.
- [28] Mushafqatar.com. (2015) Mushaf qatar. [Online]. Available: <http://www.mushafqatar.com/index.php?group=gather>
- [29] Almaany.com. (2015) Almaany. [Online]. Available: <http://www.almaany.com/quran-b/>
- [30] B. Grn, J. Kepler, U. Linz, K. Hornik, and W. W. Wien, "topicmodels: An r package for fitting topic models," *Journal of Statistical Software*, vol. 3, no. 8, 2011.

Localisation of Information and Communication Technologies in Cameroonian Languages and Cultures: Experience and Issues

Mathurin Soh

Department of Mathematics and Computer Science
University of Dschang, Cameroon
Dschang, Cameroon

Jean Romain Kouesso

Department of African Studies,
University of Dschang
Dschang, Cameroon

Laure Pauline Fotso

Department of Computer Science,
University of Yaounde 1, Cameroon
Yaounde, Cameroon

Abstract—In this paper, we tackle the problem of adapting Information and Communication Technologies (ICTs) in local languages of Cameroon. The objectives are to reduce the digital and language divides, and to pave the way for the usage of such technologies to local populations who don't understand this technological language. We first discuss and highlight several concerns about the localisation of ICTs. Afterwards, we address some challenges and issues to computerize cultural and linguistic features, and indigenous knowledge (IK) for national languages and cultures in Cameroon. As case study, we describe our experience in localising an open source editor for the Yemba language, within the of Rural Electronic Schools in African Languages Project. Because Cameroonian languages are based on the same basic alphabet, this qualitative research is extensible to other languages.

Keywords—Culture, Digital divide, ICTs, Language divide, Localisation, National language.

I. INTRODUCTION

Despite the rapid adoption of Information and Communication Technologies (ICTs) such as software, web sites and cell phones in day-to-day activities, the greatest challenge remains the adoption of these technologies by the local populations in Africa and in Cameroon in particular. Cameroon is endowed with about 250 local languages[1][2]. Unfortunately, the languages of communication in these technologies are all except these local languages. As a consequence, the local populations see ICTs as gadgets designed for the few who understand the inbuilt languages of communication.

The majority of Africans, mainly those in the rural areas, can only communicate in local languages, and hence cannot use a computer. They barely manage to master the keyboard of a telephone. Those who understand the language of the technology are able to use them, while the illiterates in the inbuilt language are unable to fully enjoy the benefits that the technology brings[4]. In [5], the authors advocate the writing and adapting of a software system to a specific culture. They emphasize that, successful software systems must be written; so that adapting them to a culture can be done easily. This is the idea of internationalisation/localisation in which software is built in such a way that localising it to another language and other cultural preferences can be done easily, possibly at runtime by reading a user's profile. In [6], it appears that ICTs are adaptable cultural western products. Therefore, they

include the ideologies of the target languages and cultures while influencing their users. To correct this bias, [7] and [8] think that we should make computing technology available, understandable, and participable for everyone regardless of culture, gender, age, income, language, degree of disability, or ethnicity. But, [9] think that the non-moulded user in the designer's culture must not be part of a bandwagon process, from imported models of the western society. In fact, all universals such as ICTs are after all used locally.

The rest of this paper is organized as follows. Section II addresses the Internationalisation/Globalisation/Localisation processes. Section III is devoted to an overview of Cameroonian national languages and cultures, and their presence in ICTs. It is focused on a specific Cameroonian language i.e Yemba. Section IV is related to the methodology of our field experience within the Rural Electronic Schools in African Languages Project, specifically on the Yemba language. We focus on the localisation process in the Cameroonian national language and culture, the challenges and the issues. We also discuss and highlight several concerns about the localisation of Information and Communication Technologies for the national languages and cultures of Cameroon. Section V and VI deal with the integration of Yemba cultural features in ICTs and computational results respectively.

II. INTERNATIONALISATION/GLOBALISATION/LOCALISATION

Internationalisation (I18N in abbreviated form) is defined as the process of developing applications that can easily be converted to operate in different cultural or linguistic environments [10].

Globalisation (G11N in abbreviated form) is the process of designing and developing applications that are meant to be used in multiple cultures.

Localisation, of course, has several definitions relating to the adaptation of computer applications and/or the content of computing to the linguistic and cultural realities of a particular country, region, or national community [11]. Localisation (L10N in abbreviated form) is the process of converting applications to operate in a specific cultural environment, which extends beyond the local language to aspects such as beliefs, customs and ethics of a society. For [12], the localisation process reasonably consists of two main stages. The first step

is the translation of language resources to reflect the local language. In this stage, all language resources and features such as menus, commands, help texts, etc. are translated to the local language. The second step in localisation adjusts software to local cultural habits. Here, the application is adapted to reflect local customs. If the application is already internationalised, this stage may be unnecessary, since changes such as special sorting algorithms, may have already been met during the development stage. For symbolic features such as the currency or the comma delimiter, the application may simply inherit configuration from the operating system. Localisation thus, is the process of customizing your application for a given culture and locale. Localisation, Internalisation and Globalisation are illustrated in Figure 1.

Thus, localisation is both a linguistic and software technology problem. It is a linguistic task because the translation requirement is not simply the substitution of one body of text by another. It is also a linguistic problem because many software packages capture and manipulate text that has been supplied by the users. Among these software packages are word processors and database management systems. In using these packages we are frequently required to match text. What constitutes an acceptable match depends upon the language. Localisation is also a software technology problem because we must be able to organize the software so that the linguistic components be isolated and can easily be replaced. This leads to the consideration of how standard software packages like window management systems, word-processors and database management systems are constructed where the assumptions about a particular natural language and culture are embedded. In the quest for wider computer access, many authors have stressed the need to accommodate users whose first language is not English. We believe localisation can contribute to the reduction of digital divide and language divide.

III. CAMEROONIAN NATIONAL LANGUAGES AND LOCALISATION

The writing of Cameroonian national languages is based on the General Alphabet of Cameroon Languages [3]. The Cameroonian alphabet is a subset of the International Phonetic Alphabet [13]. The General Alphabet of Cameroon Languages has permitted to write many Cameroonian national languages. Without loss of generality, this work is based on the Standard of Yemba (SY) spoken in the west region of Cameroon [14]. Yemba is the language of trade, education, mass communication and general everyday interaction between Yemba people, whatever their dialects of the language might be. According to the World Atlas of Languages Structures (WALS), Yemba national language WALS coordinates are: 5 25' N, 10 5' E [15].

Prior to the localisation of ICT, Yemba language and culture had never been widely used in the domain of modern technology in general, and computer technology in particular. [18] presents a website dedicated to Yemba language learning tools. This website contains a Yemba Wikipedia, a very basic Yemba online editor, and offers a possibility to translate common English, German, Czech, Spanish, Italian and Chinese words and expressions into Yemba. A vocabulary of computer terminology in Yemba language was addressed within the RESAL project[16]. Selected terms from this vocabulary is shown in Figure 2. This primary work paves the way to the

development and localisation of a word processing in Yemba language. Hence, even though these earlier efforts provided useful background materials, the localisation of word processing tools for Yemba language and culture demands more than mere translation of computer terms into Yemba[19]. The project necessarily demanded the creation of Yemba equivalent terms, which involves application of scientific strategies and principles for technical-term creation.

IV. METHODOLOGY AND DESIGN

The localisation results in solving many issues from linguistic aspects to hardware and software ones, including Cameroonian main cultural features to be integrated to the localized editor environment. We identify some of the problems and addressed them.

A. Terminological issues

Language problems are often caused by terminology. Whenever English language software is translated to a local language, decisions are taken on mapping from English terms to local terms. Inevitably, some measure of arbitrariness is attached to this procedure. In consequence, some aspects of localised software may appear stranger to the local audience than the English (foreign language) original. This goes some way toward explaining why many users when faced with a choice between a localized (fully translated) application and an English-language original, express a preference for the latter [17].

B. Some Yemba cultural features

Yemba calendar has eight days and each day has a meaning related to activities that are reserved or inspired by the historical facts. No specific day correspond to the beginning of the week as compared to Gregorian calendar, where Monday marks the beginning of the week. These days are successively named Efaa, Njla, Nga, Mbuwa, Mbulo, Meta, Mbukh, Mbucu. Per week, each village has one or two sacred day(s) during which some activities are forbidden.

C. Linguistic issues

The development of Cameroonian languages in the direction of technological development has suffered over the years due to the use of English as the language of technology. Hence, many of the English terms used in nowadays word processors do not have corresponding Cameroonian terms. It is necessary therefore to develop terms that can convey the meanings of the original English terms to the Cameroonian user.

D. Hardware issues

The localized word processor must support the character set of Cameroonian languages and must be configured to present numbers and other values in the local format. Localizing a word processor might require adding a new spell checker that recognizes words in the local language. We use the Keyboard GoingKompuyta (Figure 3) to make support of the character set of Cameroonian languages.

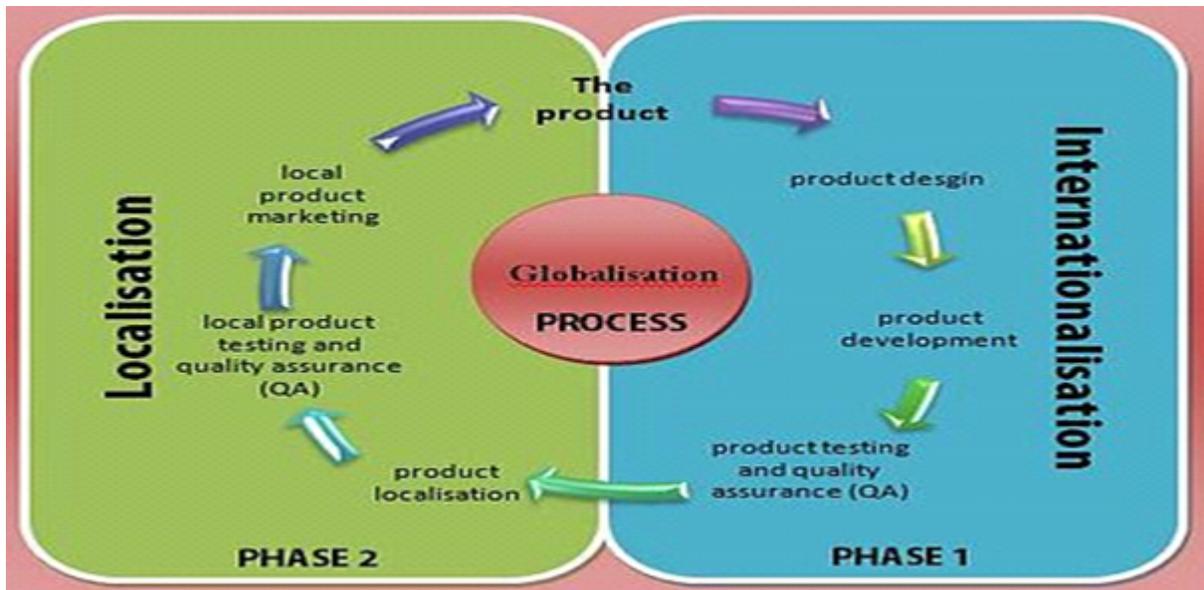


Fig. 1: The globalization process (based on a chart from the LISA website)

E. Software issues

There are varying degrees of localisation. Yet there are no obvious criteria for guiding the appropriate level of localisation. Evidence from our study of localised word processors shows that original menu shortcuts (such as Ctrl+C for 'Copy') are consistently retained rather than changed to accord with localised menu commands. In [20], the assumptions underlying this decision are obscure but may presume that retaining shortcut consistency across localised versions is beneficial to local users. Even this strategy can lead to anomalous results. Shortcut keys are often mnemonics for the English command names, e.g., Ctrl+N for 'New', Ctrl+O for 'Open' and Ctrl+S for 'Save', and Ctrl+P for 'Print'. When mapped to a localised Yemba version of the 'File' menu, these mnemonics are inappropriate yet this set of shortcut keys from the English context are retained. This contrasts with the other set of application shortcut keys. Alt+F (mnemonic in English for 'File') invokes the File menu in such examples illustrate a fundamental tension within localisation efforts, the need to change interface characteristics while attempting to maintain consistency [20].

V. INTEGRATING YEMBA CULTURAL FEATURES

To integrate Yemba cultural schemes into the localized editor (Figure 4), we need to tackle the shallow level of localisation defined by [21] as composed of the following areas:

- Colour schemes
- Pictures and images
- Sounds
- Historical data
- Hand signals
- Symbols

- Product names and acronyms.

All these issues have different meaning in a different cultural context. But to the best of our knowledge, these cultural features are not formalized at all, except the date, time and days representation issues.

VI. COMPUTATIONAL RESULTS

Our experience required expertise from three key areas: (i) linguistics, (ii) language technology, (iii) computer science. Command keys that have international status are not subject to localisation. Their commands have a mnemonic connection to the visual appearance of the letter and are not related to the usage of the command name in some languages [20].

Word processors vary considerably, but all word processors support some basic features. Our localized editor supports the following basic features:

- insert text: Allows to insert text anywhere in the document.
- delete text: Allows to erase characters, words, lines, or pages as easily as you can cross them out on paper.
- cut and paste : Allows to remove (cut) a section of text from one place in a document and insert (paste) it somewhere else.
- copy: Allows to duplicate a section of text.
- page size and margins :Allows to define various page sizes and margins, and the word processor will automatically readjust the text so that it fits.
- search and replace: Allows to direct the word processor to search for a word or a phrase. You can also direct the editor to replace one group of characters with another everywhere that the first group appears.

Word in English	Localized Word in Yemba
File	Ŋkaŋa menu
New	Eshwī
Open	Letsō'
Save	Lepā
Save as...	Lepā le ...
Page layout	Atana
Print	Lefōk
Quit	Letō
Edition	levho
Cancel	lepikne
Cut	Lezā'
Paste	Levet
Copy	Lelōk
Delete	Lepik
Search	Lefa'
Search the next	Lefa'a pááté
Replace	lekumne
Go to	Leguə
Select all	Letsō' ŋkwa
Hour/date	Nihū/ale'é álā'
Format	Efa'
Font	Ŋkia
Display	Lepete
Status bar	Ŋkaŋa eshū
Help	letswiite
Help contents	Menu atsiite
About	Á ne

Fig. 2: Selected Yemba terms for the editor

- word wrap : The word processor automatically moves to the next line when you have filled one line with text, and it will readjust the text if you change the margins.
- print: Allows to send a document to a printer to get hardcopy. Our localized word processor supports additional features that enable a user to manipulate and format documents in more sophisticated ways. These full features are:
- file management: Many word processors contain file management capabilities that allow you to create, delete, move, and search for files.
- font specifications: Allows to change fonts within a document. For example, you can specify bold, italics, and underlining. Most word processors also let you change the font size and even the typeface.
- spell checker : A utility that allows to check the spelling of words. It will highlight any words that it

does not recognize.

- WYSIWYG (what you see is what you get): With WYSIWYG, a document appears on the display screen exactly as it will look when printed.

The use of this editor will faced many other factors shown in [22] who proved that other essential aspects intervene such as dialectal variations, inter-comprehension level between neighbored languages, the writing systems used,the orthography and the terminology.

VII. CONCLUSION

Most ICTs are designed and developed by researchers and designers who unintentionally apply their cultural values and systems of thought while designing and developing computer applications like word processors. This results in that, users who are culturally different from the researchers and designers might have difficulty to use these computer applications. In this work, we have localised an open source editor in the Yemba language. This can contribute to fill the gap between computer programs designers/developers, and end-users whose language is Yemba. In fact, it will be of use within the RESAL project and mostly in rural primary schools for the teaching of/in local languages. The localized editor is extensible to other Cameroonian and African languages; thus offering to speakers of these languages a tool to easy production of documents in their language.

ACKNOWLEDGMENTS

We would like to thank the many people who have been helpful in giving their time and assistance in the knowledges reported in this paper. We are pleased to acknowledge the contribution of translation and terminology made by Jean Claude Gntedem and Albert Tsopmejo. The authors would also like to thank all the reviewers of this article and the extremely valuable suggestions they made to improve it.

REFERENCES

- [1] C. Binam Bikoi, *Cartographie administrative des langues du Cameroun*, CERDOTOLA N1, Yaounde, 2012.
- [2] M.P. Lewis, F.S. Gary and D.F. Charles, *Ethnologue: Languages of the world*, Eighteenth edition, N1, M.P. Lewis, F.S. Gary and D.F. Charles (eds), Dallas; Texas: SIL International, 2015, Online version: <http://www.ethnologue.com>, [accessed on december 15, 2015].
- [3] M. Tadjadjeu and E. Sadembouo, *Alphabet Gnral des Langues Camerounaises*, Collection PROPELCA N1, Universit de Yaound, M. Tadjadjeu and E. Sadembouo (ds), 34p. 1984.
- [4] F. Wolff, *Effecting change through Localization: Localization guide for free and open source software*, IDRC, Canada, 2011.
- [5] B. Keller, M. Prez-Quiones and R. Vatrappu, *Cultural issues and opportunities in Computing Education*, 9th International Conference on Engineering Education, RIE 14, July 23-28, 2006.
- [6] L.E. Dyson, *Cultural Issues In The Adoption Of Information And Communication Technologies By Indigenous Australians*. In Sudweeks, F. and Ess, C. (eds). Proceedings Cultural Attitudes towards Communication and Technology 2004, Murdoch University, Australia, 58-71, 2004.
- [7] E. Sutinen and al. *Is Universal Usability Universal Only to Us?*, CUU 03, November 10-11, 2003, Vancouver BC, Canada. Copyright 2003 ACM 1-58113-000-0/00/0000, 2003.
- [8] M. Tedre, E. Sutinen, E. Kahkonen and P. Kommers, *Ethnocomputing: ICT in cultural and social context*, Communications of the ACM, 49 (1). pp. 126-130. ISSN 0001-0782, 2006.



Fig. 3: Keyboard GoingKompyuta for Cameroonian national languages

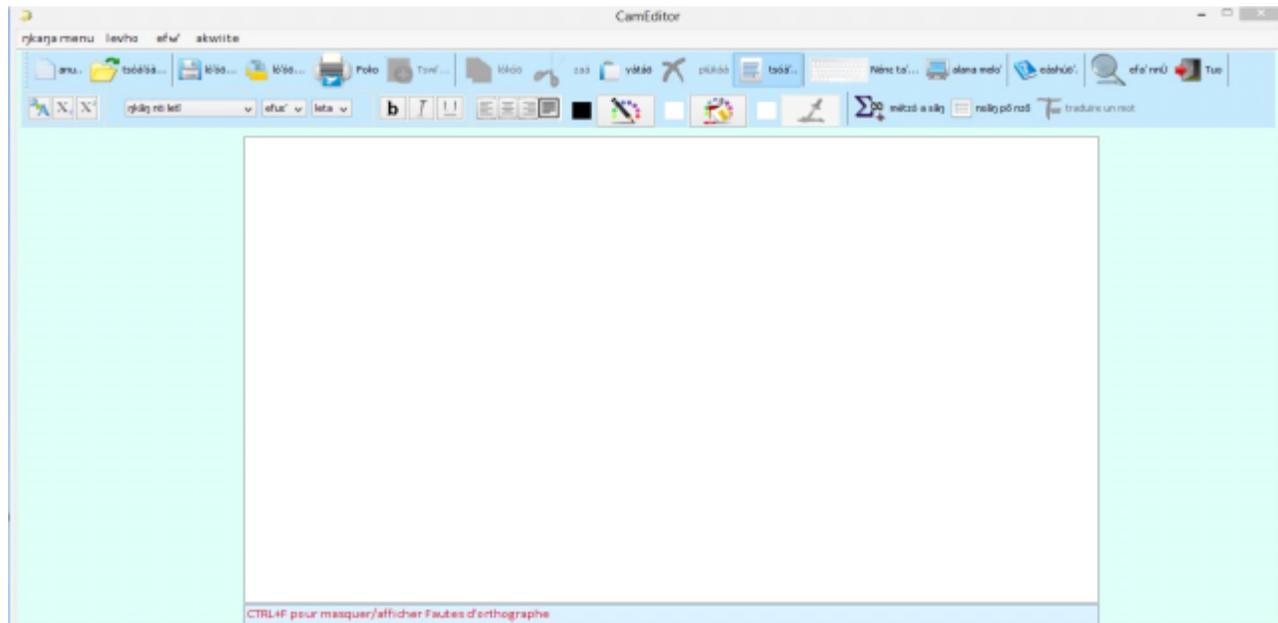


Fig. 4: A bilingual (French and Yemba) main interface of the localised editor

- [9] J.R. Kouesso and M. Soh, *Relever le dfi du dveloppement informatique de l'enseignement des langues camerounaises : les leons de lexprience ERELA au Cameroun* , Communication AGIS 11, 02-05 december 2011, Addis Abeba, Ethiopia,2011.
- [10] P. Russor and S. Boor, *How fluent is your interface? Designing for international users*, Proceedings of InterCHI'93, ACM Press, pp. 342-347, 1993.
- [11] D. Osborn, *African Languages and Information and Communication Technology: Localising the Future?*, Localization Focus, June, 2005.
- [12] S. W.Q. Jaffry and U.R. Kayani, *FOSS Localization : A Solution for the ICT Dilemma of Developing Countries*, 9th International IEEE MultiTopic Conference, December 24-25, 2005. NUCES-FAST Karachi, Pakistan, INMIC05, 2005.
- [13] Wikipedia, *Cameroonian languages*, http://en.wikipedia.org/wiki/Cameroonian_language[accessed on august 15,2015].
- [14] J.R.. Kouesso, *Variation dialectale et standardisation de l'orthographe du yemba*, PhD thesis, University of Yaound, Faculty of Letters and Human Science, 2009.
- [15] WALS, *World atlas of languages structures*, http://wals.info/languoid/lect/wals_code_yem,[accessed on august 16,2015].
- [16] L. Ngoumamba and J.C. Gnintedem, *Lexique informatique plurilingue franais, nuasue et yemba*, in the book *Ecoles Rurales Electroniques en Langues Africaines: Exprimentation au Cameroun et orientation politique panafricaine* , Editions L'Harmattan, pages 253-260, ISBN : 978-2-343-05029-4, EAN : 9782343050294, 2015.
- [17] G. Weir and G.. Lepouras, *Localisation and linguistic anomalies*, Universal Access In HCI: Towards an Information Society for All, Proceedings of HCI International '2001 (the 9th International Conference on Human-Computer Interaction), New Orleans, USA, August 5-10, 2001, Volume 3.
- [18] Karewa, *Yemba learning tools*, <http://www.karewa.com/aleco/>, [accessed on august 17,2015].
- [19] M. Soh and J.R. Kouesso, *Aspects de la localisation des Technologies de l'Information et de la Communication pour l'enseignement des langues camerounaises*, in the book *Ecoles Rurales Electroniques en Langues Africaines: Exprimentation au Cameroun et orientation politique panafricaine* , Editions L'Harmattan, pages 149-161, ISBN : 978-2-343-05029-4, EAN : 9782343050294, 2015.
- [20] G. Grigas, T. Jevsikova and A. Strelkauskyt, *Localization Issues of Software Shortcut Keys* , The International Journal of Localization, Vol.11 Issue 1, 2011.
- [21] R. Schler, *The Cultural Dimension in Software Localization*,

Localization Reader 2003- 2004. Selected articles from Localization Focus and Multilingual Computing Technology 58, 2003.(online), [http://www.Localization.ie/resources/Research/ELECT/Consortium/ContentFiles/00-11 %20LR-S.pdf](http://www.Localization.ie/resources/Research/ELECT/Consortium/ContentFiles/00-11%20LR-S.pdf)[accessed 15 August 2015].

- [22] D. Osborn, *African languages in a digital age: challenges and opportunities for indigenous language computing*, HSRC press, ISBN (soft cover) 978-0-7969-2249-6, 2010.

Real-Time Talking Avatar on the Internet Using Kinect and Voice Conversion

Takashi Nose
Graduate School of Engineering
Tohoku University
Sendai City, Japan

Yuki Igarashi
Graduate School of Engineering
Tohoku University
Sendai City, Japan

Abstract—We have more chances to communicate via the internet. We often use text/video chat, but there are some problems, such as a lack of communication and anonymity. In this paper, we propose and implement a real-time talking avatar, where we can communicate with each other by synchronizing character's voice and motion from ours while keeping anonymity by using a voice conversion technique. For the voice conversion, we improve accuracy of the voice conversion by specializing to the target character's voice. Finally, we conduct subjective experiments and show the possibility of a new style of communication on the internet.

Index Terms—Talking avatar; Voice conversion; Kinect; Internet; Real-time communication

I. INTRODUCTION

Typical examples for human communication via internet are text, voice, and video chatting. Users of the text chat input text with interfaces such as a keyboard and easily communicate to each other. However, the text-based communication has difficulty in expressing emotions and intentions correctly, which sometimes leads to misunderstanding of the user's internal state. In addition, the lack of the real-time communication is often stressful. On the other hand, the video chat has an advantage in communicating both linguistic and para-linguistic information through the facial expression and the speaking style [1], [2], [3]. Although the video chat is the most advanced and rich communication tool of the chat systems, one of the biggest problems in the use of the audio and video information is the lack of anonymity, and we must choose an appropriate tool depending on the situation.

In this paper, we present a real-time talking avatar system in which the user's motion and speech are reflected to the avatar in real-time. The system enables us to communicate to each other via the network without directly conveying the user's personal information. The system of the real-time talking avatar consists of the two technologies as follows:

- Voice conversion: Converting the speaker characteristics of the user to that of the avatar using neural network in real-time [4]
- Synchronization skeleton: Capturing the user's motion and reflecting the motion to the avatar in real-time using Kinect [5]

In the voice conversion, we focus on the character of the avatar, virtual singer *Hatsune Miku*. We conduct an experiment

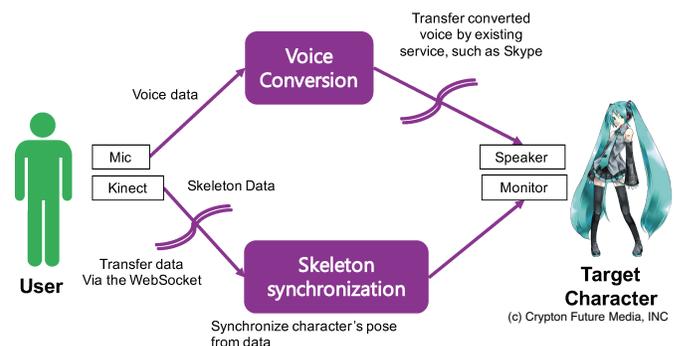


Fig. 1: Overview of the proposed real-time talking avatar system via internet.

where the speech parameters extracted from the input speech is averaged using moving average to improve the reproducibility of the robotic voice of the character.

We develop the real-time talking avatar system using the voice conversion and synchronization skeleton techniques, which enables the users to anonymize their voice and facial information. We conduct subjective evaluations and compare the proposed system to the conventional text and video chat systems in terms of the anonymity, entertainment, and communicability.

The rest of this paper is organized as follows: Section II overviews the real-time talking avatar system proposed in this paper. Section III introduces the voice conversion depending on the target character to improve the conversion performance based on neural networks, and the objective experimental evaluations for the voice conversion part is shown in Section IV. Section V explains how to control the character's motion using Kinect. The total subjective evaluation is conducted in Section VI and the result is discussed. Finally, Section VII summarizes this study and shortly discusses the remaining issues as conclusions.

II. SYSTEM OVERVIEW

We use two sets of Kinect for Windows and microphone, and a PC to control them and network environment to realize the real-time communication using the proposed talking avatar.

The synchronization of the user's motion including hand gestures is achieved by acquiring the skeleton data of the user and by reflecting the data to the character model of the avatar. In this study, we use Kinect v2 and Skeletal Tracking of Kinect SDK v2 [6] to acquire the user's motion data. The Kinect is able to obtain the position data of twenty-five joints per one user.

To reflect the user's motion, only the rotation parameters of each joint and the position of center part of pelvis, which is given as SpineBase in Kinect SDK, are extracted and transmitted to the client user using a communication protocol, e.g., WebSocket. The client system receives the transmitted data and maps them to each joint and position of the character model. Finally, the avatar image having similar pose to the user is outputted to the display.

In the voice conversion, the speech of the user is recorded using a microphone, and the speech parameters, i.e., spectral and excitation parameters, are extracted and converted to those of the target character using a neural network that is one of the nonlinear mapping techniques. The speaker-adapted speech is obtained by synthesizing speech from the converted parameters. The speech data after the voice conversion is outputted through a virtual sound device using several windows APIs. By introducing the virtual device, we can use the converted voice as a source of existent voice conference applications such as Skype. Figure 1 shows the overview of the proposed system.

III. CHARACTER-DEPENDENT VOICE CONVERSION

Voice conversion is a technique for changing the input speaker's voice characteristics to that of another speaker (target speaker) while keeping the linguistic information. In this study, the target speaker is not a person but an avatar, i.e., virtual singer Hatsune Miku as shown in Figure 2 to increase the entertainment factor in the communication.

A. Character's voice for avatar

We use a voice of a Japanese famous virtual character, Hatsune Miku of singing voice synthesis software *VOCALOID* [7] developed by YAMAHA [8]. Figure 2 shows an illustration of Hatsune Miku. Recently, the singing voice of Hatsune Miku is used for many users and is becoming much popular in the community cite such as Youtube and Niconico [9] in Japan [10]. By using *VOCALOID*, users can easily synthesize singing voice by inputting melody and lyrics. We prepare the parallel speech of human and Hatsune Miku using *VOCALOID* and use the synthesized speech for the training of the voice conversion. In this study, we chose the Miku as the target speaker to take the friendly feeling into account. Figure 3 shows an example of controlling character model of Hatsune Miku using Kinect v2.

B. Voice conversion based on neural networks

Figure 4 shows an overview of the voice conversion part, In the voice conversion, we use neural networks to map the spectral features of the input speech of a certain user to



Fig. 2: Virtual character Hatsune Miku of a singing voice synthesizer VOCALOID.

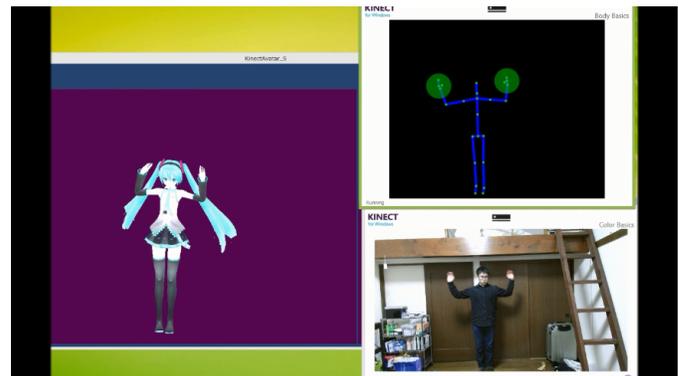


Fig. 3: Example of controlling character model of Hatsune Miku using Kinect v2.

those of the target character [4]. By using neural network, the multi-dimensional feature vectors consisting of the spectral parameters are mapped in a non-linear form. Since the relation of corresponding frames between source and target speakers is highly non-linear, the neural network is known to be effective compared to the traditional mapping technique based on Gaussian mixture model (GMM) [11], [12], [13] that has been widely used in the study of voice conversion [14]. The process of the voice conversion from the user's voice to the Hatsune Miku's voice is as follow:

- Prepare parallel speech data of two speakers uttering the same sentences.
- Extract mel-cepstral coefficients and fundamental frequency (F0) data using Speech Signal Processing Toolkit (SPTK) [15]
- Perform dynamic time warping [16] to align the frames of the spectral features of the two speakers.
- Train neural networks that maps source speaker's features to those of the target speaker and obtain a set of weight

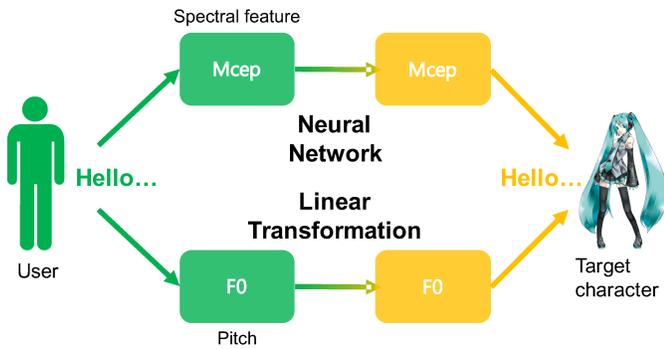


Fig. 4: Overview of the voice conversion part using neural networks for spectral mapping and linear transformation for pitch conversion.

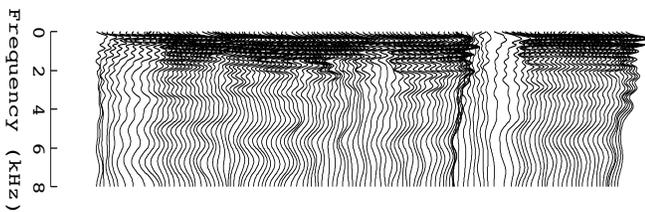


Fig. 5: Running spectrum extracted from the speech /yobou/ uttered by Hatsune Miku in Japanese.

parameters.

- Convert the spectral features of the input speech of the user using the trained neural networks.
- Convert the log F0 parameter using affine transformation so that the mean and variance parameters become the same between the two speakers.
- Synthesize speech from spectral and F0 parameters.

For the F0 conversion, we use affine transformation defined by

$$y = (x - \mu_x) / \sigma_x * \sigma_y + \mu_y \quad (1)$$

where μ_x and σ_x are global mean and variance of the source speaker, respectively, and μ_y and σ_y are global mean and variance of the target speaker, respectively.

Figures 5 and 6 show examples of the running spectrum and the F0 sequence of synthetic speech of Hatsune Miku. As is shown in the figure, the trajectory of the spectral envelope and log F0 is smooth, which is different from those of the human speech. To utilize this property, we apply smoothing filter by moving average for the spectral and log F0 parameter sequences after the voice conversion process.

IV. VOICE CONVERSION EXPERIMENTS

A. Experimental conditions

For the experiments, we used a hundred sentences, i.e., subsets A and B of the ATR phonetically-balanced Japanese sentences [17]. A male non-professional speaker uttered the sentences. We used fifty sentences of the subset A for the

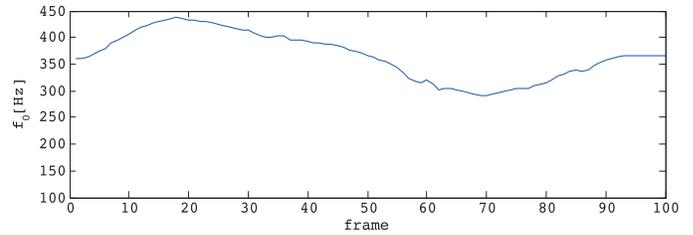


Fig. 6: Log F0 sequence extracted from the speech /yobou/ uttered by Hatsune Miku in Japanese.

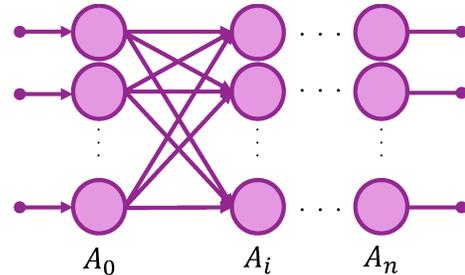


Fig. 7: Example of neural network.

training of the neural networks in the voice conversion and used fifty sentences of the subset B for the testing. As for the structure of the neural network, we fixed the number of the hidden layers to one to achieve the real-time processing of voice conversion though we might improve the performance by increasing the number of the layers under the condition where we have a sufficient amount of training data. From a preliminary experiment, we set the number of units of the hidden layer to fifty. The numbers of units for input, hidden, and output layers were 25, 50, 25, respectively.

The 0th to 24th mel-cepstral coefficients were extracted using mcep command of SPTK where we set the window length to 25 ms and the frame shift to 5 ms. The log F0 was extracted using pitch command with the RAPT algorithm [18] with the same frame shift. To create the parallel speech of Hatsune Miku, we generated log F0 sequences using Open JTalk [19]. The log F0 sequences were then quantized into several levels with a semitone interval for each mora and the Miku's voice corresponding to the training text was generated using VOCALOID. As a result, we created the Miku's speech data of subset A of the ATR sentences.

B. Results and analysis

We used mel-cepstral distance as the objective measure of spectral reproducibility. The mel-cepstral distance of the n -th frame is defined as

$$d(n) = \sum_{k=1}^M (c_n^{(t)}(k) - c_n^{(s)}(k)) \quad (2)$$

where $c_n^{(t)}(k)$ and $c_n^{(s)}(k)$ is the k -th mel-cepstral coefficient of n -th frame of source and target speakers, respectively. We calculated the mel-cepstral distance between the original

TABLE I: Average mel-cepstral distance between source and target speakers before and after the voice conversion.

	Average (dB)	Min. (dB)	Max. (dB)
Before	16.025	14.312	17.319
After	9.356	8.251	10.368

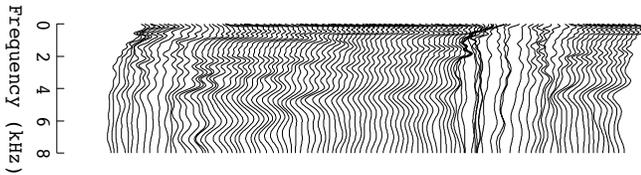


Fig. 8: Running spectrum before voice conversion.

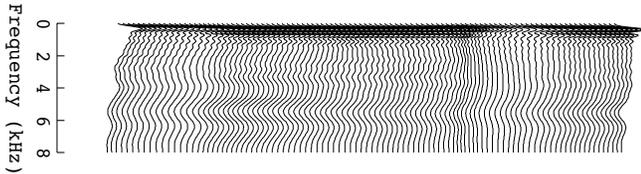


Fig. 9: Running spectrum after voice conversion.

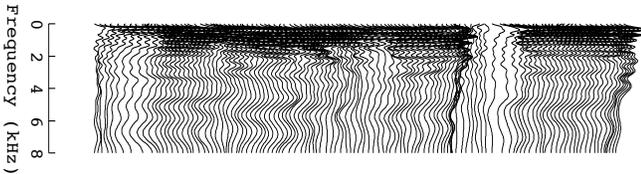


Fig. 10: Running spectrum of original speech of the target speaker (Hatsune Miku).

speech of the target speaker and the converted speech. For comparison, we also calculated the mel-cepstral distance of speech between the source (before conversion) and the target speakers. Table I shows the average, minimum, and maximum values of the mel-cepstral distance for the test data.

From the table, it is seen that there was large spectral difference between the source and the target speaker but the difference became smaller after the voice conversion with spectral mapping based on neural networks. Figures 8 and 9 show examples of running spectra before and after the voice conversion. For the reference, we also show the running spectrum of the original speech in Figure 10.

To evaluate the effect of smoothing after the voice conversion, we also calculated the mel-cepstral distance between converted speech and the target speaker's speech by changing the width for the moving average. Table II show the result. From the table, we found that the mel-cepstral distance decreases by introducing moving average is smallest when the number of frames is six. This result indicates that the smoothing operation by moving average well captures the voice property of the target character.

Next, we evaluated the effect of the smoothing by moving average for the log F0 sequence with subjective evaluation. We conducted a listening test where subjects listened to the con-

TABLE II: Width (# of frames) in moving average and mel-cepstral distance.

# of frames	Average (dB)	Min. (dB)	Max. (dB)
0	9.356	8.251	10.368
1	9.288	8.225	10.303
2	9.316	8.273	10.357
3	9.228	8.182	10.332
4	9.199	8.201	10.215
5	9.202	8.271	10.216
6	9.179	8.316	10.168
7	9.189	8.295	10.345
8	9.211	8.382	10.347
9	9.274	8.412	10.406
10	9.271	8.506	10.359

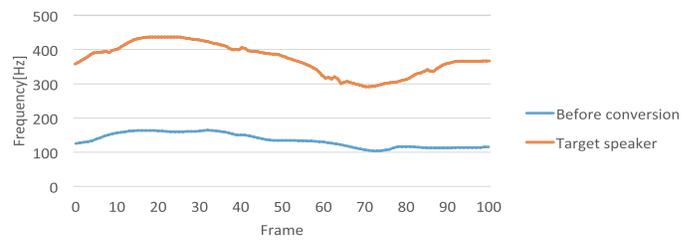


Fig. 11: F0 contours of the source speaker (before conversion) and the target speaker.

verted speech samples and rated the speaker similarity using five-point scale: 5 (similar), 4 (a little similar), 3 (undecided), 2 (a little dissimilar), and 1 (dissimilar). Table III shows the result.

From the table, we found that the speaker similarity substantially degrades when the F0 conversion is not applied. Figures 11 and 12 show examples of F0 contours before and after the F0 conversion, respectively. The original F0 contour of the target speaker is also shown in the figures. From the figures, it is seen that the F0 contour of the source speaker became closer to that of the target speaker even when the target speaker is a virtual character, Hatsune Miku.

In contrast to the case of spectral features, the smoothing was not effective in the case of F0, and over-smoothing also degrades the speaker similarity. A possible reason of the degradation is that the F0 smoothing by moving average does not take the mora units and their average pitch into account. Since the speech of Hatsune Miku was synthesized using VOCALOID, the F0 became flat within each mora. To improve the conversion performance, we explicitly use the mora information for the input speech by forced alignment technique. However, it is not easy to use the alignment information in the real-time application. Hence, this is a remaining problem in this study.

V. SYNCHRONIZATION TO THE CHARACTER'S MOTION

In this study, we need to acquire skeleton information of the user to synchronize the motion of the user to that of the character model. We use Kinect for Windows v2 that is a

TABLE III: Subjective evaluation results comparing converted speech with moving average of log F0 to the converted speech without moving average.

# of frames	Similar	A little similar	Undecided	A little dissimilar	Dissimilar
No conversion	0	0	0	0	5
0	1	3	1	0	0
5	0	4	1	0	0
10	0	0	4	1	0
15	0	0	0	2	3
20	0	0	0	1	4

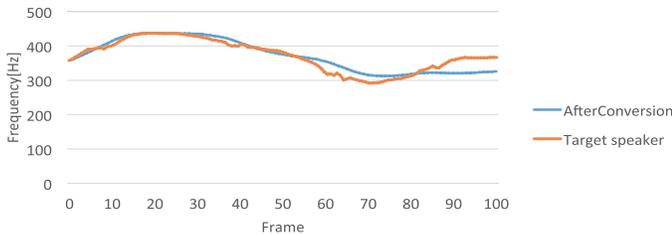


Fig. 12: F0 contours of the source speaker (after conversion) and the target speaker.

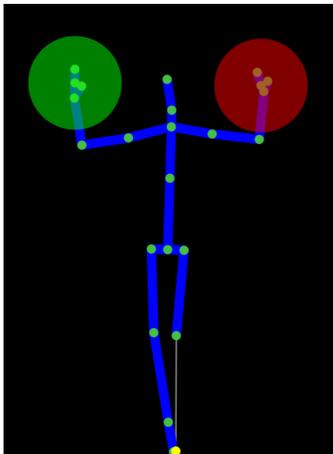


Fig. 13: Example of skeleton detection by Kinect v2.

motion sensor device developed by Microsoft corporation. By using this sensor device, we can obtain the information of twenty-five joints per one person, and we can also obtain the position and rotation information of each joint. Kinect for Windows v2 has better performance than the conventional Kinect sensor in terms of the accuracy and resolution, and the face recognition and facial expression analysis are supported. However, we only utilize the function of the synchronization skeleton in this study.

The synchronization skeleton is realized by acquiring rotation information of each joint of the user, calculating the rotation between the joints, and synchronizing with the rotation of each joint of the character model. By applying these processing, even when there is a large difference of body size and body characteristics between the user and the target character,



Fig. 14: Example of controlling MMD model of Hatsune Miku by Kinect v2.

mapping is not sensitive to the difference, and natural mapping is achieved. In addition, to deal with the longitudinal and lateral movement, first we set the position of the calibration to the origin and calculate the relative position of the pelvis of the center, and apply the position to the character model. This enables the system to display the character as if the character was walking to the direction corresponding to the user's walking motion to the same direction.

We used Unity [20], which is an environment for the programming easily handling the character model, for the implementation of the above functions. For the character model, we used Lat-style Miku Ver.2.31 [21] that is a 3D CG model of Hatsune Miku for MikuMikuDance (MMD) [22]. Figures 13 and 14 show an example of the mapping from the user's skeleton information to the character's motion.

VI. DEMONSTRATION EXPERIMENTS OF REAL-TIME TALKING AVATAR

On the basis of the experimental results in the previous sections, we implemented a real-time talking avatar system. To demonstrate the effectiveness of the system, we conducted an experiment where five subjects used the system for the communication to a reference person.

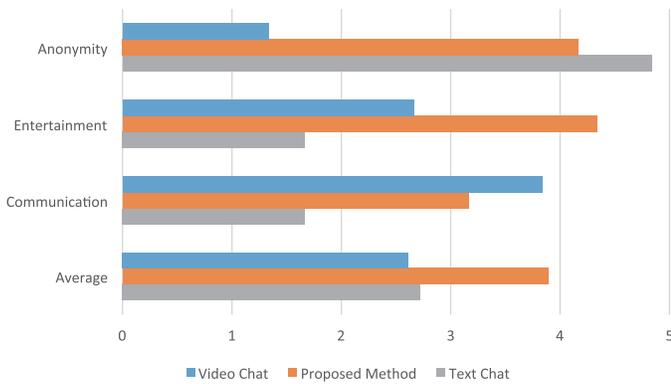


Fig. 15: Comparison of the effectiveness of the proposed real-time talking avatar to the other communication systems.

A. Experimental procedure

Users were naive and did not know the conversation partner, and the communication through the system was performed in the following order.

- 1) Text chat
- 2) Real-time talking avatar (the proposed system)
- 3) Video chat

Users evaluated the three systems in terms of the following three criteria using five-point scale: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad).

- Anonymity: Whether the user directly recognizes the partner through the communication
- Entertainment: How the user enjoys the communication
- Communication: How the conversation with the system continues smoothly

The real-time talking avatar is proved to be effective if the proposed system outperforms the video chat in terms of Anonymity and outperforms the text chat in terms of Entertainment and Communication criteria.

B. Experimental results

Figure 15 shows the results. From the aspect of Anonymity, the proposed system is slightly worse than text chat but is substantially better than the video chat. In the experiment, we found that the user sometimes obtained the information about the partner's personality and gender through the gesture and habit of the user even though the motion and the voice were conveyed through the character model. However, it is difficult to completely separate the information related to the motion and personality, and hence the slight degradation of anonymity would be acceptable in the real communication of the most of users. As for the degree of the entertainment, the proposed system gave higher score than both the text and video chat systems. One of the reasons is that the attractive target character instead of a real person enhances the pleasantness in the conversation, which is the main purpose of this work. The advantage of our system is that the user can choose the target character so as to be fun for him/her. In the factor of



Fig. 16: Example of the conversation using the real-time talking avatar system.

Communication, the real-time conversation of the proposed system made the score much higher than the text chat and the score was close to that of the video chat. The communication performance would be improved by the advance of Kinect and motion and face tracking SDKs.

The above results are summarized as follows. The proposed real-time talking avatar system has the intermediate property between the text chat and the video chat, and the average score is highest of three systems, which indicates that our system is the more balanced and attractive system than the conventional text and video chat systems.

VII. CONCLUSIONS

In this paper, we presented a novel communication tool via internet, where the communication style is different from the conventional text and video chat systems. The most attractive point of our system is that both of the anonymity and entertainment factors are achieved at a sufficient level while keeping the smoothness of the communication in real-time. The system utilizes Kinect-based motion capturing and processing and a voice conversion technique, and the user can choose the favorite character and voice to anonymize him/herself. In the voice conversion part, the property of the target speaker was taken into account, and we showed that the smoothing operation using moving average increase the the spectral reproducibility when the width was appropriately set.

In the future work, the performance improvement in the voice conversion is important. Especially, the noise robustness is highly required in the real environment in our daily life. Also the use of facial information including emotional expressions is beneficial for more advanced human communication with high anonymity and security.

ACKNOWLEDGMENT

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 15H02720 and Step-QI School of Tohoku University.

REFERENCES

- [1] C. Neustaedter and S. Greenberg, "Intimacy in long-distance relationships over video chat," in *Proc. the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 753–762.
- [2] O. Boyaci, A. G. Forte, and H. Schulzrinne, "Performance of video-chat applications under congestion," in *Proc. 11th IEEE international symposium on multimedia*, 2009, pp. 213–218.
- [3] J. Scholl, P. Parnes, J. D. McCarthy, and A. Sasse, "Designing a large-scale video chat application," in *Proc. the 13th annual ACM international conference on Multimedia*, 2005, pp. 71–80.
- [4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP. IEEE*, 2009, pp. 3893–3896.
- [5] Microsoft Kinect for Windows, <http://www.microsoft.com/en-us/kinectforwindows/>.
- [6] Tracking Users with Kinect Skeletal Tracking, <https://msdn.microsoft.com/ja-jp/library/jj131025.aspx>.
- [7] H. Kenmochi and H. Ohshita, "Vocaloid–commercial singing synthesizer based on sample concatenation," pp. 4011–4010, 2007.
- [8] YAMAHA Corporation, <http://www.yamaha.com/>.
- [9] Niconico, <http://www.nicovideo.jp/>.
- [10] M. Hamasaki, H. Takeda, and T. Nishimura, "Network analysis of massively collaborative creation of multimedia contents: case study of hatsune miku videos on nico nico douga," in *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*, 2008, pp. 165–168.
- [11] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.
- [12] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1, 1998, pp. 285–288.
- [13] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.
- [14] Y. Stylianou, "Voice transformation: a survey," pp. 3585–3588, 2009.
- [15] The SPTK working group, "Speech Signal Processing Toolkit (SPTK)," <http://sp-tk.sourceforge.net/> (2015.9.24).
- [16] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.
- [17] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.
- [18] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, pp. 495–518, 1995.
- [19] Open JTalk, <http://open-jtalk.sourceforge.net/>.
- [20] Unity, <http://unity3d.com/>.
- [21] Lat-style Miku Ver.2.31, <https://bowlroll.net/file/30199>.
- [22] T. Yoshikawa, "Miku miku dance starter pack," 2010.