

Volume 4 Issue 6

June 2013



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



www.ijacsa.thesai.org



W H E R E W I S D O M S H A R E S

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS



THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org



Editorial Preface

From the Desk of Managing Editor...

It is our pleasure to present to you the June 2013 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, 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.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 4 Issue 6 June 2013
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)
©2013 The Science and Information (SAI) Organization

Editorial Board

Dr. Kohei Arai – Editor-in-Chief

Saga University

Domains of Research: Human-Computer Interaction, Networking, Information Retrievals, Optimization Theory, Modeling and Simulation, Satellite Remote Sensing, Computer Vision, Decision Making Methodology

Dr. Ka Lok Man

Xi'an Jiaotong-Liverpool University (XJTLU)

Domain of Research: Computer Science and Microelectronics

Dr. Sasan Adibi

Research In Motion (RIM)

Domain of Research: Security of wireless systems, Quality of Service

Dr. Zuqing Zuh

University of Science and Technology of China

Domains of Research : Optical Communication Systems, Optical network architecture and design, Next generation Internet, Signal processing, Broadband access network, such as cable access (DOCSIS) networks, passive optical networks (PON), fiber to the home (FTTH), Energy-efficient network and green technologies

Dr. Sikha Bagui

University of West Florida

Domain of Research: Database, database modeling, ER diagrams, XML data, web databases, data mining, association rule mining, data preprocessing

Dr. T. V. Prasad

Lingaya's University

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

Dr. Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Data Mining, Database, Web-based Application, Mobile Computing

Reviewer Board Members

- **A Kathirvel**
Karpaga Vinayaka College of Engineering and Technology, India
- **A.V. Senthil Kumar**
Hindusthan College of Arts and Science
- **Abbas Karimi**
I.A.U_Arak Branch (Faculty Member) & Universiti Putra Malaysia
- **Abdel-Hameed A. Badawy**
University of Maryland
- **Abdul Wahid**
Gautam Buddha University
- **Abdul Hannan**
Vivekanand College
- **Abdul Khader Jilani Saudagar**
Al-Imam Muhammad Ibn Saud Islamic University
- **Abdur Rashid Khan**
Gomal University
- **Aderemi A. Atayero**
Covenant University
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
Menoufia University, Egypt
- **Ajantha Herath**
University of Fiji
- **Ahmed Sabah AL-Jumaili**
Ahlia University
- **Akbar Hossain**
- **Albert Alexander**
Kongu Engineering College,India
- **Prof. Alcinia Zita Sampaio**
Technical University of Lisbon
- **Amit Verma**
Rayat & Bahra Engineering College, India
- **Ammar Mohammed Ammar**
Department of Computer Science, University of Koblenz-Landau
- **Anand Nayyar**
KCL Institute of Management and Technology, Jalandhar
- **Anirban Sarkar**
National Institute of Technology, Durgapur, India
- **Arash Habibi Lashakri**
University Technology Malaysia (UTM), Malaysia
- **Aris Skander**
Constantine University
- **Ashraf Mohammed Iqbal**
Dalhousie University and Capital Health
- **Asoke Nath**
St. Xaviers College, India
- **Aung Kyaw Oo**
Defence Services Academy
- **B R SARATH KUMAR**
Lenora College of Engineering, India
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Badre Bossoufi**
University of Liege
- **Balakrushna Tripathy**
VIT University
- **Basil Hamed**
Islamic University of Gaza
- **Bharat Bhushan Agarwal**
I.F.T.M.UNIVERSITY
- **Bharti Waman Gawali**
Department of Computer Science & information
- **Bremananth Ramachandran**
School of EEE, Nanyang Technological University
- **Brij Gupta**
University of New Brunswick
- **Dr.C.Suresh Gnana Dhas**
Park College of Engineering and Technology, India
- **Mr. Chakresh kumar**
Manav Rachna International University, India
- **Chandra Mouli P.V.S.S.R**
VIT University, India
- **Chandrashekhara Meshram**
Chhattisgarh Swami Vivekananda Technical University
- **Chao Wang**
- **Chi-Hua Chen**
National Chiao-Tung University
- **Constantin POPESCU**
Department of Mathematics and Computer Science, University of Oradea
- **Prof. D. S. R. Murthy**
SNIST, India.
- **Dana PETCU**
West University of Timisoara
- **David Greenhalgh**

- University of Strathclyde
- **Deepak Garg**
Thapar University.
 - **Prof. Dhananjay R.Kalbande**
Sardar Patel Institute of Technology, India
 - **Dhirendra Mishra**
SVKM's NMIMS University, India
 - **Divya Prakash Shrivastava**
EL JABAL AL GARBI UNIVERSITY, ZAWIA
 - **Dr.Dhananjay Kalbande**
 - **Dragana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational sciences
 - **Driss EL OUADGHIRI**
 - **Firkhan Ali Hamid Ali**
UTHM
 - **Fokrul Alom Mazarbhuiya**
King Khalid University
 - **Frank Ibikunle**
Covenant University
 - **Fu-Chien Kao**
Da-Y eh University
 - **G. Sreedhar**
Rashtriya Sanskrit University
 - **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh
 - **Ghalem Belalem**
University of Oran (Es Senia)
 - **Gufran Ahmad Ansari**
Qassim University
 - **Hadj Hamma Tadjine**
IAV GmbH
 - **Hanumanthappa.J**
University of Mangalore, India
 - **Hesham G. Ibrahim**
Chemical Engineering Department, Al-Mergheb University, Al-Khoms City
 - **Dr. Himanshu Aggarwal**
Punjabi University, India
 - **Huda K. AL-Jobori**
Ahlia University
 - **Iwan Setyawan**
Satya Wacana Christian University
 - **Dr. Jamaiah Haji Yahaya**
Northern University of Malaysia (UUM), Malaysia
 - **Jasvir Singh**
Communication Signal Processing Research Lab
 - **Jatinderkumar R. Saini**
- S.P.College of Engineering, Gujarat
- **Prof. Joe-Sam Chou**
Nanhua University, Taiwan
 - **Dr. Juan José Martínez Castillo**
Yacambu University, Venezuela
 - **Dr. Jui-Pin Yang**
Shih Chien University, Taiwan
 - **Jyoti Chaudhary**
high performance computing research lab
 - **K Ramani**
K.S.Rangasamy College of Technology, Tiruchengode
 - **K V.L.N.Acharyulu**
Bapatla Engineering college
 - **K. PRASADH**
METS SCHOOL OF ENGINEERING
 - **Ka Lok Man**
Xi'an Jiaotong-Liverpool University (XJTLU)
 - **Dr. Kamal Shah**
St. Francis Institute of Technology, India
 - **Kanak Saxena**
S.A.TECHNOLOGICAL INSTITUTE
 - **Kashif Nisar**
Universiti Utara Malaysia
 - **Kavya Naveen**
 - **Kayhan Zrar Ghafoor**
University Technology Malaysia
 - **Kodge B. G.**
S. V. College, India
 - **Kohei Arai**
Saga University
 - **Kunal Patel**
Ingenuity Systems, USA
 - **Labib Francis Gergis**
Misr Academy for Engineering and Technology
 - **Lai Khin Wee**
Technischen Universität Ilmenau, Germany
 - **Latha Parthiban**
SSN College of Engineering, Kalavakkam
 - **Lazar Stosic**
College for professional studies educators, Aleksinac
 - **Mr. Lijian Sun**
Chinese Academy of Surveying and Mapping, China
 - **Long Chen**
Qualcomm Incorporated
 - **M.V.Raghavendra**
Swathi Institute of Technology & Sciences, India.
 - **M. Tariq Banday**
University of Kashmir

- **Madjid Khalilian**
Islamic Azad University
- **Mahesh Chandra**
B.I.T, India
- **Mahmoud M. A. Abd Ellatif**
Mansoura University
- **Manas deep**
Masters in Cyber Law & Information Security
- **Manpreet Singh Manna**
SLIET University, Govt. of India
- **Manuj Darbari**
BBD University
- **Marcellin Julius NKENLIFACK**
University of Dschang
- **Md. Masud Rana**
Khunla University of Engineering & Technology,
Bangladesh
- **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
- **Messaouda AZZOUZI**
Ziane AChour University of Djelfa
- **Dr. Michael Watts**
University of Adelaide, Australia
- **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in
Vranje
- **Miroslav Baca**
University of Zagreb, Faculty of organization and
informatics / Center for biomet
- **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
- **Mohammad Talib**
University of Botswana, Gaborone
- **Mohamed El-Sayed**
- **Mohammad Yamin**
- **Mohammad Ali Badamchizadeh**
University of Tabriz
- **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science &
Technology
- **Mohd Helmy Abd Wahab**
Universiti Tun Hussein Onn Malaysia
- **Mohd Nazri Ismail**
University of Kuala Lumpur (UniKL)
- **Mona Elshinawy**
Howard University
- **Monji Kherallah**
University of Sfax
- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**
Universiti Teknologi Malaysia UTM
- **Dr. Murugesan N**
Government Arts College (Autonomous), India
- **N Ch.Sriman Narayana Iyengar**
VIT University
- **Natarajan Subramanyam**
PES Institute of Technology
- **Neeraj Bhargava**
MDS University
- **Nitin S. Choubey**
Mukesh Patel School of Technology
Management & Eng
- **Noura Aknin**
Abdelamlek Essaadi
- **Om Sangwan**
- **Pankaj Gupta**
Microsoft Corporation
- **Paresh V Virparia**
Sardar Patel University
- **Dr. Poonam Garg**
Institute of Management Technology,
Ghaziabad
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **Pradip Jawandhiya**
Jawaharlal Darda Institute of Engineering &
Techno
- **Rachid Saadane**
EE departement EHTP
- **Raghuraj Singh**
- **Raj Gaurang Tiwari**
AZAD Institute of Engineering and Technology
- **Rajesh Kumar**
National University of Singapore
- **Rajesh K Shukla**
Sagar Institute of Research & Technology-
Excellence, India
- **Dr. Rajiv Dharaskar**
GH Rasoni College of Engineering, India
- **Prof. Rakesh. L**
Vijetha Institute of Technology, India
- **Prof. Rashid Sheikh**
Acropolis Institute of Technology and Research,
India
- **Ravi Prakash**
University of Mumbai
- **Reshmy Krishnan**
Muscat College affiliated to stirling University.U
- **Rongrong Ji**
Columbia University

- **Ronny Mardiyanto**
Institut Teknologi Sepuluh Nopember
- **Ruchika Malhotra**
Delhi Technoogical University
- **Sachin Kumar Agrawal**
University of Limerick
- **Dr.Sagarmay Deb**
University Lecturer, Central Queensland
University, Australia
- **Said Ghoniemy**
Taif University
- **Saleh Ali K. AlOmari**
Universiti Sains Malaysia
- **Samarjeet Borah**
Dept. of CSE, Sikkim Manipal University
- **Dr. Sana'a Wafa Al-Sayegh**
University College of Applied Sciences UCAS-
Palestine
- **Santosh Kumar**
Graphic Era University, India
- **Sasan Adibi**
Research In Motion (RIM)
- **Saurabh Pal**
VBS Purvanchal University, Jaunpur
- **Saurabh Dutta**
Dr. B. C. Roy Engineering College, Durgapur
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Sergio Andre Ferreira**
Portuguese Catholic University
- **Seyed Hamidreza Mohades Kasaei**
University of Isfahan
- **Shahanawaj Ahamad**
The University of Al-Kharj
- **Shaidah Jusoh**
University of West Florida
- **Shriram Vasudevan**
- **Sikha Bagui**
Zarqa University
- **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
- **Slim BEN SAOUD**
- **Dr. Smita Rajpal**
ITM University
- **Suhas J Manangi**
Microsoft
- **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia
- **Sumit Goyal**
- **Sunil Taneja**
Smt. Aruna Asaf Ali Government Post Graduate
College, India
- **Dr. Suresh Sankaranarayanan**
University of West Indies, Kingston, Jamaica
- **T C. Manjunath**
HKBK College of Engg
- **T C.Manjunath**
Visvesvaraya Tech. University
- **T V Narayana Rao**
Hyderabad Institute of Technology and
Management
- **T. V. Prasad**
Lingaya's University
- **Taiwo Ayodele**
Lingaya's University
- **Tarek Gharib**
- **Totok R. Biyanto**
Infonetmedia/University of Portsmouth
- **Varun Kumar**
Institute of Technology and Management, India
- **Vellanki Uma Kanta Sastry**
SreeNidhi Institute of Science and Technology
(SNIST), Hyderabad, India.
- **Venkatesh Jaganathan**
- **Vijay Harishchandra**
- **Vinayak Bairagi**
Sinhgad Academy of engineering, India
- **Vishal Bhatnagar**
AI&R, Govt. of NCT of Delhi
- **Vitus S.W. Lam**
The University of Hong Kong
- **Vuda Sreenivasarao**
St.Mary's college of Engineering & Technology,
Hyderabad, India
- **Wei Wei**
- **Wichian Sittiprapaporn**
Mahasarakham University
- **Xiaoqing Xiang**
AT&T Labs
- **Y Srinivas**
GITAM University
- **Yilun Shang**
University of Texas at San Antonio
- **Mr.Zhao Zhang**
City University of Hong Kong, Kowloon, Hong
Kong
- **Zhixin Chen**
ILX Lightwave Corporation
- **Zuqing Zhu**
University of Science and Technology of China

CONTENTS

Paper 1: A multi-scale method for automatically extracting the dominant features of cervical vertebrae in CT images
Authors: *Tung-Ying Wu, Sheng-Fuu Lin*

PAGE 1 – 8

Paper 2: Evolutionary approach to optimisation of the operation of electric power distribution networks
Authors: *Jan Stępień, Sylwester Filipiak*

PAGE 9 – 16

Paper 3: Expected Reliability of Everyday- and Ambient Assisted Living Technologies
Authors: *Frederick Steinke, Tobias Fritsch, Andreas Hertzner, Helmut Tautz, Simon Zickwolf*

PAGE 17 – 22

Paper 4: Modeling the Cut-off Frequency of Acoustic Signal with an Adaptive Neuro-Fuzzy Inference System (ANFIS)
Authors: *Y. NAHRAOUI, E.H. AASSIF, R.LATIF, G.Maze*

PAGE 23 – 33

Paper 5: The quest towards a winning Enterprise 2.0 collaboration technology adoption strategy
Authors: *Robert Louw, Jabu Mtsweni*

PAGE 34 – 39

Paper 6: Face Recognition System Based on Different Artificial Neural Networks Models and Training Algorithms
Authors: *Omaima N. A. AL-Allaf, Abdelfatah Aref Tamimi, Mohammad A. Alia*

PAGE 40 – 47

Paper 7: Image Blocks Model for Improving Accuracy in Identification Systems of Wood Type
Authors: *Gasim, Kudang Boro Seminar, Agus Harjoko, Sri Hartati*

PAGE 48 – 53

Paper 8: A Strategy for Training Set Selection in Text Classification Problems
Authors: *Maria Luiza C. Passini, Katusca B. Estébanez, Graziela P. Figueredo, Nelson F. F. Ebecken*

PAGE 54 – 60

Paper 9: Study of the capacity of Optical Network On Chip based on MIMO (Multiple Input Multiple Output) system
Authors: *S.Mhatli, B.Nsiri, R.Attia*

PAGE 61 – 65

Paper 10: Face Recognition as an Authentication Technique in Electronic Voting
Authors: *Noha E. El-Sayad, Rabab Farouk Abdel-Kader, Mahmoud Ibraheem Marie*

PAGE 66 – 71

Paper 11: Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework
Authors: *Roobaea AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew*

PAGE 72 – 91

Paper 12: Proposed Multi-Modal Palm Veins-Face Biometric Authentication
Authors: *S.F.Bahgat, S. Ghoniemy, M. Alotaibi*

PAGE 92 – 96

Paper 13: Micro Sourcing Strategic Framework for Low Income Group

Authors: Noor Habibah Arshad, Siti Salwa Salleh, Syaripah Ruzaini Syed Aris, Norjansalika Janom, Norazam Mastuki

PAGE 97 – 105

Paper 14: A New Algorithm to Represent Texture Images

Authors: Silvia María Ojeda, Grisel Maribel Britos

PAGE 106 – 111

Paper 15: Image and Video based double watermark extraction spread spectrum watermarking in low variance region

Authors: Mriganka Gogoi, Koushik Mahanta, H.M.Khalid Raihan Bhuyan, Dibya Jyoti Das, Ankita Dutta

PAGE 112 – 116

Paper 16: A Framework for Creating a Distributed Rendering Environment on the Compute Clusters

Authors: Ali Sheharyar, Othmane Bouhali

PAGE 117– 123

Paper 17: Integrating Social Network Services with Vehicle Tracking Technologies

Authors: Ahmed ElShafee, Mahmoud ElMenshawi, Mena Saeed

PAGE 124 – 132

Paper 18: An Efficient Approach for Image Filtering by Using Neighbors pixels

Authors: Smrity Prasad, N.Ganesan

PAGE 133 – 138

Paper 19: A Comparative Study of Three TDMA Digital Cellular Mobile Systems (GSM, IS-136 NA-TDMA and PDC) Based On Radio Aspect

Authors: Laishram Prabhakar

PAGE 139 – 143

Paper 20: Format SPARQL Query Results into HTML Report

Authors: Dr Sunitha Abburu, G.Suresh Babu

PAGE 144– 148

Paper 21: A Comprehensive Evaluation of Weight Growth and Weight Elimination Methods Using the Tangent Plane Algorithm

Authors: P May, E Zhou, C. W. Lee

PAGE 149– 156

Paper 22: Exploiting the Role of Hardware Prefetchers in Multicore Processors

Authors: Hasina Khatoun, Shahid Hafeez Mirza, Talat Altaf

PAGE 157– 167

Paper 23: Improving Assessment Management Using Tools

Authors: Shang Gao, Jo Coldwell-Neilson, Andrzej Goscinski

PAGE 168– 173

Paper 24: Data fusion based framework for the recognition of Isolated Handwritten Kannada Numerals

Authors: Mamatha.H.R, Sucharitha Srirangaprasad, Srikantamurthy K

PAGE 174 – 182

Paper 25: Designing a Markov Model for the Analysis of 2-tier Cognitive Radio Network

Authors: Tamal Chakraborty, Ili Saha Misra

PAGE 183 – 192

Paper 26: A Fuzzy Rule Based Forensic Analysis of DDoS Attack in MANET

Authors: Ms. Sarah Ahmed, Ms. S. M. Nirkhi

PAGE 193 – 197

Paper 27: A comparative study of Image Region-Based Segmentation Algorithms

Authors: Lahouaoui LALAOUI, Tayeb MOHAMADI

PAGE 198 – 206

Paper 28: Automated Classification of L/R Hand Movement EEG Signals using Advanced Feature Extraction and Machine Learning

Authors: Mohammad H. Alomari, Aya Samaha, Khaled AlKamha

PAGE 207 – 212

Paper 29: Case Study of Named Entity Recognition in Odia Using Crf++ Tool

Authors: Dr.Rakesh ch. Balabantaray, Suprava Das, Kshirabdhii Tanaya Mishra

PAGE 213 – 216

Paper 30: TX-Kw: An Effective Temporal XML Keyword Search

Authors: Rasha Bin-Thalab, Neamat El-Tazi, Mohamed E.El-Sharkawi

PAGE 217 – 226

Paper 31: Correlated Topic Model for Web Services Ranking

Authors: Mustapha AZNAG, Mohamed QUAFAROU, Zahi JARIR

PAGE 227– 239

Paper 32: Development of Copeland Score Methods for Determine Group Decisions

Authors: Ermatifa, Sri Hartati, Retantyo Wardoyo, Agus Harjoko

PAGE 240 – 242

Paper 33: New electronic white cane for stair case detection and recognition using ultrasonic sensor

Authors: Sonda Ammar Bouhamed, Imene Khanfir Kallel, Dorra Sellami Masmoudi

PAGE 243 – 255

Paper 34: Watermarking in E-commerce

Authors: Peyman Rahmati, Andy Adler, Thomas Tran

PAGE 256 – 265

Paper 35: A Novel Software Tool for Analysing NTR File System Permissions

Authors: Simon Parkinson, Andrew Crampton

PAGE 266 – 272

Paper 36: Probabilistic Distributed Algorithm for Uniform Election in Triangular Grid Graphs

Authors: El Mehdi Stouti, Ismail Hind, Abdelaaziz El Hibaoui

PAGE 273– 282

Paper 37: Correlated Topic Model for Web Services Ranking

Authors: Mustapha AZNAG, Mohamed QUAFAROU, Zahi JARIR

PAGE 283– 291

Paper 38: Wideband Parameters Analysis and Validation for Indoor radio Channel at 60/70/80GHz for Gigabit Wireless Communication employing Isotropic, Horn and Omni directional

Authors: E. Affum, E.T. Tchao, K. Diawuo, K. Agyekum

PAGE 292 – 297

Paper 39: Wideband Parameters Analysis and Validation for Indoor radio Channel at 60/70/80GHz for Gigabit Wireless Communication employing Isotropic, Horn and Omni directional

Authors: E. Affum, E.T. Tchao, K. Diawuo, K. Agyekum

PAGE 298 – 306

A multi-scale method for automatically extracting the dominant features of cervical vertebrae in CT images

Tung-Ying Wu

Department of Electrical and Computer Engineering
National Chiao Tung University
Hsinchu City, Taiwan (R.O.C)

Sheng-Fuu Lin

Department of Electrical and Computer Engineering
National Chiao Tung University
Hsinchu City, Taiwan (R.O.C)

Abstract—Localization of the dominant points of cervical spines in medical images is important for improving the medical automation in clinical head and neck applications. In order to automatically identify the dominant points of cervical vertebrae in neck CT images with precision, we propose a method based on multi-scale contour analysis to analyzing the deformable shape of spines. To extract the spine contour, we introduce a method to automatically generate the initial contour of the spine shape, and the distance field for level set active contour iterations can also be deduced. In the shape analysis stage, we at first coarsely segment the extracted contour with zero-crossing points of the curvature based on the analysis with curvature scale space, and the spine shape is modeled with the analysis of curvature scale space. Then, each segmented curve is analyzed geometrically based on the turning angle property at different scales, and the local extreme points are extracted and verified as the dominant feature points. The vertices of the shape contour are approximately derived with the analysis at coarse scale, and then adjusted precisely at fine scale. Consequently, the results of experiment show that we approach a success rate of 93.4% and accuracy of 0.37mm by comparing with the manual results.

Keywords—cervical spine; active contour; curvature scale space; turning angle.

I. INTRODUCTION

Anatomical landmarks and dominant points of cervical vertebrae are of considerable importance for many applications on orthopedics, neurology, and radiation therapy planning. In many researches about computer-aided techniques, the geometric characteristics of anatomical features were mentioned to be utilized on the applications like image-based surgical guidance and operation planning [1]. However, because of the anatomical variation between patients and the complexity of medical images, automatic analysis and information collection in computerized tomography (CT) images are still challenging tasks.

In [1], Lee *et al.* proposed a method to automatically locate the lumbar spine pedicles in CT images by referencing the canal boundaries for pedicle screw. But for cervical vertebrae operation planning, more landmarks are required [7]. Dominant feature points of cervical vertebrae include transverse foramens, spinous processes, and corners of lateral facets, etc. In order to automatically find the dominant features points in cervical vertebrae, Rochies and Winter proposed researches about detection of anatomical landmarks and dominant points by matching feature sets derived from 2D wavelet and Gabor transform in CT and MRI images [8-9]. The

proposed methods used the graph matching algorithm to perform a global search, and the similarity of two feature sets was utilized to localize dominant points. However, the method was less adaptive to morphological deformation and the performance on accuracy was not satisfied. Besides, automatic recognition of spine shape is not only used for surgery applications mentioned above, but also an issue of importance of computer-aided diagnosis (CAD). As the growth of the volume of medical images with the progress of medical imaging techniques, it becomes exhausting to inspect all the data in detail manually. Therefore, CAD system is introduced to improve medical automation and the medical data can be pre-processed automatically and then provide information for assisting diagnosis. The anatomical structure around the spine such as soft tissues, muscles, and glands can be regarded as a planar adjacent anatomical space, so the relative location of anatomical structure is able to be inferred according to anatomical knowledge [10].

The anatomical landmarks detected in images are able to be set as the starting points of image segmentation for automatic diagnosis and navigation. In addition, the points of dominant features in 2D slices can also be seeds for shape modeling, 3D reconstruction and registration. In neck CT images, the cervical vertebrae are significant landmarks for medical application but to automatically extract the precise feature points from the complex images is still a challenging task. In this paper, we propose a method to automatically find the feature points of cervical spines in CT slices as shown in Fig 1 based on geometric analysis in companion with anatomical knowledge.

In Section 2, the geodesic active contour model referencing the gradient information is mentioned to extract the cervical spine. We moreover propose a method to automatically set the initial conditions including initial contour and initial distance field for active contour iteration. Then in Section 3, the shape contour extracted from the previous step is coarsely segmented and modeled with a proper scale by means of the curvature scale space (CSS) analysis, and multi-scale geometric analysis is then applied to identify the dominant feature points at each segment. As shown in Fig.1, the dominant points proposed to extract include the vertices at both sides of vertebral body, near transverse foramens and pedicles (points 1 and 2), the corners of the facets (points 3 and 4) and the corners of spinous process. Furthermore, the vertebral body and lamina regions can be further inferred based on the four determined dominant points. In Section 4, the experiment is carried out on 250 neck CT slices, and the points found out are finally examined with

the points pointed out by the clinical experts to evaluate the success rate and accuracy. In Section 5, we discuss the result and conclude the work.

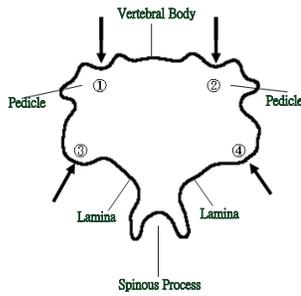


Fig. 1. Typical cervical spine shape with dominant features labeled.

II. VERTEBRAE EXTRACTION

As shown in Fig. 2, the cervical spine locates at the center of the neck and appears in high brightness, and there are various textured soft tissues around the cervical spine. The air path in relatively low brightness is adjacent to the cervical spine.

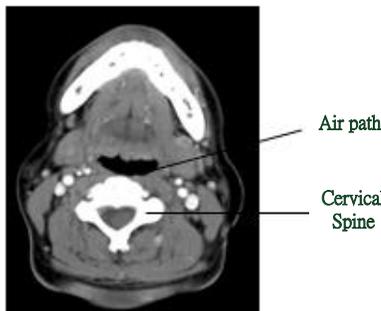


Fig. 2. A cervical CT image and the relative location of the air path and the cervical spine.

In order to extract the spines in CT images, there have been several segmentation methods proposed, including model-based segmentation, adaptive thresholding, multi-scale canny edge detection and active contour algorithm [1, 10-15]. Among these methods, gradient-based methods are considered to perform better accuracy than gray level thresholding because the magnitude and direction of the gradient can be used to accurately locate the edges. Besides, in order to segment deformable objects, active contour methods are considered as an effective method to generate continuous boundaries. Geodesic active contour (GAC) is an active contour model (ACM) based on the relation between active contour and the computation of geodesic or minimum distance field [16-17]. The initial contour deforms and gradually converges toward the region boundaries through iterations controlled by the gradient-based stopping function with updating the distance field [16]. In many applications of medical image segmentation, the initial contours were manually placed near the targets for more effective converging properties and less computation. However, manual placement is not appropriate for automatic

segmentation. In this section, we describe a method about automatic placement of the initial contours for delineating the cervical spines and the initial distance field for GAC computation. The method is summarily shown as the schematic diagram in Fig 3.

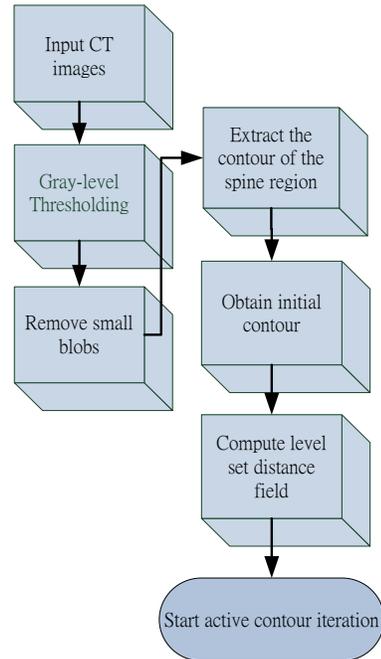


Fig. 3. Schematic flow diagram of the proposed method

A. Geodesic Active Contour

The main idea of the active contour model is to minimize the energy associated to the contour derived from the image gradient and the contour curvature. In order to formulate the energy for GAC computation, a contour C in an image is represented by the parametric vector equation:

$$C(t) = (x(t), y(t)) \quad (1)$$

Hence, the energy function of the GAC contour model comprising the internal and external energy terms can be described as the terms below:

$$E(C(t))_{geo} = \alpha \int_0^1 C'(t)^2 dt - \lambda \int_0^1 g(|\nabla I(C(t))|)^2 dt \quad (2)$$

$$= \int_0^1 E_{int}(C(t)) + E_{ext}(C(t)) dt$$

where

$$g(x) = \frac{1}{\sqrt{1+x^2}}, \quad (3)$$

t is the arc length parameter, E_{int} is the internal energy while E_{ext} is the external energy of contour C . Let C_0 be the initial contour for active contour iterations and $g(x)$ denotes a monotonically stopping function which conducts the contour converge toward the boundary points based on the direction and magnitude of gradient. In order to deform an initial contour

towards local minima points of the energy function in the image, the steady state solution is given by

$$\frac{\partial C}{\partial t} = g(I)\kappa\vec{N} - (\nabla g \cdot \vec{N})\vec{N}, \quad (4)$$

where κ represents the curvature derived from the equation:

$$\kappa(t) = \frac{\dot{x}(t)\ddot{y}(t) - \dot{y}(t)\ddot{x}(t)}{(\dot{x}(t)^2 + \dot{y}(t)^2)^{3/2}}, \quad (5)$$

where $\dot{\cdot}$ means the derivative, i.e., $\dot{x}(t) = \frac{dx(t)}{dt}$ and \vec{N} is

the unit inward normal. The curve evolves according to the steepest-descent method to deform the initial curve C_0 based on the curvature and gradient. This geodesic problem can be solved by introducing the level-sets approach [16-17]. In level set formulation, the contour C is regarded as the zero level-set of a function, so a contour can be represented as a distance map measuring the minimum distance from a point to the contour. Therefore, the curve evolution (4) can be represented by

$$\frac{\partial u}{\partial t} = g(I)\kappa|\nabla u| + \nabla g \cdot \nabla u \quad (6)$$

where u is a signed distance field of a contour C , and C can be regarded as the zero level-set in u with u_0 denoting the initial distance field. The level set method evolves a contour by propagating the wave front. The fronts move ahead with a velocity V and arrival time T , and the level set front propagation equation is given by [18]

$$|\nabla T|V = 1. \quad (7)$$

The distance map u is iteratively updated by means of computing the narrow band near the existing front and solving the propagation equation to bring new pixels into the narrow band. The curve evolution operation keeps until the front does not move or the number of iterations approaches the limit. However, in practice, GAC iterations require an initial contour C_0 in companion the corresponding distance field u_0 , and the initial condition significantly affects the result.

B. Automatic cervical vertebrae extraction

In order to obtain C_0 and u_0 for GAC, we start with thresholding the original CT images and the region of gray level higher than the threshold is set to 1 or 0 otherwise. Because the spine regions in a CT image appear in uniformly higher brightness than other regions, the threshold should be chosen higher than the result from Otsu thresholding method [19]. However, other regions with high brightness as the spine region would possibly be found, like mandible bones and carotids.

Fortunately, the air path shows distinguishable darkness in cervical CT images and anatomically locates nearby the cervical spine. Therefore, after the morphological operation is proceeded to remove the regions of small area in the binary image, the large blob nearest to the air path is selected. Erosion operation with a 3×3 mask is then involved in to extract the external boundary by collecting the removed elements. The

approximate contour of a spine as a result could be sketched out and arranged into a series. The initial contour can be set closed to the real boundary, so it is an effective initial contour for active contour iteration. Different from traditional methods which simply place a circle or a square around the target, the initial distance field can be generated based on geometric relation. In order to build the distance field with an arbitrary contour C_0 , the field need to be derived by distance transform which measured the minimum distance from a point to C_0 . The distance transform can be determined by

$$Dist(x, y) = \min_{(i, j) \in C_0} (\sqrt{(x-i)^2 + (y-j)^2}), \quad (8)$$

and

$$Dist(x_n, y_n) = 0, \text{ if } (x_n, y_n) \in C_0.$$

The pixel inside the closed contour C_0 is set negative. The initial distance field u_0 can be computed by

$$u_0(x, y) = \begin{cases} Dist(x, y) \times -1, & \text{if } (x, y) \text{ inside } C_0 \\ Dist(x, y), & \text{else.} \end{cases} \quad (9)$$



Fig. 4. (a) An extracted initial contour (b) the corresponding distance field of (a) representing with gray level. The brighter means the point has longer distance to the contour.

III. SHAPE ANALYSIS AND FEATURE IDENTIFICATION

In this section, we describe a coarse-to-fine method to deal with the deformable shape. The spine shape contour is at first coarsely segmented and modeled with CSS, and the dominant feature points on the segmented curves are then figured out by analyzing the detail features in fine scales. The method in this section is demonstrated in the summarizing diagram in Fig.5.

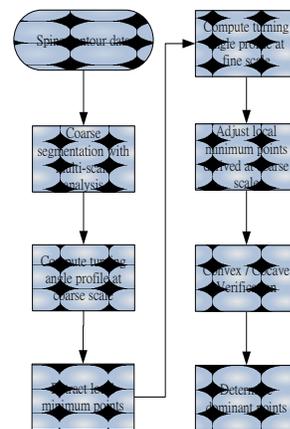


Fig.5. The schematic diagram of our algorithm of extracting the dominant feature points by analyzing the shape contour.

A. Coarse contour segmentation with CSS

After shape contours are extracted from the original images, the dominant points on the contour which are meaningful for deformable shape recognition need to be figured out. For convenience and symmetry, the point, which is closest to the center of air path, is assigned as the starting point of the closed contour. For shape registration and recognition, geometric points invariant over rotation, scaling and partial occlusion are in considerable importance for the deformable shapes. In general, curvature is a significant property of curves, and the local maximum points or zero-crossing points of curvature profile are considered as meaningful points of the shape [20-22]. However, local maximum points of the curvature, which could correspond to the corners or vertices of the contour, are too sensitive and easily affected by noise. Because zero-crossing points demonstrate the intersection between a concave contour segment and the adjacent convex segment, as the scale gets higher, the neighboring zero-crossing points also gradually merge together.

Zero-crossing points of the curvature are more adaptive geometric features for deformable shape analysis. CSS is a multi-scale method of collecting zero-crossing points of curvature of a closed contour derived from each scale and has been proven as an effective method for shape description and matching over scaling, rotation, partial occlusion and deformation [23-26]. The curvature scale space image (CSSI) is a binary two-dimensional image that records the position of inflection points of the curve convoluted by different-scaled Gaussian filters. In CSSI, along the horizontal axis is the normalized arc length of the contour from 0 to 1; along the vertical axis is the scale parameter. As the standard deviation of Gaussian functions varies from small to large, the contour is gradually blurred while details are gradually eliminated. The multi-scale curvature can be computed by the following equations.

$$k(t, \sigma) = \frac{\dot{X}(t, \sigma)\ddot{Y}(t, \sigma) - \dot{Y}(t, \sigma)\ddot{X}(t, \sigma)}{(\dot{X}(t, \sigma)^2 + \dot{Y}(t, \sigma)^2)^{3/2}}, \quad (10)$$

where

$$X(t, \sigma) = x(t) \otimes h(t, \sigma) \quad (11)$$

$$Y(t, \sigma) = y(t) \otimes h(t, \sigma), \quad (12)$$

where

$$h(t, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}, \quad (13)$$

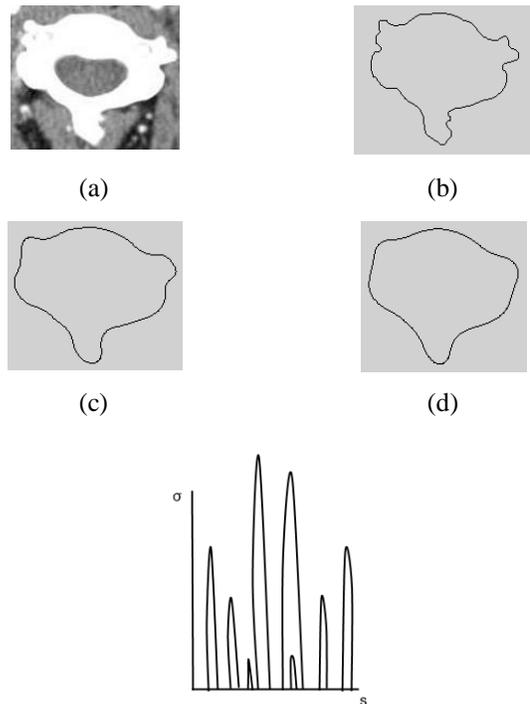
\otimes is the convolution operator. Function $h(t, \sigma)$ denotes a zero-mean Gaussian function with kernel size parameter σ and σ is also referred as the scale parameter. If the curve with smoothing parameter has a zero-crossing point at location s on scale σ , we set $CSSI(\sigma, s) = 1$, or $CSSI(\sigma, s) = 0$ otherwise.

As a result, a contour can be represented by a CSSI image with several CSS contours corresponding to segments of the shape contour in it, and the CSS contours are constructed by zero-crossing points at different scales. Each CSS contour in the CSSI represents a concave or convex segment of the corresponding shape contour. In the CSSI derived from a typical cervical spine shape, four significant contours of the largest σ are depicted and correspond to four segments of a spine shape contour. With the other four segments squeezed between every two of the fours corresponding to the four significant contours in CSSI, the spine shape contour can be mainly segmented into eight periods of curve. Because a zero-crossing point is regarded as a breakpoint of a concave segment and another convex one, a shape contour can be divided into several meaningful segments by localizing the zero-crossing points.

Each shape contour has its peculiar arrangement of zero-crossing points. Ming *et al.* proposed a CSS-based method for pattern matching by comparing the CSSI line by line [23], because the zero-crossing points at each scale can be recognizable features of shape contours. In order to effectively separate the shape contour of spine into the eight main segments, an appropriate standard deviation value of Gaussian filter σ , which is related to scale needs to be chosen. Let σ_n denote the peak of the CSS contour having the n th highest σ in CSSI. The threshold for choosing the analysis scale in this paper is determined by

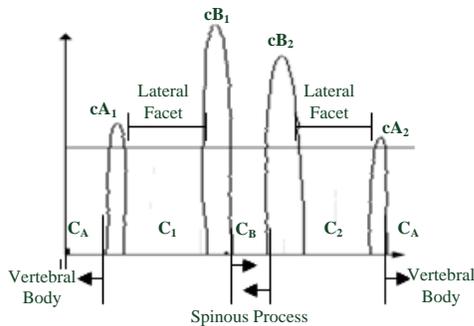
$$\sigma_{analysis} = \begin{cases} 0.5\sigma_1, & \text{if } 0.5\sigma_1 < \sigma_4 \\ 0.75\sigma_4, & \text{else.} \end{cases} \quad (14)$$

Only the contours with peaks higher than $\sigma_{analysis}$ are considered for the following steps.

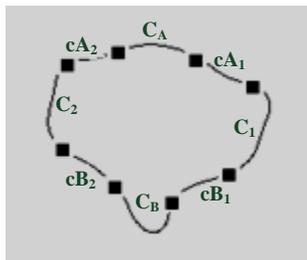


(e)

Fig.6. (a) is the original image. (b) is the extracted contour. (c) is the contour after convolving with the Gaussian filter of $\sigma=15$. (d) contour after convolving Gaussian filter of $\sigma=30$.(e) the CSSI of the contour in (b)



(a)



(b)

Fig.10. (a) CSSI of a cervical contour (b) The corresponding segments of contour labeled in (a).

Fig.10 demonstrates a spine shape model based on CSS. It can be observed that a spine shape contour can be segmented into several CSS contours, and there are four main CSS contours labeled as cA_1 , cA_2 , cA_2 and cB_2 in two symmetric pairs in a CSSI of a spine shape contour. cA_1 , cA_2 , cA_2 and cB_2 are corresponding to the four contour segments A_1 , A_2 , A_2 and B_2 as shown in Fig.10(b) respectively. The segment corresponding to the spinous process locates at the period labeled as C_B between the pair of CSS contours (cB_1 , cB_2) near the middle of the horizontal axis. Also, the segment C_A corresponding to the vertebral body can be deduced by referencing another two apparent CSS contours (cA_1 , cA_2) at the both sides of the starting (end) point. The segments squeezed by (cA_1 , cB_1) and (cA_2 , cB_2), which are corresponding to the facets at both lateral sides of the spine, are labeled as C_1 and C_2 respectively. Eight zero-crossing points among (cA_1 , cA_2) and (cB_1 , cB_2) can be extracted at scale $\sigma_{analysis}$, and the eight points are used to separate the spine contour into eight segments. The apparent segments corresponding to the vertebral body, the lateral facets and the spinous process, as a result can be coarsely indicated in CSSI.

In order to find the two main symmetric pairs from the original CSSI as Fig.6(e), the symmetry property is utilized. (cA_1 , cB_1) and (cA_2 , cB_2) are symmetric against the starting point of contour, which is contour point closest to the air path

center at C_A . From the CSS contours with the highest σ , every four CSS contours such as (cA_1 , cA_2) and (cB_1 , cB_2) in Fig.10 are extracted at a time. The contour segments within each two CSS contour pairs are extracted, such as the segments corresponding to the periods of (cA_1 , C_1 , cB_1) and (cB_2 , C_2 , cA_2). The extracted segments are measured for their curve similarity, and high similarity means high symmetry of the contour segments. The four CSS contours in CSSI corresponding to the two contour segments having the highest curve similarity are considered as the four main CSS contours of a spine shape and labeled based on the model. The similarity of two curves is estimated by measuring the difference the two curves. Let f_1 and f_2 be two curves with N uniform sampling points

$$f_1 = \{(x_1(n_1), y_1(n_1))\} \quad (15)$$

$$f_2 = \{(x_2(n_2), y_2(n_2))\} \quad (16)$$

where t_1 and t_2 are integers, (x_1, y_1) and (x_2, y_2) are the points belonging to f_1 and f_2 and $f_1(0) = f_2(0)$, $f_1(N-1) = f_2(N-1)$. The dissimilarity D of two curves is evaluated by measuring the distance of the points between two curves.

$$D(f_1, f_2) = \max(h(f_1, f_2), h(f_2, f_1)) \quad (17)$$

where $D(f_1, f_1) = 0$ and

$$h(A, B) = \frac{1}{N} \sum_{n_1 \in A} \min_{n_2 \in B} (A(n_1), B(n_2)) \quad (18)$$

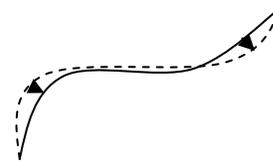


Fig.11 Measurement of the dissimilarity of two curves.

The lower dissimilarity of two curves means the higher similarity.

B. Shape analysis and dominant point identification

After coarse segmentation, dominant points or landmarks can be finely determined by analyzing the contour based on geometrical properties of each segment. Curvature is an important property of a curve, and has been widely used for precisely finding vertices of shape. However, it is so sensitive to noise and small variation of the contour that it is not appropriate for deformable objects or objects in complex images like cervical CT slices. Turning angle is another useful geometric property of curves for comprehending the local variation [27-29]. Xu *et al.* applied the turning angle property for curve evolution in automatic spine shape analysis [30]. The bending angle of two adjacent line segments is computed and normalized for evaluating the contribution to the whole shape. "Included angle" was defined in [1] to figure out the sharp convex characteristics between two adjacent elements. In our research, to evaluate the curve segments, bending angle profile along the whole contour is calculated by the following equation:

$$\theta(t,d) = \cos^{-1} \left(\frac{\mathbf{V}_1(t,d) \bullet \mathbf{V}_2(t,d)}{\|\mathbf{V}_1(t,d)\| \times \|\mathbf{V}_2(t,d)\|} \right) \quad (19)$$

where \mathbf{V}_1 and \mathbf{V}_2 are vectors,

$$\mathbf{V}_1(t,d) = \langle x(t+d) - x(t), y(t+d) - y(t) \rangle, \quad (20)$$

$$\mathbf{V}_2(t,d) = \langle x(t-d) - x(t), y(t-d) - y(t) \rangle, \quad (21)$$

where \bullet means the inner product operator, $\|\cdot\|$ means the norm of a vector and d is the scale parameter. The value of turning angle is closed to but not larger 180 in degrees at the period without acute variation. On the contrary, vertices may locate at the points of local extreme value. As the scale parameter d grows from small to large, only salient vertices can be preserved and more noise is eliminated. Therefore, a multi-scale method is introduced in this work. The turning angle profile of a contour is at first derived at a higher scale d_1 , and the local minimum points recognized as dominant points are coarsely localized. Then, the points are adjusted by referencing the nearest minimum points derived at a finer scale d_2 , as indicated in Fig.12. For suppressing undesired noise, the turning angle profiles convolve with a Gaussian function before extracting the local minimum points.

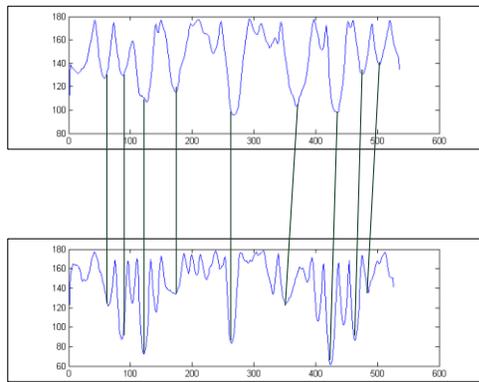


Fig.12. Turning angle profiles of a cervical spine contour with different scales and the some corresponding vertices at the two scales are pointed out.

However, the vertices can not be directly discriminated as a concave or a convex one by θ but it is necessary for verifying the dominant shape features at each segment. After points of local extreme are localized, each extracted point is recomputed for its curvature by (11) at proper scale associated with d_1 . If the curvature is positive, it means that the angle is convex, or concave otherwise. As shown in Fig. 1, points 1 and 2 are two concave vertices and the segment within point 1 and point 2 indicates the vertebral body. In order to locate the vertices with precision, three adjacent segments of the contour corresponding to the period A_1 , C_A and A_2 in Fig.7 are introduced to compute the geometric information. The two concave vertices result in two local minimum points at both ends of the corresponding segment of turning angle profile. Moreover, the arc segment of the vertebral body can also be accurately determined within these two vertices. Besides, the spinous process is considered to locate at the period C_B , which is the middle segment of the whole shape contour. The apparent angles within this period

are important because the angles may denote the corners of spinous processes or bifurcation which are apparent landmarks of cervical anatomy. The corners of spinous processes are convex angles, and if bifurcation exists, there will be another concave angle between the two convex corners. Fig.12 shows the convex and concave vertices in the cervical spine denoting with different marks. In Fig.1, the facet corners at points 3 and 4 are convex vertices with angle larger than 90 in degree. The “corners” are not only important for nerve root injection operations [2, 4], but also for determining the position of lamina periods. In addition to the points labeled in Fig.1, at the middle of lateral facets C_1 and C_2 , there are sometimes two concave vertices near the facet corners as the points denoted with crosses in Fig.12.

If existing, the concave vertices are also dominant points near the foramen and spine pedicles as points 1 and 2. The region inside the spine contour between the point 1 or 2 with the adjacent lateral facet concave vertex could be inferred as the spine pedicles denoted with dotted line in Fig.12.

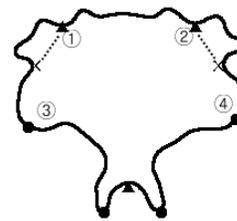


Fig.12. The typical cervical spine shape and \blacktriangle denotes concave vertices and \bullet denotes convex vertices. \times denote the concave vertices near foramen.

IV. EXPERIMENT AND RESULT

All experiments were carried out on 250 cervical CT slices without distinction of sex. The CT slices were acquired with a pixel size 0.78 mm and with a thickness of 3.0 mm. Each image contains various pathologies at the cervical region and treatment, e.g., radiation therapy or biopsy was needed. The CT images for experiments were chosen from the database of Cathay General Hospital.

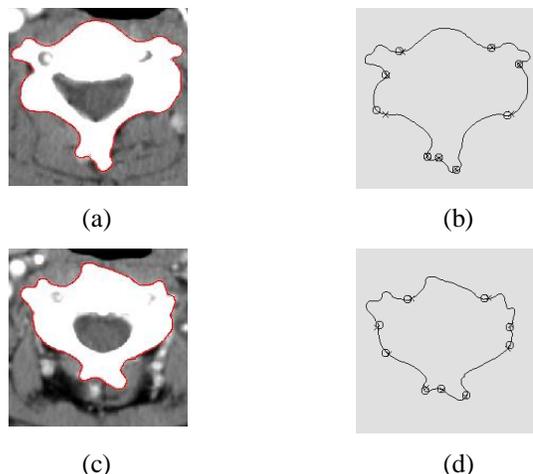


Fig.13. (a)(c) Results of automatically contouring. (b)(d). Results of figuring out the dominant points of (a) and (c). \circ denotes the dominant points derived from our algorithm and \times denotes the dominant points pointed out by the experts.

We evaluated the proposed overall framework with two criteria: success rate and accuracy. The success rate was defined as the relative number of dominant points that localized at acceptable positions, and the accuracy was defined by measuring the distance of the points derived from the algorithm with the dominant points drawn by the clinical experts. The accuracy Acc is calculated as follows:

$$Acc = \frac{1}{M} \sum_{i=1}^M \|S_i - R_i\| \quad (22)$$

where M is the number of points recognized as success. S_i is the pixel of the i th point recognized as success derived from our algorithm and R_i is the closest reference point drawn by clinical experts.

The average success rate was 93.4%, and the average success rate per vertebra was within the range 70%-100%. The coarse scale of turning angle d_1 could affect the average success rate from 81.7% to 93.4% in our experiment. The reason is that corners with less curvature are blurred at higher scales such that the detailed information is suppressed. The average accuracy is 0.37 mm while the fine scale of turning angle d_2 is 15.

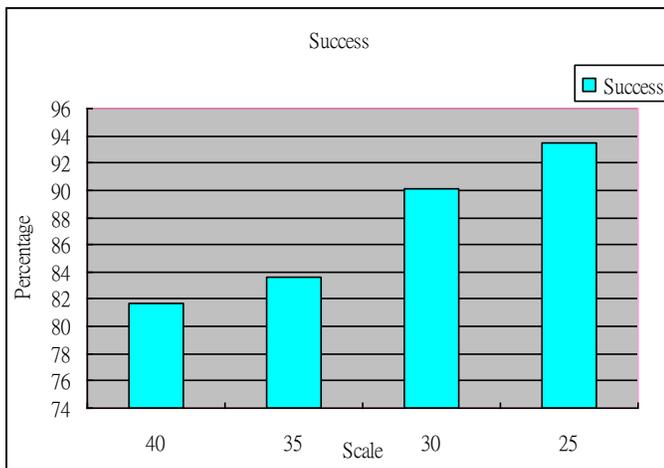


Fig. 14. The success rate with different coarse scale d_1 of 40, 35, 30 and 25.

V. DISCUSSION

The experiment results show that the success rate and accuracy are affected by the scale chosen for extract the dominant vertices. The scale relies on the resolution of the CT images, and if the resolution gets higher, the analyzing scale can also get larger. The prominent segments of vertebrae could be adopted as the landmarks for modeling, registration and clinical evaluation. The method we proposed for automatically extract the spine contour can effectively sketch the main contour of the vertebrae without manually setting the initial contour. Unsatisfied results are mostly caused by wrongly

contouring, and might occur at the bright blobs connected with the spine without obvious edges. Fortunately, scale-based segmentation could overcome partial deformation. Furthermore, in continuous CT slices, the derived contour can also be the initial contour of the adjacent slices and the distance field for iteration can also be deduced. This is also an extended application of the proposed method in this paper to 3D scene. Turning angle is the main idea for geometric analysis in our study, and dominant points can be recognized by finding the local extreme points of turning angle profile. The vertices with low curvature, such as the facet corners and vertebral body vertices have lower accuracy and success rate than the vertices with sharper angle like spinous processes. The corresponding results derived in adjacent slices could be collected for adjusting, so the performance can be improved. For the purpose of modeling, the dominant points can be applied for building the model mentioned in [14] and [15], and the polygonal approximation can also be deduced. Points 1 and 2 can be used for segmenting the vertebral body, and points 3 and 4 can be used for determine the facet corners. Besides, the dominant points on spinous processes and bifurcation are important landmarks on C7 spine, and can be accurately figured out in this paper. In [1], the accuracy defined by MDCP was 0.14mm with pixel size in-plane ranging from 0.233 to 0.309mm. Comparing with points 3 and 4, which are also landmarks for screw insertion operation, our algorithm performed MDCP accuracy in 0.35mm with pixel size in-plane of 0.78mm. We believe that the detected dominant points are capable for operation assist and the accuracy can also be improved if the image resolution can be higher.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we propose a method for automatically extracting the shape features of cervical vertebrae in CT images. With shape analysis, dominant points of the extracted contour can be figured out. The major contribution of the work is that the proposed method can automatically segment the dominant feature points of shape or landmarks used for operation guiding, therapy planning and model registering. Many proposed models of vertebrae can also be built with more precision by implemented the proposed framework. The framework can also be applied in other the thoracic vertebrae and lumbar vertebrae with adjusting the shape model based on anatomical knowledge. Future works include not only extending the proposed method to other spines, but also building an interactive system for aiding surgery and treatment planning.

REFERENCES

- [1] J. Lee, S. Kim, Y. S. Kim and W. K. Chung, "Automated segmentation of the lumbar pedicle in CT images for spinal fusion surgery", IEEE Trans. Biomed. Eng., 58, pp.2051-63, 2011.
- [2] A.S. Reddy, D. Dinobile, J.E. Orgeta and N. Peri, "Transoral approach to CT-guided C2 interventions", Pain Physician, vol. 12, pp.253-258, 2009.
- [3] D.H. Lee, S.W. Lee, S.J. Kang, C.J. Hwang, N.H. Kim, J.Y. Bae, Y.T. Kim, C.S. Lee and K. Daniel Riew, "Optimal entry points and trajectories for cervical pedicle screw placement into subaxial cervical vertebrae", Eur. Spine J., vol. 20, pp.905-911, 2011.
- [4] W. Peh, "CT-guided percutaneous biopsy of spinal lesions", Biomed. Imaging Inter. J. vol.2, e25, 2006.

- [5] Hanaoka S, Fritscher K, Schuler B, Masutani Y, Hayashi N, Ohtomo K and Schubert R, "Whole vertebral bone segmentation method with a statistical intensity-shape model based approach", *Medical Imaging: Proceedings of the SPIE*, vol.7962, pp.796242-796242-14, 2011
- [6] J. Yao, "Automated spinal column extraction and partitioning", *ISIB*, pp. 390-393, 2006.
- [7] S. Ludwig, D. Kramer, R. Balderston, A. Vaccaro, K. Foley and T. Albert, "Placement of pedicle screws in the human cadaveric cervical spine, Comparative accuracy of three techniques", *Spine*, vol. 25, pp.1655-1667, 2000.
- [8] B. Roeschies and S. Winter, "Feature Processing for Automatic Anatomical Landmark Detection Using Reservoir Networks" In: *Proceedings of Bildverarbeitung für die Medizin (BVM)*, pp.277-281,2009.
- [9] B. Roeschies and S. Winter, "Detection of Vertebrae in CT slices by Bunch Graph Matching", In: *Proceedings of the European Congress for Medical and Biomedical Engineering*, vol.22, pp.2575-2578, 2008.
- [10] C. Teng, L.G. Shapiro and I. Kalet, "Automatic segmentation of neck CT images", In: *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, Salt Lake City, Utah 2006, pp.442-445.
- [11] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in CT images", *Medical Image Analysis*, vol.13, pp.471-482, 2009.
- [12] J. Ma, L. Lu, Y. Zhan, X. Zhou, M. Salganicoff and A. Krishnan, "Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model", *Med. Image Comput. Comput. Assist. Interv.*, vol.13, pp.19-27, 2010.
- [13] B.V. Ginneken, A.F. Frangi, J.J. Staal, B.M. ter Haar Romeny and M.A. Viergever, "Active shape model segmentation with optimal features", *IEEE. Trans. Med. Imaging.*, vol.21, pp.924-933, 2002.
- [14] L.R. Long, S. Antani, D.J. Lee, D.M. Krainak and G. Thoma "Biomedical information from a national collection of spine X-rays: film to content-based retrieval", *SPIE Medical Imaging 2003: PACS and Integration Medical Information Systems*, vol. 5003, pp.70-84, 2003.
- [15] X. Han, M.S. Hoogeman, P.C. Levendag, L.S. Hibbard, D.N. Teguh, P. Voet, A.C. Cowen and T.K. Wolf "Atlas-based auto-segmentation of head and neck CT images", *Med. Image Comput. Comput. Assist. Interv.*, vol.11, pp.434-41,2008.
- [16] V. Caselles, R. Kimmel and G. Sapiro, "Geodesic Active Contours", *International Journal of Computer Vision*, vol.22, pp.61-79, 1997.
- [17] V. Caselles, F. Catte and Coll, T and F. Dibos, "A geometric model for active contours", *Numerische Mathematik*, vol.66, pp.1-31, 1993.
- [18] J. Sethian, "A Fast Marching Level Set Method for Monotonically Advancing Fronts," *Proc. Nat'l Academy of Science*, vol. 93, pp. 1591-1694, 1996.
- [19] N. Otsu "A threshold selection method from gray-level histograms", *IEEE Trans. Sys., Man., Cyber.*, vol. 9, 62-66, 1979.
- [20] W.Y. Wu, "An adaptive method for detecting dominant points", *Pattern Recognition*, vol. 36, pp. 2231-2237, 2003.
- [21] Rosenfeld and E. Johnston, "Angle detection on digital curves", *IEEE Trans. Compu.*, vol. 22, pp.875-878, 1973.
- [22] W.Y. Wu and M.J. Wang, "Detecting the dominant points by the curvature-based polygonal approximation", *CVGIP: Graphical Model Image Process*, vol. 55, pp.79-88, 1993.
- [23] C. Ming, P. Wonka, A. Razdan and J. Hu, "A New Image Registration Scheme Based on Curvature Scale Space Curve Matching", *The Visual Computer*, vol. 23, pp. 607-618, 2007.
- [24] F. Mokhtarian and A. Mackworth, "Scale-based description and recognition of planar curves and two-dimensional shapes", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.8, pp. 34-43, 1986.
- [25] S. Abbasi, F. Mokhtarian and J. Kittler, "Curvature scale space image in shape similarity retrieval", *Multimedia Systems*, vol.7, pp.467-476, 1999.
- [26] F. Mokhtarian, "Curvature Scale Space Representation for Robust, Silhouette-Based Object Recognition with Occlusion", *IEEE Trans. on Pattern Analysis and Machine Intelligence - PAMI*, vol. 17, no. 5, pp. 539-544, 1995
- [27] "Technical Summary of Turning Angle Shape Descriptors", Proposed by IBM, ISO/IEC JTC1/SC29/WG11/P162, 1999.
- [28] H.Nishida, "Structural feature indexing for retrieval of partially visible shapes", *Pattern Recognition*, vol.35, pp.55-67, 2002.
- [29] C. Zibreira and F. Pereira, "A Study of Similarity Measures for a Turning Angles-based Shape Descriptor", *Instituto de Telecomunicações, Figueira da Foz, Portugal*, April 2001.
- [30] X. Xu, D.J. Lee, S. Antani and L. R. Long, "A Spine X-Ray Image Retrieval System Using Partial Shape Matching", *IEEE Trans. on Information Technology in Biomedicine*, vol.12, pp.100-108, 2008.

Evolutionary approach to optimisation of the operation of electric power distribution networks

Jan Stępień

University of Technology Kielce
Electrical Engineering Automatics and Computer Science
Faculty Kielce, Poland

Sylwester Filipiak

University of Technology Kielce
Electrical Engineering Automatics and Computer Science
Faculty Kielce, Poland

Abstract—An idea of using a classifying system and co-evolutionary algorithm for operation support of electric power distribution systems operators has been presented in the paper. The method proposed by the author of the work is typified by the short time of designating the most rational post breakdown configurations in complex electric power Medium Voltage distribution network structures. It is the use by the classifying system working with the co-evolution algorithm that enables the effective creation of substitute scenarios for the Medium Voltage electric power distribution network. The method drawn up may be used in current systems managing the work of distribution networks to assist network operators in taking decisions concerning connection actions in supervised electric power systems.

Keywords—evolutionary algorithms; distribution power networks; electric breakdown

I. INTRODUCTION

Distribution networks faults can cause power failure for many users, what can be a reason of major economical losses. In the literature on the distribution systems operation the problem of power delivery recovery in case of the network failure is one of the very important aspects of a proper operation of the distribution systems. Planning a restoration service for distribution systems is a critical task for dispatchers in a control center. Restoration attempts to supply an ample amount of power to nonfaulty out-of-service areas for as many customers as possible while safely operating the distribution system. Reconfiguration is the process of changing the open/closed status of switches and is done for volt/var support, loss reduction, load balancing and restoration. Reconfiguration for restoration is a combinatorial problem involving searching an enormous space of solutions. The problems with integer variables are NP hard, meaning no known algorithm exists to solve these problems in polynomial time. However, reconfiguration for restoration problem is both NP hard and NP and hence belongs to the class of NP complete problems. For such kind of problems, the solution time increases with an increase in the number of integer variables. However, the solution time generally depends on the formulation.

The aim of the research is to develop a method that uses the classifying system and co-evolutionary algorithm to determine, for the assumed conditions, the most profitable distribution network configuration. The important feature of the method is the possibility to form the substitute network configuration with the use of information coming from the simulated network

operation states, where the information on reliability parameters of the network or exploitation periods of the network elements can be also exploited.

II. METHODS USED

First, Many approaches have been proposed to solve the restoration problem from different perspectives. For instance, researchers [1, 2] incorporated dispatcher's experience and operating rules into an expert system to assist the dispatcher. Related investigations formulated the restoration problem as an optimization problem to minimize the number of unserved customers [3, 4]. This problem has been approached using heuristics [5, 6, 7] mathematical programming [8], meta-heuristics (genetic algorithms, tabu search, simulated annealing) [9, 10] and expert systems [11].

In works [12, 13, 14] are presented methods concerning the use revolution algorithms drawn up to resolve multi-criteria problems in optimising electric power networks. These methods concern the development of specialised means of coding, reproduction methods based on domination and also use of co-evolutionary approaches. Several evolutionary algorithms have been developed to deal with distribution system reconfiguration problems [15, 16, 17, 18]. Although the obtained results have been encouraging, the majority of evolutionary algorithms still demand high running time when applied to large-scale distribution systems. In [14], it was shown that the tree encoding (data structure) used is a critical factor for the performance of evolutionary algorithms applied to such large distribution systems. Other critical aspects of distribution systems are the genetic operators that are implemented. Generally these operators do not generate radial configurations [19]. In order to improve the performance obtained by evolutionary algorithms in distribution system reconfiguration problems, a tree encoding based on graph chains, called graph chains representation, and its corresponding genetic operators were developed in [14]. These operators produce only radial configurations. Although the requirement of a radial configuration is common for distribution system reconfiguration problems, it makes the network modeling more difficult to efficiently reconfigure distribution systems.

In the article the author presented the results of his works concerning the method drawn up using the system classifying cooperation with the co-evolutionary algorithm, in order to assist the work of electrical energy distribution systems operators. The elaborated method uses the classifying system to

determine, for the assumed conditions, the most profitable distribution network configuration. The important feature of the method is the possibility to form the substitute network configuration with the use of information coming from the simulated network operation states, where the information on reliability parameters of the network or exploitation periods of the network elements can be also exploited.

Cooperation of the co-evolutionary algorithm with the classification system (drawn up by the author of the work) enables significant reduction of the classification time (reduces the iterative calculation process on average by 40 %), which is significant from the practical point of view in the application of this method in current systems of distribution network operation management. The application of a classification system to the analysed task also enables improvement of the effectiveness of the performance process of designating the scenario of the substitute network configurations. Improvement of the efficiency of the network configuration designation process is obtained using the sought information (with use of the announcement creation process), in the collections of classifiers to create sub-populations of solutions for the co-evolutionary algorithm, which would be used to search for the collection of Pareto-optimal solutions. The process of creating a collection of classifiers describing the substitute network configuration was performed by the author supported by the theoretical genetic basics of self-teaching system. Classifiers may be created (for analysed network structure) for the most probable break down situations, which arise from regarding the stage of choice of the simulated break down situations (in the analysed network) reliability characteristics and the usage durations of network elements. The result of the works performed is the drawing up of an effective method enabling the rapid designation of substitute network configurations, also for very complex network structures. The method may be used in information systems for the current operation management of electric power distribution networks..

III. A METHOD USING THE CLASSIFIER SYSTEM

A. Evaluating system in the elaborated method

The classifier system is a system that learns the syntactic simple rules in order to co-ordinate its actions in any environment and includes the three basic components [20, 21]: rule and message system, evaluating system, evolutionary algorithm. In the classifier system the information from the outer environment is processed into the messages of a given format. The messages are further placed on the message list, where they can activate the classifiers.

In the elaborated method (based on the classifier system idea) known procedures, performing message processing or classifier evaluation have been used. Certain modifications resulting from the specificity of the considered problem have been introduced:

- the message about the fault is described in the form: a list with numbers of not supplied nodes, and a list with numbers of fault elements :

$\langle message \rangle ::= (numbers\ of\ not\ supplied\ nodes) + (numbers\ of\ fault\ elements)$

- in the classifier notation following syntax has been taken into account in the notation actually used:

$\langle classifier \rangle ::= \langle condition \rangle : \langle message \rangle$
 $\langle classifier \rangle ::= \langle numbers\ of\ not\ supplied\ nodes + numbers\ of\ fault\ elements \rangle : \langle post-fault\ configuration \rangle$

With regard to the specific character of the analysed task (concerning breakdown of elements in the network structure) the author has drawn up a modified announcement processing procedure (describing network break down situations). In the suggested method the announcement processing process and the evaluation of classifiers is divided into two stages described below.

Step 1 consists of the search in the collections of classifiers for such, for which the conditions are compatible with the announcement describing the existing network breakdown. Comparison of the announcement (containing information about damaged network elements and of network nodes deprived of current supply) with the conditions of classifiers enables the search and activation of classifiers containing coded information about network configurations, in which there are no damaged elements. Conformity of the announcement with the conditions of the classifiers in the first stage of the suggested method is defined on the basis of comparison of the numbers of network nodes without power supply recorded in the announcement, with the numbers of network nodes recorded in the first part of the classifiers (which corresponded to the concealed zeros on the appropriate positions of communication code tracts and classifier). After searching for classifiers conforming to the announcement the evaluation takes place of the so-called offer of these classifiers. The classifier distinguished by the highest offer was next used as a following announcement.

Step 2 concerns the search for classifiers whose conditions will be according to the announcement of the classifiers designated in stage 1. The conformity of the announcement with the conditions of the classifiers in this stage was defined on the basis of the differences between the power supply routes of the chosen line sections (from the list of network nodes deprived of power supply) with the configuration recorded in the announcement and classifiers.

With regard for these specific natures of the analysed task the author suggested a two-part description of the announcement (describing the breakdown situation of the network). The first part of the announcement is recorded as a length of zero-ones relating to the number of elements equal to the number of network line sections of the analysed network. Value 1 on the defined position corresponding to the number of the network length, indicates length with power supply, and 0 indicates network node without power supply as a result of breakdown. The second part of the announcement contains information about the damaged elements and also information about the configuration of network elements. For the description of this part of the announcement the author introduced the following marking notation: 0 - means damaged element, 1 - means actually used element, # - means element remaining in reserve. Below is showing an example of the process of creating announcements for the breakdown status of

the electric power network system (composed of a small number of elements), the structure of which is reflected in graphic form on drawing 1. For the network graph from drawing 1, the case is examined of a breakdown on line 14.

The performance of the process of creation of announcements enables search in the collection of classifiers for information, the use of which assists the process implemented by the co-evolutionary algorithm.

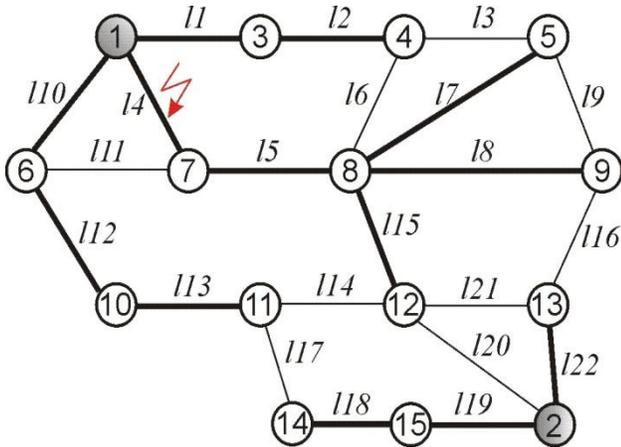


Fig. 1 Graph of the analysed distribution network

The announcement describing the considered breakdown status was described as follows:

message 1

111101000110111 | 11#0#####1#11####11##1

The sought for classifiers in the first stage (fig.2):

classifier 1

111101100110111 | 11110####11#1111##11##1

classifier 2

111101111111111 | 11111#0#1#1#1#1##11##1

In the initial part of calculations a message on the fault occurring in the network is being read. Procedures verifying the matching between the classifiers and the generated message are performed subsequently and then the classifiers are assessed.

The strength S of the classifier, which has shown the best bid in the so-called auction process, is increased by the reward given by the system. Simultaneously its strength is decreased by the value of the bid given by the classifier. The rule and message procedures perform the process of classifiers checking and evaluation, in aspect of using the information contained in them for solving the problematic situations. This allows for appointing of the group of classifiers containing the useful information on the searched post-fault network configuration.

The bid of the best classifier increases the strength of other active classifiers proportionally to their bids. Moreover, the strength of all the active classifiers is decreased by a certain, determined value. The effective bid value has been calculated in a following way [20, 21]:

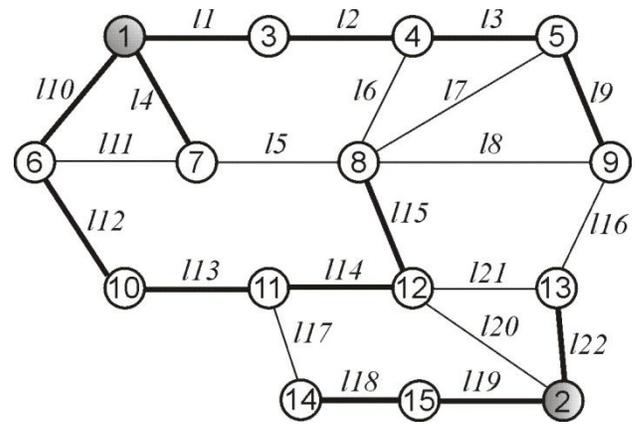


Fig. 2 Graph of the analysed distribution network for classifier 1

$$S_i(t+1) = S_i(t) - c_{bid} \cdot S_i(t) - c_{tax} \cdot S_i(t) + r(t) \quad (1)$$

$$B_i = c_{bid} \cdot (e_{bid1} + e_{bid2} \cdot Sp_i) \cdot S_i \quad (2)$$

$$EB_i = c_{bid} \cdot (e_{bid1} + e_{bid2} \cdot Sp_i) \cdot S_i + e_{br} \quad (3)$$

where: B_i - bid value of the i -th classifier, EB_i - effective bid value of the i -th classifier, Sp_i - specificity of the i -th classifier, S_i - strength of the i -th classifier, c_{bid} - investment coefficient ($c_{bid}=0,1$), e_{bid1} , e_{bid2} - coefficients of the classifier linear specificity function ($e_{bid1}=0,65$, $e_{bid2}=0,35$), e_{br} - random value generated with the use of a normal distribution generator, c_{tax} - turnover tax coefficient $c_{tax}=001$, r - coefficient of reward paid for the best classifier $r=2$.

B. Co-evolutionary algorithm

To modifications of the evolutionary algorithm enabling solutions of multi-criteria tasks are counted among others the application of the co-evolutionary approach. Application of the co-evolutionary algorithm to the analysed task creates m population; in each of them the adaptation function is defined on the basis of another component quality indicator vector. After successive performance (population supplementation with new elements), and through renewed reproduction, these populations are connected, and then were again divided so that each population elements may attain an unlimited population. The sought-after solution is the Pareto-optimal collection of solutions.

To encode the individuals representing various network configuration variants in a form of a sparse graph, the bequest of chromosomes in the form of a vector of inversion has been assumed. Each component of the vector of inversion, corresponding to the number of the graph node, is equal to the number of the supplying node. A well-known roulette selection method on the remaining fractional part has been used as a selection method. Two specialised reconfiguration operators have been used in the algorithm to create new solutions (crossover probability $pk=0,95$, mutation probability $pm=0,15$). In the presented calculation method creating of new variants of the analysed problem solutions has been realised according to the following procedure:

- 1) Selection of two network configuration variants from the current population (recorded in the vectors of inversion),

- 2) Node selection from the list of nodes with no supply,
- 3) Rewriting of the supply routes of the formerly selected node from the vector selected in step 1 to the auxiliary table,
- 4) Roulette selection of the node from the created table,
- 5) Rewriting of the further part of the supply route from the second vector, starting from the node selected in step 4, to the second of the selected vectors.

The aim of using of this kind of operator, creating new variants of distribution network configuration, was to examine the change variants effectiveness in the part of the networks close to the supply points, as well as in parts of the analysed network system affected by failures.

In order to obtain proper solutions following limiting constraints resulting from technical requirements for proper operation of the distribution network have been taken into account:

- Not exceeding of the maximum transmission currents of the line sections,
- Not exceeding of the allowable voltage drops in the network nodes supply routes,

On the base of the source data [18, 20] and own research following values of significant parameters of the calculation system have been assumed in the calculation procedures: number of classifiers $n=200$, crowding factor for classifier population $cs=3$.

C. Assumed optimization criteria

Following criteria have been assumed substantial for the optimisation problem of post-fault network configuration:

- Minimisation of the number of switching activities leading to obtaining a substitute network configuration:

$$\text{Min}_j u_1(X_j) = n_j - n_0 \quad \text{Where } j = 1, 2, \dots, m \quad (4)$$

where: X_j – vector containing information on the j -th variant of the distribution network configuration, m – number of solution variants, n_j – number of switching activities, n_0 – number of switching activities in the basic configuration.

- Maximal level of reliability of supply of electric power to recipients:

$$\text{Min}_j u_2(X_j) = \min\{\max(1 - p_{ik})\} \quad (5)$$

Where: p_{ik} – coefficient of reliability of supply track i -th w of recipient node designated for k -th of year of cable exploitation,

- Minimisation of the voltage deviation in the network nodes:

$$\text{Min}_j u_3(X_j) = \max_i \left(\frac{U_i}{U_N} \cdot 100 \right) \quad (6)$$

Where: U_N – distribution network nominal voltage, U_i – voltage value in the i -th user node of the network,

- Minimisation of the power load degree coefficient of the found group of the most loaded network elements.

$$\text{Min}_j u_4(X_j) = \max_k \frac{\sum_{i=1}^n P_{\max, i}}{n} \quad (7)$$

Where: k – the number of power supply route network nodes of the reception network, n – the number of the most heavily loaded network elements.

- Minimisation of the technical losses in the distribution systems:

$$\text{Min}_j u_5(X_j) = \min\left\{ \sum_{i=1}^g (\Delta P_i + k_e \cdot \Delta Q_i) \right\} \quad (8)$$

Where: g – number of sections underload in given network configuration variant, ΔP_i – value of loss of power in i -th network section, k_e – electric power equivalent of passive power.

The assumed membership functions used for the main variables description have been defined as follows:

$$u_{f_i}(X) = \begin{cases} 1, & \text{if } f_i(X) \leq f_i^{\min} \\ \left(\frac{f_i^{\max} - f_i(X)}{f_i^{\max} - f_i^{\min}} \right), & \text{if } f_i^{\min} < f_i(X) \leq f_i^{\max} \\ 0, & \text{if } f_i^{\max} < f_i(X) \end{cases} \quad (9)$$

IV. CASE STUDIES

The pre-analysed calculation problem concerns the designation of the supply network configuration for the breakdown operations statuses of the network, arising from damaged network elements, their loading and also the exceeding of permissible voltage deviations in network line sections. Considered breakdown status busbars main supply station number 6. The consequence of this breakdown is lack of power supply for a significant part of the network nodes. The sought-after solution is a substitute network configuration enabling the restoration of power supplies to as great a number of recipient network nodes as possible. The considered task is a multi-criteria optimisation task. In the presented graph (fig. 3) the filled nodes symbolise the main supplying points, whereas the bold branches symbolise the elements taking part in the load transfer.

For the below considered breakdown situation in the analysed network, which is composed of 556 network nodes the author accepted the abbreviated description of announcements and also classifiers. The abbreviated description however contains instead of the zero-one tract (part of the first announcement) the numbers of line sections deprived of power, whereas as part of the second announcement the numbers of damaged elements are given. In the elaborated calculation model a so-called vector of inversion has been used for the network configuration description. As a result of the accomplishment of the first stage of the process the announcement creation for the searched for classifier is

shown in table 1. In the column relating to network configurations noted in the inversion vector only the initial and final elements of this vector are noted.

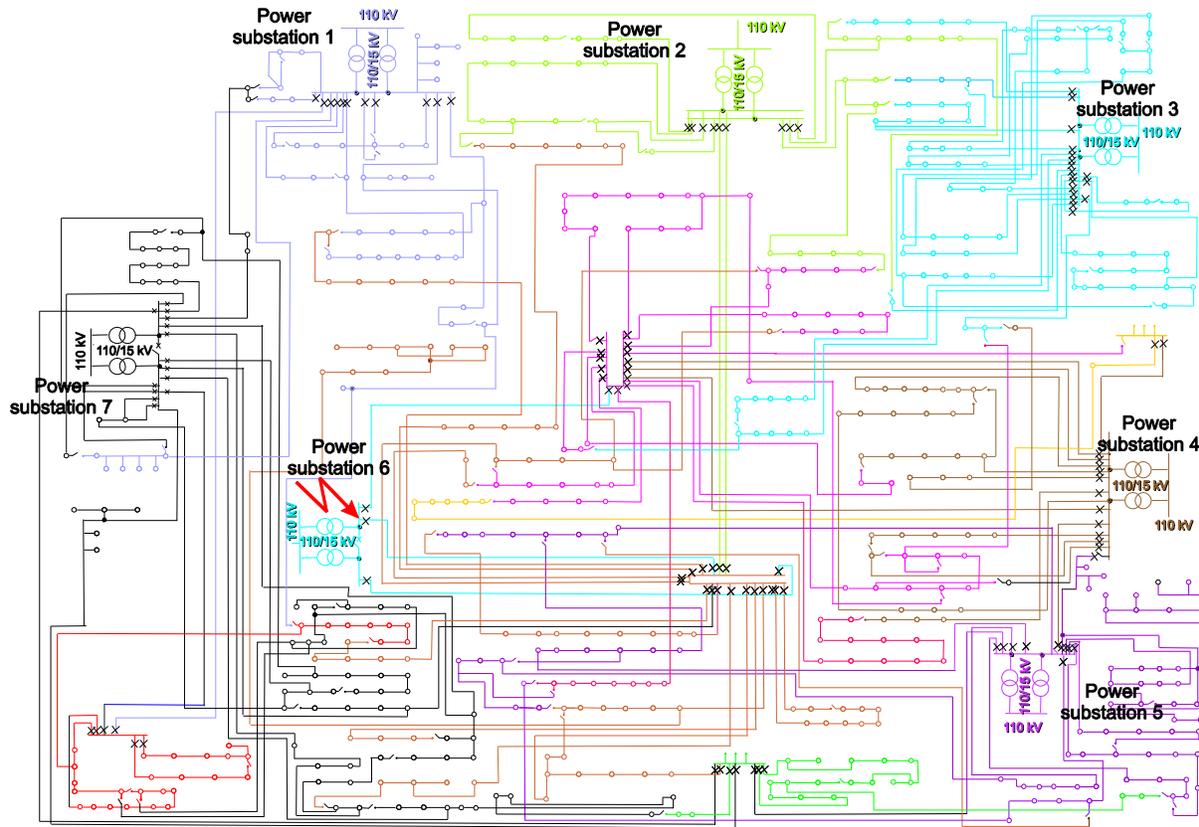


Fig. 3 Diagram of the analysed system of the medium voltage distribution network

TABLE I. ACTIVE CLASSIFIERS

No	condition: (numbers of non-supplied nodes) and (numbers of fault elements)	answer of system (recorded in the vectors of inversion)	S_i – strength of the i -th classifier	Bid	S_i – strength of the i -th classifier
1.	(8,22,28,29,31,73,178,195,291...,188) and (6_8)	x,x,x,x,x,x,x,x, 4,6,5,1,13,14, ... 554,7	$S_1 = 10$	$B_1 = 0,813$ $EB_1 = 0,802$	$S_1 = 10 + 0,622$
2.	(9,16,20,21,57,80,81,101,117, ..., 363) and (6_9)	x,x,x,x,x,x,x,x, 6,2,5,7,13,14, ... 554,7	$S_2 = 10$	$B_2 = 0,823$ $EB_2 = 0,831$	$S_2 = 10 + 2$
3.	(29, 73, 72, 71, 70, 69, 68) and (8_29)	x,x,x,x,x,x,x,x, 6,6,5,1,13,14, ... 554,7	$S_2 = 10$	$B_2 = 0,639$ $EB_2 = 0,637$	$S_3 = 10 + 0,211$

According to the idea of classifying systems through the process of announcement creation, then follows the evaluation of the revealed classifiers, which consists of the calculation of the so-called offer of the classifiers being the measure of their suitability to resolve the analysed task. As a result of the performance of the process of the creation and evaluation of announcements executed in the first stage, as

the classifier with the best offer is defined as classifier number 2. This classifier served to create the announcement of the following stage of classifiers search.

TABLE II. ACTIVE CLASSIFIERS AFTER PROCESS EVALUATING

No	condition: (numbers of non-supplied nodes) and (numbers of fault elements)	answer of system (recorded in the vectors of inversion)	S_i – strength of the i -th classifier	Bid	S_i – strength of the i -th classifier
1.	(99,16,20,21,57,80,81,101,117, ..., 363) and (2_9)	x,x,x,x,x,x,x,6,2, 5,7,13,14,15,16,9 ... 554,7	$S_1 = 10$	$B_1 = 0,941$ $EB_1 = 0,975$	$S_1 = 10 + 2$
2.	(10,413,414,419,432,412,415, ..., 438) and (7_10)	x,x,x,x,x,x,x,6,2, 5,7,13,14,15,16,9 ... 554,7	$S_2 = 10$	$B_2 = 0,712$ $EB_2 = 0,732$	$S_2 = 10 + 0,975$

As a result of the performance of the announcement creation process and the evaluation of active classifiers, information was obtained, which might be used to create a population (size 100) of solution variants subsequently created by the co-evolutionary algorithm. This algorithm is based simultaneously on 5 subpopulations, from which each evaluation was the basis for another adaptation function

(dependencies 4 to 8). The sought-after solution in this case is a collection of solutions in the form of alternative configurations of the analysed network. The course of the process designating the best solutions in subpopulation 3 and 4 is shown in drawings 4 and 5. Figures 4b and 5b show the realisation of the calculation process, which used information for active classifiers.

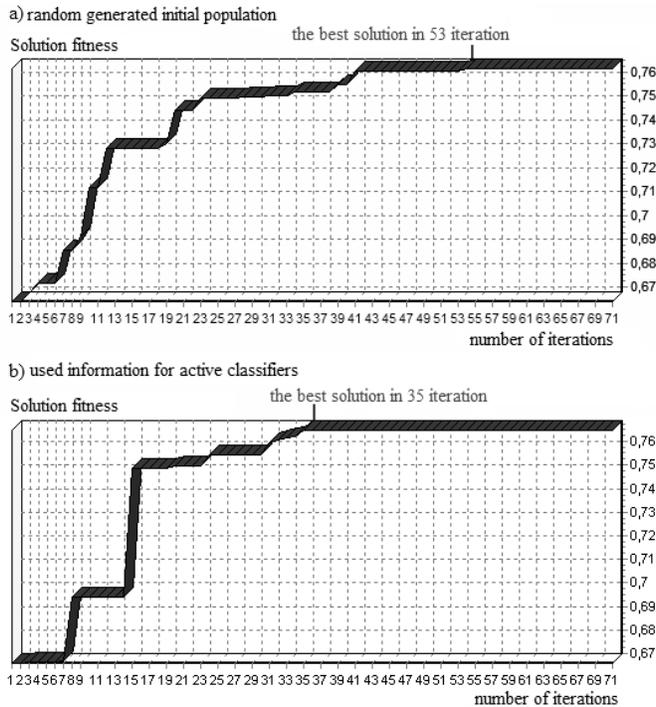


Fig. 4 Example of best solutions fitness waveform in subpopulation number 4

Cooperation of the co-evolutionary algorithm with the classification system enables significant reduction of time of obtainment of solutions (reduces the iterative calculation process on average by 40 %), which is significant from the practical point of view in the application of this method in current systems of distribution network operation management.

Information on the best solutions in subpopulations no. 3 is shown in graphic form on drawing 6. As a solution to the task of designating a substitute network configuration in the event of a breakdown of the analysed distribution network, obtained with the use of co-evolution algorithm the best solution variants is accepted from 5 subpopulations. Information concerning the best obtained solutions is shown in table 3.

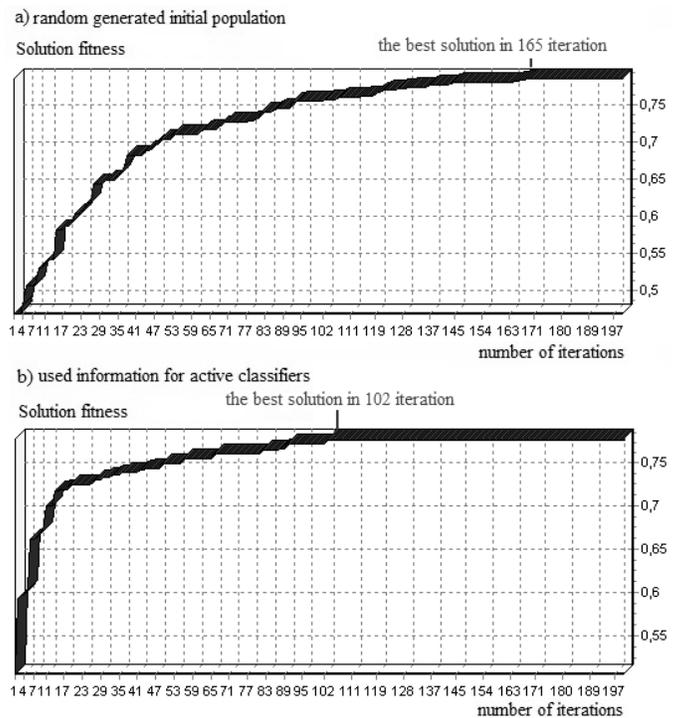


Fig. 5 Example of best solutions fitness waveform in subpopulation number 4

TABLE III. RESULTS ARE SHOWN OF CALCULATIONS FOR SOUGHT-AFTER POST BREAKDOWN CONFIGURATION OF ANALYSED NETWORK

No	Results for subpopulation 1	Results for subpopulation 2	Results for subpopulation 3	Results for subpopulation 4	Results for subpopulation 5
1	$L_{pz} = 4$ $u_1(X) = 0,893$	$L_{pz} = 8$ $u_1(X) = 0,671$	$L_{pz} = 7$ $u_1(X) = 0,707$	$L_{pz} = 10$ $u_1(x) = 0,641$	$L_{pz} = 12$ $u_1(x) = 0,619$
2	$p = 0,998031$ $u_2(x) = 0,896$	$p = 0,998381$ $u_2(x) = 0,998$	$p = 0,998152$ $u_2(x) = 0,943$	$p = 0,997752$ $u_2(x) = 0,790$	$p = 0,998261$ $u_2(x) = 0,985$
3	$\delta U = 2,23 \%$ $u_3(x) = 0,647$	$\delta U = 1,22 \%$ $u_3(x) = 0,752$	$\delta U = 1,18 \%$ $u_3(x) = 0,765$	$\delta U = 1,31 \%$ $u_3(x) = 0,734$	$\delta U = 1,22 \%$ $u_3(x) = 0,752$
4	$\Delta P = 2895 \text{ kW}$ $u_4(x) = 0,635$	$\Delta P = 2679 \text{ kW}$ $u_4(x) = 0,667$	$\Delta P = 2675 \text{ kW}$ $u_4(x) = 0,670$	$\Delta P = 2561 \text{ kW}$ $u_4(x) = 0,740$	$\Delta P = 2654 \text{ kW}$ $u_4(x) = 0,682$
5	$k_{obc} = 0,669$ $u_5(x) = 0,743$	$k_{obc} = 0,584$ $u_5(x) = 0,886$	$k_{obc} = 0,577$ $u_5(x) = 0,898$	$k_{obc} = 0,771$ $u_5(x) = 0,669$	$k_{obc} = 0,544$ $u_5(x) = 0,955$

In cooperation of the co-evolution algorithm with the classifying system after performance of the calculation process the **best** solutions obtained from particular subpopulations the solutions are written into the classifier collection. The choice of the final solution variant depends upon the decision maker decider, who in this instance may be the operator managing the operation of the electric power Medium Voltage distribution network.

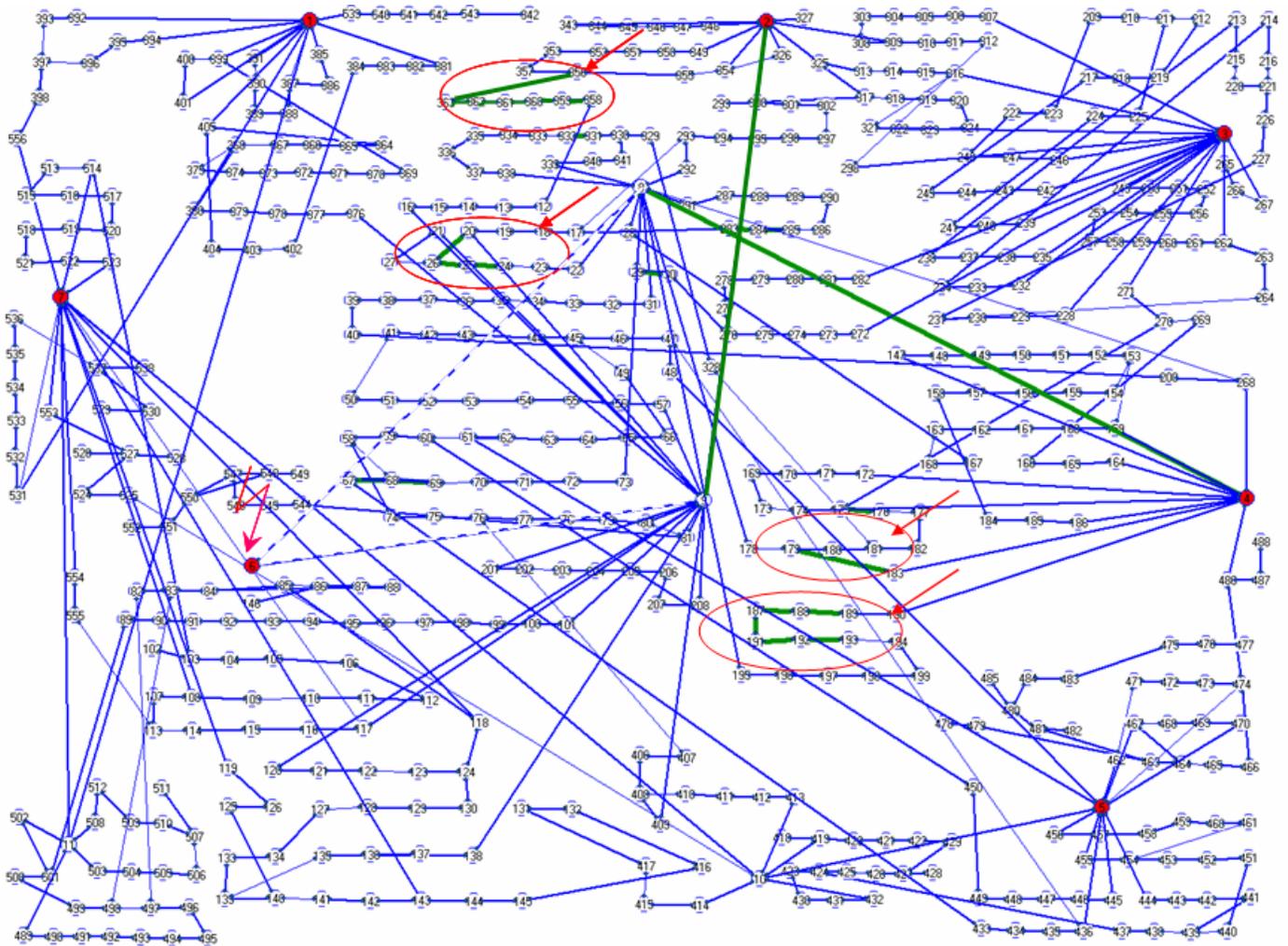


Fig. 6 Graph of the analysed distribution network with network configurations, being the best solution obtained in subpopulation 3

V. CONCLUSION

This article describes the development of this type of calculation methods, simultaneously containing their own innovative solution proposals concerning the application of a classification system working with the co-evolutionary algorithm. The calculations performed for the mapped real system of the medium voltage municipal distribution network of 556 nodes have given satisfactory results, confirming the adequate direction of the research. On the base of the results obtained so far the authors assume that the results can be further used in creation of decisive procedures for complex power electric systems management, taking the fault operation states into special consideration.

The method proposed by the author of the work is typified by the short time of designating the most rational post breakdown configurations in complex electric power Medium Voltage distribution network structures.

It is the use by the classifying system working with the co-evolution algorithm that enables the effective creation of substitute scenarios for the Medium Voltage electric power distribution network. The method drawn up may be used in

current systems managing the work of distribution networks to assist network operators in taking decisions concerning connection actions in supervised electric power systems.

VI. FUTURE SCOPE

The proposed method fulfils the stated assumptions. It is typified by a short time of designation of substitute post breakdown configurations of configurations of complex medium voltage electric power distribution network structures. The drawn up method makes use of the classifying system cooperating with the co-evolutionary algorithm, which enables effective creation of substitute medium voltage distribution network configurations, for various network breakdown statuses.

It is thus a method based on systems that teach themselves. The method may be used in current systems and managing work on distribution networks to assist operators in making decisions concerning connection actions in supervised electric power systems.

The results obtained within the research performed, in the form of drawn up procedures to create the most effective configuration of the work of network appliances, may be used

as elements of large very complex information systems used in intelligent electric power distribution networks. Application of integrated information systems assists distribution network management to optimise decision processes of system operator and in effect achieve the required level of services, by the minimisation of expenditure on network maintenance and appliances in a condition guaranteeing the supply of electric power of the required parameters.

REFERENCES

- [1] C. C. Liu, S. J. Lee, S. S. Venkata, An expert system operational aid for restoration and loss reduction of distribution system. *IEEE Trans. on Power Delivery*, vol. 3, 1988, pp. 619-629.
- [2] Y. Y. Hsu, M. Huang, Distribution system service restoration using a heuristic search approach. *IEEE Trans. on Power Delivery*, vol. 7, 1992, pp. 734-740.
- [3] Y. Fukuyama, H. D. Chiang, Parallel genetic algorithm for service restoration in electric power distribution systems. *Electric Power & Energy Systems*, vol. 18, no. 2, 1996, pp. 111-119.
- [4] K. N. Miu, H. D. Chiang, B. Yuan, Fast service restoration for large-scale distribution systems with priority customers and constraints. *IEEE Trans. on Power Systems*, vol. 13, no. 3, Aug. 1998, pp. 789-795.
- [5] L. Morelato, A. J. Monticelli, Heuristic search approach to distribution system restoration. *IEEE Trans. Power Delivery*, vol. 4, Oct. 1989, pp. 2235-2241.
- [6] S. Wu, K. L. Tomsovic, C. S. Chen, A heuristic search approach to feeder switching operations for overload, faults, unbalanced flow and maintenance. *IEEE Trans. Power Delivery*, vol. 6, Oct. 1991, pp. 1579-1586.
- [7] D. Shirmohammadi, Service restoration in distribution networks via network reconfiguration. *IEEE Trans. Power Delivery*, vol. 7, Apr. 1992, pp. 952-958.
- [8] K. L. Butler, N. D. R. Sarma, R. Prasad, Network reconfiguration for service restoration in shipboard power distribution systems. *IEEE Trans. Power Systems*, vol. 16, Nov. 2001, pp. 653-661.
- [9] Y. Hsiao, C. Chien, Enhancement of restoration service in distribution systems using a combination fuzzy-GA method. *IEEE Trans. Power Systems*, vol. 15, Nov. 2000, pp. 1394-1400.
- [10] S. Toune, H. Fudo, T. Genji, Y. Fukuyama, Comparative study of modern heuristic algorithms to service restoration in distribution systems. *IEEE Trans. Power Delivery*, vol. 17, Jan. 2002, pp. 173-181.
- [11] C. Chao-Shun, C-H. Lin, T. Hung-Ying, A rule-based expert system with colored petri net models for distribution system service restoration. *IEEE Trans. Power Systems*, vol. 17, Nov. 2002, pp. 1073-1080.
- [12] S. Khushalani, J.M. Solanki, N.N. Schulz, Optimized Restoration of Unbalanced Distribution Systems. *IEEE Transactions on Power Systems*, no. 22, Issue 2, 2007, p. 624-630.
- [13] Y. Kumar, B. Das, J. Sharma, Multiobjective, Multiconstraint Service Restoration of Electric Power Distribution System With Priority Customers. *IEEE Transactions on Power Delivery*, no. 23, Issue 1, 2008, p. 261-270.
- [14] C. B. Delbem, A. C. P. L. F. Carvalho, N. G. Bretas, Main chain representation for evolutionary algorithms applied to distribution system reconfiguration. *IEEE Trans. Power Systems*, vol. 20, no. 1, Feb. 2005, pp. 425-436.
- [15] K. Nara, A. Shiose, M. Kitagawa, T. Ishihara, Implementation of genetic algorithm for distribution systems loss minimum reconfiguration. *IEEE Trans. Power Systems*, vol. 7, no. 3, Aug. 1992, pp. 1044-1051.
- [16] L. Augugliaro, F. R. Sanseverino, Multiobjective service restoration in distribution networks using an evolutionary approach and fuzzy sets. *Elect. Power Energy Syst.*, vol.22, 2000, pp. 103-110.
- [17] J. Stępień Z. Madej, Evaluation of structural redundancy effects in medium voltage cable networks. *Rynek Energii No 4(83)*, 2009, pp. 55-62
- [18] J. Stępień: Changes in demand structure of energy carriers with the use of waste heat and renewable energy. *Rynek Energii Issue: 5, OCT 2008*, p. 58-62
- [19] J. Stępień, Evaluation of structural redundancy effects in medium voltage cable networks. *Przegląd Elektrotechniczny Volume: 84 Issue: 4 2008*, p. 128-131.

Expected Reliability of Everyday- and Ambient Assisted Living Technologies – Results From an Online Survey

Frederick Steinke
Humboldt-Universität zu Berlin
Berlin, Germany

Andreas Hertzner
Universität Augsburg
Augsburg, Germany

Tobias Fritsch
Universität Heidelberg
Heidelberg, Germany

Helmut Tautz
Technische Hochschule Ingolstadt
Ingolstadt, Germany

Simon Zickwolf
Ludwig-Maximilians-Universität München
Munich, Germany

Abstract—To receive valuable information about expected reliability in everyday technologies compared to Ambient Assisted Living (AAL) technologies, an online survey was conducted including five everyday (train, dishwasher, navigation system, computer, mobile phone) and three AAL (stove, window, floor sensors) technologies. The age range of the 206 participants (109 men; 97 female) was from 14 to 88 years (mean=38.0). The descriptive analysis indicates expected reliabilities of more than 90% for most technologies. Only train punctuality is considered as less reliable with a mean expected reliability of 86%. Furthermore, by using t-tests it can be shown that the three AAL technologies are expected to have a higher reliability than the everyday technologies. Additionally, a sample split at the age of 50 years indicates that elderly participants expect that technologies have a higher reliability than younger participants do. Using these findings, in a next step an experiment with different reliability levels of AAL technologies will be designed. This differentiation will be used to measure the influence of reliability on trust and intention to use in context of Ambient Assisted Living.

Keywords—Ambient Assisted Living; Elderly People; Expected Reliability; Online Survey; Technology

I. INTRODUCTION

Elderly people are an interesting target group for companies to sell their products because of their rising percentage among the worldwide population [1]. Due to the fact, that this target group is often financial strong and has a higher income, there is more money left for consumption [2]. For this case, science deals for several decades with the research of new technologies to support people in their own home [3]. As a result, different concepts have entered the market [4][5][6][7][8]. On this basis, dependencies between different technologies and variables like age are a very important subject for research.

In consequences of the demographic change, the proportion of elderly people, who would like to spend an independent life at home, is increasing. The market for technology-supported systems for the use at home is growing because of the physical effects of older people [9]. Therefore, concepts like Ambient Assisted Living are getting more attractive in the last years. With the aid of sensors and actuators within the framework of an intelligent platform- the time older people can live independently can be extended. By the use of AAL it is possible to use pervasive devices for integrating them into a reliable environment for the elderly. Ambient Intelligence enables automatic services which are dependent on the need of the user and can be seen as essential part of AAL [10]. By means of summarizing and demand-oriented analyzing of sensor data, an individualization of care as well as nursing services is possible [5]. Product designers of such technologies have to consider a lot of different factors in the process of development to design a marketable solution. As one of these factors, the reliability of the technique in general is a crucial point [11][12].

For that reason the present study discusses the expected reliabilities for different technologies (including AAL) by the users. The following brief description should help to gain a better structural overview. The background section includes former studies which demonstrate the importance of knowing the expected reliability of users and brings it in context with AAL. In the next step, a description about the methodology for getting the required information by an online questionnaire and the sample details are presented. Afterwards, the results by means of a descriptive analysis, correlation analysis, t-tests, and analyses of variances are underlined. The correlation analysis shows the dependences between expected reliability of users on different technologies and other variables like age and gender. With the aid of paired-sample t-tests, the expectations about reliabilities of different technologies are shown. An

Omitted Least Squares (OLS) regression illustrates expected reliability by AAL-related and everyday-technologies in connection with other variables (e.g., gender, age).

Finally, a conclusion section describes the participants' expectations between reliability of AAL and everyday technologies. Additionally, recommendations for the designing of AAL technologies are given. For this reason, the following research question is answered: *What differences exist regarding the expected reliability of everyday vs. Ambient Assisted Living Technology?* A brief review about the limitations concerning the sample and test execution as well as further research needs round off the survey.

II. BACKGROUND

Knowledge about the reliability of different technologies expected by the users is highly important as reliability is a crucial component of the technology adaption decision [13]. In order to address potential users more appropriately, a view on their expected reliability on different technologies is the object of this study. According to the Oxford Dictionary, reliability means the "consistently good in quality or performance" and therefore, the ability to be trusted [14]. Thus, it is very important for a product or service to meet the users' expected reliabilities to get their trust for using it.

"Expected", in this context, could be understood as fulfilling the personal requirements of each user about the system-functionality [15]. Consequently, expected reliability can be seen as a pre-condition for building user-trust [16]. Some former studies already described the relationship between reliability of technologies and their consequences in different ways. One example is a test which examined the influence of trust and etiquette in high-criticality automated systems. In this study, user performance was much better when the automation reliability was higher and good automation etiquette also contributed to a better performance [17]. AAL systems could be seen as high-criticality automated systems as well wherefore reliability of the system will be important. In another study [18] groups were divided into younger (age 20-45) and older (60-80) adults. Nevertheless, it was also obvious that both groups will begin to appropriately use the systems if they work in a proper way. Especially older adults are willing to change their behavior when the system was useful and they can trust on their reliability [18].

In order to get the required information from the user, different types of sensors are in the field. Radio-frequency identification (RFID), motion detectors, heat, and pressure sensors for example are in use to send up information to the system for the purpose of doing the right actions at the right time. It is possible to switch off the oven if somebody had forgotten to do that or to do an emergency call after a person slumped on the floor because of a qualm. To realize these life-saving measures it is very important to have a reliable interaction between sensor and actuator [19]. Furthermore, it is already possible to transmit physiological and psychological information about the user. With the help of sensors attached to the user's body and video cameras and microphones it is also manageable to get a pattern of respiration and features of facial expression [20]. A disadvantage for the user due to the physical

and social discomfort by wearing such devices could be reduced by advances in miniaturization of the devices.

However, a study revealed that fixed attachment of sensors in the accommodation was considered to be more reliable than attachment to clothing or on/ in the body. In addition, reliability and ease of use were also assessed as highly important as a basis for trust in AAL technology [21].

To check how different technologies are evaluated, the present study is conducted to analyze differences between everyday used technology and AAL technology.

III. METHODS

To gather information about whether the expected reliability (ER) in a working system differs between technologies and between younger and older people, several scenarios from daily life are considered in an online questionnaire. To recruit participants, emails were sent out to students and to acquaintances of the authors requesting for participation and for forwarding the email (mainly to persons older than the age of 50).

The survey was conducted on a three-week period in January 2013 using the web page "oFb – der onlineFragebogen". The first part of the questionnaire contained 14 questions regarding eight scenarios.

Five scenarios were queried with one question each, dealing with the topics *train punctuality*, *dishwasher functioning*, *navigation system functioning*, *computer functioning*, and *mobile phone functioning*. These scenarios are considered as everyday technologies in the paper at hand. Participants had to decide on how reliable they believed the technologies worked. The reliability scale ranged from "70% or less" to "100%" in steps of 5%. To answer the questions a 7-point Likert scale was used.

Following, three technical assistance scenarios in the context of AAL were examined, queried by three questions each. These scenarios are similar to those queried in [22] and dealing with AAL technologies as well as sensor devices. One scenario considered the possibility to turn off the stove via an application on a tablet computer. A second scenario considered a situation where sensor technology detects a person lying on the floor and automatically calls an ambulance. The third scenario again dealt with actively executing a computer application, but this time the application enables the person to open or close the windows via the Internet. Participants had to evaluate their expected reliability, i.e. how well the technological instruments described would perform. The reliability scale again was set from "70% or less" to "100%".

The order of the scenario-based questions was the following: The different scenarios from everyday- and AAL technologies were queried in random order but the three questions of each AAL scenario were queried together.

Additionally to these reliability related questions, participants had to answer socio-demographic questions about their age, gender, and living condition as well as whether they possessed a smartphone, and about their computer, and tablet computer experience.

A Sample

In the three-week period, 251 persons started the survey and 206 persons finished it. The following analysis will only consider individuals that finished the survey and will refer to them as “participants.” 52.9 percent of the participants were men and 47.1 percent were women. The average age was 38.0 years (SD=17.0) and the median age was 29. The exact distribution in eight age-categories is shown in figure 1.

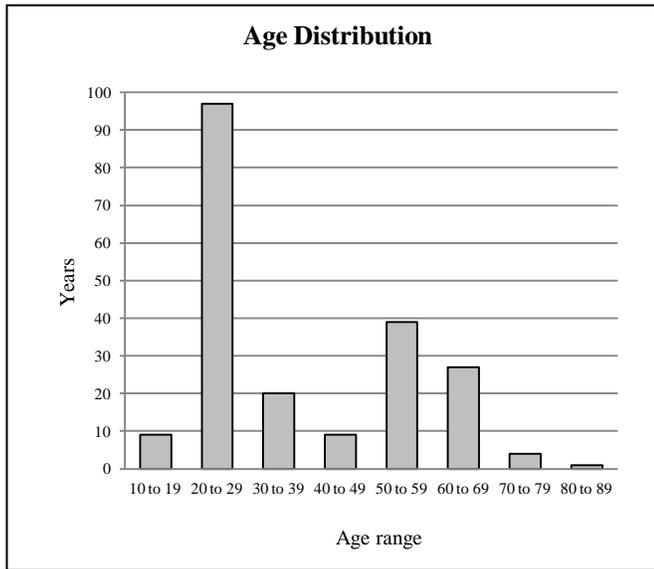


Fig. 1. Illustration of participant age distribution.

The reason for this age distribution is the fact that students were addressed and that they were asked to forward the mail to two persons who were older than 50.

27 percent of the participants were living alone (m=33%, f=21%) and 63 percent were in possession of a smartphone (m=67%, f=60%). The average participant has been using PCs for 16.9 years and tablet PCs for 0.45 years. The average usage time of PCs is 37.4 hours per week, that of tablet PCs 1.8 hours per week. On the other hand, those people using tablet PCs, i.e. 24.8% of the sample, on average use it 7.1 hours per week.

The following table illustrates these descriptive variables while differentiating between male and female participants.

TABLE I. SAMPLE DESCRIPTIVES

	Total Sample	Male	Female
N	206	109	97
Age	37.97	35.38	40.35
Living alone	27%	33%	21%
Smartphone possession	64%	67%	60%
Computer experience (years)	15.77	16.89	14.52
Computer usage (hours per week)	37.40	41.50	32.80
Tablet experience (years)	0.45	0.44	0.45
Tablet usage (hours per week)	1.81	1.58	2.07

IV. RESULTS

This section will present the results from the online survey and will mainly focus on the differences across the technologies. First, a descriptive analysis is executed showing basic information about the results of the survey. Second, a correlation analysis is executed to indicate important relationships. Third, t-tests and analyses of variances are considered. Fourth, a regression section concludes the results section. Fifth, the sample is split into two groups dependent on the participants’ age to analyze differences between older and younger participants.

A Descriptive analysis

The following graph illustrates the mean of expected reliabilities of the technologies.

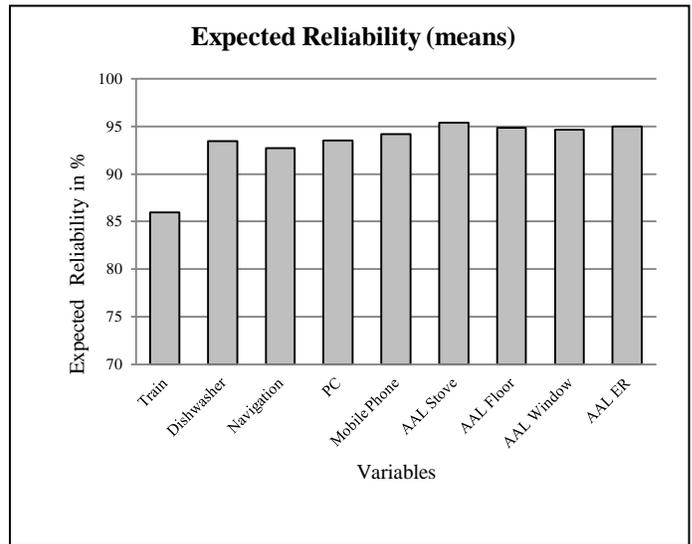


Fig. 2. Illustration of mean expected reliabilities.

As can be seen in the graph, all mean expected reliabilities exceed 85% or even 90%. One technology (stove) even exceeds the 95% level of expected reliability. Furthermore, the graph indicates that the expected reliability of the train scenario is lower than the expected reliabilities of the other technologies and that the AAL-related technologies (stove, floor, and window, together summarized in the variable AAL ER) are expected to have a higher reliability than all other technologies. The mean expected reliabilities as well as their standard deviations are illustrated in Table II.

TABLE II. COMPARISON OF DIFFERENT TECHNOLOGIES REGARDING EXPECTED RELIABILITY

Scenario	Mean	Standard Deviation
Train	0.860	0.094
Dishwasher	0.934	0.057
Navigation System	0.927	0.062
Computer	0.935	0.053
Mobile Phone	0.942	0.052
Stove	0.953	0.052
Floor sensor	0.949	0.062
Window	0.947	0.057
AAL technologies (combined)	0.949	0.048

B Correlation analysis

The correlation analysis reveals several important aspects regarding the relationship between the different technologies. First, positive correlations exist between the expected reliabilities for all different technologies (correlations ranging from 0.211 to 0.690), meaning that participants who consider one technology to be reliable tend to consider another technology reliable as well.

With respect to the participant's gender no clear result can be drawn regarding the expectations of the technologies' reliabilities except for the window and stove technology, where being a woman is negatively correlated with expected reliability (significant correlations of -0.154 for the window and -0.159 for the stove technology).

The age of the participants, on the other hand, was positively correlated with the expected reliabilities. The corresponding correlations range from 0.178 to 0.366, all significant on at least the 5% level. This indicates that elderly participants expect the technologies to have a higher reliability than younger participants do.

The possession of a smartphone is negatively correlated with the expectation of reliability. Additionally, the number of years a participant has been using computers is positively correlated with his opinion of the reliability of the AAL technologies. Contrary to the other technologies (non-significant correlations of 0.019 to 0.102) the correlations with the technologies stove (0.272), floor (0.216), and window (0.214) are significant at the 1% level.

On the other hand, neither living alone nor the variables regarding the tablet experience (in years), the weekly usage of computers, or that of tablet computers have clear relationships with expected reliability.

C Analysis using t-tests

The analysis of the means using a one-sample t-tests as well as paired sample t-tests showed significant differences in the perception of the different technologies. The mean expected reliabilities are significantly greater than 80% but differ across technologies.

The one-sample t-tests show that while the train scenario is expected to have the lowest reliability, i.e. significantly lower than 90% but higher than 80%, all other technologies show expected reliabilities significantly higher than 90%.

The paired-sample t-tests again indicate that the reliability of the punctuality of trains is significantly lower than that of all other technologies ($p < 0.001$ for all technologies). The reliabilities of the AAL technologies, on the other hand, are expected to be the most reliable technologies. The respective expected reliabilities of the stove, window, and floor technologies are significantly higher than those of the train, dishwasher, navigation system, and PC technologies and non-significantly higher than the expected reliability of the mobile phone technology.

D Regression analyses

An OLS regression with the mean of all AAL-related expected reliabilities as the dependent and the non-AAL-

related technologies (train, dishwasher, navigation system, PC, mobile phone) as well as gender, age, living environment, and smartphone possession indicates the following: The expected reliabilities of the train ($p < 0.05$) and navigation system ($p < 0.01$) technologies as well as the age of the participant ($p < 0.05$) and being male ($p < 0.005$) have a significant positive effect on the expected reliability of the AAL technologies, the other variables are not influential.

E Sample Split at age of 50

To analyze possible differences between younger and older participants, a split is executed at the age of 50. The group of "younger" persons (below the age of 50) consisted of 135 participants and the group of "older" persons (at least of age 50) of 71 participants. The following figure shows the mean expected reliabilities differentiated between younger and older participants.

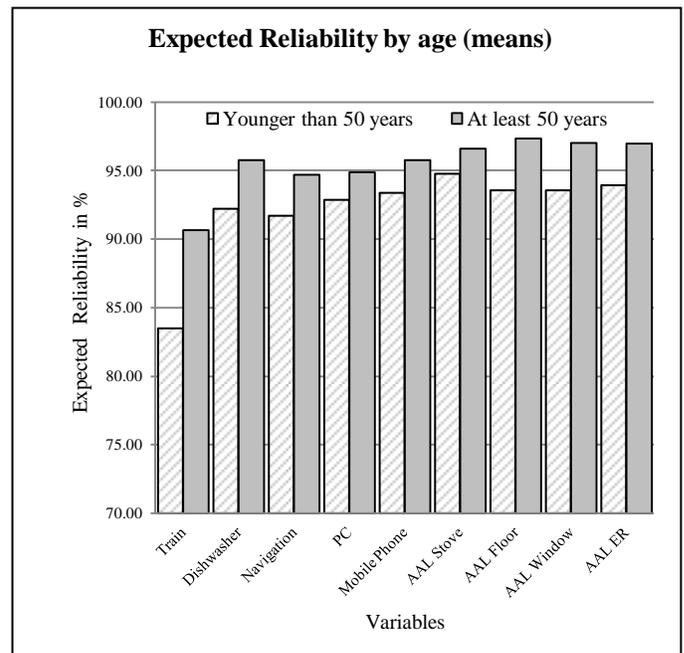


Fig. 3. Comparison of younger and older participants.

The graph shows that younger participants of the survey on average reported lower expected reliabilities than older participants for all technologies. In the train scenario, younger participants expected the punctuality to be less than 85% whereas older participants reported an average value of more than 90%. In the other everyday technologies as well as in the AAL scenarios younger participants reported average values between 90 and 95% while participants of at least 50 years of age reported average values of close to or even exceeding 95% reliability.

Sample t-tests comparing all reliability variables show significant differences for all expected reliabilities. They also show that the two groups differ with respect to mobile phone possession, computer experience, and computer usage per week. The following table shows the mean differences between younger and older participants with respect to expected reliability.

TABLE III. MEAN DIFFERENCES OLDER VS. YOUNGER

Variable	Mean Difference*	Significance
Train	0.071	<0.001
Dishwasher	0.036	<0.001
Navigation system	0.030	<0.005
Computer	0.020	<0.01
Mobile phone	0.024	<0.005
Stove	0.018	<0.05
Floor sensor	0.037	<0.001
Window	0.035	<0.001
AAL technologies (combined)	0.028	<0.001

* Represents the difference between older and younger participants

V. DISCUSSION

The study describes the differences of expected reliabilities between everyday- and AAL technologies in dependency of user-age to gain deeper insights about differences between young and older users (cut-off age: 50 years) regarding the expected reliabilities of the technologies considered.

The results of the study show that the participants of the study expected all reliabilities of the technologies to be far from the minimum value (70%) but often close to the maximum value (100%). This indicates high expectations with respect to the technologies queried in the survey. Furthermore, the results indicate differences in expected reliability among different technologies as well as among different groups of age.

As mentioned above, two important aspects exist with respect to the differentiation between technologies. First, the expected reliability of trains is evaluated significantly lower than that of all other technologies. That points to negative sentiments of participants towards the German railway system. Second, the scenarios regarding AAL technologies were evaluated as being more reliable than the everyday technologies. This shows high expectations of people towards such high-criticality systems compared to technologies which people are used to, such as mobile phones and dishwashers.

The study conducted analyzed the expected reliabilities of five everyday technologies, as well as three AAL-related technologies. The results indicate a reliability expectancy of more than 90% for all technologies except for the train scenario. The reliability of train schedules (i.e. train punctuality) was expected to be close to 86%. The comparison between these two types of technologies revealed significant differences with respect to the expected reliabilities. AAL technologies are expected to show a higher reliability than the everyday technologies.

The comparisons between younger and older participants further revealed that older persons, i.e. persons of at least 50 years of age, expect all technologies to be more reliable than younger persons do. In connection with AAL technology, the age of the participants as well as the gender male shows a significant positive effect on the expected reliability. This is an important result considering the target group of AAL products, namely elderly people. For producers of AAL technologies it is

very important to know about the expectations elderly persons have regarding the reliability of high-criticality systems.

VI. CONCLUSION

The central statements of this study regarding expected reliability of different technologies subject to the age of the participants should be taken into account for providers of AAL technologies. The results already treated in the discussion section give interesting insights which should be considered for addressing the target group for AAL products. The combination of the findings leads to the conclusion that an extraordinary high reliability of AAL technologies is surely one of their crucial points for the acceptance by the users. In order to get deeper insights for the acceptance and therefore, market success of supporting systems for the elderly, further researches regarding the remaining crucial point should be conducted. On the basis of this knowledge more detailed requirements as a part of a high-quality specification of such systems could be made. This would ensure a respectable fundament for the subsequent product engineering.

VII. DATA LIMITATIONS

Several limitations exist with respect to our sample. First, acquiring participants through personal contacts and students might not lead to a representative sample of the German population. Instead, our participants might be younger, more educated, and they might have more interest in and more knowledge of information technology. Second, since our questionnaire consisted of one question each for all non-AAL-related technologies but three questions each for the AAL-related ones, comparisons between these two types of technologies have to be evaluated with care.

Third, due to the usage of a 7-point scale and, thus, a lower limit of 70% reliability for all technologies, outliers were made impossible. This generates a problem regarding the average expected reliabilities because participants expecting the reliabilities to be below 70% probably would have chosen the lower limit instead of their true beliefs.

Fourth, a split dividing the sample into two groups with a cut point of 50 years of age does not reveal two groups of equal size. Instead, our sample of "elderly" people consists of 71 persons while the sample of "younger" people consists of nearly twice as much participants (n= 135).

VIII. FURTHER RESEARCH

With respect to the expected reliability as well as the acceptance of AAL technologies, further research is necessary to evaluate the success of AAL in the future and – in the long run – to develop possible market entry strategies. The study at hand focuses on the aspect of expected reliabilities and compared AAL to everyday technologies. It made possible a first evaluation of subjective differences between technologies. Additionally, it enabled an analysis of differences between persons of different ages. So general research of age-related effects could be added or developed regarding the issue of AAL.

Further research, nevertheless, is needed to evaluate the influence of reliability on trust and intention to use. Since these two aspects have a significant influence on whether consumers

buy a product, this can hint producers towards important aspects they have to consider when designing AAL technologies. Furthermore, in a next step an experiment with different reliability levels of AAL technologies will be designed. This differentiation will be used to measure the influence of actual reliability on trust, intention to use, and other variables.

ACKNOWLEDGMENT

This research was supported by grants from the German Federal Ministry of Education and Research (BMBF). It is part of the project SMILEY (Smart and Independent Living for the Elderly) supported by BMBF under contract 01FC10004.

REFERENCES

- [1] UNDESA, "World population ageing 2009," Department of Economic and Social Affairs: Population Division, New York, 2010.
- [2] D. Stroud, "Don't fall for the 50-plus blind spot," *Brandweek*, vol. 47 (10), 2006.
- [3] W. Paulus, J. Hilbert, and W. Potratz, "ICT for housing," in N. Malanowski, M. Cabrera (Eds.): *Information and Communication Technologies for Active Ageing. Opportunities and Challenges for the European Union*. IOS Press. Amsterdam, 2009.
- [4] BMBF, "Assistenzsysteme im Dienste des älteren Menschen," Steckbriefe der ausgewählten Projekte in der BMBF-Fördermaßnahme „Altersgerechte Assistenzsysteme für ein gesundes und unabhängiges Leben – AAL“, URL: <http://www.aal-deutschland.de/deutschland/dokumente/projektportrats-aal.pdf> (last checked 19.04.2013).
- [5] Fraunhofer (2011): „Zuhause Daheim: Das Projekt JUTTA.“ URL: <http://www.inhaus.fraunhofer.de/de/Geschaeftsfelder/Health-und-Care/jutta.html> (last checked 19.04.2013).
- [6] W. Heusinger, „Das intelligente Haus - Entwicklung und Bedeutung für die Lebensqualität,“ Frankfurt am Main: Lang, 2005.
- [7] S. Solaimani, H. Bouwman, and M. de Reuver, "Smart home: aligning business models and providers processes; a case survey," 21st Australian Conference on Information Systems Aligning Business Models and Providers Processes, Brisbane, 2010.
- [8] J. Botia, A. Villa and J. Palma, "Ambient assisted living system for in-home monitoring of healthy independent elders", *Expert Systems with Application*, vol. 39, 2012, p. 1.
- [9] Commission of the European Communities (2005). Green paper "Confronting demographic change: a new solidarity between the generations". p. 10.
- [10] H. Sun, V. de Florio, and N. Gui, "Promise and challenges of Ambient Assisted Living," *Sixth International Conference on Information Technology: New Generations*, 2009.
- [11] D. A. Wiegmann, A. Rich, and H. Zhang, "Automated diagnostic aids: The effects of aid reliability on users' trust and reliance," *Theoretical Issues in Ergonomics Science*, vol 2 (4), 2001, pp. 352–367.
- [12] A. Keller, and S. Rice, "System-wide versus component-specific trust using multiple aids," *The Journal of General Psychology*, vol. 137 (1), 2010, pp. 114-128.
- [13] T. Bahmanziari, J. Pearson, and L. Crosby, "Is trust important in technology adoption? A policy capturing approach," *Journal of Computer Information Systems*, vol. 43 (4), 2003, pp.46-54.
- [14] Oxford Dictionary, URL: <http://oxforddictionaries.com/definition/english/reliable>, (last checked 19.04.2013).
- [15] Oxford Dictionary, URL: <http://oxforddictionaries.com/definition/english/reliable>, (last checked 26.04.2013)
- [16] E. N. H. Montague, W. W. Winchester, and B. M. Kleiner, "Trust in medical technology by patients and healthcare providers in obstetric work systems," *Behaviour & Information Technology*, vol. 29 (5), 2010, pp. 541-554.
- [17] R. Parasuraman, and C. Miller, "Trust and etiquette in high-criticality automated systems," *Communications of the Association for Computing Machinery*, vol 47 (4), 2004, pp. 51-55.
- [18] J. Sanchez, G. Calcaterra, and Q. Q. Tran, "Automation in the home: The development of an appropriate system representation and its effects on reliance," *Proceedings of the Human Factors and Ergonomics Society 49th annual meeting*, 2005.
- [19] A. Munoz, E. Serrano, A. Villa, M.Valdes and J. Botia, "An approach for representing sensor data to validate alerts in ambient assisted living", *Sensors 2012*, 2012
- [20] M.S. Bartlett, G. Littlewort, I. Fasel., & J.R. Movellan, „Real time face detection and facial expression recognition: development and application to human computer interaction.,” In *Proceedings of the Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction at the 2003 Conference on Computer Vision and Pattern Recognition* (pp. 53-58).
- [21] F. Steinke, T. Fritsch, D.Brem and S.Simonsen, "Requirement of AAL systems-older perons' trust in sensors and characteristics of AAL technologies," *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, Heraklion, Crete, Greece, 2012.
- [22] F. Steinke, A. Ingenhoff, and T. Fritsch, "Personal remote support in Ambient Assisted Living – experimental investigation on trust and intention to use of elderly people," unpublished.

Modeling the Cut-off Frequency of Acoustic Signal with an Adaptative Neuro-Fuzzy Inference System (ANFIS)

Y. Nahnraoui, E.H. Aassif,

LMTI, Faculté des sciences, Université Ibn
Zohr, Agadir, Maroc

G.Maze

LOMC, Université du Havre,
Institut Universitaire de Technologie,
Place Robert Schuman, 76610 Le Havre,
France

R.LATIF

ESSI, Ecole Nationale des Sciences
appliquées, Université Ibn Zohr,
Agadir, Maroc

Abstract—An Adaptative Neuro-Fuzzy Inference System (ANFIS), new flexible tool, is applied to predict the cut-off frequencies of the symmetric and the anti-symmetric circumferential waves (S_i and A_i , $i=1,2$) propagating around an elastic aluminum cylindrical shell of various radius ratio b/a (a : outer radius and b : inner radius). The time-frequency of Wigner-Ville and the proper modes theory are used in this study to compare and valid the frequencies values predicted by the ANFIS model. The useful data, of the cut-off frequencies $(ka)_c$, are used to train and to test the performances of the model. These data are determined from the values calculated using the proper modes theory of resonances and also from those determined using the time-frequency images of Wigner-Ville. The material density, the radius ratio b/a , the index i of the symmetric and the anti-symmetric circumferential waves, and the longitudinal and transverse velocities of the material constituting the tube, are selected as the input parameters of the ANFIS model. This technique is able to model and to predict the cut-off frequencies, of the symmetric and the anti-symmetric circumferential waves, with a high precision, based on different estimation errors such as mean relative error (MRE), mean absolute error (MAE) and standard error (SE). A good agreement is obtained between the output values predicted using the propose model and those computed by the proper modes theory.

Keywords—ANFIS; time-frequency; SPWV; Acoustic scattering, acoustic circumferential waves; cut-off frequency;cylindrical shell.

I. INTRODUCTION

In a previous studies [1, 2], we have analysed the acoustic signal scattered by a thin elastic tube immersed in water using the time-frequency representation of Wigner-Ville. The Wigner-Ville image obtained in these analyses allowed to determine the cut-off frequency, of the anti-symmetric circumferential waves A_1 propagating around the aluminum cylindrical shell of different radius ratio b/a . These analyses permitted also to determine, form the time-frequency image, the thickness of elastic cylindrical shell.

Many studies, theoretical and experimental, showed that acoustic resonances of a cylindrical shell are related to its physical and geometrical properties. Conversely, starting from the resonances of circumferential waves we can characterize

material constituting a cylindrical shell the geometry of which is known [1-6].

The resonances of the symmetric and the anti-symmetric circumferential waves (S_i and A_i , $i=0, 1, 2, \dots$: index of the mode) are observed on the spectrum of the acoustic pressure backscattered by the cylindrical shell [7]. Apart from the specular reflection, the backscattered pressure field results mainly from the interactions of different kinds of creeping waves that generate resonances in the spectrum. The resonance frequencies of the circumferential waves (S_i and A_i) essentially depend on the radius ratio b/a . Using the proper modes theory, we can determine the cut-off frequencies of the symmetric and the anti-symmetric circumferential waves (S_i and A_i , $i=1,2$) for a aluminium cylindrical shell with different radius ratio b/a . One of the most important points is find out some parameters that carry most of the information available from the response of the cylindrical shell. Such parameters may be found from the velocity dispersion of the circumferential waves (S_i and A_i), since it is directly related to the geometry and to the physical properties of the shell.

Different methods have been proposing for analyse of the circumferential waves propagating around the cylindrical shell which includes temporal analysis [7, 9], spectrum analysis [5, 7], parametric time-frequency analysis [10-16], wavelet transform [19-20] and neural networks [21-22].

The present paper is especially concerned with the soft computing technique such as fuzzy logic system. The adaptative neuro-fuzzy inference system (ANFIS) is selected and applied to predict the cut-off frequencies of the symmetric and the anti-symmetric circumferential waves (S_i and A_i , $i=1,2, \dots$) for cylindrical shell of various radius ratio b/a that cannot be measured experimentally. The cut-off frequencies obtained from the computed values using the proper modes are used as data in the ANFIS model. In experiments, the time-frequency representation of Wigner-Ville of the acoustic signal backscattered by cylindrical shell is calculated. The ANFIS model and the Wigner-Ville technique are tools for the statistical analysis, making possible the construction of a model of behavior starting from a certain number of examples. The model is able to predict the cut-off frequencies of the symmetric and the anti-symmetric circumferential waves (S_i and A_i , $i=1, 2, \dots$) for aluminum cylindrical shell of various

radius ratio b/a . The radius ratios used, in this paper, are between 0.4 and 0.99. The cut-off frequencies values determined using the ANFIS model are compared with those determined from the time-frequency images of Wigner-Ville to validate the robustness of the model proposed. In this study, we have use three aluminum cylindrical shell of various radius ratio b/a (0.9, 0.95 and 0.97). These examples are used to evaluate the performance and robustness of the ANFIS model and make a comparison with the analysis of time-frequency Wigner-Ville to determine the dimensional radius ratio of the cylindrical shell studied.

II. BACKSCATTERING RESPONSE FROM A CYLINDRICAL SHELL

A. Acoustic scattering by an air-filled cylindrical shell

The analysis of acoustic signals scattered by an air-filled cylindrical shell immersed in water is a topic that has received large attention for several years [1-8]. In previous studies, the characterization of the scattering problem is mainly performed in the frequency domain. The module of the backscattered pressure in the faraway field, called “form function”, by the cylindrical shell can be derived directly from a computational model [7, 11].

This module is also called a backscattered spectrum. Apart from the specular reflection, the backscattered pressure field results mainly from the interactions of different kinds of creeping waves that generate “resonances” in the spectrum. These resonances are in relation with the symmetric and anti-symmetric circumferential waves (S0, A1, S1, S2, A2, ...).

The scattering of an infinite plane wave by an air-filled cylindrical shell of radii ratio b/a is investigated through the solution of the wave equation and the associated boundary conditions.

Fig. 1 shows the cylindrical coordinate orientation and the direction of a plane wave incident on an infinitely long cylindrical shell in a fluid medium. The fluid (1) inside the shell has a density of ρ_1 and propagation velocity c_1 . In general, the outer fluid (2) will be different and is described by the parameters ρ and c . The parameters for the two fluids outside and inside the shell are given in Table 1.

The axis of the cylindrical shell is taken to be the z -axis of the cylindrical coordinate system (r, θ, z) . Let a plane wave incident on an infinite cylindrical shell with air-filled cavity (fluid 2), be submerged in water (fluid 1), see figure 1.

The backscattered complex pressure P_{diff} by a cylindrical shell in a faraway field ($r \gg a$, we have neglected the diffraction of waves and one receives only the part backscattered of the complex pressure field) is the summation of the incident wave, the reflective wave ①, surface waves tell shell waves ② (whispering Gallery, Rayleigh, ...) and Scholte waves (A) ③ connected to the geometry of the object (figure 2). The waves ② and ③ are the circumferential waves. For these waves one distinguishes the waves A, the symmetric waves S0, S1, S2 and the anti-symmetric waves A1, A2.

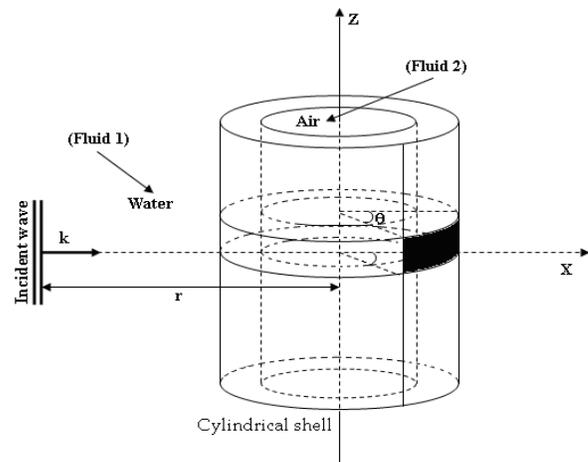


Fig.1. Geometry used for formulating the sound backscattering from a cylindrical shell

The general form of the backscattered pressure field at normal incidence can be expressed as [7-8, 23]

$$P_{diff}(\omega) = P_0 \sum_{n=0}^{\infty} \frac{D_n^{(1)}(\omega)}{D_n(\omega)} H_n^{(1)}(kr) \quad (1)$$

Where $\omega = 2\pi f$ is the angular frequency, k the wave number with respect to the wave velocity in the external fluid and P_0 the amplitude of the plane incident wave. $D_n^{(1)}(\omega)$ and $D_n(\omega)$ are determinants computed from the boundary conditions of the problem (continuity of stress and displacement at both interfaces). The function $H_n^{(1)}$ is the Hankel function of the first kind.

The module of the backscattered complex pressure in a faraway field is called form function. This function is obtained by the relation [7-8, 23]

$$|P_{diff}(\omega)| = \frac{2}{\sqrt{\pi k r}} \left| \sum_{n=0}^{N_{max}} \varepsilon_n (-1)^n \frac{D_n^{(1)}(\omega)}{D_n(\omega)} \right| \quad (2)$$

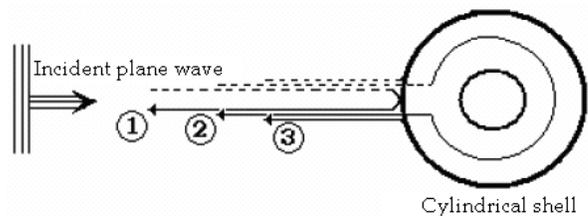


Fig.2. Mechanisms of the formation of echoes showing the specular reflection ① and shell waves ② and Scholte wave (A) ③.

where is the Neumann factor ($\varepsilon_n = 1$, if $n = 0$; $\varepsilon_n = 2$, if $n > 0$), $k = \omega/c$ is the incident wave number and c is the phase velocity in water.

The physical parameters used in the calculation of the backscattered complex pressure are illustrated in table I.

TABLE I. PHYSICAL PARAMETERS

	Density ρ (kg/m ³)	Longitudinal Velocity c_L (m/s)	Transverse Velocity c_T (m/s)
Aluminum	2790	6380	3100
Water	1000	1470	-
Air	1.29	334	-

The figure 3 shows the module of the backscattered complex pressure in function of the reduced frequency ka (without unit) given by :

$$ka = \frac{\omega a}{c} = \frac{2\pi}{c(1-\frac{b}{a})} f d \quad (3)$$

Where $d=a-b$ is the thickness of a cylindrical shell and f is the frequency of resonance of a wave in Hz.

The temporal signal response $P(t)$ of a cylindrical shell is computed by taking the Inverse of Fourier Transform of the module of the backscattered complex pressure:

$$P(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} h(\omega) P_{diff}(\omega) e^{-i\omega t} d\omega \quad (4)$$

Where $h(\omega)$ is a smoothing window.

The succession of shell resonances (corresponding to frequency of resonances) in the spectrum of the figure 3 is connected with the propagation of acoustic circumferential waves: Scholte wave (A) and shell waves (S0, A1, S1, S2, A2, ...). The temporal signal backscattered by an Aluminum cylindrical shell is obtained by the Inverse Transform Fourier

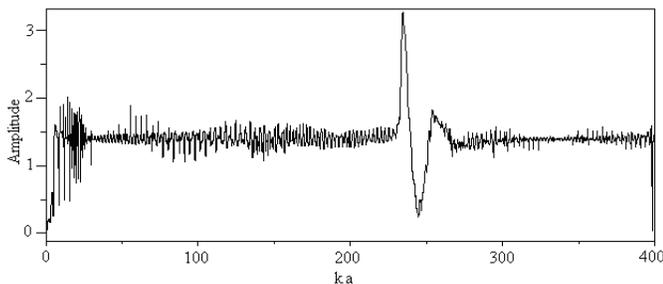


Fig.3. Module of the backscattered complex pressure for an infinite aluminum cylindrical shell with air-filled cavity of radii radio $b/a=0.95$

of the module of the backscattered complex pressure using the equation (4). The figure 4 presents this signal and shows the specular reflection ① (large amplitude and short duration) and several wave packets ② and ③ associated with different circumferential waves (A, S0, A1, S1, S2, A2, ...). The observation of this signal shows a succession of components more or less distinct that one seeks then to identify. The different echoes finish by overlapping and in these conditions, the identifications and measures of arrival times of echoes (this time depends on the radii of the tube a and b) become difficult, perhaps impossible. This constitutes a major

disadvantage of the temporal approach. An important feature of the acoustic circumferential waves is the velocity dispersion that leads to a time spreading of wave packets.

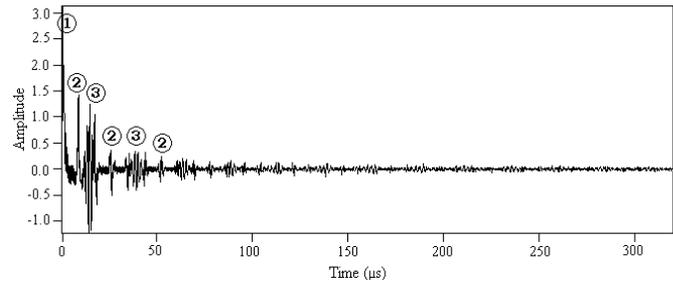


Fig.4. Signal backscattered by an aluminum cylindrical shell with air-filled cavity, $b/a=0.95$ (Specular reflection echo ①, shell waves echoes ② and Scholte wave echo (A) ③).

B. Dispersion and cut-off frequency determined using the proper modes theory

An important feature of the circumferential waves is the velocity of dispersion that leads to a time spreading of wave packets (shell of waves). In the case of the circumferential waves ② for instance, the dispersion velocity is significant and the time spreading is much more important than in the case of ③. Resonances that appear on the backscattered pressure field are linked to the propagation of circumferential waves around the tube. One finds the wave of Scholte (A) and the waves of shell (S0, A1, S1, S2, A2).

The group velocity of circumferential waves is estimated from the resonance frequencies, using the proper modes theory, that correspond to the circumferential waves. The calculation of the resonance frequencies of these waves have been made by the cancellation of the determinant D given by [8]:

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix} \quad (5)$$

where the 16 non-vanishing elements d_{ij} can all be determined from the boundary conditions of the problem, and they have all been listed elsewhere [2]. The resolution of the equation $D=0$ allows to determine the different proper modes for each type of the symmetric and the anti-symmetric circumferential waves (S_i and A_i , $i=0, 1, 2, \dots$). Once frequencies of resonances are determined, we calculate the difference Δka between two successive resonance frequencies. The group velocity of the symmetric and the anti-symmetric circumferential waves for each frequency is given by [8]:

$$c_g = c \Delta ka \quad (6)$$

where Δka the gap between two successive resonances.

The figure 5 shows the evolution of the group velocity in function of the (ka) for different waves.

Starting from the similitude that exists between the circumferential waves in the case of a thin elastic tube and the Lamb waves in the case of a plaque of the same thickness, it is possible to use the classical relations on the Lamb waves to ascend to the value of the reduced cut-off frequency of circumferential waves in the case of a tube [2-5, 14-15, 24-25].

In the case of a thin plaque, the cut-off frequencies of the symmetric and anti-symmetric Lamb waves are given by [2, 14]:

$$(fd)_c = \begin{cases} m_s c_T \\ (m_s + \frac{1}{2})c_L \end{cases} \quad (7)$$

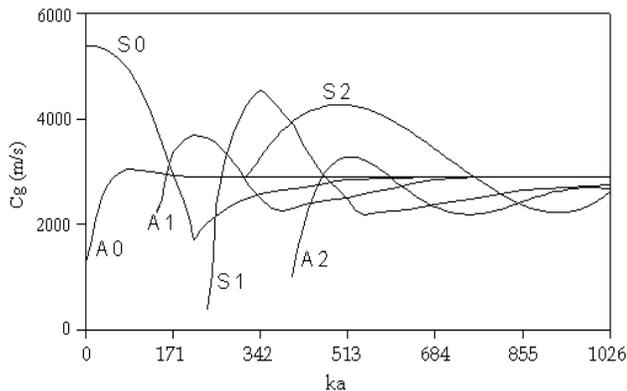


Fig.5. Dispersion velocity of the different circumferential waves of an aluminum cylindrical shell of radii ratio $b/a=0.95$

$$(fd)_c = \begin{cases} m_a c_L \\ (m_a + \frac{1}{2})c_T \end{cases} \quad (8)$$

where c_T and c_L are transverse and longitudinal velocities of the material constituting the cylindrical shell. The indices s and a on the integers m number indicating symmetric and anti-symmetric modes of plate vibrations respectively.

The cut-off frequencies, of the symmetric and anti-symmetric circumferential waves, are determined by exploiting the equations 3, 7 and 8 [2, 14]:

$$(ka)_c = \frac{2\pi}{c(1-\frac{b}{a})} \begin{cases} m_s c_T \\ (m_s + \frac{1}{2})c_L \end{cases} \quad (9)$$

$$(ka)_c = \frac{2\pi}{c(1-\frac{b}{a})} \begin{cases} m_a c_L \\ (m_a + \frac{1}{2})c_T \end{cases} \quad (10)$$

where m_s and m_a (integers numbers) are the symmetric and anti-symmetric modes of circumferential waves respectively.

For the symmetric modes $S1$ and $S2$ the cut-off frequencies values are calculated from the equations (11) and (12) respectively:

For $S1$ mode:

$$(ka)_c^{S1} = \frac{2\pi}{c(1-\frac{b}{a})} \cdot c_T \quad (11)$$

For $S2$ mode:

$$(ka)_c^{S2} = \frac{\pi}{c(1-\frac{b}{a})} \cdot c_L \quad (12)$$

For the anti-symmetric modes $A1$ and $A2$ the cut-off frequencies values are calculated from the equations (13) and (14) respectively:

For $A1$ mode:

$$(ka)_c^{A1} = \frac{\pi}{c(1-\frac{b}{a})} \cdot c_T \quad (13)$$

For $A2$ mode:

$$(ka)_c^{A2} = \frac{3\pi}{c(1-\frac{b}{a})} \cdot c_T \quad (14)$$

The calculated values, using the equations (11) to (14), of the cut-off frequencies of the symmetric and anti-symmetric circumferential waves $A1$, $S1$, $S2$ and $A2$ are given in table II.

TABLE II. CUT-OFF FREQUENCIES VALUES OF DIFFERENT CIRCUMFERENTIAL WAVES FOR ALUMINUM CYLINDRICAL SHELL OF VARIOUS RADIUS RATIOS

Cylindrical shell	Cut-off frequencies $(ka)_c$			
	Mode $A1$ $m_a=0$	Mode $S1$ $m_s=1$	Mode $S2$ $m_s=0$	Mode $A2$ $m_a=1$
$b/a=0.9$	66.21	132.43	136.28	198.65
$b/a=0.95$	132.43	264.87	272.56	397.30
$b/a=0.97$	220.72	441.45	454.26	662.17

III. DISPERSION ANALYSIS USING TIME-FREQUENCY IMAGES

The analysis of the returned echoes has traditionally been done in the frequency domain, and later in the time domain. A recent processing technique that seems to be gaining acceptance is to work in the combined time-frequency domain. Usually, projections of these three-dimensional surfaces are shown in the two-dimensional time-frequency plane. This evolution can be extracted from the echoes and displayed in as much detail as is feasible. Among the large number of existing time-frequency representations, some authors [1-2, 8, 10-16] have proposed to use the Smoothed Pseudo Wigner-Ville. The

choice of this particular distribution results from its interesting properties in terms of acoustic applications [1-2, 8, 12, 14-15].

A. Theoretical fundamentals

The Wigner-Ville distribution (WVD) of the real signal $x(t)$ is defined by [4, 8, 11-14] :

$$WV_x(t, \nu) = \int_{-\infty}^{+\infty} x(t + \frac{\tau}{2}) x^*(t - \frac{\tau}{2}) e^{-i2\pi\nu\tau} d\tau \quad (15)$$

Time-frequency smoothing can then be applied to reduce the amplitude of these spurious terms. It can be achieved by using the Smoothed Pseudo Wigner-Ville (SPWV) [1-2]:

$$SPWV_x(t, f) = \int_{-\infty}^{+\infty} \left| h\left(\frac{\tau}{2}\right) \right|^2 \int_{-\infty}^{+\infty} g(t-u) x(u + \frac{\tau}{2}) \times x^*(u - \frac{\tau}{2}) \exp(-2j\pi f\tau) dud\tau \quad (16)$$

The smoothing windows $g(t)$ and $h(t)$ are introduced into the SPWV definition in order to allow a separate control of interferences either in time (g) or in frequency (h).

B. Dispersion and cut-off frequency using SPWV

Scattering from a finite object provides many interesting subjects for analysing the circumferential waves. For example, one of the challenging problems is how to determine the shape and physical properties of an object thanks to the SPWV. This technique appears to be a very useful tool for such a task, as it is able to represent a given signal simultaneously in time and frequency domains.

The resonances brought into evidence on the scattered complex pressure (figure 3) are linked to the propagation of circumferential waves: Scholte waves (A) and shell waves ($S0, A1, S1, S2, A2$) in the case of a cylindrical shell with light thickness. In this study, one is interested only in the symmetric and the anti-symmetric circumferential waves (S_i and $A_i, i=1,2,\dots$). According to this spectrum (figure 3), the reduced frequencies scale in which appears the symmetric and the anti-symmetric waves (S_i and $A_i, i=1,2$) are illustrated in the table III.

TABLE III. RANGE FREQUENCIES OF CIRCUMFERENTIAL WAVES FOR ALUMINUM CYLINDRICAL SHELL OF RADII RADIO $b/a=0.95$

	Range frequencies (ka)
Anti-symmetric wave $A1$	130 – 200
Symmetric wave $S1$	260 – 340
Symmetric wave $S2$	270 – 350
Anti-symmetric wave $A2$	> 390

Figures 6, 7 and 8 represent the time-frequency images for the anti-symmetric circumferential wave $A1$ for aluminium cylindrical shell of various radius ratio b/a . When the time augments, the trajectory associated to anti-symmetric wave $A1$, for each case (figures 6, 7 and 8), tends to an asymptotic value which equal the cut-off frequency $(ka)_c$ of this wave.

Using the proper modes theory, this frequency is calculated by the equation (13). More precisely, this cut-off frequency is the intersection point of the asymptotic trajectory of the anti-symmetric wave $A1$ and the axis of frequencies (figures 6, 7 and 8). The values of the cut-off frequency $(ka)_c$ obtained from these images are presented in table IV. This table presents also those values computed with the proper modes theory (equation 13). We notice that the cut-off frequencies determined from the time-frequency images are in good concordance with those computed from proper modes theory (PMT).

TABLE IV. COMPARISON BETWEEN THE CUT-OFF FREQUENCIES VALUES COMPUTED THEATRICALY AND DETERMINED FROM SPWV IMAGES FOR ANTI-SYMMETRIC CIRCUMFERENTIAL WAVE $A1$

Cylindrical shell	Cut-off frequencies $(ka)_c$	
	Computed using PMT	Determined using SPWV
$b/a=0.9$ (figure 8)	66.21	66.0±0.3
$b/a=0.95$ (figure 7)	132.43	132.0±0.3
$b/a=0.97$ (figure 6)	220.72	221.0±0.2

IV. MATERIALS AND METHOD

A. Fuzzy Inference System

Fuzzy logic is an extension of Boolean logic that allows intermediate values between “True” and “False”. In this approach the classical theory of binary membership in a set, is modified to incorporate the memberships between “0” and “1”. The fuzzy models are means of capturing humans expert knowledge about the process, in terms of fuzzy (if-then) rules.

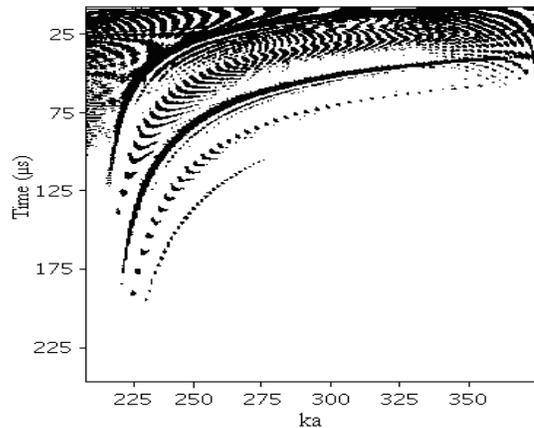


Fig.6. SPWV of backscattered signal for the first aluminum cylindrical shell of radii radio $b/a=0.97$ (Anti-symmetric circumferential wave $A1, 220 < ka < 375$)

The fuzzy inference system (FIS) can initialize and learn linguistic and semi-linguistic rules; hence it can be considered as direct transfer knowledge, which is the main advantage of fuzzy inference systems over classical learning systems and Neural Networks [26-28]. Often the rules of the fuzzy system are designated a priori and the parameters of the membership functions are adapted in the learning process from input-output data sets.

Basically, a fuzzy inference system is composed of five functional blocks, shown in Figure 9, as follows [26-28]:

1) A rule base containing a number of fuzzy if-then rules. All the uncertainties, non linear relationships, or model complications are included in the descriptive fuzzy inference procedure in the form of if-then statements. In general, a fuzzy if-then rule has two constituents; first the if part and the second the then part; which are called premise and consequent, respectively. The general form of a fuzzy if-then

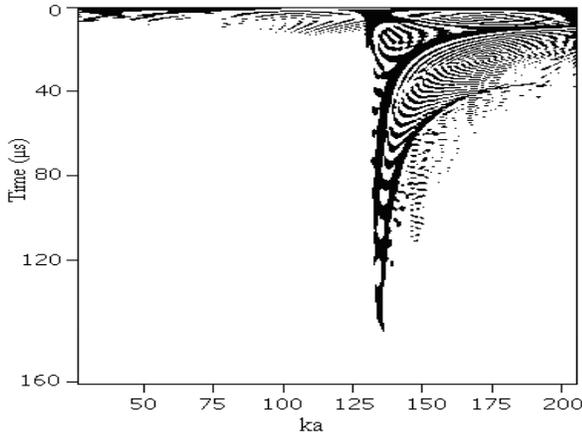


Fig.7. SPWV of backscattered signal for the second aluminum cylindrical shell of radii ratio $b/a=0.95$ (Anti-symmetric circumferential wave $A1$, $130 < ka < 200$)

rule is as follows; Rule: if Z is A then f is B.

- 2) A database, which defines the membership functions of the fuzzy sets used in the fuzzy rules.
- 3) A decision-making unit, which performs the inference operations on the rules.

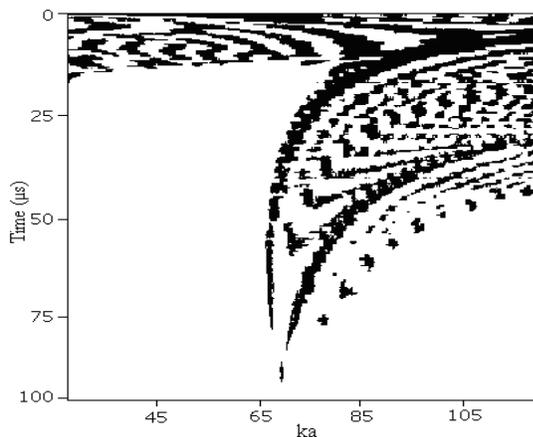


Fig.8. SPWV of backscattered signal for the third aluminum cylindrical shell of radii ratio $b/a=0.9$ (Anti-symmetric circumferential wave $A1$, $65 < ka < 120$)

- 4) A fuzzification inference, which transforms the crisp inputs into degree of match with linguistic values.
- 5) A defuzzification inference, which transforms the fuzzy results of the inference into a crisp output.

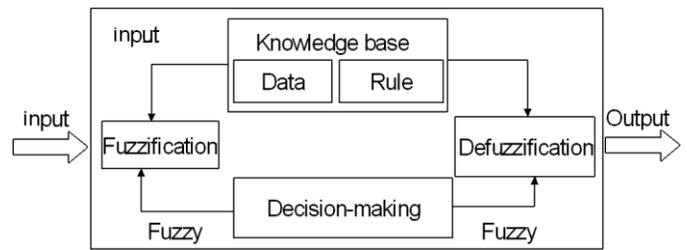


Fig.9. Bloc diagram for a fuzzy Inference System

Several types of FIS have been proposed in the literature[29], which, vary due to differences between the specification of the consequent part and the defuzzification schemes. This paper incorporates one of these types, the so-called Takagi and Sugeno FIS [30], to propose a systematic scheme for the development of fuzzy rules using the input/output data sets.

A typical fuzzy rule in a sugeno fuzzy model has the format:

$$\text{If } x \text{ is } A \text{ and } y \text{ is } B \text{ then } z = f(x, y)$$

where A and B are fuzzy sets in the antecedent; $z = f(x, y)$ is a crisp function in the consequent. Usually $f(x, y)$ is a polynomial in the input variable x and y , but it can be any other functions that can appropriately describe the output of the system within the fuzzy region specified by the antecedent of the rule. When $f(x, y)$ is a first order polynomial, we have the first-order sugeno fuzzy model. When f is a constant, we then have the zero-order Sugeno fuzzy model. Consider first-order Sugeno fuzzy inference systems which contain two rules:

$$\text{Rule 1: if } x \text{ is } A_1 \text{ and } y \text{ is } B_1 \text{ then } f_1 = p_1x + q_1y + r_1.$$

$$\text{Rule 2: if } x \text{ is } A_2 \text{ and } y \text{ is } B_2 \text{ then } f_2 = p_2x + q_2y + r_2.$$

Weighted averages are used in order to avoid complexity in defuzzification processes. Figure 10 illustrates graphically the fuzzy reasoning mechanism to derive an output f from a given input vector (x, y) . The firing strengths ω_1 and ω_2 are usually obtained as the product of the membership grades in the premise part, and the output f is the weighted average of each rule's output. To facility the learning of the sugeno fuzzy model, into the framework of adaptative networks we can compute gradient vectors systemically. The resultant network architecture is called Adaptive Neuro Fuzzy Inference system (ANFIS).

B. Adaptive neuro-fuzzy inference system architecture

The Adaptive Network-based Fuzzy Inference System (ANFIS) is developed by Jang in 1993 [26]. This model use neuro-adaptive learning techniques, which are similar to those of neural networks. Given an input/output data set, the ANFIS can construct a Fuzzy Inference System whose membership function parameters were adjusted using a hybrid algorithm learning that is a combination of Last Square estimate and the gradient descent back-propagation algorithm or other similar optimisation technique. This allows Fuzzy system to learn from the data they are modelled.

For simplicity, we assume the fuzzy inference system with two input, x and y with one response f . From the first-order Sugeno fuzzy model, a typical rule set with two fuzzy if-then rules can be expressed as below. The corresponding equivalent ANFIS architecture is as shown in figure 11. The system architecture consists of five layers, namely; fuzzy layer, product layer, normalized layer, fuzzy layer and total output layer. The following section in depth the relationship between the input and output of each layer in ANFIS.

Layer 0: It consists of plain input variable set.

Layer 1: It is the fuzzy layer. Each node in this layer generates a membership grade of a linguistic label. For instance, the node function of the i^{th} node may be generalized bell membership function:

$$\mu_{A_i} = \frac{1}{1 + \left[\frac{x - c_i}{a_i} \right]^{b_i}} \quad (17)$$

where x is the input to node i ; A_i is the linguistic label (small, large, etc.) associated with this node; and $\{ a_i, b_i, c_i \}$ is the parameter set that changes the shapes of the membership function. Parameters in this layer are referred to as the premise parameters.

Layer 2: The function is T-norm operator that performs the firing strength of the rule, e.g., fuzzy conjunctive AND and OR. The simplest implementation just calculates the product of all incoming signals.

$$\omega_i = \mu_{A_i}(x)\mu_{B_i}(y) \quad , i=1,2 \quad (18)$$

Layer 3: Every node in this layer is fixed and determines a normalized firing strength. It calculates the ratio of the ratio of the j^{th} rule's firing strength to the sum of all rules firing strength.

$$\bar{\omega}_i = \frac{\omega_i}{\omega_1 + \omega_2} \quad , i=1,2 \quad (19)$$

Layer 4: The nodes in this layer are adaptive are connected with the input nodes and the preceding node of layer 3. The result is the weighted output of the rule j .

$$\bar{\omega}_i f_i = \omega_i(p_i x + q_i y + r_i) \quad (20)$$

where $\bar{\omega}_i$ is the output of layer 3 and $\{ p_i, q_i, r_i \}$ is the parameter set. Parameters in this layer are referred to as the consequent parameters.

Layer 5: This layer consists of one single node which computes the overall output as the summation of all incoming signals.

$$\text{Overall Output} \quad \sum_i \bar{\omega}_i f_i = \frac{\sum_i \omega_i f_i}{\sum_i \omega_i} \quad (21)$$

The constructed adaptive network in figure 11 is functionally equivalent to a fuzzy inference system in figure

10. The basic learning rule of ANFIS is a combination of last squar error and the back-propagation gradient descent, which calculates error signals (the derivative of the squared error with respect to each node's output) recursively from the output layer backward to the input nodes. This learning rule is exactly the same as the back-propagation learning rule used in the common feed-forward neural networks.

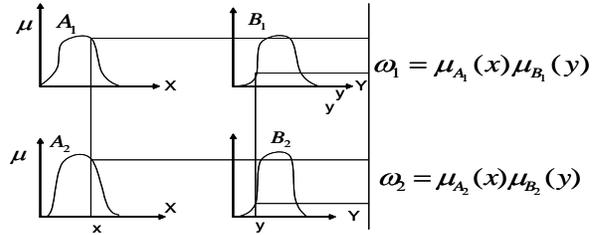


Fig.10. First-order Sugeno fuzzy model

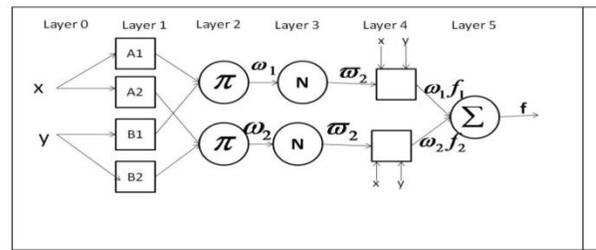


Fig.11. ANFIS architecture

V. COLLECTION OF DATA

The conception of the fuzzy logic model requires the determination of the relevant entries that have a significant influence on the required model. In this work, a data base is collected to involve and test the performance of the model starting from the results obtained by the time-frequency of Wigner-Ville method then supplemented by data resulting from the proper modes theory of the circumferential waves. The density of material, the radius ratio, the index of the anti-symmetric and symmetric circumferential waves, and longitudinal and transverse velocities, of the material constituting the cylindrical shell, are retained like relevant entries of the model because these parameters characterize the cylindrical shell and the types of circumferential waves propagating around this one. The cut-off frequency $(ka)_c$, of the anti-symmetric and symmetric circumferential waves (S_i and A_i , $i=1,2$) for an aluminum cylindrical shell with different radius ratios b/a , constitutes the output of fuzzy system. The collected data for the training and validation phases of the fuzzy logic system model are represented in tables I and II. For example, for aluminum cylindrical shell, the density is 2700 kg/m^3 , the transverse velocity is 3100 m/s and the longitudinal velocity is 6380 m/s . For the anti-symmetric circumferential wave $A1$ the cut-off frequency is 132.43 for a radius ratio b/a equal to 0.95 .

VI. RESULTS AND DISCUSSION

The performance of ANFIS models for training and testing data sets were evaluated according to statistical criteria such as, coefficient of correlation R , MAE , MRE , SE , and root mean square error ($RMSE$). The selection of different models is done

comparing the errors of the ANFIS configuration, calculating the MAE, the MRE and the SE of the cut-off frequency. The coefficient of correlation R and the determination R^2 of the linear regression are used like performance measures of the model between the predicated and the desired output. The different error measures and the coefficient of correlation are given by the following relations:

$$MAE = \frac{1}{n} \sum_{i=1}^n |D_i - P_i| \quad (22)$$

$$MRE = \frac{1}{n} \sum_{i=1}^n \frac{|D_i - P_i|}{D_i} \quad (23)$$

$$R = 1 - \frac{\sum_{i=1}^n (D_i - P_i)^2}{\sum_{i=1}^n (D_i - P_m)^2} \quad (24)$$

$$SE = \frac{\sqrt{\sum_{i=1}^n (D_i - P_i)^2}}{n - 1} \quad (25)$$

where n is the number of data, P_i and D_i is the predicted and desired of cut-off frequency respectively and P_m is the mean of predicted values.

The coefficient of correlation is a commonly used statistic and provides information on the strength of linear relationship between the observed and the computed values. The training and testing performances of ANFIS models are given in figures 12 to 15.

The analysis is repeated several times. Indeed, the error values are measured for each ANFIS architecture based on the number of rules and the type of the membership function used. In this work we tried to play on the number of rules and the number of epochs we have observed that the error values of our models decrease more than the number of rules, and the number of epochs is increases. The results of the measured errors are presented in figures 12 to 15 for each circumferential wave (A1, S1, S2 and A2). Tables V to VI show that the results obtained by the fuzzy system method are in good agreement with those determined from the results calculated using the proper modes theory of resonances, and they are better to those determined manually from the time-frequency of Wigner-Ville images (Table V).

TABLE V. RESULTS OF THE CUT-OFF FREQUENCIES OF MODE A1 OBTAINED BY THE ANFIS MODEL, THE PROPER MODES THEORY AND BY THE TIME-FREQUENCY OF WIGNER-VILLE IMAGES

Cylindrical shell	Cut-off frequencies (ka) _c		
	Computed using PMT	Determined using ANFIS	Determined using SPWV
b/a=0.9 (figure 8)	66.21	66.16	66.0±0.3
b/a=0.95 (figure 7)	132.43	132.59	132.0±0.3
b/a=0.97 (figure 6)	220.72	221.32	221.0±0.2

TABLE VI. RESULTS OF THE CUT-OFF FREQUENCIES OF DIFFERENT MODES OBTAINED BY THE ANFIS MODEL AND THE PROPER MODES THEORY PMT FOR THE CYLINDRICAL SHELL

	(ka) _c computed and determined					
	ANFIS	PMT	ANFIS	PMT	ANFIS	PMT
	b/a=0.9		b/a=0.95		b/a=0.97	
Mode A1	66.16	66.21	132.59	132.43	221.32	220.72
Mode S1	132.32	132.43	265.17	264.87	446.64	441.45
Mode S2	136.16	136.28	272.87	272.56	459.60	454.26
Mode A2	198.51	198.65	398.49	397.30	668.95	662.17

The results of the different error measures and the coefficient of correlation (MRE, MAE, SE and R) are given in the table VII. And also are illustrated on the Figs. 12a to 15a. So, it is interest to use the approach of the Fuzzy Logic. The best configuration is found for a network with 13 rules. The predicted values are traced according to the desired values in the figures 12 to 15 ((a), (b), and (c)). The results show the good agreement between the predicted and the desired values of the cut-off frequency. The coefficient of determination R^2 for this optimal configuration is 1 (Figs. 12(a) to 15 (a)). Figs. 12 to 15 (a and b) show that the cut-off frequency increases rapidly when the radius ratio b/a of the cylindrical shell tends to one. The evolution of the mean quadratic errors (RMSE) of training during the training phase is illustrated on Fig. 16.

TABLE VII. RESULTS OF THE DIFFERENT ERROR MEASURES AND THE COEFFICIENT OF CORRELATION (MRE, MAE, SE AND R) WITH 13 RULES

Error measures	Mode A1	Mode S1	Mode S2	Mode A2
MAE	0.03 ka	0.08 ka	0.07 ka	0.07 ka
MRE	$0.8 \cdot 10^{-3}$ ka	$0.8 \cdot 10^{-3}$ ka	$0.8 \cdot 10^{-3}$ ka	$0.5 \cdot 10^{-3}$ ka
SE	$9 \cdot 10^{-3}$ ka	10^{-2} ka	$2 \cdot 10^{-2}$ ka	10^{-2} ka
R=R ²	1	1	1	1

VII. CONCLUSION

The main aim of this work was to train an ANFIS model to predict cut-off frequency with the minimum of input data. Results show that the trained model can be used as an alternative way in the modelling behaviour system. This fuzzy logic model taking into account some characteristics of the tube is developed in order to predict the cut-off frequency for various types of circumferential waves A₁, S₁, S₂, A₂. In this article, this model is applied to aluminum tubes. This model can be used to predict the evolution of the group and phase velocities according to the frequency. It also can constitute a help for the estimate of various parameters of a tube starting from the characteristics of which it is disposed.

The use of the fuzzy logic does not present any approximation as in the case of the natural modes method which assimilates the tubes to the plates with the same thickness and that is not sullied with errors as in the case of the time-frequency representations of Wigner-Ville that determines the cut-off dimensionless frequency manually

starting from the time-frequency image. This article can be used as a new tool for characterization of an elastic tube. This method allows one to determine automatically and with good

precision the reduced cut-off frequency of an antisymmetric wave propagating around the tube. The R^2 value in fig is about 1, which can be considered as very satisfactory.

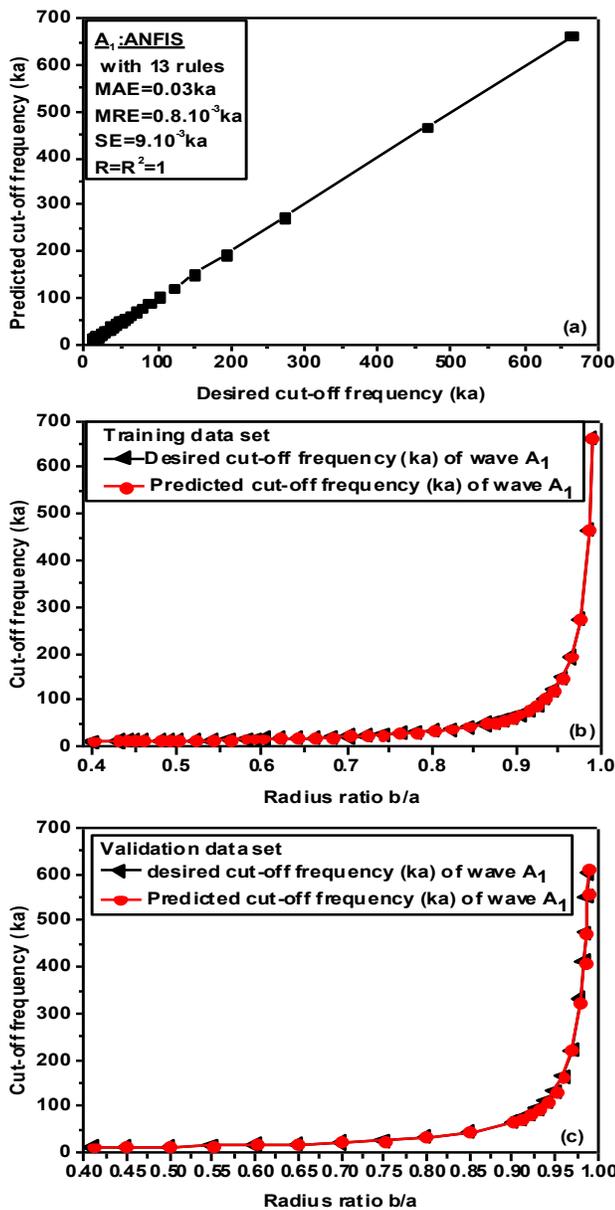


Fig.12. (a) Correlation of desired versus ANFIS values of cut-off frequency of anti-symmetric wave A_1 with validation data set, (b) Cut-off frequency as a function of radius ratio of aluminum cylindrical shell on training data set and (c) Cut-off frequency as a function of radius ratio of an aluminum cylindrical shell on validation data set

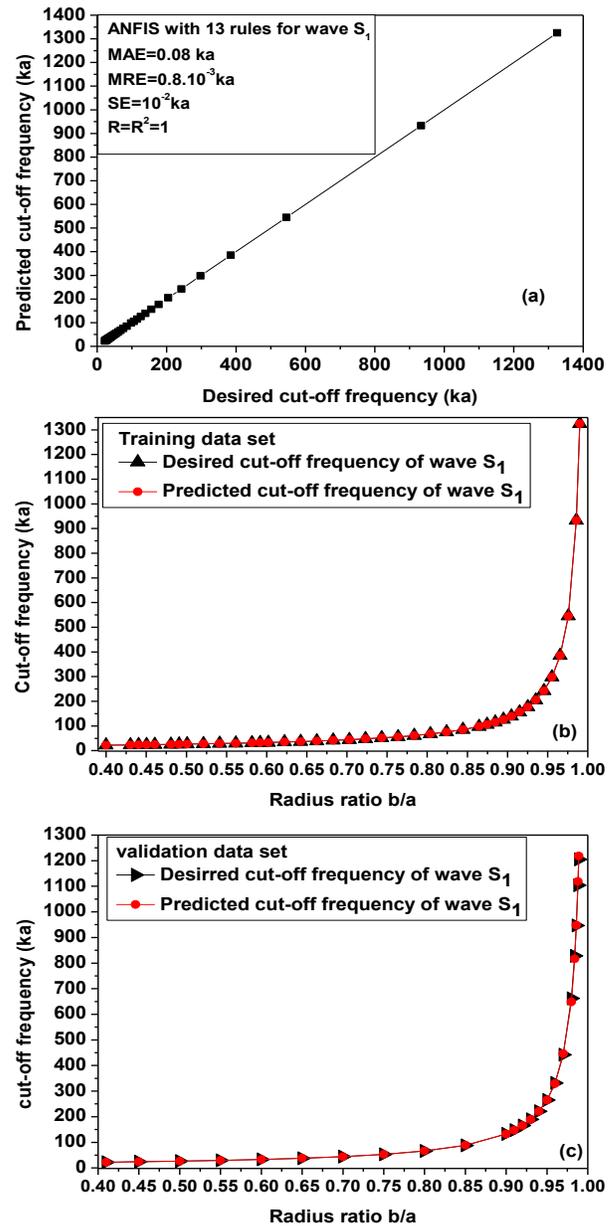


Fig.13. (a) Correlation of desired versus ANFIS values of cut-off frequency of Symmetric wave S_1 with validation data set, (b) Cut-off frequency as a function of radius ratio of aluminum cylindrical shell on training data set and (c) Cut-off frequency as a function of radius ratio of an aluminum cylindrical shell on validation data set

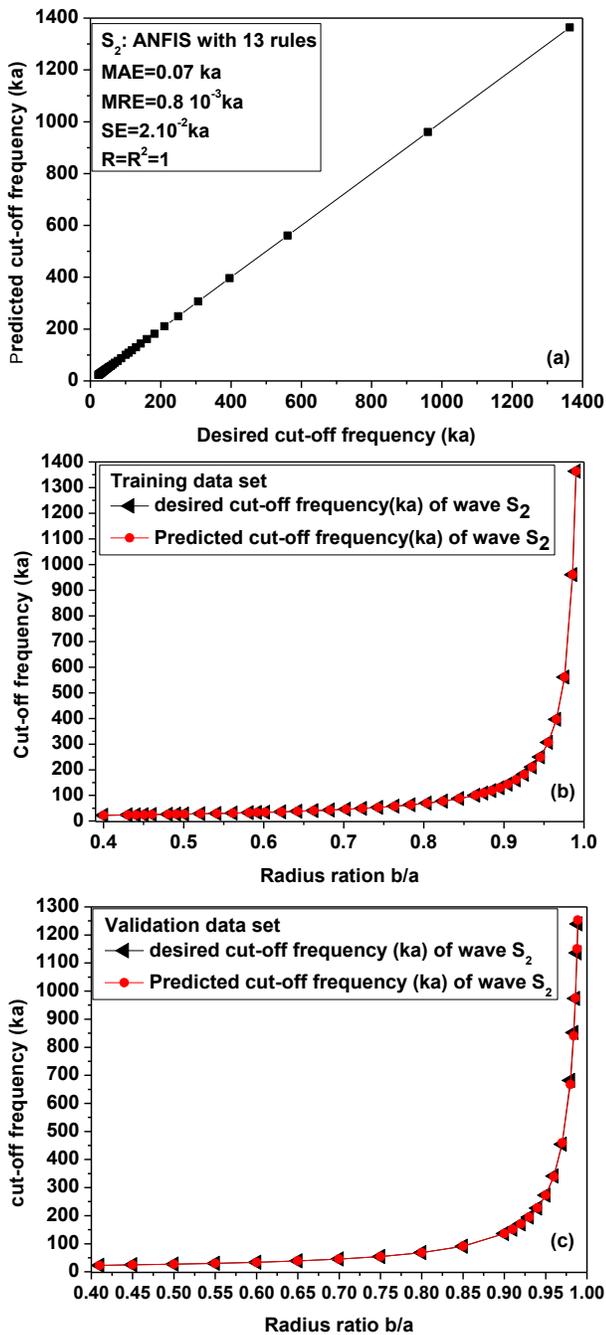


Fig.14. (a) Correlation of desired versus ANFIS values of cut-off frequency of Symmetric wave S_2 with validation data set, (b) Cut-off frequency as a function of radius ratio of aluminum cylindrical shell on training data set and (c) Cut-off frequency as a function of radius ratio of an aluminum cylindrical shell on validation data set

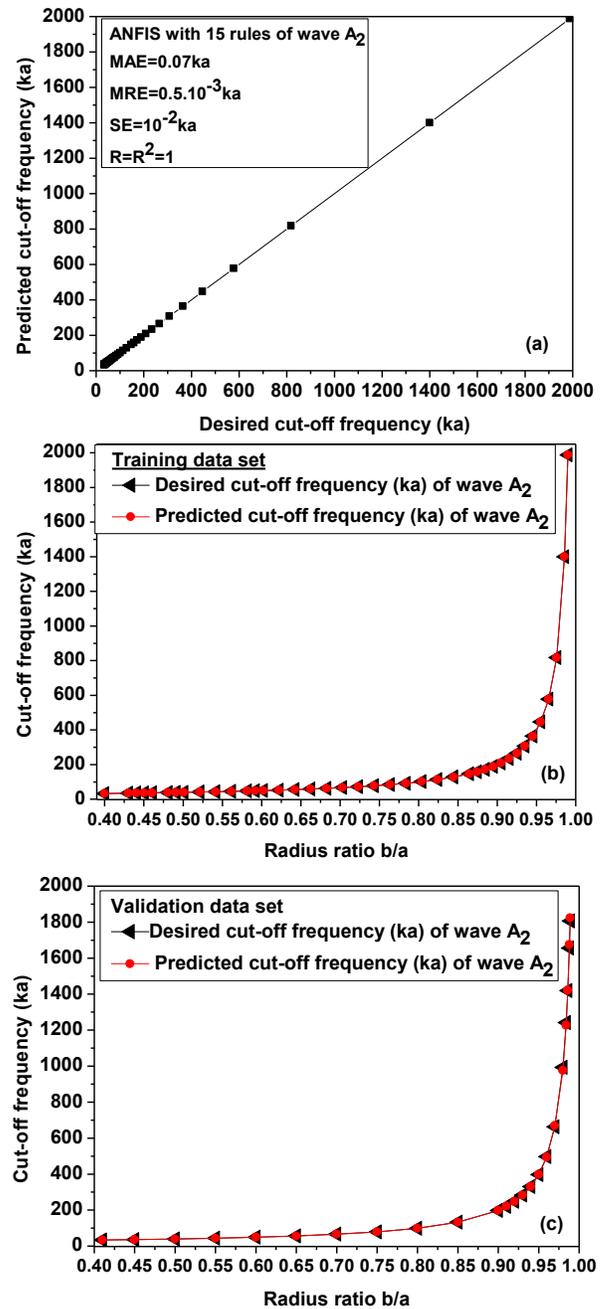


Fig. 15. (a) Correlation of desired versus ANFIS values of cut-off frequency of Anti-symmetric wave A_2 with validation data set, (b) Cut-off frequency as a function of radius ratio of aluminum cylindrical shell on training data set and (c) Cut-off frequency as a function of radius ratio of an aluminum cylindrical shell on validation data set

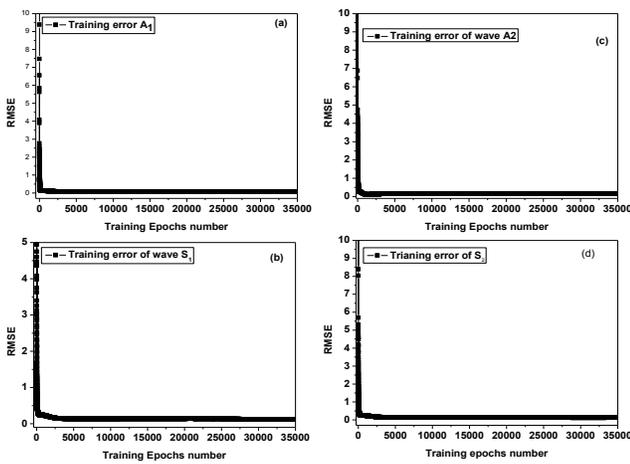


Fig.15. Visualization of errors of training and testing as a function of the number of iteration for an ANFIS to four entries and tree rules. Anti-symmetric and Symmetric waves (a) A1, (b) S1, (c) S2, (d) A2

VIII. REFERENCES

[1] R. Latif, E. Aassif, A. Moudden, D. Decultot, B. Faiz, and G. Maze, "Determination of the cut-off frequency of an acoustic circumferential wave using a time-frequency analysis," *J. NDT&E Int.*, vol. 33, pp. 373-376, 2000.

[2] R. Latif, E. Aassif, M. Laaboubi, G. Maze, "Détermination de l'épaisseur d'un tube élastique à partir de l'analyse temps-fréquence de Wigner-Ville", *Acta Acustica united with Acustica*, Vol. 95, pp. 253-257, (2009)

[3] M. Talmant, J. L. Izbecki, G. Maze, G. Quentin J. Ripoche. "External wave resonance on thin cylindrical shells". *J. Acoustique*, 4, pp. 509-523(1991).

[4] L. Haumesser, D. Décultot, F. Léon, and G. Maze, "Experimental identification of finite cylindrical shell vibration modes", *Journal of the Acoustical Society of America*, Vol. 111, 5, pp. 2034-2039, (2002)

[5] J. D. N. Cheeke, X. Li, and Z. Wang, "Observation of flexural Lamb waves (A0 mode) on water-filled cylindrical shells", *Journal of the Acoustical Society of America*, Vol. 104, 6, pp. 3678-3680, (1998).

[6] P. L. Marston and N. H. Sun, "Backscattering near the coincidence frequency of a thin cylindrical shell: Surface wave properties from elastic theory and an approximate ray synthesis," *J. Acoust. Soc. Amer.*, vol. 97, pp. 777-783, 1995.

[7] G. Maze, "Acoustic scattering from submerged cylinders. MIIR Im/Re: Experimental and theoretical study," *J. Acoust. Soc. Amer.*, vol. 89, pp. 2559-2566, 1991.

[8] R. Latif, E. Aassif, G. Maze, A. Moudden, B. Faiz, "Determination of the group and phase velocities from time-frequency representation of Wigner-Ville", *Journal of Non Destructive Testing & Evaluation International*, Vol.32, 7, pp. 415-422, (1999)

[9] G. Kaduchak, C. S. Kwiatkowsi, and P. L. Marston, "Measurement and interpretation of the impulse response for backscattering by a thin spherical shell using a broad bandwidth source that is nearly acoustically transparent," *J. Acoust. Soc. Amer.*, vol. 97, pp. 2699-2708, 1997.

[10] P. Flandrin, *Temps-fréquence*. Paris: Hermès, 1993.

[11] E. Aassif, R. Latif, D. Decultot, G. Maze, B. Faiz, and A. Moudden, "Time-frequency analysis of the complex pressure scattered by immersed tubes," in *3rd Int. Conf., Acoust. Vibratory Surveillance Methods Diagnostic Techniques, Centre Technique des Industries Mécaniques, CETIM-Senlis, France*, Oct. 13-15, 1998, pp. 471-480.

[12] R. Latif, E. H. Aassif, A. Moudden, and G. Maze, "Analyse des caractéristiques acoustiques d'une plaque élastique immergée dans l'eau à partir de l'image temps-fréquence," *Acta Acustica/ Acustica*, vol. 92, pp. 549-555, 2006.

[13] S. F. Morse and P. L. Marston, "Backscattering of transients by tilted truncated cylindrical shells: Time-frequency identification of ray contributions from measurements," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1289-1294, 2002.

[14] R. LATIF, E. AASSIF, G. MAZE, D. DECULTOT, A. MOUDDEN, B. FAIZ, "Analysis of the circumferential acoustic waves backscattered by a tube using the time-frequency representation of Wigner-Ville", *Journal of Measurement Science and Technology*, Vol. 11, 1, pp. 83-88, (2000).

[15] R. LATIF, E. AASSIF, A. MOUDDEN, B. FAIZ, "Caractérisation ultrasonore d'un matériau élastique à partir de l'analyse de l'image temps-fréquence de Wigner-Ville", *Acta Acustica united with Acustica*, Vol. 89, pp. 253-257, (2003).

[16] D. H. Hughes and P. L. Marston, "Local temporal variance of Wigner's distribution function as a spectroscopic observable: Lamb wave resonances of a spherical shell," *J. Acoust. Soc. Amer.*, vol. 94, pp. 499-505, 1993.

[17] R. Latif, E. Aassif, M. Laaboubi, G. Maze, "Determination of group velocity using the time-varying Autoregressive TVAR", *Conference international IEEE ISSPA, Sharjah UAE (2007)*.

[18] R. Latif, M. Laaboubi, E. Aassif, A. Moudden, G. Maze, "Determination of the cut-off frequency of the anti-symmetric circumferential waves using the time-varying autoregressive TVAR", *3rd International Symposium on Communications, Control and Signal Processing IEEE ISCCSP'08*, pp. 357-361, St Julians Malta, 12-14 March (2008)

[19] T. Onsay and A. G. Haddow, "Wavelet transform analysis of transient wave propagation in a dispersive medium", *J. Acoust. Soc. Am.* Vol. 95, pp. 1441-1449, (1994)

[20] K. Kishimoto, H. Inoue, M. Hamada and T. Shibuya, "Time-frequency analysis of a dispersive waves by means of wavelet transform", *ASME J. Appl. Mech.* Vol. 62, pp. 841-846, (1995)

[21] A. Dariouchy, E. Aassif, G. Maze, R. Latif, D. Decultot, M. Laaboubi, "Prediction of the Acoustic Pressure Backscattered by a Steel Tube Using Neural Networks Approach", *International Symposium on Computational Intelligence and Intelligent Informatics ISCIII'07, Agadir Morocco*, 28-30 March (2007)

[22] A. Dariouchy, E. Aassif, G. Maze, D. Décultot, A. Moudden, "Prediction of the acoustic form function by neural network techniques for immersed tubes", *J. Acoust. Soc. Am.* Vol. 124, 2, pp. 1018-1025 (2008)

[23] G. Maze, J. Ripoche, A. Derem, J. L. Rousselot, "Diffusion d'une onde ultrasonore par des tubes remplis d'air immergés dans l'eau", *Acustica*, vol. 55, pp. 69-85, (1984).

[24] Younho Cho, "Estimation of ultrasonic guided wave mode conversion in a plate with thickness variation", *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions*, Vol. 47, 3, pp. 591-603, (2000).

[25] Atkinson D., Hayward G., "Embedded acoustic fibre wave guides for Lamb wave condition monitoring", *Ultrasonics Symposium, IEEE*, Vol. 1, pp. 699-702, (1999).

[26] Jang J-SR. ANFIS: Adaptive-network-based fuzzy inference systems. *IEEE Trans Syst Man Cybern* 1993; 23(3):665-85.

[27] S. Horikawa, T. Furuhashi, and Y. Uchikawa. *IEEE Trans. Neural Networks* 3:801-806, 1992.

[28] H. Ishibuchi, R. Fujioka, and H. Tanaka. *IEEE Trans. Fuzzy Systems* 1:85-97, 1993.

[29] Lee CC. Fuzzy logic in control system: Fuzzy logic controller—part I and part II. *IEEE Trans Syst Man Cybern* 1990; 20(2):404-35.

[30] Sugeno M. Industrial applications of fuzzy control. New York: Elsevier Ltd.; 1985.

The quest towards a winning Enterprise 2.0 collaboration technology adoption strategy

Robert Louw, Jabu Mtsweni

School of Computing, University of South Africa, Pretoria, 0001

Abstract—Although Enterprise 2.0 collaboration technologies present enterprises with a significant amount of business benefits; enterprises are still facing challenges in promoting and sustaining end-user adoption. The purpose of this paper is to provide a systematic review on Enterprise 2.0 collaboration technology adoption models, challenges, as well as to provide emerging statistic approaches that purport to address these challenges.

The paper will present four critical Enterprise 2.0 adoption elements that need to form part of an Enterprise 2.0 collaboration technology adoption strategy. The four critical elements were derived from the ‘SHARE 2013 for business users’ conference conducted in Johannesburg, South Africa 2013, as well as a review of the existing literature. The four adoption elements include enterprise strategic alignment, adoption strategy, governance, and communication, training and support.

The four critical Enterprise 2.0 adoption elements will allow enterprises to ensure strategic alignment between the chosen Enterprise 2.0 collaboration technology toolset and the chosen business strategies. In addition by reviewing and selecting an appropriate adoption strategy that incorporates governance, communication and a training and support system, the enterprise can improve its ability towards a successful Enterprise 2.0 adoption campaign.

Keywords--Web 2.0; Enterprise 2.0; collaboration; technology adoption; adoption strategy; critical adoption elements

I. INTRODUCTION

Web 2.0 technologies have made significant advancements in providing users with the tools required to adopt and promote a culture of enterprise collaboration. Compared to its predecessor, Web 1.0, Web 2.0 represents a paradigm shift in the way in which people share, contribute and distribute content [21],[11].

The term ‘Web 2.0’, is generally used interchangeably with the term ‘Enterprise 2.0’ [3]. However, there is a clear distinction between the two terms. Ramirez-Medina [16], states that the term ‘Enterprise 2.0’ is the application of Web 2.0 technologies within the enterprise environment, in order to allow employees to collaborate, share ideas, communicate and generate content. The term ‘collaboration’ within the Enterprise 2.0 context, can be defined as a process whereby two or more individuals, groups or enterprises work together to achieve a common goal [9]. Although enterprises are increasingly investing in Enterprise 2.0 collaboration technology toolsets to facilitate knowledge sharing, enterprise communication and collaboration, many enterprises are still facing significant

challenges pertaining to end-user adoption. The adoption process is often faced with end-user resistance resulting in a lengthy adoption process.

A market research conducted by the Association for Information and Image Management (AIIM) in 2009 on enterprises operating within United States, Canada, United Kingdom, Ireland and Europe concluded that 50% of enterprises were unable to justify a return on investment (ROI) in Enterprise 2.0 collaboration technology tools, 43% lacked a full understanding of the capabilities of Enterprise 2.0 collaboration technologies, and 40% identified corporate culture as an major stumbling block [6].

AIIM conducted a follow up market research survey in 2011 on enterprises operating within North America and Europe, in which 451 of their AIIM community network members responded. Their research findings concluded that reluctance of staff to contribute is one of the major barriers towards Enterprise 2.0 collaboration technologies adoption. Secondly, a lack of top management participation had increased from 26% in 2010 to 36% in 2011 [5].

Against this background, the focus of this paper is to identify the critical adoption elements required in order to formulate a successful Enterprise 2.0 collaboration technology adoption strategy. The remainder of this paper consists of three sections. The first section entitled: ‘State of the art’, presents an overview of the existing literature, including Enterprise 2.0 collaboration toolsets, adoption models as well as a review on existing Enterprise 2.0 collaboration technology adoption challenges. The ‘Discussion’ section, presents a comparison overview between the various Enterprise 2.0 adoption models as well as adoption strategies. The remainder of this section presents the findings of the ‘SHARE 2013 for business users’ conference, conducted in Johannesburg, South Africa 2013, in which the findings are expressed into four critical Enterprise 2.0 adoption elements. Lastly, a conclusion is presented as well as proposed future research projects.

II. STATE OF THE ART

A. Enterprise 2.0

McAfee (2006) [4], was the first to coin the term “Enterprise 2.0” defining it as “the use of emergent social software platforms within companies, or between companies and customers”. Based on this definition, Enterprise 2.0 can be regarded as a platform of services that can be applied within and outside the enterprise environment in order to stimulate enterprise collaboration.

Enterprise 2.0 allows enterprises to leverage Web 2.0 technologies to harness collective intelligence through participation. In addition Enterprise 2.0 collaboration technology toolsets, present significant benefits to an enterprise, by fostering collaboration between employees, suppliers, partners and customers and ultimately contributing towards enterprise intellectual capital and knowledge [2].

B. Enterprise 2.0 collaboration toolsets

A number of Enterprise 2.0 collaboration technology toolsets exist within the market. Gartner annually produces an Enterprise Content Management (ECM) magic quadrant analysis of Enterprise 2.0 collaboration technology toolsets. The magic quadrant analysis consists of four quadrants as depicted in Figure 1 and these include:

Leaders. Leaders refer to vendors who have established themselves as market leaders within a selected market space. Leaders can be described as vendors who consistently achieve financial performance and growth. In essence, they can be described as the best-of-breed within a selected market space.

Challengers. Challengers offer good functionality, however they still lack the vision and execution ability of vendors within the leaders quadrant.

Visionaries. Visionaries offer similar capabilities as other vendor leader toolsets; however, they have less ability to execute than vendors operating within the leader and challengers quadrants.

Niche players. Niche players typically focus on specific elements of enterprise content management technology toolsets. This quadrant generally includes vendors still maturing their enterprise content management toolset.

Gartner identifies the following Enterprise 2.0 collaboration technology toolset leaders, they include IBM WebSphere, Oracle WebCenter, Microsoft SharePoint, OpenText, EMC, OpenText and Hyland Software. Figure 1 depicts the latest Granter Enterprise Content Management magic quadrant analysis conducted in 2012 [18]. The leaders are briefly described in Table 1.

TABLE I. GARTNER ENTERPRISE CONTENT MANAGEMENT LEADERS (2012) [18]

Enterprise 2.0 collaboration technology toolset	Toolset overview
IBM WebSphere	The IBM WebSphere Portal Enterprise 2.0 collaboration toolset was one of the first collaboration toolsets to enter the market. A number of large enterprises have invested in the IBM WebSphere toolset due to it is highly scalability nature, especially around other IBM toolsets.
Oracle Web Center	The Oracle WebCenter collaboration toolset embeds a number of Web 2.0 collaboration technology tools such as content management, enterprise search, and social software collaboration and communication services. The biggest differentiator of the Oracle WebCenter collaboration toolset, is Oracles commitment to a highly Software Oriented Architecture (SOA) solutions.
Microsoft SharePoint	The latest version of Microsoft SharePoint, Microsoft SharePoint 2013 encapsulates a number of Web 2.0 technologies, allowing knowledge workers to create, collect, organize and collaborate on various forms of content within a web-based environment.
OpenText	OpenText are regarded as the leaders within the Enterprise Information Management (EIM) market space. There toolsets are highly optimized for content management and content searching. However, they lack the social and collaboration elements compared to the other toolsets within the leader’s quadrant.
EMC	EMC have focused their research and development efforts in providing a cloud based content management solution, known as EMC OnDemand. The EMC OnDemand service allows enterprises to conduct end-to-end content management without investing in any infrastructure.
Hyland Software	Hyland software mostly provides services to medium size enterprise customers in North and South America. The biggest differentiator of the Hyland software collaboration toolset is its ability to integrate with other Information systems.



Source: Gartner (October 2012)

Fig. 1. Granter Enterprise Content Management magic quadrant (2012)

C. Technology adoption models

Enterprise 2.0 collaboration technologies require user acceptance and participation to be successful [13]. It is therefore important to conduct a review of the existing adoption theories and models.

The ‘diffusion of innovations’ theory first proposed by Rogers (2003) [8] is highly regarded as one of the more popular technology adoption theories. The ‘diffusion of innovations’ theory comprises four main elements that either promote individual and enterprise acceptance or rejection towards a technology toolset.

The first element ‘innovation’ refers to the perceived newness characteristics of a technology toolset, the prospects

of new benefits towards the individual and enterprise. The second element 'communication channels' is the process whereby participants generate and share content with one another to achieve a mutual understanding. The third element 'time' relates to the rate at which individuals and enterprises adopt a technology toolset. Lastly, the fourth element 'social system', can be described as a set of interrelated units that encourage a joint problem-solving culture to accomplish a common goal.

According to Rogers [8], the innovation-decision process can be described as "an information-seeking and information-processing activity, where an individual is motivated to reduce uncertainty about the advantages and disadvantages of an innovation". Figure 2 depicts the innovation-decision process, which consists of five sequential steps, namely knowledge, persuasion, decision, implementation and confirmation.

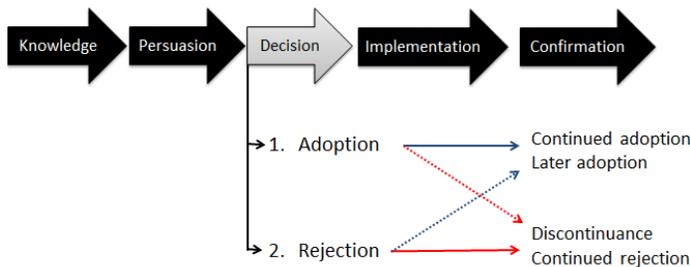


Fig. 2. Five stages of the Innovation-Decision Process [8]

- **Knowledge.** Within the knowledge stage, individuals address the question. What innovation is and how it works? The knowledge gained helps motivate individuals to learn more about the innovation, and thereby promoting adoption.
- **Persuasion.** Within the persuasion stage, the individual forms either a positive or a negative attitude towards the innovation. The individual forms his or her attitude towards the innovation based on the knowledge gained.
- **Decision.** Within the decision stage, the individual chooses either to adopt or reject the innovation. The individual may make a decision to continue to adopt the innovation or discontinuance the decision to adopt the innovation, implying to reject the innovation after adopting it. The individual may also decide to continue to reject the innovation, or to adopt the innovation at a later stage.
- **Implementation.** During the implementation stage the innovation is put into practice. Innovation brings about change, thus the implementation stage has some degree of uncertainty. It is important that during this stage, the implementer makes use of technical assistance in order to bring about change in the enterprise.
- **Confirmation.** Within the confirmation stage, the individual seeks support based on his or her decision. Depending on the support provided towards adoption, the innovation may lead to continued adoption, or discountenance of the innovation.

Another well-known technology adoption model, the Technology Acceptance Model (TAM), was first developed by Davis [12] in 1986 and has been extensively studied in terms of information system (IS) adoption. TAM adopts two primary perspectives towards the use of new technology, namely perceived usefulness and perceived ease of use. The TAM model is based on the assumption that the easier the technology is to use, the greater the acceptance and use of the technology will be [12]. The TAM model was later extended to include two additional perspectives, the social influence process and the cognitive instrument process, which could also influence the perceived usefulness of technology [23].

Although the technology acceptance model addressed the perceived usefulness and ease of use of a technology toolset, it did not address the benefits and costs associated in investing in a technology toolset. The value-added model (VAM) does however address these two elements. VAM is based on the cost-benefit trade-off approach, which weighs the perceived benefits against the costs of achieving those benefits [14].

Research conducted on the VAM model concludes that if the perceived benefits of Enterprise 2.0 collaboration technologies outweigh the costs (i.e. financial investment, risks/information leakage, loss of control of the system, ethical issues, etc.), there will be a positive attitude towards adopting Enterprise 2.0 technologies [22].

Although the technology adoption models presented above have been applied and tested during the last few decades, in a number of Information System (IS) selection processes as well as implementations, addressing elements such as perceived ease of use, identifying the underlying costs and benefits, identifying end-user and enterprise attitude towards technology acceptance or rejection. They do not address the end-user motivation elements required to sustain Enterprise 2.0 technology adoption within an enterprise.

An Enterprise 2.0 collaboration technology adoption strategy requires a well-defined governance framework, which should be aligned and be supportive of the enterprises underlying business strategy. In addition, the technology adoption models presented above do not address communication, training and support frameworks required to assist end-users to transition towards Enterprise 2.0 collaboration technology adoption.

Based on this background, it is important to review the challenges currently experienced by enterprises, when adopting Enterprise 2.0 collaboration technologies, prior to formulating a conclusion on the critical adoption elements required within an Enterprise 2.0 collaboration technology adoption strategy.

D. Enterprise 2.0 adoption challenges

The challenges associated with the adoption and promotion of Enterprise 2.0 collaboration technologies can be grouped in terms of either technological or organisational challenges. Table 2 provides a review of the existing literature; suggesting that Enterprise 2.0 collaboration technology adoption challenges can be grouped into the following five categories:

TABLE II. ENTERPRISE 2.0 COLLABORATION TECHNOLOGY ADOPTION CHALLENGE CATEGORIES

Enterprise 2.0 collaboration technology adoption challenge category	Challenge overview
Enterprise change	Users have established repetitive routines in using certain technologies on a daily basis, for example electronic email, and find it difficult to change or adapt to new forms and ways of using technology. Enterprise 2.0 collaboration technologies require a radical change within the work environment, organisational structures and business processes [17], [4].
Enterprise culture	Culture plays a significant role in technology adoption. Enterprise 2.0 collaboration technologies require a culture that promotes innovation, collaboration and participation [17], [7].
Technology interest	If there is no clear vision or strategic direction in terms of why a new type of technology should be used, it will lead to a low adoption rate. The vision, goals and benefits of Enterprise 2.0 collaboration technologies need to be communicated and clearly understood by all enterprise end-users [17], [7].
Technology complexity	Often end-users are faced with technology complexities, such as information overload, lack of user interface consistency resulting in cognitive constraints. In some cases technology complexity is as a result of poor technology design, however in most cases, it is as a result of a lack of user awareness and training. A business and technical support structure needs to be available to address end-user concerns and suggestions [20].
Enterprise security	Information security and intellectual capital protection is vital to any enterprise. In addition, any technology that could expose an organization to vulnerability or loss of information might be disregarded or restricted. This contradicts the very nature of Enterprise 2.0 collaboration technologies, which promote information sharing and social collaboration [10], [17].

The five adoption challenge categories suggest that an Enterprise 2.0 collaboration technology adoption strategy should incorporate a governance framework that addresses roles and responsibilities, in order to define ownership and acceptable usage.

In addition, a governance framework should also incorporate an effective change management process, in order to facilitate change within the selected enterprise. The chosen change management process needs to be conducive towards the enterprises underlying culture in order to be effective.

In addition to a well-structured governance framework, an effective communication plan, training and support structure is required. As with most Information System (IS), Enterprise 2.0 collaboration technology toolsets also require functionality, process and procedural information to be communicated to the target end-user community. The end-user community needs to be informed and made aware of the values as well as the guiding principles in the utilization of a selected Enterprise 2.0 collaboration technology toolset.

III. DISCUSSION

Having reviewed and examined the Enterprise 2.0 collaboration technology adoption model as well as challenges previously studied. The next section presents an alternative perspective towards formulating an Enterprise 2.0 collaboration technology adoption strategy.

The perspective is based on a systematic review of existing literature as well as the content and views expressed by experts at the ‘SHARE 2013 for business users’ conference in Johannesburg, South Africa 2013. Based on the information gathered, the research suggest that an Enterprise 2.0 collaboration technology adoption strategy should incorporate the following four critical adoption elements, as depicted in Figure 3. The four critical adoption elements will be described in the following four sections.

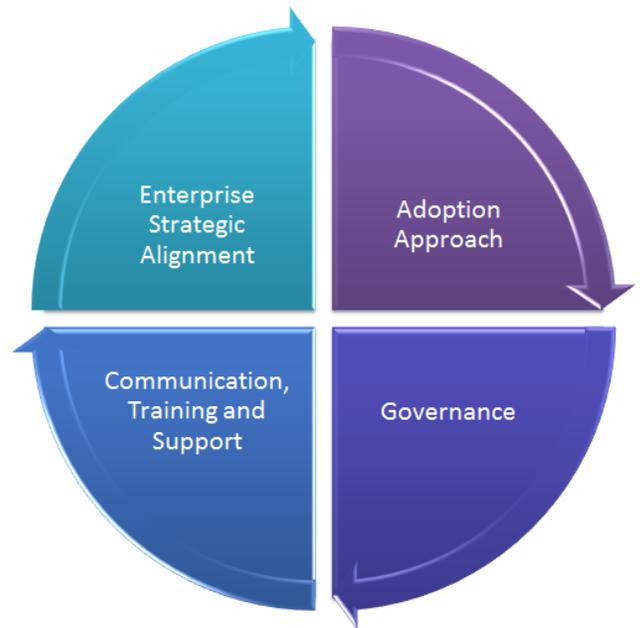


Fig. 3. Four critical Enterprise 2.0 collaboration technology adoption elements

A. Enterprise strategic alignment

As with most enterprise investments, either in people or technology, the investment needs to compliment the enterprises chosen strategic direction. Only once the enterprise strategic direction and vision is well known, can an Enterprise 2.0 collaboration technology toolset be selected. The strategic vision should be translated into business drivers, which in turn serve as business functional requirements.

In addition, an information architecture assessment needs to be performed. The information architecture should address the type of content and relationships between content that needs to be hosted and distributed by an Enterprise 2.0 collaboration technology toolset.

Once the information architecture and business functional requirements have been established, an Enterprise 2.0 collaboration technology toolset gap analysis needs to be conducted. This will assist in selecting an appropriate toolset

that can address the information architecture as well as business functional requirements [20], [15].

B. Adoption approach

A review of the existing literature suggests that the best path towards Enterprise 2.0 collaboration technology adoption is to adopt a hybrid approach. The top-down element, provides guidance, support and adherence towards the strategic objectives, while the bottom-up element allows for autonomy to explore and create content, thus improving participation.

The adoption approach needs to compliment the enterprises underlying culture. In addition, a hybrid adoption approach can assist in bringing about change within the enterprise, vital towards sustaining end-user participation [1], [2].

C. Governance

As with most information systems, Enterprise 2.0 collaboration technologies require governance. An Enterprise 2.0 collaboration technology governance framework needs to be established and maintained. The governance framework needs to compliment the enterprises strategic objectives as well as clearly define the roles and responsibilities in relation to participation.

In addition, the governance framework needs to incorporate a clear decision making authority. The decision making authority should formulate the Enterprise 2.0 collaboration technology roadmap, training and communication program as well as promote end-user participation.

Although a governance framework is vital towards a successful Enterprise 2.0 collaboration technology adoption strategy, it should not be a barrier towards end-user participation. De Hertogh et al. [19], suggests that a governance framework should also incorporate the following four grounding principles:

- **The empowerment principle.** End-users should be given sufficient autonomy to explore and master Enterprise 2.0 collaboration technology toolsets. The novelty of Enterprise 2.0 collaboration technologies sparks the curiosity and enthusiasm of end-users to adopt the technology toolset.
- **The processes principle.** Enterprise 2.0 collaboration technologies present enterprises with the ability to improve on or rather automate certain business process elements. End-users should be granted sufficient autonomy to exploit these business benefits.
- **The collaboration principle.** Top- and middle management should be wary of limiting too much access as this will have a direct impact on end-users ability to contribute and distribute content for collaboration purposes.
- **The people and culture principle.** Continuously stimulate, guide and convince potential participants of the business value of Enterprise 2.0 collaboration technologies. Training and awareness should form a critical element of the chosen governance strategy and implementation plan.

D. Communication, training and support

As with most enterprise information systems, in order to gain participation, end-user awareness and support structures are required. It is important to address the 'What is in it for me?' question when establishing end-user awareness. The more exposure end-users gain from the chosen Enterprise 2.0 collaboration technology toolset, pertaining to its capabilities, the more likely effective end-user adoption will occur.

An Enterprise 2.0 collaboration technology adoption strategy should also incorporate a formal communication plan. The communication plan needs to address the frequency of communication, type of content and end-user audience who needs to be informed.

In addition, a training and support structure needs to be established. The training program needs to incorporate both online training content as well as workshop training sessions to allow for questions and answers not addressed by the available online or printed training content. [20].

The four critical adoption elements could allow enterprises to facilitate change towards adoption, as well as assist in gaining and sustaining end-user participation. Moreover, the elements should help guide the underlying Enterprise 2.0 collaboration technology implementation team as well as toolset supporting teams in formulating a communication plan, governance framework, training plan and acceptable usage policies and procedures.

IV. CONCLUSION

This paper presented a systematic review of the existing literature pertaining to Enterprise 2.0 collaboration technology adoption models as well as the underlying Enterprise 2.0 collaboration technology adoption challenges.

Although a number of technology adoption models have been studied during the last few decades, in relation to end-user acceptance and participation of Information systems (IS) including Enterprise 2.0 collaboration technologies. The technology adoption models reviewed, the Diffusion of innovations theory, Technology Acceptance Model (TAM) and the Value-added model (VAM) do not address the motivation and sustainability elements required by an Enterprise 2.0 collaboration technology toolset.

An alternative perspective towards formulating an Enterprise 2.0 collaboration technology adoption strategy was presented. In which four critical adoption elements were suggested. The four critical adoption elements include enterprise strategic alignment, adoption strategy, governance, and communication, training and support, which should form part of any Enterprise 2.0 collaboration technology adoption strategy.

The four critical adoption elements were derived based on a systematic review of the existing literature as well as interviews conducted with leading experts within the Enterprise 2.0 collaboration field, who presented at the SHARE 2013 for business users' conference conducted in Johannesburg, South Africa 2013.

The experts interviewed have been exposed to a number of Enterprise 2.0 collaboration technology implementation projects as well as assisting enterprises in North America and South Africa, in formulating their underlying adoption strategies. The findings were analyzed based on interview notes as well as literature content from the conference and available academic repositories.

Although the four critical adoption elements could greatly facilitate end-user adoption, future research is required in order to assess the extent to which these four critical adoption elements should be encapsulated into an adoption strategy.

REFERENCES

- [1] A. Barron and D. Schneckenberg, "A theoretical framework for exploring the influence of national culture on Web 2.0 adoption in corporate contexts.", *The Electronic Journal Information Systems Education*, Vol. 15(2), pp. 176-186, 2012
- [2] A. Bruno, P. Marra and L. Mangia, "The Enterprise 2.0 adoption process: a participatory design approach.", *Advanced Communication Technology (ICACT)*, 2011 13th International Conference on. IEEE pp. 1457-1461, 2011
- [3] A.P. McAfee, "Enterprise 2.0: New collaborative tools for your organization's toughest challenges.", Harvard Business School Press, 2009
- [4] A.P. McAfee, "Enterprise 2.0: The Dawn of Emergent Collaboration.", *Management of Technology and Innovation*, Vol. 47(3), 2006
- [5] D. Miles, "Social Business Systems - success factors for Enterprise 2.0 applications.", *AIIM Industry Watch Report*, pp. 1-25, 2011
- [6] D. Miles, "Collaboration and Enterprise 2.0: Work-meets-play or the future of business?", *AIIM Industry Watch Report*, pp. 1-30, 2009
- [7] D. Riedl and F. Betz, "Intranet 2.0 Based Knowledge Production An Exploratory Case Study on Barriers for Social Software.", *eKNOW 2012: The Fourth International Conference on Information, Process, and Knowledge Management*, 2012
- [8] E.M. Rogers, "Diffusion of innovations - 5th edition.", New York: Free Press, 2003
- [9] E. Turban, T. Liang and S.P.J. Wu, "A framework for adopting collaboration 2.0 tools for virtual group decision making." *Group decision and negotiation*, Vol. 20(2), pp. 137-154, 2011
- [10] F. Almeida, "Web 2.0 Technologies and Social Networking Security Fears in Enterprises.", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3(2), pp. 152-156, 2012
- [11] F. Fuchs-Kittowski, N. Klassen, D. Faust and J. Einhaus, "A Comparative Study on the Use of Web 2.0 in Enterprises.", *Proceedings 9th International Conference on Knowledge Management and Knowledge Technologies*, Graz, pp. 372-378, 2009
- [12] F.D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology.", *MIS Quarterly*, pp. 319-340, 1989
- [13] F.H. Alqahtani, J. Watson and H. Partridge, "Users adoption of web 2.0 for knowledge management: position paper.", *Proceedings of the International Conference on Information Management and Evaluation*. Academic Publishing Limited, pp. 19-29, 2010
- [14] H.W. Kim, H.C. Chan and S. Gupta, "Value-based adoption of mobile internet: an empirical investigation.", *Decision Support Systems*, Vol. 43(1), pp. 111-126, 2007
- [15] J. Willinger, "Enterprise 2.0 and SharePoint: What's the buzz about?", Paper presented at the SHARE 2013 for business users conference, South Africa. Available at: <http://www.shareconference.com/za> [Accessed on: 2 May 2013]
- [16] J.A. Ramirez-Medina, "Enterprise 2.0 Readiness Index." *Management of Engineering & Technology*, 2009. PICMET 2009. Portland International Conference on. IEEE, pp. 2677-2684, 2009
- [17] M.H. bin Husin and P.M. Swatman, "Removing the barriers to Enterprise 2.0.", *Technology and Society (ISTAS)*, 2010 IEEE International Symposium on. IEEE, 2010, pp. 275-283, 2010
- [18] R. Gilbert, K. Shengda, K. Chin, G. Tay and H. Koehler-kruener, "Magic Quadrant for Enterprise Content Management." 2012. Available at <http://www.gartner.com/technology/reprints.do?id=1-1CKSZ07&ct=121021&st=sg> [Accessed on 10 February 2013]
- [19] S. De Hertogh, S. Viaene and G. Dedene, "Governing Web 2.0.", *Communications of the ACM*, Vol. 54(3), pp. 124-130, 2011
- [20] S. Hanley, "SharePoint Governance – Love it or hate it.", Paper presented at the SHARE 2013 for business users conference, South Africa. Available at: <http://www.shareconference.com/za>
- [21] [Last accessed: 2 May 2013]
- [22] S. Murugesan, "Understanding Web 2.0.", *IT Professional*, Vol. 9(4), pp. 34-41, 2007
- [23] T.C. Lin, C.L. Lee and J.C. Lin, "Determinants of Enterprise 2.0 adoption: A value-based adoption model approach.", *Information Society (i-Society)*, 2010 International Conference on. IEEE, pp. 12-18, 2010
- [24] V. Venkatesh and F.D. Davis, "A theoretical extension of the technology acceptance model: four longitudinal field studies.", *Management science*, Vol. 46(2), pp. 186-204, 2000

AUTHORS PROFILE

Robert Louw is a part time student at the University of South Africa (UNISA). Robert is currently studying towards a Master of Science degree in Computing. His research interests are in collaboration technologies, adoption frameworks and enterprise architectures.

Dr. Jabu Mtsweni is a Senior Lecturer and Researcher at the University of South Africa (UNISA). He received his PhD in Computer Science. His main research interests are in Internet technologies ranging from Web technologies, Web services, and Cloud computing.

Face Recognition System Based on Different Artificial Neural Networks Models and Training Algorithms

Omaima N. A. AL-Allaf

Assistant Professor,
Dept. of CIS, Faculty of Sciences &
IT, Al-Zaytoonah University of
Jordan, P.O. Box130, Amman
(11733), Jordan

Abdelfatah Aref Tamimi

Associate Professor,
Dept. of CS, Faculty of Sciences &
IT, Al-Zaytoonah University of
Jordan, P.O. Box130, Amman
(11733), Jordan

Mohammad A. Alia

Assistant Professor,
Dept. of CIS, Faculty of Sciences &
IT, Al-Zaytoonah University of
Jordan, P.O. Box130, Amman
(11733), Jordan

Abstract— Face recognition is one of the biometric methods that is used to identify any given face image using the main features of this face. In this research, a face recognition system was suggested based on four Artificial Neural Network (ANN) models separately: feed forward backpropagation neural network (FFBPNN), cascade forward backpropagation neural network (CFBPNN), function fitting neural network (FitNet) and pattern recognition neural network (PatternNet). Each model was constructed separately with 7 layers (input layer, 5 hidden layers each with 15 hidden units and output layer). Six ANN training algorithms (TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAINGD, and TRAINGD) were used to train each model separately. Many experiments were conducted for each one of the four models based on 6 different training algorithms. The performance results of these models were compared according to mean square error and recognition rate to identify the best ANN model. The results showed that the PatternNet model was the best model used. Finally, comparisons between the used training algorithms were performed. Comparison results showed that TrainLM was the best training algorithm for the face recognition system.

Keywords—Face Recognition; Backpropagation Neural Network (BPNN); Feed Forward Neural Network; Cascade Forward; Function Fitting; Pattern Recognition

I. INTRODUCTION

Human Face represents complex, multidimensional, meaningful visual motivation. It is difficult to develop a computational model for face recognition. Building good computer system similar to human ability to recognize faces and overcome humans' limitations is regarded as a great challenge [1]. The human ability to recognize faces has several difficulties such as: similarity between different faces; dealing with large amount of unknown human faces; expressions and hair can change the face; and also face can be viewed from number of angles in many situations. A good face recognition system must be robust to overcome these difficulties and generalize over many conditions to capture the essential similarities for a given human face [2]. A general face recognition system consists of many processing stages: face detection; facial feature extraction; and face recognition. Face detection and feature extraction phases could run simultaneously [3].

In the recent years, artificial neural networks (ANN) were used largely for building intelligent computer systems related to pattern recognition and image processing [4]. The most popular ANN model is the backpropagation neural network (BPNN) which can be trained using backpropagation training algorithm (BP) [5]. Many literatures related to face recognition system which based on different approaches such as: Geometrical features; Eigenfaces; Template matching; Graph matching; and ANN approaches [6]. The obtained recognition rates from these studies are different and based on: used approach; used database; and number of classes.

Different ANN models were used widely in face recognition and many times they used in combination with the above mentioned methods. ANN simulates the way neurons work in the human brain. This is the main reason for its role in face recognition. Many researches adopted different ANN models for face recognition with different recognition rates and mean square error (MSE). Therefore, there is a need to identify the ANN model for face recognition systems with best recognition results. The objective of this research is to develop a face recognition system based on using 4 different ANN models: feed forward Backpropagation neural network (FFBPNN), cascade forward Backpropagation neural network (CFBPNN), function fitting (FitNet), and pattern recognition (PatternNet). Each one of these models was constructed separately with 7 layers (input, 5 hidden layers and output layer) architectures. Each model was trained separately with six different training algorithms.

The research includes the following sections: Section II includes related literature; Section III includes details about ANN architectures and training algorithms; Section IV explains research methodology; Section V includes implementation steps of the face recognition system; Section VI includes the experimental results; and finally Section VII concludes this work.

II. RELATED LITERATURE

ANN has the ability to adjust its weights according to the differences it encounters during training [7]. Therefore, we focused in this research on literature studies which based on ANN models especially BPNN. Dmitry and Valery (2002) [8]

proposed ANN thresholding approach for rejection of unauthorized persons. They studied robustness of ANN classifiers with respect to false acceptance and false rejection errors.

Soon and Seiichi (2003) [9] presented face recognition system with incremental learning ability that has one-pass incremental learning and automatic generation of training data. They adopted Resource Allocating Network with Long-Term Memory (RANLTM) as a classifier of face images. Adjoudj and Boukelif (2004) [10] designed a face recognition system using ANN which can trained several times on various faces images. While Volkan (2003) [11] developed a face authentication system based on: preprocessing, principal component analysis (PCA), and ANN for recognition. Normalization illumination, and head orientation were done in preprocessing stage. PCA is applied to find the aspects of face which are important for identification.

Weihua and WeiFu (2008) [12] suggested a face recognition algorithm based on gray-scale. They applied ANN to the pattern recognition phase rather than to the feature extraction phase to reduce complexity of ANN training. Also Mohamed, et. al. (2006) [13] developed BPNN model to extract the basic face of the human face images. The eigenfaces is then projecting onto human faces to identify unique features vectors. This BPNN uses the significant features vector to identify unknown face. They used ORL database. While Latha et al (2009) [14] used BPNN for face recognition to detect frontal views of faces. The PCA is used to reduce the dimensionality of face image. They used Yale database and calculated acceptance ratio and execution time as a performance metrics.

Raman and Durgesh (2009) [15] used single layer feed forward ANN approach with PCA to find the optimum learning rate that reduces the training time. They used variable learning rate and demonstrate its superiority over constant learning rate. They test the system's performance in terms of recognition rate and training time. They used ORL database. Abul Kashem et. al (2011) [16] proposed a face recognition system using PCA with BPNN. The system consists of three steps: detecting face image using BPNN; extraction of various facial features; and performing face recognition. And Shatha (2011) [17] performed face recognition by 3D facial recognition system using geometrics techniques with two types of ANN (multilayer perceptron and probabilistic). At the end, Taranpreet (2012) [18] proposed face recognition method using PCA with BP algorithm. The feature is extracted using PCA and the BPNN is used for classification.

III. ARTIFICIAL NEURAL NETWORKS

FFBPNN consists of many layers as in BPNN. The first layer is connected to ANN inputs. Each subsequent layer has connections from preceding layer. The final layer produces ANN output. BPNN and FFBPNN can be trained using BP algorithm. The BP includes the following equations [19][20]:

$$U_k(t) = \sum_{j=1}^n w_{jk}(t) \cdot x_j(t) + b_{ok}(t) \quad (1)$$

$$Y_k(t) = \varphi(U_k(t)) \quad (2)$$

Where,

$x_j(t)$: input value of j at time-step t ,

$w_{jk}(t)$: weight assigned by neuron k to input j at time t ,

φ : nonlinear activation function,

$b_k(t)$: the bias of k -neuron at time t , and

$y_k(t)$: output from neuron k at time t .

The process is repeated for all entries of time series and yields an output vector y_k . The training process includes weight adjustments to minimize the error between network's desired and actual output using an iterative procedure. Output y_k is compared with target output T_k using Eq.3 as an error function:

$$\delta_k = (T_k - y_k) y_k (1 - y_k) \quad (3)$$

The error is given by Eq.4 for neurons in the hidden layer:

$$\delta_k = y_k (1 - y_k) \sum \delta_k w_k \quad (4)$$

Where δ_k is the error term of the output layer and w_k is the weight between the hidden and output layers. The error is then propagated backward from the output layer to input layer to update the weight of each connection as follows [20]:

$$w_{jk}(t+1) = w_{jk}(t) + \eta \delta_k y_k + \alpha (w_{jk}(t) - w_{jk}(t-1)) \quad (5)$$

Where, η is the learning rate, and α is a momentum variable, which determines the effect of past weight changes on the current direction of movement.

Another ANN is CFBPNN and it is similar to FFBPNN but it includes a connection from input and every previous layer to following layers. Additional connections can improve the speed at which ANN learns the desired relationship. FitNet also presents a type of FFBPNN, which is used to fit an input output relationship. While PatternNet is a feed forward network that can be used for pattern recognition problems and can be trained to classify inputs according to target classes. The target data for PatternNet consist of vectors with all values equal to 0 except for 1 in element i , where i is the class they represent.

The BP training algorithm used to train FFBPNN and other ANN models requires long time to converge. Therefore, many optimization training algorithms were suggested and described in details in Neural Network Toolbox™ User's Guide R2012a [21]. The equations of all algorithms are the same except they differs in changing weight values.

A. Learning Algorithms

The optimization training algorithms adjusted the ANN weights and biases to minimize the performance function and to reduce errors as possible. Here, mean square error (MSE) is used as a performance function of the suggested face recognition system and it is minimized during ANN training. MSE represents the difference between the desired output and actual output.

In this research, six optimization ANN training algorithms were used to train the Four models separately to identify the model with the best results for the face recognition system [20][21]:

- Levenberg-Marquardt algorithm (TRAINLM)
- TRAINBFG algorithm
- Bayesian regularization algorithm (TRAINBR)
- TRAINCGF algorithm
- Gradient descent algorithm (TRAINGD)
- Gradient descent with momentum (TRAINGDM)

IV. RESEARCH METHODOLOGY

In this research, 4 ANN models (FFBPNN, CFBPNN, FitNet and PatternNet) were used separately for the face recognition system. Each one of these models was constructed separately with 7 layers (input, 5 hidden layers and output layer). Fig.1 shows 7-layer FFBPNN with 15 neurons in each hidden layer. Fig.1 can be used also to describe FitNet and PatternNet separately. The only difference in these ANN models is in training functions. Fig. 2 shows the 7 layers CFBPNN with 15 neurons in each hidden layer.

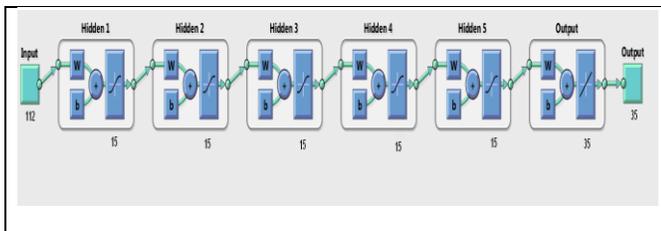


Fig. 1. Feed Forward Backpropagation Neural Network with 7 layers.

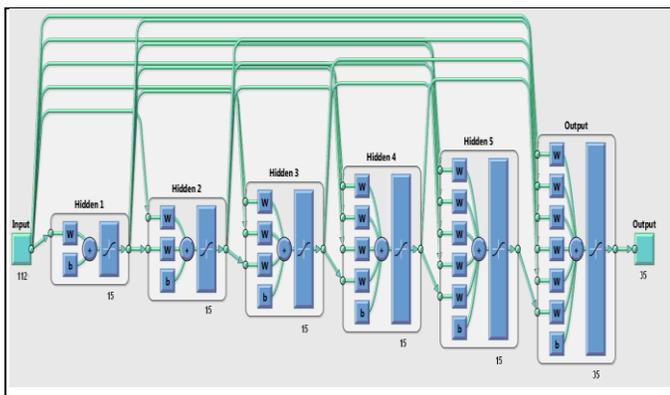


Fig. 2. Cascaded Feed Backpropagation Neural Network with 7 layers.

A. Training/Testing Samples

In this research, the training and testing samples were taken from the Oral face database (Olivetti Research

Laboratory) [22]. This database contains a set of faces taken between April 1992 and April 1994 at the Olivetti Research Laboratory in Cambridge, UK. There are 10 different images of 40 distinct persons. For each person, the images were taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images are taken against a dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for some side movement). All images are stored in ORL in PGM format with resolution 92×112 , 8-bit grey levels. Fig.3 shows samples from Oral face database for 6 persons.

As training samples, 350 face images (each with 92×112 dimension) were taken for 35 persons each with 10 samples. Each one of these images ($92 \times 112 = 10304$) is normalized and segmented into many blocks each with dimensions $8 \times 14 = 112$. This segmentation ($(92 \times 112) / (8 \times 14)$) will result in 92 sub images (blocks) for each face image. Each one of these samples (block) is with size 112. Therefore the number of input layer units is 112.

Whereas the number of output layer units is 35 to recognize these 35 persons. At the same time, the total number of training samples = number of images used in training process (350) multiplied by number of sub images (blocks) for each image (92) and this is equal to 32200. These samples are used in the face recognition system training process.

As testing samples, firstly, we select 50 random images from training samples for 5 persons each with 10 samples. Secondly, 50 (92×112) images were selected from Oral face database (which are not used in training) for 5 persons each with 10 samples. Each one of the 50 selected images (92×112) is divided into blocks of dimension 8×14 to obtain 92 blocks for each image. Therefore, the total number of testing samples for the 50 randomly selected face images is equal to 4600.

B. ANN Architecture

Fig.4 shows the architecture of the suggested ANN model for face recognition system and it consists of 7 layers (input layer, 5 hidden layers each with 15 hidden units and finally output layer). The input layer represents the face sub image (block) as system input. The number of input layer neurons depends on sub image dimensions (8×14) and here it is equal to 112.

Finally, the output layer returns the output vector. The number of output layer neurons depends on the problem nature and here it depends on the number of classes used in the face recognition training process. Since 350 images of 35 different persons were adopted, the number of classes is equal to 35 and hence, this is the number of output layer neurons.

class no.	Images belong to each class which related to one person
1	
2	
6	
7	
18	
22	

Fig. 3. Samples from Oral face database for 8 persons.

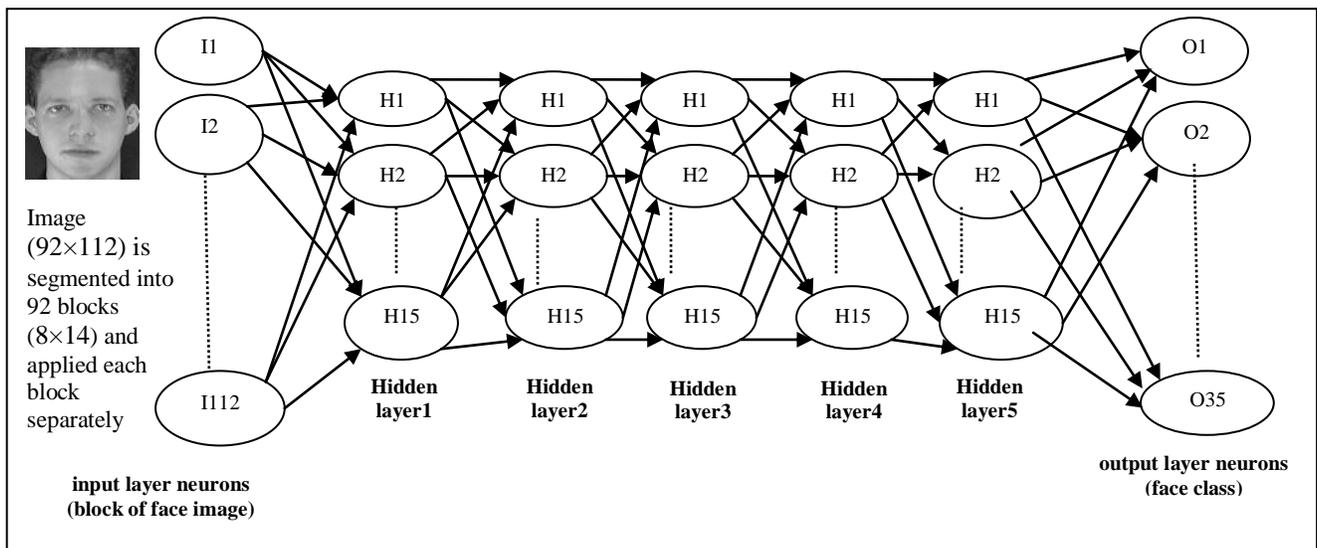


Fig. 4. Feed Forward BackPropagation Neural Network with 7Layers.

V. IMPLEMENTATION OF FACE RECOGNITION

A MathLab used to write an ANN training and testing face recognition system. This section includes the main steps of training and testing process.

A. Steps of ANN Training

Steps required to train the ANN model for face recognition system are as follows:

- 1) Initialize the ANN model weights and bias unit.
- 2) Initialize learning rate, momentum variable and threshold error with very small value like 0.0000001.

3) Initialize 35 classes: class for each person. Each class containing 10 faces images of one person.

4) Classification process: Initialize 35 target vectors one vector for each face class: vector = t1, t2... t35. All bits of vector1 are 0 except the first bit is 1. All bits of vector2 are 0 except the second bit is 1, and so on for other vectors.

5) Initialize the target output vector for each input vector (block 8x14) of the 350 face images. As example, the target output for face image blocks related to the 10 face images of the first person is equal to vector1. And, the target output for face image blocks related to the 10 face images of second

person is equal to vector2. And so on for each one of the 10 face images related to remaining 33 persons.

6) Apply steps of the selected training algorithm (LM, BFG, BR, CGF, GD, GDM) to train the ANN model. Apply input vector; compute outputs of each layer to find the actual output vector. Calculate the ANN error and according to this error the training is stopped or repeated again by adjusting the ANN weights. These operations repeated until we get ANN total error equal to threshold error to stop training process. Fig.5 shows these steps.

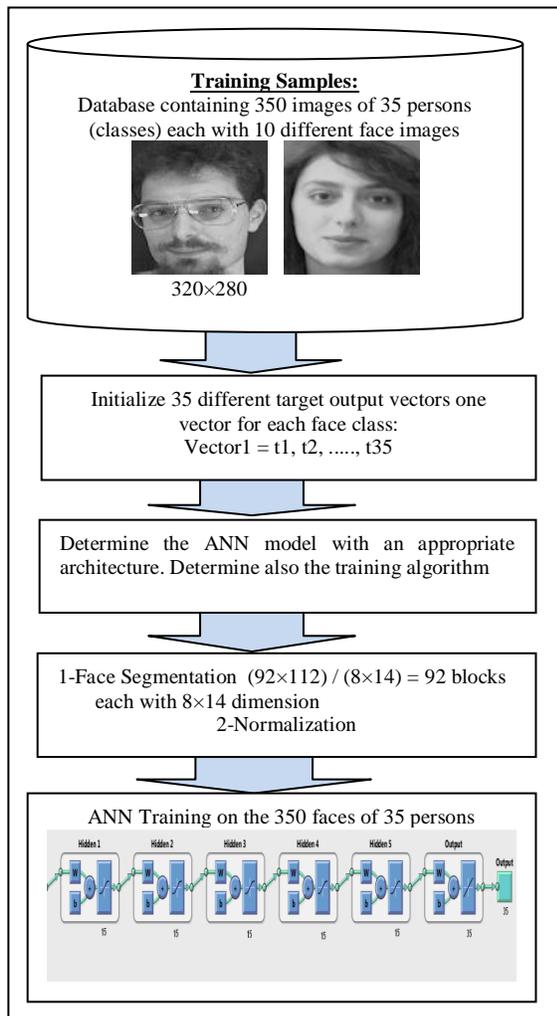


Fig. 5. ANN Training for the Face Recognition System

B. Steps of ANN Testing

The steps required to test the ANN model for face recognition system are as follows:

- 1) Apply one face block 8×14 to input layer neurons.
- 2) Compute the output of all layers in the ANN according to the steps required by training algorithm which was used in ANN training process until finding the outputs of output layer neurons.
- 3) Check if output of output layer neurons (output vector) is the same as one of the 35 classes (it's computed MSE is too

small), then ANN is recognized the block. And if the computed MSE of the ANN output is large, then the ANN is not recognized this block. Fig.6 shows these steps.

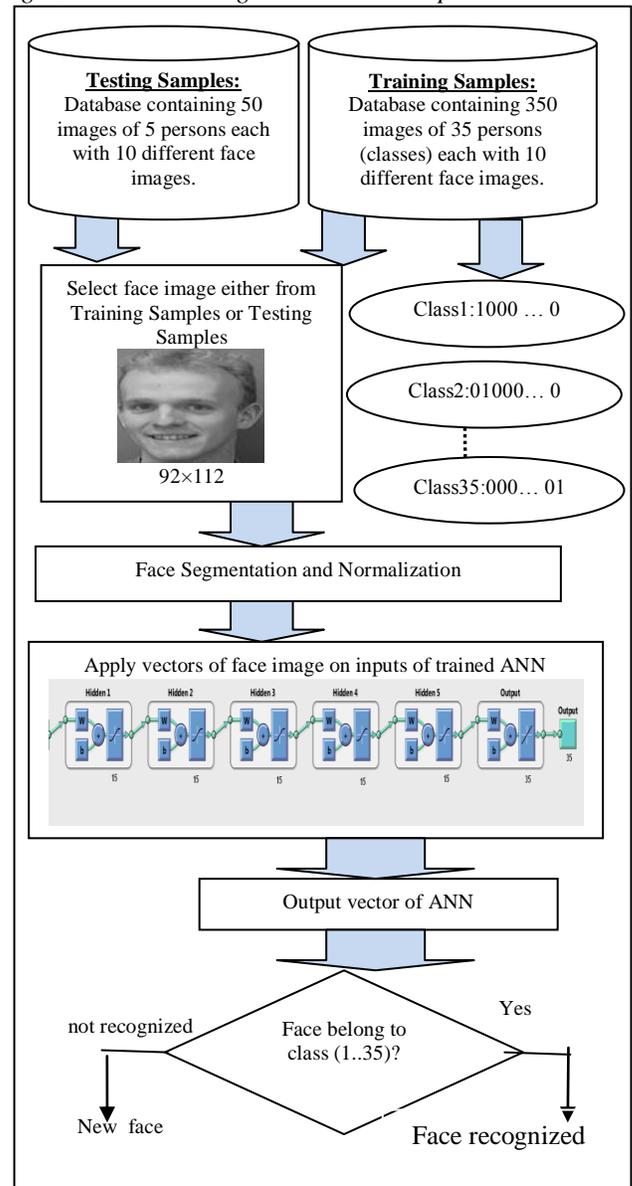


Fig. 6. ANN Testing for the Face Recognition System

VI. EXPERIMENTAL RESULTS

A MathLab was used to write the simulation program of training/testing of each one of the Four models (FFBPNN, CFBPNN, FitNet and PatternNet). The architecture of each model consists of 7 layers: input; 5 hidden layers each with 15 units; and output layer. The training data includes 350 (92×112) face images for 35 persons each with 10 samples were selected from Oral face database (Olivetti Research Laboratory) [22]. Here, we used the Mean Square Error (MSE), peak signal to noise ratio (PSNR) and recognition rate (RR) to evaluate the performance of ANN model for face recognition system.

Many experiments were conducted to examine the ANN model with best results of training and testing processes for the face recognition system. Many experiments were based on adopting different number of hidden layers (2, 3, 5, 7 and 9). Other experiments were based on adopting different numbers of neurons in each hidden layer (5, 10, 15, 20, 25 and 30). The best results were obtained from using 5 hidden layers each with 15 hidden units because we noticed from experiments that increasing number of hidden layers and number of hidden units will result in increasing the training time.

A. Results of Training Process

To determine the performance of each one of the 4 models, experiments were conducted by training these models separately each with 6 training algorithms. TABLE I shows MSE values of the 4 models.

TABLE I. IMPACT OF ANN MODELS ON MSE

Table with 5 columns: Algorithm, FFBP, CFBP, FitNet, PatternNet. Rows include TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAIINGD, and TRAIINGDM.

From TABLE I, we noted that the lower values of MSE are obtained for these models when LM training algorithm was used and these values were ranged between 0.003 and 0.09. Also the lowest MSE values were obtained from the PatternNet model. Also, we calculated the number of iterations needed for training process for each experiment. The ANN model required more number of iterations when we increased the number of hidden layer neurons. Therefore we used only 15 hidden units in each hidden layer for each model. TABLE II shows the number of iterations required to train the 4 models with 6 algorithms.

TABLE II. IMPACT OF ALGORITHMS ON NUMBER OF ITERATIONS

Table with 5 columns: Algorithm, FFBP, CFBP, FitNet, PatternNet. Rows include TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAIINGD, and TRAIINGDM.

From TABLE II, PatternNet required lowest number of iterations (21) for the training process especially when it was trained using TRAINLM algorithm. But PatternNet required 41 iterations when it was trained using TRAINCGF algorithm. Also, TABLE II shows that the other ANN models require more iterations when they were trained using TRAINLM training algorithm.

B. Results of ANN Testing Process

In testing process, we mentioned earlier in sub section A in section IV that the number of samples used in testing process is 50 images for 5 persons each with 10 different samples. Also sub section B in section V includes the main steps of testing process. The testing process includes two parts. Firstly, the 4 ANN models were tested using 50 face images which randomly selected from the training samples (samples which were used earlier in training process). The lowest values of MSE were obtained from PatternNet model as shown in TABLE III. Also TABLE III shows that the lowest values of all models were obtained from using TRAINLM training algorithm.

TABLE III. MSE FOR TESTING 50 TRAINED IMAGES

Table with 5 columns: Algorithm, FFBP, CFBP, FitNet, PatternNet. Rows include TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAIINGD, and TRAIINGDM.

TABLE IV shows the recognition rates related to testing the 4 models using 50 randomly selected trained images. Best values of recognition rate were obtained from PatternNet model trained using TRAINLM algorithm.

TABLE IV. RECOGNITION RATE FOR TESTING 50 TRAINED IMAGES

Table with 5 columns: Algorithm, FFBP, CFBP, FitNet, PatternNet. Rows include TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAIINGD, and TRAIINGDM.

Also TABLE V shows the PSNR of the testing process related to the trained 5 models using 50 randomly selected trained images. Best values of PSNR were obtained from PatternNet model. At the same time, TRAINLM algorithm results in best values of recognition rates for all models.

TABLE V. PSNR FOR TESTING 50 TRAINED IMAGES

Table with 5 columns: Algorithm, FFBP, CFBP, FitNet, PatternNet. Rows include TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAIINGD, and TRAIINGDM.

Secondly, the 4 ANN models were tested with 50 testing untrained samples (i.e. images which were not used in training

process). The MSE values obtained from this testing were very high because the 4 ANN models were not recognized these testing images.

The lowest MSE values were obtained from using PatternNet model which was trained using TRAINLM algorithm as shown in TABLE VI.

TABLE VI. MSE FOR TESTING 50 UN-TRAINED IMAGES

Algorithm	FFBP	CFBP	FitNet	Pattern Net
TRAINLM	0.88	0.85	0.76	0.61
TRAINBFG	0.87	0.81	0.75	0.72
TRAINBR	0.8	0.81	0.72	0.69
TRAINCGF	0.87	0.85	0.89	0.9
TRAIINGD	0.84	0.87	0.91	0.94
TRAIINGDM	0.89	0.91	0.95	0.96

Finally, TABLE VII shows the recognition rate for the testing process of the 4 models on untrained images. Therefore the values of recognition rates in TABLE VII are not high.

TABLE VII. RECOGNITION RATE FOR TESTING 50 UNTRAINED IMAGES

Algorithm	FFBP	CFBP	FitNet	PatternNet
TRAINLM	22.1	22.4	22.44	22.7
TRAINBFG	23.4	23.2	23.8	24
TRAINBR	21.5	21.6	21.8	22.2
TRAINCGF	19.2	19.6	19.3	19.7
TRAIINGD	19.5	19.7	19.8	20.2
TRAIINGDM	20.4	20.6	20.7	21.2

VII. CONCLUSION

In this research, we presented a face recognition system using Four feed forward ANN models (FFBPNN, CFBPNN, FitNet and PatternNet) and 6 training methods. Each one of the 4 models was constructed with 7-layer architecture. This face recognition system consists of two parts: training and testing. Six ANN optimization training algorithms (TRAINLM, TRAINBFG, TRAINBR, TRAINCGF, TRAIINGD, and TRAIINGDM) were used to train each of the constructed ANN models separately.

The training and testing samples of the suggested face recognition system were taken from The ORL Database of Faces [22]. As training samples, we selected 350 face images (92×112) from ORL database which belong to 35 persons each with 10 different samples. As testing samples (untrained images), we selected 50 images (92×112) from ORL database which belong to 5 persons each with 10 different samples.

A set of experiments were conducted to evaluate the performance of the suggested face recognition system by calculating the MSE, number of iterations, recognition rate and PSNR. This was done using 4 different ANN models and 6 different optimization algorithms. The results showed that the lowest values of MSE and number of iterations were resulted from the PatternNet model. The best results of the PatternNet model were obtained when this model was trained

using the Levenberg Marquardt training algorithm (TRAINLM).

Future work may include a survey of other techniques related to face recognition systems and comparing their results with those presented in this paper. These comparisons will be based on many factors like: recognition rate, PSNR, algorithm complexity, ANN learning time and number of iterations required for training and so on.

ACKNOWLEDGEMENT

The authors would like to thank Al-Zaytoonah University of Jordan, Amman, Jordan, for sponsoring the scientific research.

REFERENCES

- [1] Turk M. and Pentland A., 1991. "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, vol.3, pp.71–86.
- [2] Jonathan Howell A. and Hilary B., 1996. "Face Recognition using Radial Basis Function Neural Networks", *Proceedings of British Machine Vision Conference*.
- [3] Zhao W., et al, 2003. "Face recognition: A literature survey", *ACM Computing Surveys*, Vol.35, No.4, December, pp. 399–458.
- [4] Dan W. Patterson, 1996. "Artificial Neural Networks, Theory and Applications", *Singapore: Prentice Hall*.
- [5] Hamid B. and Mohamad R. M., 2009. "A learning automata-based algorithm for determination of the number of hidden units for three-layer neural networks", *International Journal of Systems Science*, vol.40, no.1, 101–118, Jan.
- [6] Steve Lawrence, et. al., 1997. "Face Recognition: A Convolutional Neural Network Approach", *IEEE Transactions on Neural Networks, Special Issue on Neural Networks and Pattern Recognition*, vol.8, no.1, pp.98–113.
- [7] Hossein S., Mahdi R., and Hamid D., 2008. "Face Recognition Using Morphological Shared-weight Neural Networks", *World Academy of Science, Engineering and Technology*, vol.45, pp.555- 558.
- [8] Dmitry B. and Valery S., 2002. "Access Control By Face Recognition Using Neural Networks And Negative Examples", *The 2nd International Conference on Artificial Intelligence*, pp.428-436, 16-20Sep, Crimea, Ukraine.
- [9] Soon Lee Toh and Seiichi O., 2003. "A Face Recognition System Using Neural Networks with Incremental Learning Ability", *Proceeding of the 8th Australian and New Zealand Conf. on Intelligent Information Systems*, pp.389-394.
- [10] Adjoudj R. and Boukelif A., 2004. "Artificial Neural Network-Based Face Recognition", *First International Symposium on Control, Communications and Signal Processing*, pp.439-442.
- [11] Volkan A., 2003. "Face Recognition Using Eigenfaces and Neural Networks", *Master of Science Thesis*, The Graduate School of Natural And Applied Sciences, The Middle East Technical University, Dec.
- [12] Weihua W. and WeiFu W., 2008. "A Gray-Scale Face Recognition Approach", *Second International Symposium on Intelligent Information Technology Application*, 978-0-7695-3497-8/08, IEEE computer society, DOI 10.1109/IITA.2008.101.
- [13] Mohamed R., et. al. 2006. "Face Recognition using Eigenfaces and Neural Networks", *American Journal of Applied Sciences*, vol.2, no.6, pp.1872-1875, ISSN 1546-9239, Science Publications.
- [14] Latha P., Ganesan L.& Annadurai S., 2009. "Face Recognition using Neural Networks", *Signal Processing: An International Journal*, vol.3, issue.5, pp.153–160.
- [15] Raman B. and Durgesh Kumar M., 2009. "Face Recognition Using Principal Component Analysis and Neural Network", *Journal of Technology and Engineering Sciences*, Vol.1, No.2.
- [16] Mohammad A. K., et, al. 2011. "Face Recognition System Based on Principal Component Analysis (PCA) with Back Propagation Neural Networks (BPNN)", *Canadian Journal on Image Processing and Computer Vision*, vol.2, no.4, April.
- [17] Shatha K. Jawad, 2011. "Design a Facial Recognition System Using Multilayer Perceptron and Probabilistic Neural Networks Based Geometrics 3D Facial", *European Journal of Scientific Research*,

- ISSN:1450-216X, vol.60, no.1, pp.95-104, Euro Journals Publishing, Inc. 2011, <http://www.eurojournals.com/ejsr.htm>
- [18] [18] Taranpreet S. R., 2012. "Face Recognition Based on PCA Algorithm", *Special Issue of International Journal of Computer Science & Informatics (IJCSI)*, ISSN:2231-5292, vol.2, No1.2, pp.221- 225.
- [19] Haykin S., *Neural Networks: A Comprehensive Foundation* 2nd edition, Prentice-Hall Inc., 1999. <http://www.statsoft.com/textbook/stneunet.html>
- [20] H. K. Elminir, Y.A. Azzam, and F. I. Younes, "Prediction of Hourly and Daily Diffuse Fraction Using Neural Network as Compared to Linear Regression Models," *Energy*, vol.32, pp.1513-1523. 2007.
- [21] Mark Hudson Beal, Martin T. Hagan and Howard B. Demuth, *Neural Network Toolbox™ User's Guide R2012a*, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098, 2012, www.mathworks.com
- [22] AT&T (ORL): Oral face database (Olivetti Research Laboratory), <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.htm>

Image Blocks Model for Improving Accuracy in Identification Systems of Wood Type

Gasim

Faculty of Information Technology,
Multi Data Palembang Bachelor Program,
Palembang, Indonesia

Agus Harjoko

Dept. of Computer Science and Electronics,
Gadjah Mada University,
Yogyakarta, Indonesia

Kudang Boro Seminar

Dept. of Mechanical and Biosystem Engineering,
Bogor Agricultural University (IPB),
Bogor, Indonesia

Sri Hartati

Dept. of Computer Science and Electronics,
Gadjah Mada University,
Yogyakarta, Indonesia

Abstract—Image-based recognition systems commonly use an extracted image from the target object using texture analysis. However, some of the proposed and implemented recognition systems of wood types up to this time have not been achieving adequate accuracy, efficiency and feasible execution speed with respect to practicality. This paper discussed a new method of image-based recognition system for wood type identification by dividing the wood image into several blocks, each of which is extracted using gray image and edge detection techniques. The wood feature analysis concentrates on three parameters entropy, standard deviation, and correlation. Our experiment results showed that our method can increase the recognition accuracy up to 95%, which is faster and better than the previous existing method with 85% recognition accuracy. Moreover, our method needs only to analyze three feature parameters compared to the previous existing method needs to analyze seven feature parameters, and thus implying a simpler and faster recognition process.

Keywords—image processing; pattern recognition; ANN; wood identification.

I. INTRODUCTION

The identification of wood types becomes very important when it related to illegal logging, taxes, and the suitability of the product. This activity is constrained, because the experts in identification of wood are very limited in terms of amount, power, and time.

The experts usually do an initial identification with respect to the macroscopic elements (the impression of touch, smell, weight, color). If there is still doubt, then the expert will observe the microscopic elements in the cross-sectional area, radial cross-section, and cross tangent. This activity uses a magnifying glass (10x).

Its unique features can identify Wood of a particular species. These features include strength, density, hardness, odor, texture and color. Reliable wood identification usually requires the ability to recognize basic differences in cellular structure and wood anatomy.

Each species has unique cellular structure that creates differences in wood properties and ultimately determines the suitability for a particular use. Cellular characteristics provide a blueprint for accurate wood identification [2].

Wood is composed of many small cells and its structure is determined by the type, size, shape and arrangement of these cells. The structure and characteristics of wood can vary between species and within the same species. With practice, a small hand lens (10x) can be used to distinguish the different cell types and their arrangements [2].

In the previous research [5], authors used 20 types of wood, using seven characteristics of RGB image, and the six characteristics of image edge detection. This research provide 85% recognition rate.

In this paper, authors will use the image blocking to identify the type of wood, all type of woods used are similar to previous research [5].

In previous research have showed that the recognition rate varied results with a variety of methods used, include: (1) feature used is the texture analyst added RGB with enlargement 24 times, using five different types of wood [4]. (2) Feature used is the texture analyst; method used is ANFIS, and uses five types of wood [6]. (3) Next research is the comparison of rate recognition based input features with enlarged 24 times, using five different types of wood [9]. (4) The next research using 15 types of wood, texture analysts and RGB as input ANN, using ANNBP, and give the recognition rate 95% [10]. this value is enough high, due to the number of species that used only 15 types, and test data that are used most of the images are sourced from the same sample with image training.

The research that has been conducted by the authors [5], where 7 features from the gray level image and 6 features from the edge detection image, 85% could be identified for test results. Recognition rate and the features that used in this research have not been satisfactory. Therefore, the authors propose a method called *image blocking*. This method expected to reduce the number of features used, and increase the recognition rate.

II. PROPOSED METHOD

The method proposed in this research is block method, i.e. the image is divided into several part, then do extract features in each part. In this research, an image is divided into four blocks. It is related to the microscopic cross section of wood. One character has a pore structure that repeats on each particular size rectangular, although not too similar, and this is characteristic of each type's wood. In addition, the level of magnification used also affects the details of the microscopic and the area has been observed. In this paper, authors use 45 times magnification (optical). Shooting direction is perpendicular to the cross section, and the radius (rays) of the wood is vertical. The details of the steps method is shown in Figure 1.

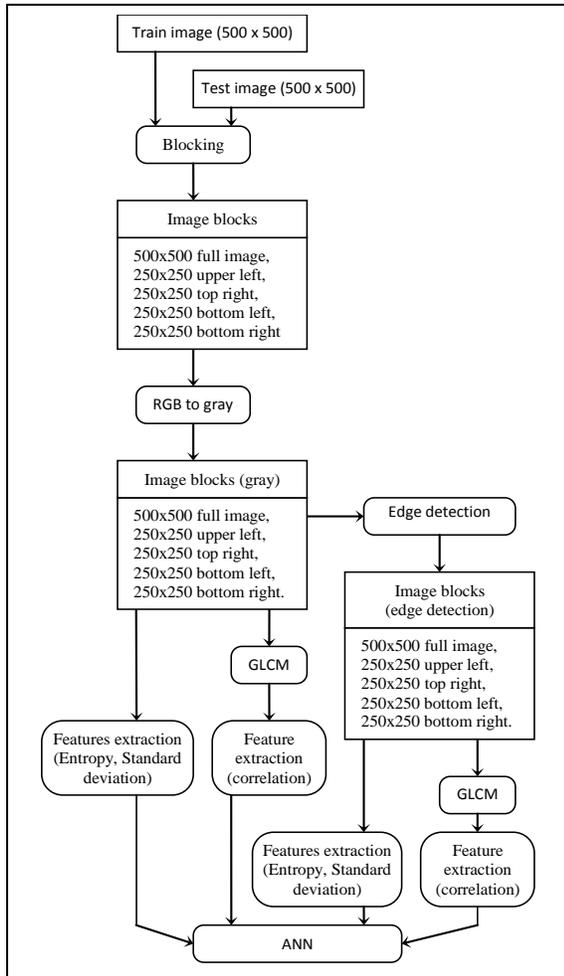


Fig.1. Work steps of the block method.

A. Training Data and Test Data

The data is a collection of images that has been cut into 500x500 pixels (Figure 5). The image acquisition is conducted using a handheld microscope 1.3 MP (Fig. 2), the vertical lines that exist in the image are the rays. Before being cropped, the image size is 1280x1024 pixels (Figure 3), that is cropped into 500x500 pixels on a good part.

That is, minimal scratches incision results, and the pore is not closed. This process uses image-editing software (Figure 4).

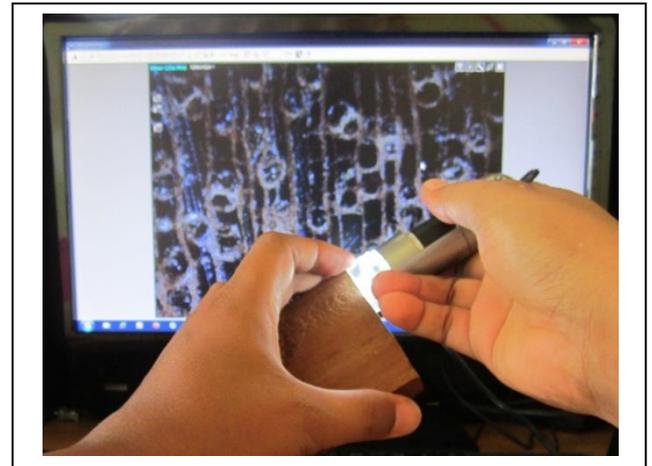


Fig.2. Capturing



Fig.3. Captured image, 1280 x 1024 pixels

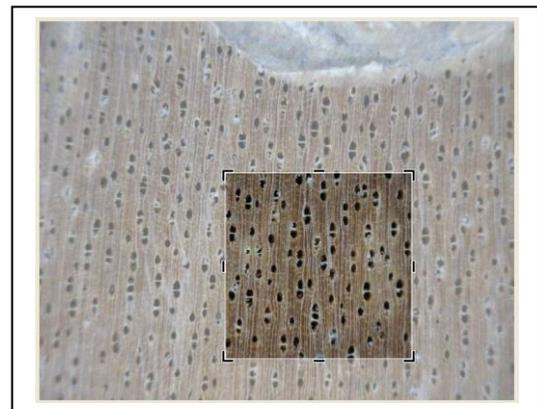


Fig.4. Cropping to 500x500

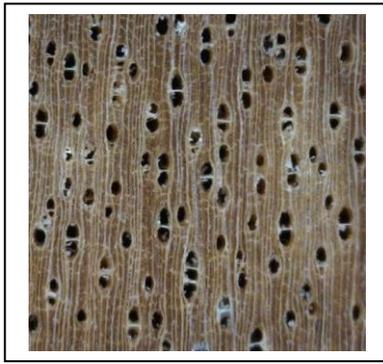


Fig.5. Image 500x500

Wood samples used in this study are presented in Table I. Imagery training consists of 20 types of wood, each type consisting of 100 images derived from some wood samples. So the total is 2,000 image training image. For the test images using five images for each type.

TABLE I. TYPE OF WOOD USED

No.	Trade Name (Scientific Name)
1.	Bakau (Rhizophora apiculata Bl.)
2.	Cenge (Mastixia trichotoma Bl.)
3.	Jabon (anthocephalus cadamba)
4.	Jabon merah (Anthocephalus macrophyllus)
5.	Kembang semangkok (Scaphium macropodium J.B.)
6.	Kruing (Dipterocarpus gracilis Bl.)
7.	Kruing (Dipterocarpus kunstleri King)
8.	Kulim (Scorodocarpus borneensis Becc.)
9.	Mempisang (Mezzetia parviflora)
10.	Meranti Kuning (Shorea acuminatissima sym)
11.	Meranti Merah (Shorea acuminata)
12.	Meranti Merah (Shorea ovalis Bl.)
13.	Meranti Putih (Shorea Javanica k.ot. val)
14.	Merawan (Hopea spp.)
15.	Merbau (Intsia bijuga O.K.)
16.	Merbau (Intsia palembanica)
17.	Mersawa (Anisoptera)
18.	Penjalin (Celtis Philippinensis)
19.	Perupuk (Lophopetalum javanicum)
20.	Rasamala (Hamamelidaceae)

B. Blocking

Blocking is the process of dividing the RGB image into four blocks, each 250x250 pixels. This method is carried out because of the texture of the cross-sectional image of the wood has a recurring trait on every particular square. Although the texture is not the same between the blocks, but it has an attachment between the turn, so it can be used as feature values.

The rules of blocking are presented in Figure 6. This process is carried out on the whole the image training and test images.

```
i = imread('500x500_01_1.jpg');  
i1 = imcrop(i,[1 1 249 249]);  
i2 = imcrop(i,[251 1 249 249]);  
i3 = imcrop(i,[1 251 249 249]);  
i4 = imcrop(i,[251 251 249 249]);
```

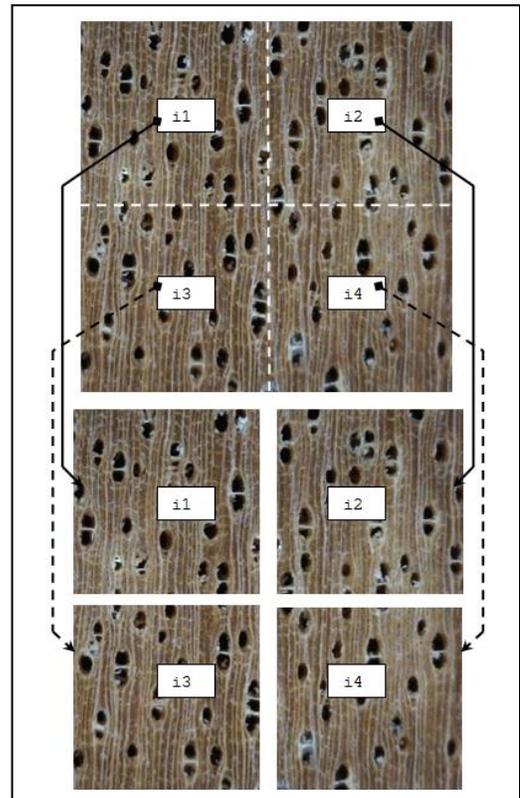


Fig.6. Blocking RGB image

C. RGB to Gray

RGB to gray is the process of converting each block RGB image into gray image. This stage is done as needed at a later stage that requires the image of a gray scale.

The `rgb2gray` converts RGB values to grayscale values by forming a weighted *sum* of the R, G, and B components [12] :

$$0.2989 * R + 0.5870 * G + 0.1140 * B \quad (1)$$

Matlab function that used to transform the RGB image into gray image is:

```
gry_i = rgb2gray(i);  
gry_i1 = rgb2gray(i1);  
gry_i2 = rgb2gray(i2);  
gry_i3 = rgb2gray(i3);  
gry_i4 = rgb2gray(i4);
```

D. Image Blocks (Gray)

Image blocks is the images converted from RGB images into gray image, which consists of the full image (500x500), top left block (250x250), the top right block (250x250), bottom left block (250x250), and the bottom right block (250x250).

E. Edge Detection

Edge detection is the process of converting each block of gray image to edge detection image. In this research, the edge detection used is canny. Canny operator is used, because it gives the expected results compared to other operators.

Edge detection is the approach used most frequently for segmenting images based on abrupt (local) changes in intensity. There are three fundamental steps performed in edge detection [7] :

- Image smoothing for noise reduction.
- Detection of edge points.
- Edge localization.

Canny's approach is based on three basic objectives [7]:

- 1) *Low error rate*
- 2) *Edge points should be well localized*
- 3) *Single edge point response*

The Matlab function for this process is :

```
cny_i = edge(gry_i , 'canny');
cny_i1 = edge(gry_i1 , 'canny');
cny_i2 = edge(gry_i2 , 'canny');
cny_i3 = edge(gry_i3 , 'canny');
cny_i4 = edge(gry_i4 , 'canny');
```

F. Image Blocks (Edge Detection)

Image blocks (edge detection) are images converted from gray-level image into image edge detection, which consists of the full image (500x500), top left block (250x250), top right block (250x250), bottom left block (250x250), and bottom right block (250x250).

G. GLCM

GLCM is a statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM) [8], also known as the gray-level spatial dependence matrix. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix [11]. From this matrix is used to calculate some statistical variables. These statistics provide information about the texture of an image. A number of texture features may be extracted from the GLCM [8].

H. Feature Extraction

Feature extraction is the process of taking a value or several values of the gray image that will used as the identity of the gray image. The feature extraction Conducted on gray image and edge detection image. The features which taken from each image are entropy, standard deviation, and correlation.

1) *Standard deviation* : The standard deviation commonly used to measure the distribution of positive and negative values of a member to the average value of all members. To calculate the standard deviation of an image, which is a two-dimensional matrix, the following steps were used :

- Calculate the total amount of values all the pixels of a two dimensional matrix $m \times n$ pixels.

$$total = \sum_{i=1}^m \sum_{j=1}^n a[i, j] \quad (2)$$

- Calculate the average values pixel matrix $m \times n$ pixels

$$mean = \frac{total}{m \times n} \quad (3)$$

- Calculate the standard deviation

$$std = \sqrt{\left(\frac{1}{(m \times n) - 1}\right) \sum_{i=1}^m \sum_{j=1}^n (a[i, j] - mean)^2} \quad (4)$$

Matlab function used is :

```
std_gray_i = std2(gry_i);
std_gray_i1 = std2(gry_i1);
std_gray_i2 = std2(gry_i2);
std_gray_i3 = std2(gry_i3);
std_gray_i4 = std2(gry_i4);
```

2) *Entropy* : Is measure the randomness of the elements of 2D matrix. The entropy is 0 when all p_{ij} 's are 0 and is maximum when all p_{ij} 's are equal :

$$ENTROPY = - \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j) \times \log(P(i, j)) \quad (5)$$

Matlab function used is :

```
entropi_gray_i = entropy(gry_i);
entropi_gray_i1 = entropy(gry_i1);
entropi_gray_i2 = entropy(gry_i2);
entropi_gray_i3 = entropy(gry_i3);
entropi_gray_i4 = entropy(gry_i4);
```

3) *Correlation* : Correlation is one of the few variables that are can generate from the GLCM [8]. This variable is used because it gives better results than some other GLCM variables.

We use the following notation [1]:

G is the number of gray levels used.

μ is the mean value of P .

μ_x, μ_y, σ_x and σ_y are the means and standard deviations of P_x and P_y . $P_x(i)$ is the i th entry in the marginal-probability matrix obtained by summing the rows of $P(i, j)$:

Correlation is a measure of how correlated a pixel is to its neighbor over the entire image. Range of values is 1 to -1, corresponding to perfect positive and perfect negative correlations (6).

$$CORRELATION = \frac{\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \{i \times j\} \times P(i, j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \quad (6)$$

The following feature extraction for gray image:

```
glcm_i_gray = graycomatrix(gry_i);
chce_i_gray = graycoprops(glcm_i_gray, {'Contrast', 'Homogeneity', 'Correlation', 'Energy'});

glcm_i1_gray = graycomatrix(gry_i1);
chce_i1_gray = graycoprops(glcm_i1_gray, {'Contrast', 'Homogeneity', 'Correlation', 'Energy'});

glcm_i2_gray = graycomatrix(gry_i2);
chce_i2_gray = graycoprops(glcm_i2_gray, {'Contrast', 'Homogeneity', 'Correlation', 'Energy'});
```

```
glcm_i3_gray = graycomatrix(gry_i3);
chce_i3_gray = graycoprops(glcm_i3_gray, {'Contrast',
'Homogeneity', 'Correlation', 'Energy'});

glcm_i4_gray = graycomatrix(gry_i4);
chce_i4_gray = graycoprops(glcm_i4_gray, {'Contrast',
'Homogeneity', 'Correlation', 'Energy'});
```

I. ANN

In this paper, authors use neural networks to identify the type of wood. It is used because is based on the results of a research journal of pattern recognition; the ANN is the best method. Information of each image for each type of wood is stored in the form of the ANN weights. Weights in the ANN will experience changes during the training period, up to the value of parameter goal is reached. To achieve the expected goal, recognizing 100% trained image, and the highest test images (95%), the ANN architecture must be the best. To get the best architecture, it was trial and error on some architecture, i.e. the number of hidden layers and number of neurons of each hidden layer. From the results of experiments on several ANN architectures, the best architecture is the 3 hidden layers, and each hidden layer has 73 neurons. While the number of input neurons is 40 neurons, four are from the full image; image comes from the four blocks. Because image the block there are 4 images derived from image the block there are 16 neurons. The image used is a gray image edge detection and image, so that the number of neurons to 40. More used the ANN architecture, presented in Table II.

TABLE II. SPECIFICATION OF NEURAL NET WORK

Characteristic	Specification
Architecture, algorithm	3 hidden layers, back propagation
Neuron input	From gray image: 1. Full image : entropy, standard deviation, correlation. 2. Image block : entropy, standard deviation, correlation. From edge detection image : 1. Full image : entropy, standard deviation, correlation. 2. Image block : entropy, standard deviation, correlation.
hidden layer	3
Neuron of hidden layer	73, 73, 73
Neuron of output	20 (Number of wood)
activation function	Sigmoid binary
goal	1e-24
learning rate	0,1
Max epoch	5000
Number of image each wood for data training	100
Number of image each wood for data testing	5

Goal value used is 1e-24, because at this value the maximum recognition rate of the test data obtained. While on a

smaller goal, i.e. 1e-32, the level recognition to the test data actually decreased. So also with the larger goal of 1e-24.

Matlab function used is :

```
net = newff(minmax(latih_N), [73,73,73,20],
{'tansig','tansig','logsig','logsig'},'traincgp')
net.trainParam.epochs = 5000;
net.trainParam.goal = 1e-24;
net.trainParam.lr = 0.1;
tic;
net_train = train(net, latih_N, target);
waktu_training = toc
```

III. RESULTS

This experiment, carried out by using 20 types of wood, with a cross-sectional image as the input image. ANN architecture used is three hidden layer, 73 neurons respectively. Tests using 100 test images. for a more complete test results can be seen in Table III.

TABLE III. TESTING RESULTS

No.	Trade Name (Scientific Name)	amount of test data.	recognizable
1	Bakau (Rhizophora apiculata Bl.)	5	5
2	Cenge (Mastixia trichotoma Bl.)	5	5
3	Jabon (anthocephalus cadamba)	5	5
4	Jabon merah (Anthocephalus macrophyllus)	5	4
5	Kembang semangkok (Scaphium macropodum J.B.)	5	5
6	Kruing (Dipterocarpus gracilis Bl.)	5	5
7	Kruing (Dipterocarpus kunstleri King)	5	5
8	Kulim (Scorodocarpus borneensis Becc.)	5	5
9	Mempisang (Mezzetia parviflora)	5	5
10	Meranti Kuning (Shorea acuminatissima sym)	5	4
11	Meranti Merah (Shorea acuminata)	5	5
12	Meranti Merah (Shorea ovalis Bl.)	5	4
13	Meranti Putih (Shorea Javanica k.ot. val)	5	5
14	Merawan (Hopea spp.)	5	5
15	Merbau (Intsia bijuga O.K.)	5	5
16	Merbau (Intsia palembanica)	5	4
17	Mersawa (Anisoptera)	5	5
18	Penjalin (Celtis Philipinensis)	5	5
19	Perupuk (Lophopetalum javanicum)	5	4
20	Rasamala (Hamamelidaceae)	5	5
	Total	100	95

The results of experiment that have been conducted on three of these features is the increasing level of recognition accuracy to 95%. Testing was conducted on 5 images of each type, so there are 100 test images.

IV. CONCLUSION

An experiment on the identification of types of wood that has been done in this research, has given better results than researches conducted previously authors. In This research has been done on the image the block method, with a combination of image blocks that is divides the image into four equal parts. It aims to reduce the number of features that are used and increasing the recognition to the types of wood.

The conclusion that authors can write is that:

- 1) the method can improve the identification of types of wood;
- 2) and the method can reduce features used in the system of identification type of wood

From the research results, there are opportunities to increase the number of the types of wood, because:

- 1) this research only use a small number of features;
- 2) there are still some combinations of blocks that have not been tested;
- 3) there is an opportunity to test the objects using another magnification level

ACKNOWLEDGMENT

The authors would like to thank : (1) STMIK Multi Data Palembang (www.mdp.ac.id); (2) the Department of Computer Science and Electronics, Gadjah Mada University (<http://mkom.ugm.ac.id>) Yogyakarta Indonesia, and Dept. of Mechanical and Biosystem Engineering, Bogor Agricultural University (IPB), Bogor, Indonesia that provides technical support for the research; (3) Department of Forestry Laboratory of Wood Anatomy Bogor Indonesia for timber sample.

REFERENCES

- [1] Albreghsen, F., "Statistical Texture Measures Computed from Gray Level Cooccurrence Matrices," Image Processing Laboratory Department of Informatics, University of Oslo, 2008.
- [2] Bond B. and Hamner P. "Wood Identification for Hardwood and Soft wood Species Native to Tennessee," <http://www.utextension.utk.edu/>, 2002.
- [3] Fausett, L., "Fundamentals Of Neural Network Architectures : Algorithm and Applications," Prectice-Hall, Inc., 1994.
- [4] Gasim, "The Design and Implementation of an Image-Based Wood Variety Recognition System Using ANN," Proceeding The 9th INTERNATIONAL CONFERENCE on QUALITY in RESEARCH (QiR).Information and Computation Engineering, ISSN : 114-1284, 2006.
- [5] Gasim, Harjoko A., Seminar KB., Hartati S. (2013), "Merging Feature Method on RGB Image and Edge Detection Image for Wood Identification", International Journal of Computer Science and Information Technologies (IJCSIT), Vol 4(1), January- February 2013, pp 188 – 193
- [6] Gasim, Hartati, S., "Arsitektur ANFIS untuk Pengenalan Kayu Berbasis Citra Cross-Section," The International Conterence on Computer and Mathematical Sciences 2010, 29 June 2010 UiTM and UGM Collaboration, Jogjakarta, 2010.

- [7] Gonzales, R. C. & R. E. Woods. "Digital Image Processing." Addison Wesley, Massachusetts, 1992.
- [8] Haralick, RM., K. Shanmugam and Itshak Dinstein. "Textural Features For Image Classification," IEEE Transaction On System, Man and Cybernetics. Vol 3, No. 6. 1973.
- [9] Harjoko, A., Gasim, "Comparison of Some Features Extraction of Wood," Proceeding The 2nd International Conference on Distributed Frameworks and Applications, Jogjakarta, 2010.
- [10] Harjoko, A., Gasim, Rulliaty, S.S., Damayanti, R., "Identification Method for 15 Names of Commercial Wood With Image of Texture Pore as an Input," Proceeding International Conference on Informatics for Development, Jogjakarta, 2011.
- [11] Mathwork Inc., "Neural Network Toolbox for Use With Matlab," The Mathwork Inc. Natick, USA, 2012.
- [12] Mathwork, "rgb2gray," www.mathworks.com/help/images/ref/rgb2gray.html, 2013.
- [13] Mathwork, "entropy," www.mathworks.com/help/images/ref/entropy.html, 2013.
- [14] Mathwork, "std2," www.mathworks.com/help/images/ref/std2.html, 2013.

AUTHORS PROFILE



Indonesia.



Gasim is a lecturer at the Faculty of Information Technology, Multi Data Palembang, Bachelor Program, Palembang, Indonesia. He is graduated as Bachelor of Computer in STMIK Bandung, Indonesia. He is received his Master of Computer at Bogor Agricultural University (IPB), Bogor, Indonesia. He is currently taking his Doctoral Program at the Departement of Computer Science and Electronics, Gadjah Mada University in Yogyakarta,

Agus Harjoko is a Associate Professor and Chair of Electronics and Instrumentation Lab. He is graduated from the Electronics and Instrumentation study program, Faculty of Mathematics and Natural Sciences, UGM Yogyakarta, Indonesia. He got M.Sc. and PhD in Computer Science from the University of New Brunswick, Canada.

Kudang Boro Seminar is Professor in Computer Technology. He is head of laboratory bioinformatics engineering, director of communications and information systems, IPB, and Honorary Member of AFITA (Asian Federation for Information Technology in Agriculture). He is graduated from Bogor Agricultural University (IPB) Bogor, Indonesia. He got M.Sc. and PhD in Computer Science at University of New Brunswick, Canada.

Sri Hartati is a Associate Professor and Chair of Computer Science Graduate Program, Gadjah Mada University, Yogyakarta, Indonesia. She got M.Sc. and PhD in Computer Science Dept, at University of New Brunswick, Canada.

A Strategy for Training Set Selection in Text Classification Problems

Maria Luiza C. Passini, Kátiusca B. Estébanez, Graziela P. Figueredo, Nelson F. F. Ebecken
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil

Abstract—An issue in text classification problems involves the choice of good samples on which to train the classifier. Training sets that properly represent the characteristics of each class have a better chance of establishing a successful predictor. Moreover, sometimes data are redundant or take large amounts of computing time for the learning process. To overcome this issue, data selection techniques have been proposed, including instance selection. Some data mining techniques are based on nearest neighbors, ordered removals, random sampling, particle swarms or evolutionary methods. The weaknesses of these methods usually involve a lack of accuracy, lack of robustness when the amount of data increases, overfitting and a high complexity. This work proposes a new immune-inspired suppressive mechanism that involves selection. As a result, data that are not relevant for a classifier's final model are eliminated from the training process. Experiments show the effectiveness of this method, and the results are compared to other techniques; these results show that the proposed method has the advantage of being accurate and robust for large data sets, with less complexity in the algorithm.

Keywords—text mining; data reduction; classification problems; feature selection

I. INTRODUCTION

Nowadays most of the information is stored electronically, in the form of text databases. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mails, and the World Wide Web.

Text mining, also known as knowledge discovery from textual databases, is a semi-automated process of extracting knowledge from a large amount of unstructured data. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of these documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining (Feldman 1995).

There are many types of statistical and artificially intelligent classifiers, as it can be seen in [1],[2]. One of the main issues in classification problems involves the choice of

good samples to train a classifier. A training set capable to represent well the characteristics of a class has better chances to establish a successful predictor.

II. OBJECTIVES

This paper proposes a new approach for addressing the training data reduction in text mining classifications problems. This new algorithm was inspired by suppression mechanisms found in biological immune systems [3]. The suppression concept is applied to the training process to eliminate very similar data instances and to keep only representative data. The propose consists in a non-statistical method to select samples for training. The main objectives of this work are to find a subset of samples for training without spending excessive processing time and to simultaneously maintain good accuracy.

In order to do this, this paper is set out as follows. The Section 2 presents a literature review of what has been done to solve the reduction problem as well as the features and problems associated to each of them. Section 3 introduces a detailed description of the algorithm proposed and the suppression mechanism. Section 4 explains the methodology used in the experiments. Finally, Section 5 points out the conclusions and gives some direction of future work.

III. PREVIOUS WORK

An important contribution in the area of data reduction for structured data (data mining) can be found in (Cano et al. 2003). In this work, the authors present a review of the main instance selection algorithms. In addition, they perform an empirical performance study that compares the classical instance selection methods with four major evolutionary-based strategies. The authors divide the instance selection methods into four sets. The first set involves techniques based on nearest neighbor (NN) rules. These techniques are Cnn [4], Enn, Renn [5], Rnn [6], Vsm [7], Multedit [8], Mcs [9], Shrink, Icf [10], Ib2 [11], and Ib3 [12]. The second set involves methods based on ordered removal. These methods are Drop1, Drop2 and Drop3 [13]. There are two methods based on random sampling that were considered, i.e., Rmhc [14] and Enns [15]. The evolutionary-based methods are the generational genetic algorithm (GGA) [17] and [17], the steady-state genetic algorithm (SGA) [18], and the CHC adaptive search algorithm [19]. The authors in [19] claim that the execution time associated with evolutionary algorithms (EAs) represents a greater cost compared to the execution time

of the classical algorithms. However, when compared to non-EAs that have a short execution time, EA-based algorithms offer more reduction without overfitting. The authors concluded that the best algorithm corresponds to the CHC, whose time is lower compared to the rest of the EAs, the probabilistic algorithms and some of the classical instance selection algorithms. The classical and evolutionary algorithms are affected when the size of the data set increases, whereas CHC is more robust. In CHC, the chromosomes select a small number of instances from the beginning of the evolution, so that the fitness function based on 1-NN has to perform a smaller number of operations. There are many other strategies in the literature [20], [21], [22], [23], [24], [25], [26] and [27].

IV. THE SUPPRESSION MECHANISM

The suppression concept for proposed algorithm SeleSup (selection by suppressor) is employed in the training set to eliminate very similar data instances and to keep those instances that are truly representative of a certain class [28]. To perform such tasks, the mechanism divides the training database into two subsets. The first subset represents the white blood cells (WBCs) or antibodies in the organism, representing the training set. The second subset represents a set of pathogens or antigens that will select the higher affinity with WBCs; hence, this method performs suppression. The algorithm starts with the idea that the system's model must identify the best subset of WBCs to recognize pathogens, i.e., the training set, and to be able to identify new pathogens that are presented.

Both antibodies and antigens were represented as vectors containing the most relevant terms of the documents. Each vector was normalized to belong to the same scale of values which is mapped to the interval [0,1]. The affinity between antibodies and antigens was determined by the cosine distance. This measure is commonly used to measure the level of similarity between two documents.

Given two vectors representing documents, *WBC* and *Pathogen*, their cosine will describe the similarity.

As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two vectors are getting closer, or more similar.

According to [28] the algorithm aims to identify the best subset of antibodies to recognize the antigens, i.e., the new training set must be able to identify new antigens. Finally, the antibody survivors are represented by an evaluation measure (fitness value) and are selected to be a part of the new reduced training set.

In other words, those WBCs able to recognize pathogens from the suppression set remain while the others are eliminated from the population. The signals for a WBC's survival are represented by a fitness variable. Each time the nearest WBC recognizes a same class-label pathogen, the survival signal is sent and the fitness is incremented. Every WBC with a fitness greater than zero is selected to be part of the new suppressed repertoire. The pseudo-code for this technique can be seen in Algorithm 1.

Algorithm 1: The Suppressive Algorithm

input: The normalised (in[0, 1]) full training data set T and the fraction f of WBCs (default f =0.9)
output: A reduced training data set T

// Initialisation phase

Shuffle T and assign $[f \cdot |T|]$ samples as WBCs (training set); the remaining samples are assigned as pathogens (suppression set);

for all the WBCs do fitness = 0;

// Suppression phase

for each pathogen p do

NearestWBC \leftarrow Find the nearest WBC with regard to p;
if NearestWBC's class = p's class then

// NearestWBC was able to recognize the pathogen

Increment the NearestWBC's **fitness** by one;

endif;

end;

// Output phase

Eliminate those WBCs whose **fitness** value is 0;

Output the set of surviving WBCs as the reduced training set T

V. EXPERIMENTAL STUDY

In this section, the experiments presented aims to evaluate the reduced training instances selected by the SeleSup algorithm in four data sets (shown in **Error! Reference source not found.**) frequently used in information retrieval research.

TABLE I. DATA SETS FEATURES

Data set	Instances Total	Number Instance Train	Number Instance Test	Number Attributes	Number Classes
Reuters-4	1337	888	449	2833	4
Reuters-10	6689	4416	2273	2833	10
Original Reuters	8250	5169	2680	2833	62
NewsGroup	18300	16470	1830	1154	20

The Reuters-21578 Text Collection contains documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark that contains 135 classes. The collection was divided it in two subsets: one consisting of the four more balanced classes, which was identified as Reuters-4, and the other consisting of the ten most frequent classes, which was identified as Reuters-10. The third datasets consists of the sixty two classes, which was identified as Reuters-Original.

The last data set, the NewsGroup (20NG) dataset contains approximately 20000 articles evenly divided among 20 Usenet

newsgroups. Over a period of time 1000 articles were taken from each of the newsgroups, which makes an overall number of 20000 documents in this collection. Except for a small fraction of the articles, each document belongs to exactly one newsgroup (Joachims 1997).

The performance of the two classification algorithms Naive Bayes and Support Vector Machine (SVM) over the resulting reduced training and test subsets of SeleSup is compared to the performance over the subsets selected by the CHC algorithm, which is based on genetic algorithms [19] and random sampling (RS) based on the reduction percentages of experiments of each algorithm.

For each one of these subsets, the algorithms SeleSup and RS of each method were run out ten times and the reduced sets of training data were submitted to the classification algorithms (Naive Bayes and Support Vector Machine). The CHC percentage reduction, obtained in just one execution, due to computational cost was adopted. The RS was run 10 times. The average was obtained as final result for each experiment.

VI. THE DATA SETS

A. REUTERS

The first experiment performed in this paper makes use of the Reuters collection (Zeidat et al. 2006; Yang et al.1996; Schapire 1990; Schapire et al. 2000; Sebastiani 2002). The Reuters-21578 collection is a collection of documents from the Reuters news agency that was released in 1987. By 1990, the collection was given to the scientific community to perform research related to text categorisation. The rights of authorship belong to Reuters Ltd. and the Carnegie Group, which promoted its free distribution for research activities. The document basis consists of 21578 Reuters articles that consist of files in the SGML language.

These documents are grouped into 22 separate files. Each document possesses several attributes that indicate different characteristics. The attributes used in this work are: Lewisplit (related to the information of the experiments done by Lewis who defines the values Test, Training and Not-Used); Oldid, which represents the identification number of the collection (before the Reuters- 21578); D, which represents the categories or classes; and Body, which presents the text content of major news. The number of documents per class varies from class "earnings" (3964 documents) to class "castor-oil" (which contains a single document). Furthermore, some documents are not associated with any of the classes, and others are associated with up to 12 of the classes.

The SGML files were transformed into XML format and were pre-processed in Microsoft Excel, joining all documents in one single file. The resulting file was considered as the format for the input file for the mining process containing a collection of 8250 records sorted into 62 categories.

Then, the usual text mining data preparation techniques were performed. From this subset it was partitioned other two subsets: Reuters-4 and Reuters-10 as explained in next section. The four more balanced and the ten most frequent classes are indicated in Table 2 and 3.

TABLE II. FOUR MORE BALANCED CLASSES OF REUTERS DATA SET.

Class name	Samples
1 - Grain	375
2- Crude	362
3- Money-fx	313
4-Trade	287

TABLE III. TEN MOST FREQUENT CLASSES OF REUTERS DATA SET.

Class name	Samples
1 - Earn	3126
2 - Acq	1744
3 - Grain	375
4 - Crude	362
5 - Money-fx	313
6 - Trade	287
7 - Interest	154
8 - Ship	150
9 - Sugar	90
10 - Coffee	88

B. Newsgroup Data

The 20 Newsgroups data set is a collection of approximately 20000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. This collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). The **Error! Reference source not found.** presents a list of the 20 newsgroups, partitioned (more or less) according to subject matter (Table 4).

TABLE IV. NEWSGROUPS CLASSES

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

C. Parameters

The parameter setting is given in Table 5 and remained constant throughout the experiments. It was used stopwords and stemming in the document preparation stage. In addition, it was performed a filter on keywords with more than 50% significance and keyword's relevance was used to generate the vector space model.

TABLE V. PARAMETER SETTING

Algorithm	Parameter	Value
SeleSup	fraction of training samples (WBCs)	0.9
Random Supression ¹	fraction of training samples	from reduction rate of SeleSup and CHC
CHC ²	population's size	50
	number of evaluations	100/1000
	alfa equilibrate factor	0.5
	percentage of change in restart	0.35
	0 to 1 probability in restart 0 to 1 probability in diverge	0.25 0.05

¹ Implementation from POLYANALYST v 6 - <http://www.megaputer.com>

² Implementation from KEEL v2.0 rev. 2010-05-13 - <http://www.keel.es>

D. Significance Test

Statistical evaluation of experimental results has been considered an essential part of validation of the new machine learning methods [29],[30]. The statistical test has the objective of reject a false null hypothesis [31].

This paper shows a comparison between nonparametric tests, Wilcoxon signed rank test [32] and Mann-Whitney test [33] for comparing of two classifiers, Naïve Bayes and SVM. [29] mentions Wilcoxon signed rank test as safe and robust non-parametric tests for statistical comparisons of classifiers.

It was used data sets with high dimension space, which demand a high processing time. So, it was chosen the training data set of the each one of the four data sets (see Table 1), which have been run on 10-fold cross-validation method to obtain a random sample of 10 results. The test is two-tailed with significance level of 0.05. The results have been obtained through the KEEL software [34], [30] and [29].

Generally when the p value is greater than 0.05, the null hypothesis is accepted resulting as no evidence that the samples are significantly different. However, if the null hypothesis is rejected ($p < 0.05$) denotes that the samples are statistically significant.

VII. RESULTS AND ANALYSIS

The first experiment was carried out in the Reuters-4 data set. This data set is characterized by balanced classes (see Table 6 and 7). The accuracy of SeleSup is just as good as results of CHC-100 and with the same data set without reduction, the results presented are very similar. The CHC-100 produces the best performance. Therefore, CHC-100 hasn't nearly as high reduction rate as SeleSup.

The CHC-1000 has a bigger reduction, but comparing with SeleSup the accuracy don't nearly produce as good results as its. In the tests, there was only one case (CHC1000) where the performance hasn't shown significantly different.

TABLE VI. RESULTS FOR REUTERS-4 DATA SET

Reuters - 4	Reduction (%)	Naïve Bayes Accuracy Test (%)	SVM Accuracy Test (%)	Execution Time (s)
None	0.00	92.89	93.56	00:00:00
SeleSup	90.43	88.38	88.96	00:00:06
Random Sampling		88.64	89.67	00:00:00
CHC_100	77.11	93.11	92.22	00:00:04
Random Sampling		91.16	92.27	00:00:01
CHC1000	97.18	72.89	79.11	00:01:45
Random Sampling		74.53	74.71	00:00:01

TABLE VII. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR REUTERS-4 DATA SET

Reuters - 4	Mann_Whitney p-value	Wilcoxon p-value
None	1.57E-4	0.0055
SeleSup	4.39E-4	0.0055
CHC_100	2.12E-4	0.0055
CHC_1000	1.2662	0.7037

The second experiment was carried out with the Reuters-10 data set. This data set is characterized by an imbalance on its classes (see **Error! Reference source not found.**).

Therefore, as can be seen in Table 8, all the classifiers produced satisfactory results when their learning process used all the training and test data set. In addition, as expected, the same behavior occurs when suppression mechanism is applied.

The accuracy of SeleSup is just as good as results with the same data set without reduction, Random Sampling and CHC-100. The results are very similar between the classifiers. Therefore, CHC-100 has not nearly as high reduction rate as SeleSup.

It can be noticed that if the number of evaluation increases, the accuracy test of CHC-1000 decreases and consumes a high time execution (more than 50 higher). So, the CHC-1000 doesn't produce nearly as good results as SeleSup.

The results (Table 9) indicate that the Wilcoxon test is more powerful than the Mann-Whitney test according to [29].

TABLE VIII. RESULTS FOR REUTERS-10 DATA SET

Reuters - 10	Execution Time (s)	Naïve Bayes Accuracy Test (%)	SVM Accuracy Test (%)	% Reduction
None	00:00:00	92.92	93.53	0.
SeleSup	00:01:46	90.13	90.21	91.
Rand. Samp.	00:00:00	89.35	89.16	91.
CHC_100	00:58:29	91.95	91.29	77.
Rand. Samp.	00:00:01	92.00	92.06	77.
CHC1000	01:58:12	84.43	83.77	97,
Rand. Samp.	00:00:01	84.70	82.29	97.

TABLE IX. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR REUTERS-10 DATA SET

Reuters - 10	Mann_Whitney p-value	Wilcoxon p-value
None	1.57E-4	0.0055
SeleSup	1.57E-4	0.0055
CHC_100	1.57E-4	0.0055
CHC_1000	2.12E-4	0.0055

TABLE X. RESULTS FOR ORIGINAL REUTERS DATA SET

Original Reuters	Red. (%)	Naïve Bayes Accuracy Test (%)	SVM Accuracy Test (%)	Exec. Time (s)
None	0.00	83.62	87.01	00:00:00
SeleSup	91.82	78.02	78.66	00:02:30
Random Sampling		77.48	78.00	00:00:00
CHC_100	76.42	81.98	83.99	01:00:33
Random Sampling		81.54	83.57	00:00:00
CHC1000	97.12	72.61	71.83	02:43:27
Random Sampling		72.65	71.61	00:00:00

TABLE XI. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR ORIGINAL REUTERS DATA SET

Original Reuters	Mann_Whitney p-value	Wilcoxon p-value
SeleSup	1.57E-4	0.0055
CHC_100	1.57E-4	0.0055
CHC_1000	1.57E-4	0.0055

The third experiment was carried out with the Reuters Original data set. This data set is characterized by a great imbalance on its classes and high dimensionality (Table 10 and 11). SeleSup produced results almost as good as CHC-1000 in the training set, but the Reuters Original without suppression produces the best results in the test set.

It can be noticed once more that the CHC-1000 produces the best data reduction percentages, but it isn't nearly as fast as SeleSup. According to (Cano et al. 2003) the main limitation of CHC is its long processing time, which makes it difficult to apply this algorithm to very large data sets.

This experiment shows the limitations of the SVM with the larger dataset (Original Reuters) which were omitted.

Finally, the last experiment was carried out using the Newsgroup data set. This data set is an example of a very large data set with 18300 instances (see Table 12). This is the largest data set in our experiments.

The SeleSup and CHC obtained results are very similar in accuracy. In addition, the algorithm SeleSup was easily applied in this data set and its results were just as good as CHC-1000. Its processing time has been very meaningful when compared with the CHC that produces a very similar percentage of reduction (92,09% and 93,29%).

It can be observed that the RS had in general results very similar to the algorithms SeleSup and CHC, but it has a clear disadvantage of not reducing data by itself. Therefore, another algorithm has to be used to define the reduction percentage.

TABLE XII. IT IS ALSO POSSIBLE TO NOTICE THAT THERE IS NO STATISTICAL DIFFERENCE BETWEEN THE METHODS APPLIED IN THIS DATASET (TABLE 13). RESULTS FOR NEWSGROUP DATA SET

News group	Reduction (%)	Naïve Bayes Accuracy Test (%)	SVM Accuracy Test (%)	Exec. Time (s)
None	0.00	88.8	93.01	00:00:00
SeleSup	92.09	79,2	91,84	00:13:00
Random Sampling		79.5	91,18	00:00:00
CHC_100	77.01	85.1	93.55	17:12:00
Random Sampling		85.2	93.45	00:00:00

CHC_100		80.1	90.55	13:48:05
Random Sampling	93.29	78.1	90.30	00:00:00

TABLE XIII. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR NEWSGROUP DATA SET

Newsgroup	Mann_Whitney p-value	Wilcoxon p-value
None	1.57E-04	0.0055
SeleSup	1.57E-04	0.0055
CHC_100	1.57E-04	0.0055
CHC_1000	1.57E-04	0.0055

VIII. CONCLUSION

To carry out efficiently the training of classifiers of large collections of text the selection of the training set must be done carefully. If it is used an excessive number of documents the computational effort can make the task impossible. Using a very small sample leads to the inaccuracy of the classifier.

This paper presented a new method for instance selection (IS) by suppressing data in the original training set. IS can be very useful to reduce costs, improve computational performance and eliminate non-informative data. The proposed technique was designed to work together with different types of classifiers. The goal was to improve the performance related to the time spent on training without losing accuracy. This approach was inspired by the suppression mechanisms found in biological immune systems.

The experiments were conducted by testing the SeleSup algorithm in four data sets. The performance of three classification algorithms over the resulting training subsets of SeleSup was compared with the performance over the subsets selected by the CHC algorithm and random sampling (RS).

In order to test whether the algorithms' performances were significantly different or not, it was adopted a comparison between non-parametric tests Mann-Whitney U and Wilcoxon signed rank. In the tests, there were only one case where the performances haven't shown significantly different. Therefore, the statistical tests have provided strong evidence concerning the results obtained when comparing the evaluated algorithms.

The SeleSup algorithm significantly reduces the data set size. This algorithm is just as good as CHC algorithm and it offers the advantage of being faster. Then, it consumes less processing time. Although CHC has a higher reduction rate, it does not produce the best results with high dimensionality data sets and it showed high time execution. Moreover, on the contrary of CHC, the presented approach was applied to all the data sets on a less power computer, and overall, its results were better than RS.

IX. FUTURE WORK

An alternative method for performing a faster test would be inserting into the WBCs' population the pathogen-specific

WBC whose distance is the minimum distance. This technique should provide the system with the capability of keeping rare cases or rare classes in the training set.

An additional improvement to the original algorithm could be to insert some probabilistic information on the choice of the WBCs to be eliminated. The way that the mechanism works currently is deterministic with regard to data selection.

ACKNOWLEDGMENT

The authors acknowledge the support provided by CNPq, the Brazilian Research Agency, FAPERJ, the Rio de Janeiro Research Foundation and CAPES, Coordination for the Improvement of Higher Level Education.

REFERENCES

- [1] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques" Morgan Kaufmann Publishers, San Francisco, CA, 2001.
- [2] T. M. Mitchell, "Machine Learning" Mc Graw-Hill Series in Computer Science, USA, 1997.
- [3] J. Timmis, "Artificial Immune Systems: A Novel Data Analysis Technique Inspired by the Immune NetWork Theory." PhD Thesis, University of Wales, Department of Computer Science, Aberystwyth, Ceredigion, Wales, 2000.
- [4] P.E. Hart, "The condensed nearest neighbor rule" IEEE Transactions on Information Theory, 14, pp. 515-516, 1968.
- [5] D.L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data." IEEE Transaction on Systems Man Cybernetics, 2, pp.408-421, 1972.
- [6] G.W. Gate, "The reduced nearest neighbor rule." IEEE Transactions on Information Theory, 14, pp. 431-433, 1972.
- [7] D.G. Lowe, "Similarity metric learning for a variable-kernel classifier", Neural Computation, 7, pp. 72-85 1995.
- [8] P.A. Devijver and J. Kittler, "Pattern recognition: A statistical approach", Prentice-Hall International, 1982.
- [9] C.E. Broadley, "Automatic algorithm/model class selection", Proceedings of the Tenth International Machine Learning Conference, pp. 17-24.
- [10] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms". Data Mining and Knowledge Discover, 6 pp. 153-172, 2002.
- [11] D. Kibber, D.W. Aha, "Learning representative exemplars os concepts: An initial case of study." Proceedings of 4th International Machine Learning Workshop, pp. 24-30, 1987.
- [12] D.W. Aha and M.K. Albert D, "Instance based learning algorithms" Machine Learning, 6 pp. 37-66, 1991.
- [13] D.R. Wilson and T.R. Martinez, "Instance pruning techniques". In Proceedings of 14 th International Conf. Machine Learning, pp. 404-417, 1997.
- [14] D.B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms". In Proceedings os 11 th International Conference on Machine Learning, New Brunswick, NJ Morgan Kaufmann, 1994.
- [15] D.R. Wilson and T.R. Martinez, "Reduction techniques for instance-based learning algorithms". Machine Learning, 38 pp. 257-268.
- [16] D.E. Goldberg, "Genetic Algorithms in Search Optimization, and Machine Learning." Addison-Wesley longman Publishing Co., Boston, Mass, 1989.
- [17] J.H. Holland, "Adaptation in Natural and Artificial Systems". University of Michigan Press, Ann Arbor, MI, 1975.
- [18] D. Whitley, "The genitor algorithm and selective pressure: Why rank based allocation of reproductive trials ins best." In Proceedings os 3 rd Int.Conf. Gas, pp. 116-121, 1989.
- [19] J.E. Cano and M. Lozano F., "Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study". IEEE Transaction on Evolutionary Computation, 7 pp. 561-575, 2003.

- [20] A. Franco, D. Maltoni and L. Nanni, "Data pre-processing through reward-punishment editing." *Pattern Analysis and Applications*, 13, pp. 367-381, 2010.
- [21] J. Kittler, M. Hatef and J. Duin R, " On combining classifiers." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 pp. 226-239, 1998.
- [22] L. Nanni and A. Lumini, "Particle swarm optimization for prototype reduction." *Neurocomputing*, 72, pp. 1092-1097, 2009.
- [23] L. Nanni, "Experimental comparison of one-class classifiers for online signature verification", *Neurocomputing*, 69, pp. 869-875, 2006.
- [24] R. Parades and E. Vidal, "Learning Prototypes and distances: a prototype reduction technique based on nearest neighbor error minimization." *Pattern Recognition*, 39, pp. 180-188, 2006.
- [25] C. Pedreira, "Learning Vector quantization with training data selection". *IEEE TPAMI*, 18 pp. 157-162, 2006.
- [26] J. R. Cano, F. Herrera and M. Lozano F, "On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining". *Applied Soft Computing*, 6 pp. 323-332, 2006.
- [27] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98*", Tenth European Conference on Machine Learning, pp. 137-142, 1998.
- [28] G.P. Figueiredo, N.F.F. Ebecken and H.J.C. Augusto D.A. " An Immune-inspired Data Selection Mechanism for Supervised Classification, *Memetic Computing*, v. 4, pp. 135-147, 2012.
- [29] J. Demsar, "Statistical comparison of classifiers over multiple data sets." *Journal of Machine Learning Research*, 7, pp.1-30, 2006.
- [30] S. Garcia, A. Fernández, J. Luengo and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power". *Information Sciences*. DOI: 10.1016/j.ins. 2009.12.010.
- [31] S. Garcia and F. Herrera, "An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons". *Journal of Machine Learning Research*, 9, pp. 2579-2596, 2008.
- [32] F. Wilcoxon, "Individual Comparisons by Ranking Methods". *Biometrics* 1, pp. 80-83, 1945.
- [33] H.B. Mann and D.R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other". *Annals of Mathematical Statistics*, 18, pp. 50-60, 1947.
- [34] J. Alcalá-Fdez, L. Sánchez, S. García, Del, M.J. Jesus, S. Ventura, J.M. Garrell, Romero, J. Otero, C. Romero, Rivas J. Bacardit, J.C. Fernández and F. Herrera, " Keel: a software tool to assess evolutionary algorithms to data mining problems." *Soft Computing*, 13 (3), pp. 307-318, 2009.

Study of the capacity of Optical Network On Chip based on MIMO (Multiple Input Multiple Output) system

S.Mhatli¹

¹URCSE (Unité de Recherche Composants et Systèmes Electroniques), Ecole Polytechnique de Tunisie
Université de Carthage, 2078, La Marsa, Tunisie

B.Nsiri², R.Attia¹

²SYSCOM (Laboratoire Système de communication), Ecole National D'ingénieur de Tunis
Université Tunis el Manar, Tunisie

Abstract—When designing Optical Networks-On-Chip, designers have resorted to make dialogue between emitters (lasers) and receivers (photo-detectors) through a waveguide which is based mainly on optical routers called λ -router. In this paper, we propose a new method based on the Multiple Input Multiple Output concept, and we give a model of the channel propagation, then we study the influence of different parameters in the design of Optical Networks-On-Chip.

Keywords— λ -ROUTER; MIMO CHANNELS; CAPACITY; CDMA

I. INTRODUCTION

Networks-on-Chip (NoC) have recently become popular as an option for increasing the bandwidth, lowering the latency and reducing the power in chip multiprocessors.

Several network architectures have been presented in the literature to construct efficient photonic networks-on-chip, and this architecture is based on λ -router.

We present here an overview of On-chip optical networks based on (λ -router). The figure below shows an example of a network-on-chip based on an optical waveguide using optical router architecture liabilities (the λ -router) [1,2].

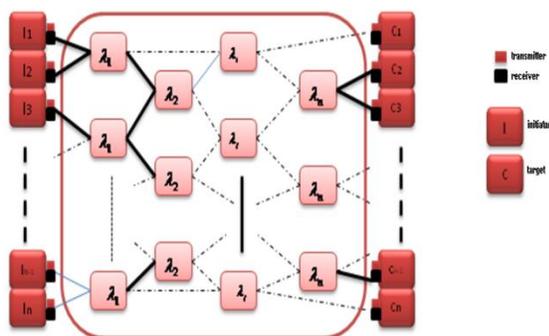


Fig. 1. On-chip optical networks based on λ -router

The basic element of the network is the λ -router [3] which consists of two basic elements:

- * Two parallel waveguides.
- * A ring cavity, square...

In this paper, we will study and present a novel ONOC (Optical Network On Chip) system based on MIMO technology; we begin in section II by the modeling the propagation's channel of the system which allows us to determine the attenuation between transmitter and receiver. Then in section III, we present the study of ONOC capacity and finally in section IV, numerical results are presented.

II. MIMO CHANNEL MODELING

In this section, we solve Maxwell's equations to determine the electromagnetic field equation that describes the outgoing laser light [4] (channel 1) and we modeling the diffuser (channel 2) which is an important component to diffuse light to all receivers.

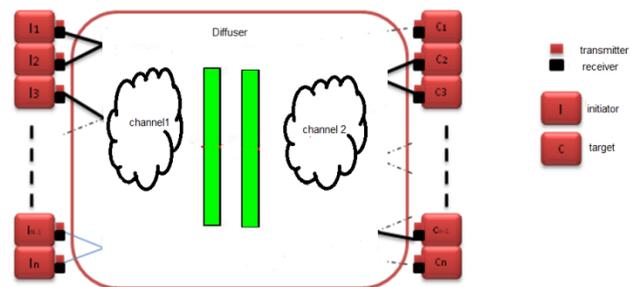


Fig. 2. On-chip optical networks based on MIMO technology

A. Channel 1 modeling

We assume that electromagnetic wave Propagates in a homogeneous medium is subject to Maxwell's equations. Thus, the equation of wave propagation in isotropic medium is:

$$\Delta \vec{E} - \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = \vec{0} \quad (1)$$

If we consider the propagation of a monochromatic electromagnetic wave frequency then we have:

$$\Delta E(x, y, z) + k^2 E(x, y, z) = 0 \quad (2)$$

The solution of this equation is :

$$I(r, z) = I_0(z) \exp\left(\frac{-2r^2}{w^2(z)}\right) \quad (3)$$

$w(z) = w_0 \sqrt{1 + \left(\frac{z}{z_R}\right)^2}$: Is a measure of the amplitude of Gaussian field with distance from the z-axis.

The percentage of energy F received by the receiver defined by:

$$F = \frac{\int_0^{\rho} I(r) ds}{\int_0^{\infty} I(r) ds} \quad (4)$$

Replacing I(r) by its value at a given point z, s we have after simplification, the function that represent the channel 1:

$$F = 1 - \exp\left(-2\left(\frac{\rho}{w}\right)^2\right) \quad (5)$$

B. Channel 2 modeling

As lasers scatters light linearly, the diffuser appears as a solution to distribute the quantity of light received at the receivers, in this case we say that the diffusion process is done in a Lambertian.

Existing broadcasters are either single surface or double surfaces [5,9].



Fig. 3. diagram of a diffuser

The diffuser is an optical component, composed of several different micro-lenses, designed in a manner that each micro-lens arranged to avoid repetition pattern, so that there's control over the distribution of diffusion and intensity profile [8].

For each diffuser:

$$I_0(\theta) = \begin{cases} \cos^p\left(\frac{\pi}{2} \frac{\theta}{\theta_0}\right), & |\theta| \leq \theta_0 \\ 0 & \text{else} \end{cases} \quad (6)$$

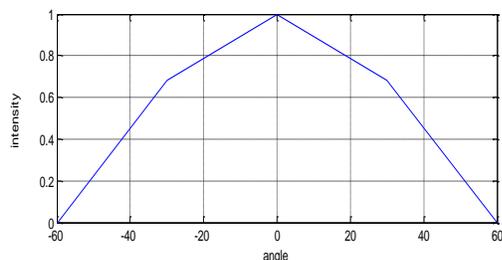


Fig. 4. Response of a diffuser

In our study we choose $\theta_0 = 60^\circ$ and $p = 0.6$.

It is a simple diffuser with low spectral efficiency [5]. In order to increase the spectral efficiency we adopt the use of two broadcasters which are placed one in front of the other as it is shown in the following figure:

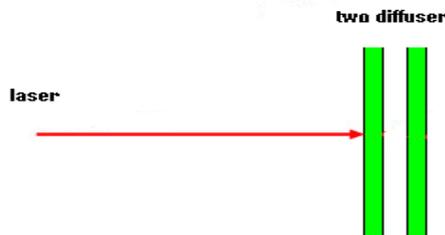


Fig. 5. diagram of two diffusers

The total scattering given by two broadcasters is given by the convolution product:

$$I(\theta) = \int_{-90}^{90} I_0(\varphi) I_0(\theta - \varphi) d\varphi \quad (7)$$

In our study we choose $\theta_0 = 60^\circ$ and $p = 0.6$.

The spectral efficiency is 70% [5] and there is an increasing scattering angle going from 120° to 180° .

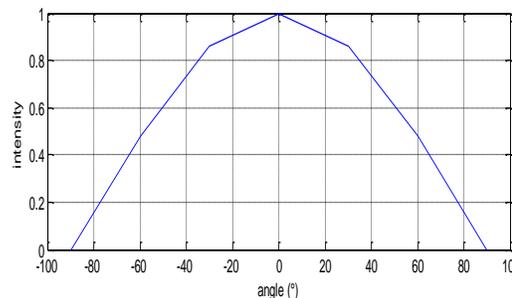


Fig. 6. Response of two diffusers

Then, the channel 2 is represented with:

$$h_{ij} = \frac{A}{d_{ij}^2} I(\theta) \cos(\varphi_{ij})$$

h_{ij} : The coefficient of the MIMO channel.

III. MIMO CHANNEL CAPACITY

To better understand and present the theoretical formulation of the MIMO channel capacity perspective, we introduce the notion of capacity based on the theory of information.

When the channel matrix is known and in reception, it was shown in [7] that the mutual information described above is

maximized if $\Phi = \frac{\sigma_e}{Nt}$, σ_e is the variance of the transmitted signal in the case where it follows a normal distribution.

The formula for the ergodic capacity will be given by [6]:

$$C = \left(\frac{1}{2}\right)E \left[\log_2 \det \left(I_{N_r} + \frac{SNR}{Nt} HH^H \right) \right] \quad (8)$$

With I_{N_r} represents an identity matrix $N \times N$ and H is the channel matrix representing the attenuation between each sub-channel.

When the channel is known at both transmitter and receiver, the optimal solution of capacity is a water-filling solution described in 1995 by Telatar [6].

To fully express the ability of this technique, we rewrite the model of the transmission system, to do this we apply the theorem of singular value decomposition of the matrix H which allows us to apply and explain the impact of one of the power allocation techniques of the MIMO system performance. So H will be as follows:

$$H = UDV^H \quad (9)$$

The received signal becomes:

$$Y = HX + N = UDV^H X + N \quad (10)$$

We can now write the system as equivalent:

$$\tilde{Y} = D\tilde{X} + \tilde{N} \quad (11)$$

With:

$$\tilde{Y} = U^H Y, \tilde{X} = V^H X, \tilde{N} = U^H N.$$

Using Rule determinant, we have:

$$\det \left(I_{N_r} + \frac{SNR}{Nt} H\Phi H^H \right) = \det \left(I_{N_r} + \frac{SNR}{Nt} \Phi H^H H \right) \quad (12)$$

By replacing the expression of H in this last formula we obtain:

$$\det \left(I_{N_r} + \frac{SNR}{Nt} \Phi H^H H \right) = \det \left(I_{N_r} + \frac{SNR}{Nt} \Phi V D^2 V^H \right)$$

Now apply the rule of determining the last formula becomes:

$$\det \left(I_{N_r} + \frac{SNR}{Nt} \Phi V D^2 V^H \right) = \det \left(I_{N_r} + \frac{SNR}{Nt} D V^H \Phi V D \right)$$

Where $\tilde{Q} = V^H \Phi V$ corresponding to the covariance matrix of the equivalent signal, the covariance matrix of the received signal becomes equivalent $A = D\tilde{Q}D$, for A diagonal must \tilde{Q} is also diagonal and in this case the expression of A becomes: $A = \tilde{Q}D^2$.

Using this latter approach of the covariance matrix of the received signal equivalent we can express the MIMO channel capacity using the technique water-filling.

Using the technique of water-filling [7], we seek to optimize the connection between the transmitter and receiver by dividing the total power transmitted on the transmit antennas in order to achieve optimal capacity offered by the system channels MIMO.

To express the optimal capacity, consider the transmission of symbols on a chain of transmission using the technique of Waterfilling such that the matrix $A = \tilde{Q}D^2$ where $\tilde{Q} = \text{diag}(a_1, a_2, \dots, a_{N_t})$.

In practice, a condition is imposed for power: the total power transmitted on all transmitters is equal to the total power transmitted:

$$\sum_{i=1}^{N_t} P_i = P_t \quad (13)$$

This condition can be written differently as follows:

$$\sum_{i=1}^{N_t} a_i = P_t \quad (14)$$

The formula of the total system capacity allocation technique using a constrained power is as follows:

$$C = \left(\frac{1}{2}\right)E \left[\sum_{i=1}^n \log_2 \left(1 + a_i \frac{SNR}{Nt} \lambda_i \right) \right] \quad (15)$$

With $n = \min(Nt, Nr)$ and λ_i is a diagonal matrix D^2 . Suppose μ constant to check the power constraint in the case of MIMO system using water-filling technique, we can write the formula for the optimal power allocation as follows:

$$a_i = \frac{Nt}{SNR} \left(\mu - \frac{1}{\lambda_i} \right)^+$$

With the symbol $(Z)^+$ means:

$$(Z)^+ = \begin{cases} Z & \text{si } Z > 0 \\ 0 & \text{si } Z \leq 0 \end{cases} \quad (16)$$

Thus, the power output of the transmitter i will be $P_i = \left(\mu - \frac{1}{\lambda_i} \right)^+$ and therefore the water-filling algorithm is optimal power allocation such that :

$$P_i = \left(\mu - \frac{1}{\lambda_i} \right)^+.$$

Water-Filling technique is to have variable power level transmitters. This change is made so that we have an optimization of the channel capacity.

The ability of such a system Water-Filling is given by the following expression:

$$C = (1/2)E \left[\sum_{i=1}^n \log_2(\mu\lambda_i)^+ \right] \quad (17)$$

Or

$$C = (1/2).E \left[\sum_{i=1}^n \log_2(1 + P_i\lambda_i) \right]$$

Since $P_i = \left(\mu - \frac{1}{\lambda_i} \right)^+$

IV. NUMERICAL RESULTS

A. Study of the capacity enhancement with water-filling

We simulate the ability of the capacity previously presented with and without Water-Filling technique to show the contribution of this technique in terms of capacity:

We simulated an optical MIMO transmission system with the use and no use of the Technical Water-Filling knowing that there's direct sight between transmitters and receivers lasers and photodiodes result of this simulation is given by the following figure:

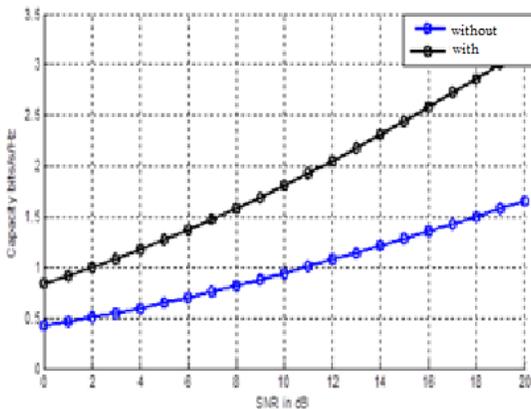


Fig. 7. Capacity = f (SNR) of a MIMO system with or without water-filling technique

This figure shows the improvement carried by Water-Filling technology on the capacity of MIMO channels in the case of sight between transmitters and receivers. Indeed, for low SNR the capacity of the transmission chain using the technique of power allocation Water-Filling is greater than other that does not use it and this increase improves if we increase the SNR more and more. For example if SNR = 12 dB, the ability of the system using the Water-Filling is 2 bits / s / Hz. However, the ability of the chain without using this technique is 1 bit / s / Hz. For large values of SNR the effect is

remarkable, there's an increase of the capacity in the area between the two curves.

We also simulated a chain of MIMO transmission in the absence of optical diffusers knowing that there's a direct sight between the transmitters and receivers.

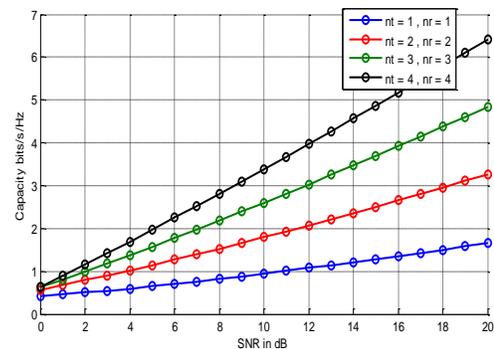


Fig. 8. Comparison of MIMO capacity without Water-Filling

This figure shows the improvement achieved by the addition of a laser on the capacity of MIMO channels in the case of sight between transmitters and receivers. Indeed, for low SNR the capacity of the transmission chain of the various systems are very close but if we increase the SNR, it gives a duplication of bit rate and this is a logical result considering that there is no overlap of data between sub-channels. For SNR = 12 dB, for example the ability of the SISO system is 1 bit / s / Hz, 2x2 MIMO system is 2 bits / s / Hz, 3x3 MIMO system is 3 bits / s / Hz, the system MIMO 4x4 is 4 bits / s / Hz.

B. Parametrics analysis of an ONOC design

We assume that we have a MIMO system as shown

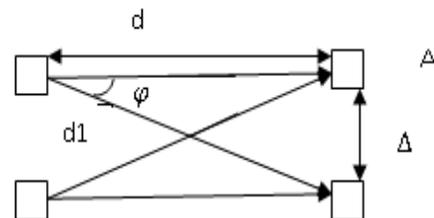


Fig. 9. Schematic of 2 x 2 MIMO systems

In our study we choose:

$$\theta_0 = 60^\circ \text{ and } p = 0.6.$$

The coefficient of the MIMO channel is given by h_{ij}

$$h_{ij} = \frac{A}{d_{ij}^2} I(\theta) \cos(\varphi_{ij})$$

From the diagram above, we have:
$$\begin{cases} d^2 + \Delta^2 = d_1^2 \\ \text{tg}(\varphi) = \frac{\Delta}{d} \end{cases}$$

A: the surface of the photo-detector is equal to $0.025mm^2$.

We assume the 2x2 MIMO system before, then we set the distance $d = 1mm$ between lasers and photo-detectors and we change different positions of photo-detectors.

We assume that all other parameters are fixed. Under the same conditions we get the following results:

$$\begin{cases} h_{11} = h_{22} \\ h_{12} = h_{21} \end{cases}$$

We simulated a 2x2 MIMO system and the result is represented by the following figure:

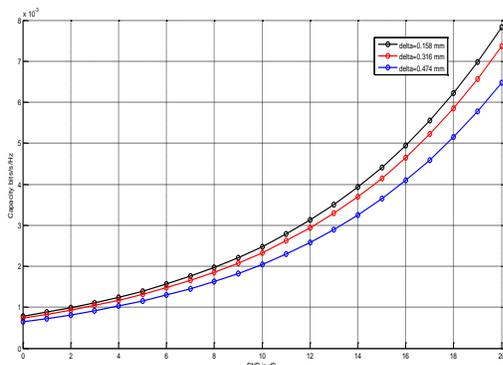


Fig. 10. Capacity = f (SNR) of a 2x2 MIMO system with variation of the result Δ

From the figure above we can see the degradation of the capacity, if we increase the distance between the photo-detector which is an expected result because when the light reaches the diffuser we find that a quantity of light is lost because spectral efficiency diffuser is 70%. As the two photo-detectors are close and there is no distance between them, we find the optimal capacity with a delta value equal to 0.158 mm, which corresponds to two photo-detectors just one after another.

We simulated a 2x2 MIMO system and we vary the distance D, we obtain the following figure:

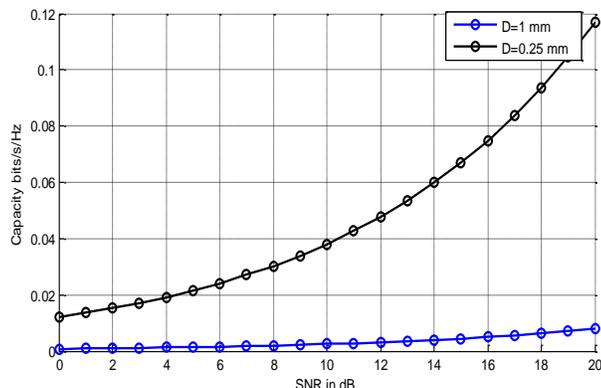


Fig. 11. Capacity = f (SNR) of a 2x2 MIMO system with variations of distance d

From the figure above we observe the degradation of the capacity if we increase the distance between the lasers and photo-detectors which is an expected result.

This figure below shows the improvement achieved by the addition of a laser on the capacity of MIMO channels. Indeed, for low SNR the capacities of the transmission chain of the various systems are very close but if we increase the SNR, there is duplication in the capacity.

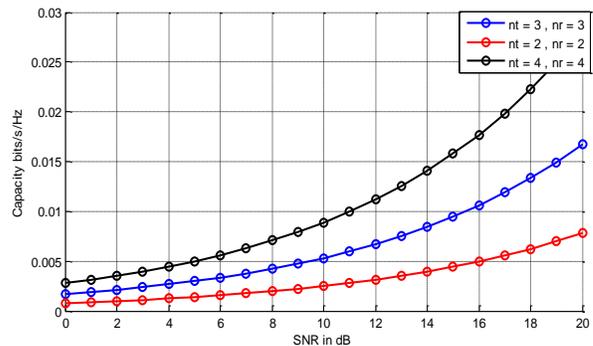


Fig. 12. Capacity = f (SNR) of the MIMO system with different technology water-filling

V. CONCLUSION

We detailed in this paper MIMO channel modeling of ONOC system based on the MIMO technologies. Then we studied the capacity with knowledge of transmission channel for both transmission and reception where we have very detailed Water-Filling technology that allows us to optimize the capacity of MIMO channels and finally we show the important parameters which enhance the capacity of ONOC system design.

After system design, we work now on the performance evaluation of this system with Code Division Multiplexing Acces code.

REFERENCES

- [1] Balac Stéphane: " Modélisation de micro-résonateurs optiques," FOTON ENSSAT de Lannion Université de Rennes 1.
- [2] Atef Allam and Ian O'Connor, Alberto Scandurra: "Optical Network-on-Chip Reconfigurable Model for Multi-Level Analysis", 978-1-4244-5309-2/10©2010 IEEE.
- [3] G. F. Fan ; R. Orobtcouk ; J. M. Fédéli : " Highly integrated optical 8x8 lambda-router in silicon-on-insulator technology: comparison between the ring and racetrack configuration ", Proc. SPIE 7719, Silicon Photonics and Photonic Integrated Circuits II, 77190F , May 17, 2010
- [4] Forget Sébastien: "Optiques des lasers et faisceaux gaussiens", Cours, Exercices et exemples d'applications.
- [5] Tasso R.M.SALES, Donald J.Schertler : "Deterministic microlens diffusers for lambertian scatter", RPC, RPC photonics, Inc, Newyork.
- [6] I. E. Telatar: "Capacity of multi-antenna Gaussian channels", Technical Memorandum, Bell Laboratories, Lucent Technologies, October 1995. Published in European Transactions on Telecommunications, 10(6):585–595, December 1999.
- [7] Wang Liejun: "An Improved Water-filling Power Allocation Method in MIMO OFDM Systems", Information Technology Journal, 10: 639-647,2011.
- [8] RPC photonics. <http://www.rpcphotonics.com/>.
- [9] Luminic optics. <http://www.luminicco.com>.

Face Recognition as an Authentication Technique in Electronic Voting

Noha E. El-Sayad

Electrical Engineering Department
Faculty of Engineering, Port-Said
University PortFouad42523,
Port-Said, Egypt

Rabab Farouk Abdel-Kader

Assistant Prof. Electrical
Engineering Department
Faculty of Engineering, Port-Said
University PortFouad 42523, Port-
Said, Egypt

Mahmoud Ibraheem Marie

Assistant Prof. In Computer and
System Engineering Department
Faculty Of Engineering, Al-Azhar
University

Abstract—In this research a Face Detection and Recognition system (FDR) used as an Authentication technique in online voting, which one of electronic is voting types, is proposed. Web based voting allows the voter to vote from any place in state or out of state. The voter's image is captured and passed to a face detection algorithm (Eigenface or Gabor filter) which is used to detect his face from the image and save it as the first matching point. The voter's National identification card number is used to retrieve and return his saved photo from the database of the Supreme Council elections (SCE) which is passed to the same detection algorithm (Eigenface or Gabor filter) to detect face from it and save it as second matching point. The two matching points are used by a matching algorithm to check wither they are identical or not. If the results of the matching algorithm are two point match then checks wither this person has the right to vote or not. If he has right to vote then a voting form is presented to him.

The result shows that the proposed algorithm capable of finding over 90% of the faces in database and allows their voter to vote in approximately 58 seconds.

Keywords—*Electronic Voting; Face Recognition; Gabor Filter; Eigenface.*

I. INTRODUCTION

Online voting system is a voting system in which the election data is recorded, stored and processed primarily as digital information and it needs to address, obtain, mark, deliver, and count ballots via computer. Therefore voter identification and authentication techniques are essential for more secure platform mechanisms to overcome vulnerabilities of the client used by the voter to cast her vote.

Security can be achieved using some of techniques of electronic voting such as *Guidelines*, only need to develop a list of instructions and then send it via email or put it on the election web page; *Bootable CD*, approach to overcome the secure platform problem was proposed by Otten (2005). She recommended developing a special voting operating system based on Knoppix. It is an operating system based on Debian that is designed to be booted and run directly from a CD or DVD; *Smart Cards as Observers*, in which an observer is a manipulation resistant piece of hardware which is owned by the voter. The idea is that the observer is not allowed to directly communicate with the Internet. All the communication needs to be forwarded by the voter; *Code*

Sheets, the idea of code sheets is that the voter gets a piece of paper together with the general election information via post where each candidate or each party is linked to a particular code. Now, in order to cast a voter the voter does not click on the candidate or party of her choice but enters the corresponding code; *Trusted Computing*, the idea is to use an appropriate security architecture based on a security kernel and on Trusted Computing elements. Such a solution is the only one that could efficiently overcome malicious software on the voting casting device as well as potential malicious voters installing malware on purpose on their device. However, currently, there are still open problems with Trusted computing itself and it is not wide-spread enough; *Individual Verifiability* [5], the idea is that you use one software to prepare a voter and a second one to verify that the vote has been properly prepared (encrypted). Plus, you can also do the verification with an offline tool

In this research, we proposed an authentication technique using a Face Detection and Recognition system in online voting to achieve the rules of Supreme Electoral Council as follow: Only eligible persons vote, No person gets to vote more than once, the vote is secret, and each (correctly cast) vote gets counted and to achieve the aims of online voting as follow: increase participation, lower the costs of running elections, and improve the accuracy of results.

In general, an FDR system starts by Interfacing with an image source for grabbing facial images, Automatic detection or manual selection of human face may be found within the scene, Manipulate (create, add, delete) a database of faces, Launching the recognition process by comparing the face previously detected with the database's faces.

The remainder of this paper is organized as follows: Section 2 is devoted to mention the previous related work. In Section 3, explain the proposed algorithms (Gabor filter and Eigenface). Experimental results and comparison between two algorithms in Section 4. Finally concluding remarks are drawn in Section 5.

II. RELATED WORK

In related research, several voter identification and authentication techniques were introduced to secure voting platforms and overcome fake voting. Some of these techniques are:

Highly Secure Online Voting System with Multi Security using Biometric and Steganography, the basic idea is to merge the secret key with the cover image on the basis of core image. The result of this process produces a stego image which looks quite similar to the cover image. The core image is a biometric measure, such as a fingerprint image. The stego image is extracted at the server side to perform the voter authentication function. It used secret message with 288 bit length. As the actual secret key is never embedded in the stego image, there will be no chance of predicting secret key from it [9].

Karlof et al., combines the verifiability definition without distinguishing universal or individual as follows: "Verifiably cast-as-intended means each voter should be able to verify his ballot accurately represents the vote he cast. Verifiably counted-as-cast means everyone should be able to verify that the final tally is an accurate count of the ballots." [11]

Online Signature Verification Using Temporal Shift Estimated by the Phase of Gabor Filter, A new online signature-verification method using temporal shift estimation is presented. Local temporal shifts existing in signatures are estimated by the differences of the phase outputs of Gabor filter applied to signature signals. An input signature signal undergoes preprocessing procedures including smoothing, size normalization and skew correction, and then its feature profile is extracted from the signature signal. A Gabor filter with the predetermined center frequency is applied on a feature profile, and phase profile is computed from the phase output. The feature profile and the phase profile are length normalized and quantized so that a signature code of fixed size is generated. The temporal shifts existing between two signatures are computed by using the differences between the phase profiles. The information about the temporal shifts is used as offsets for comparing the two feature profiles. Therefore, two kinds of dissimilarities are proposed. Temporal dissimilarity is a measure reflecting the amount of total temporal disturbance between the two signatures. The difference between the two signature profiles is computed at each corresponding point pair and is accumulated into temporally arranged feature profile dissimilarity. The decision boundary is represented as a straight line in the dissimilarity space whose two axes are the two dissimilarity measures. The slope and the position of the decision boundary are computed using the distribution of the dissimilarities among the sample signatures involved in the enrollment procedure [9], [11].

III. IMPLEMENTATION

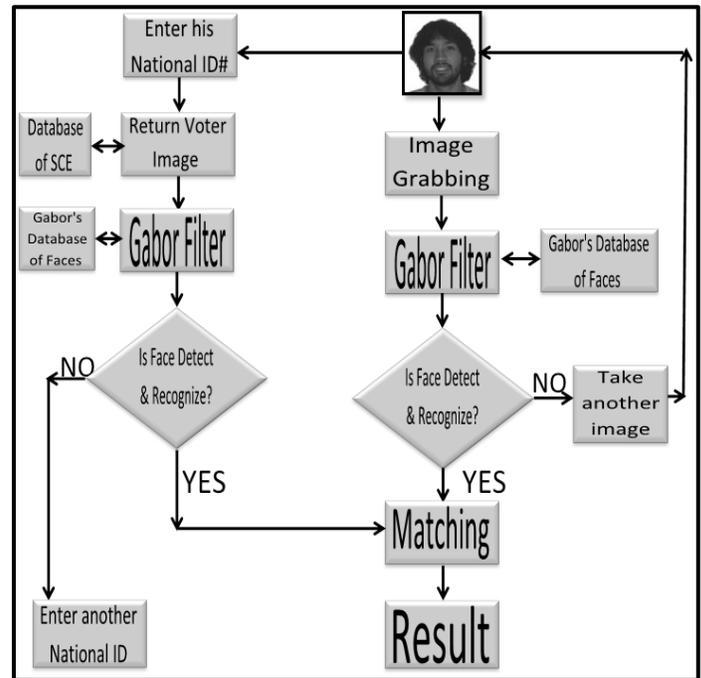
Identity authentication applied for online voting system: The Online Voting System is an online application designed to be operated by two users, the election controller or administrator and the voter.

"Identity Authentication" generally involves two stages: the first is Face Detection and Recognition, where a photo is

searched to find any face in it. Next, an image processing algorithm is applied to clean up the facial image for easier recognition. The second stage is Face Matching, where the detected face is compared to an image retrieved from the SCE database using a national ID#. A matching algorithm is applied to verify the person for both matching.

We compared between two Identity Authentication algorithms. The first algorithm is face recognition using Gabor filter and the second algorithm is Face Recognition with Eigenface.

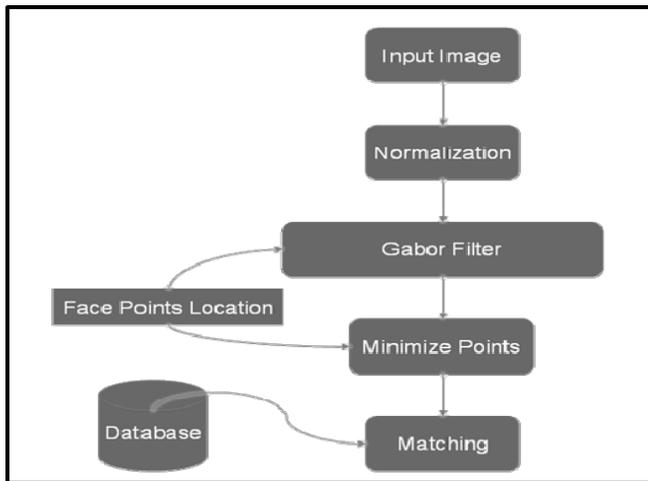
A. **First algorithm**, a schematic diagram for the Online Voting System Based on Face Recognition using Gabor Filters is show in Figure 1.



Online Voting System Based on Face Recognition using Gabor Filters.

Algorithm steps are as follows:

- 1) The voter's image which is captured using a webcam is used as the input to the face detection algorithm.
- 2) Before entering image to Gabor filters, it must be normalized by three steps as show in Figure 2.
 - a) Input image is resized to 128×128.
 - b) Pixel adjustment, in this step, Image Pixel intensities are used, such that the standard deviation of Image Pixel is one.
 - c) Borders are smoothed, across band 30 pixels wide and they are weighted by an aspect $d= 30$, where d is distance of image edge.



Normalize image

3) Gabor filter algorithm consists of 40 filter used to detect faces from the captured image; the proposed system applied different Gabor filters on the image to generate 40 images with different angles and orientation [1] [2] [8].

4) Next, maximum intensity points in each filtered image are calculated and marked as fiducial points. If the distance is minimum between these face points then system reduces the points. The next step is calculating the distances between the reduced points using distance formula. At last, the calculated distances are compared with Gabor database. If match occurs, it means that the image is recognized as a face.

5) Eq. (1) shows the major expression of Gabor wavelet as a Gaussian kernel function which is changed by sinusoid [1].

$$W(x, y) = \mathcal{F} e^{-\frac{x'^2 + y'^2}{2x^2}} (\cos(2\pi\mathcal{F}x' + \theta) - DC) \quad (1)$$

Where,

$$DC = \cos \theta e^{-2\pi\frac{e^2}{\mathcal{F}^2}}$$

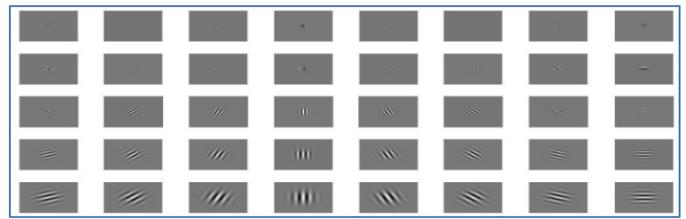
And,

$$\begin{aligned} x' &= x \cos \theta + y \sin \theta, \\ y' &= -x \sin \theta + y \cos \theta \end{aligned}$$

The factor $F=\sigma/K$ makes sure the filter spatial range of action is partial correspondingly to the central frequency f . In this equation σ is frequency of filter, θ are 8 different orientations in the filter. A discrete recognition of Eq. (1) using five dissimilar scales and eight angles are engaged and in result of it, 40 Gabor filters are acquired. In the end of Eq. (1), the term DC creates the filter DC -free.

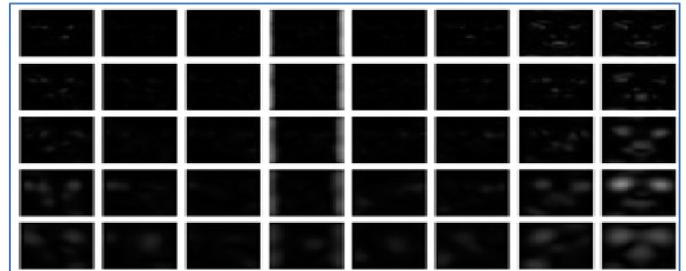
6) Then, Gabor filters are applied on the input images

a) Gabor Filters: an image of the 40 Gabor filters are shown below in figure 3.



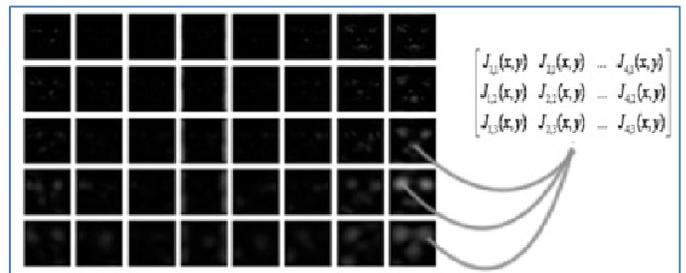
Gabor Filters

b) Results of applying Gabor Filter on image show in figure 4.



Result of applying Gabor Filter

c) Face point extraction, When the original image $I(x, y)$ multiplied with Gabor filter $g(x, y)$, a new image is acquired which is equal to $J(x, y)$. Where x is height and y is width of image show in figure 5.



Face point extraction

d) Find the maximum intensity face points using the formula in Eq. (2).

$$\sum_{i=1}^{N1} \sum_{j=1}^{N2} (\max(x_{ij})) \quad (2)$$

Where:

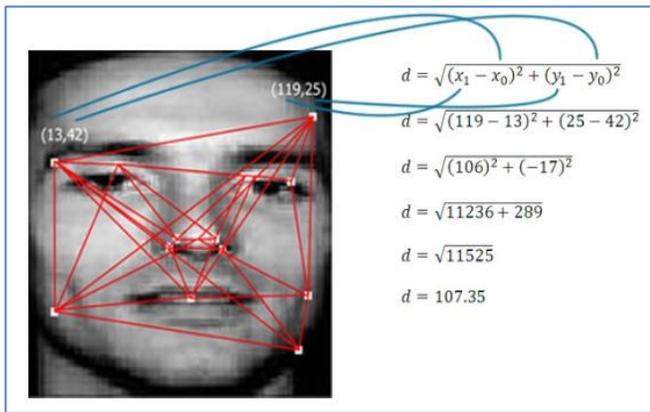
$$X = I_{ij}$$

I = Intensity at coordinate i, j

Where $N1, N2$ are the width and height of image.

e) Calculate the distance d to minimize the points as in Eq. (3).

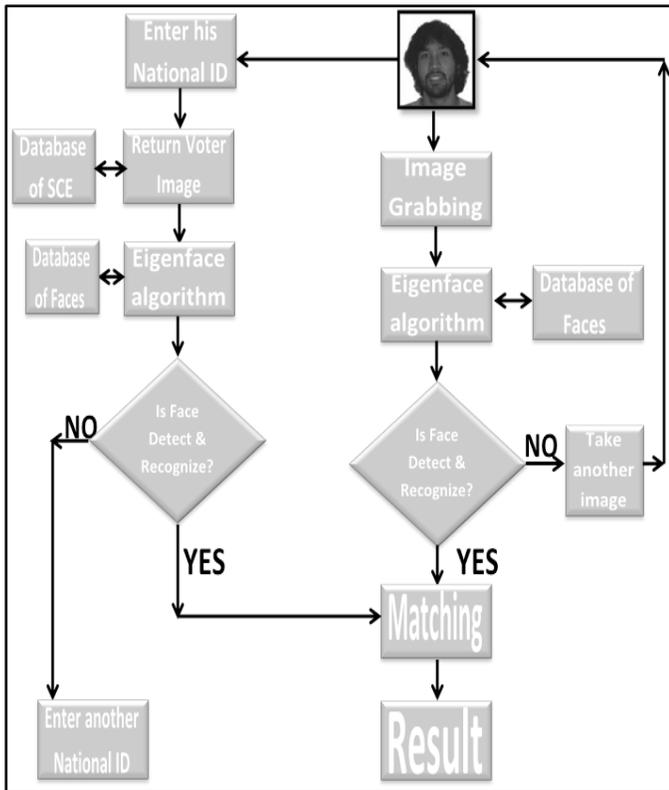
$$d = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2} \quad (3)$$



Calculate of distance d

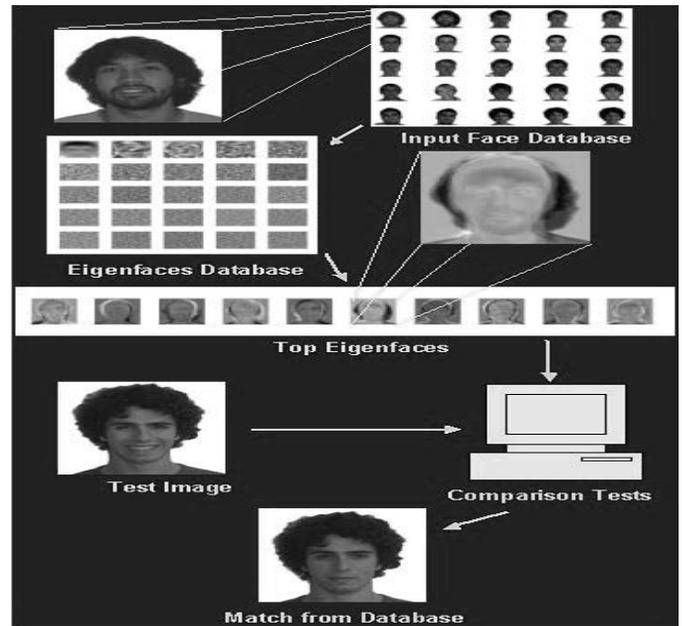
f) The distances of the selected points which show in figure 6 are compared with the database; if the distances get matched with database the face is recognized.

B. **Second algorithm**, Online Voting System Based on Face Recognition using Eigenface Filters. The suggested system is illustrated in figure 7.



Online Voting System Based on Face Recognition using Eigenface system.

We used Eigenface system to detect and recognize face from image; this system can be divided into two main segments: creation of the Eigenface basis and recognition of a new face [3] [10]. The system follows the following general flow as show in figure 8.



The Eigenface face recognition system segments.

The Eigenface technique uses much more information by classifying faces based on general facial patterns. These patterns include, but are not limited to, the specific features of the face [7]. Eigenface can be related directly to one of the most fundamental concepts in electrical engineering: Fourier analysis.

Eigenface system needs a database of known faces in which all images are the same size (in pixels), and grayscale, with values ranging from 0 to 255. Each face image is converted into a vector of length N (where, $N = \text{image width} * \text{image height}$).

We will use The ORL Database of Faces (AT&T database) [6]. This database contains 400 pictures of 40 subjects. A preview image of the database of faces is shown in figure 9.



A preview image of the Database of Faces.

Eigenface system use of Fourier analysis reveals that a sum of weighted sinusoids at differing frequencies can recompose a signal perfectly. In the same way, a sum of weighted Eigenfaces can seamlessly reconstruct a specific person's face. The algorithm calculates the average face in face space and returns the top Eigenface vectors then it uses these differences to compute a covariance matrix C for our dataset. The covariance between two sets of data reveals how much the sets correlate [4], [7] as in Eq. (4).

$$C = \frac{1}{M} \sum_{n=1}^M \Phi_n \Phi_n^T = \frac{1}{M} \sum_{n=1}^M \begin{pmatrix} var(p1) & \dots & cov(P1, PN) \\ \vdots & \dots & \vdots \\ cov(PN, P1) & \dots & var(PN) \end{pmatrix}_n = AA^T \quad (4)$$

In Equation 4,

$$A = [\Phi_1 \Phi_2 \dots \Phi_M],$$

$p_i = \text{pixel } i \text{ of face } n,$

$M = \text{the number of faces in our set.}$

The Eigenfaces that we are looking for are simply the eigenvectors of C . However, since C is of dimension N (the number of pixels in our images), solving for the Eigenfaces gets ugly very quickly. Eigenface face recognition would not be possible if we had to do this. This is where the magic behind the Eigenface system happens.

The output of the Eigenfaces system is the first point extraction of the person's face which we will be used to verify the voter's identity. The voter will enter his ID number which is used to fetch his image from SCE data base this image will be considered as the first point. The voter's image captured using a webcam is the input to the Eigenface system to detect the face from image and this will be the second point. The two points are checked for matching using pattern matching algorithm.

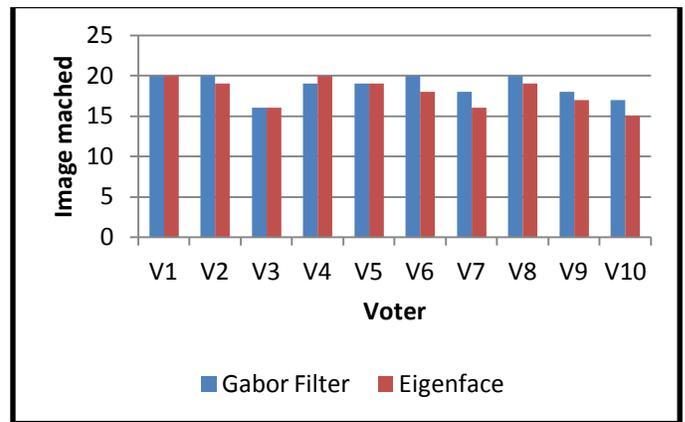
IV. ANALYSIS

To measure the performance of the two proposed systems (Gabor filter and Eigenface algorithm) the standard database and SCE data base were used. The number of matched images and the execution time were calculated and used to compare between the two proposed systems.

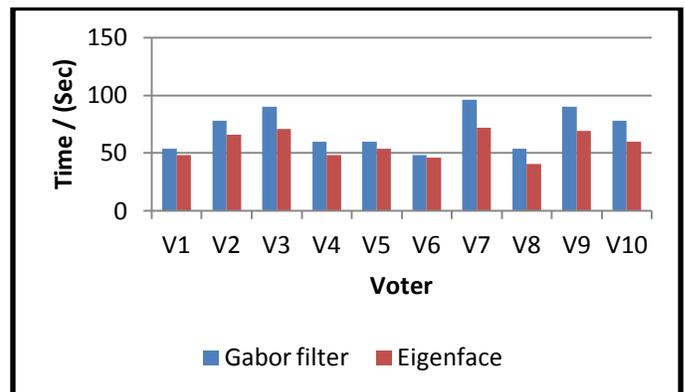
Images for ten different persons were used. For each person 20 images in different positions and different facial expressions were saved in the database. The proposed Gabor filter system was tested using 200 images. For each person 20 different images were tested. Figure 10, 11 represent the number of matched images and the average execution time respectively for both Gabor filter system and Eigenface system.

It has been observed that in the absence of face-facing webcam or in the case the of a rotated face captured image the efficiency of the Gabor filter algorithm is 80% as in the cases of voters 3, 7, 9 and 10.

While, in a face- facing webcam or a no rotated face the efficiency raises up to 100% as in the cases of voters 1, 2, 6, 8. And the efficiency of the Eigenface algorithm is 75% as in the cases of voters 10. While, in a face- facing webcam or a no rotated face the efficiency raises up to 100% as in the cases of voters 1.



Number of images matched for every voter.



Execution time for every voter.

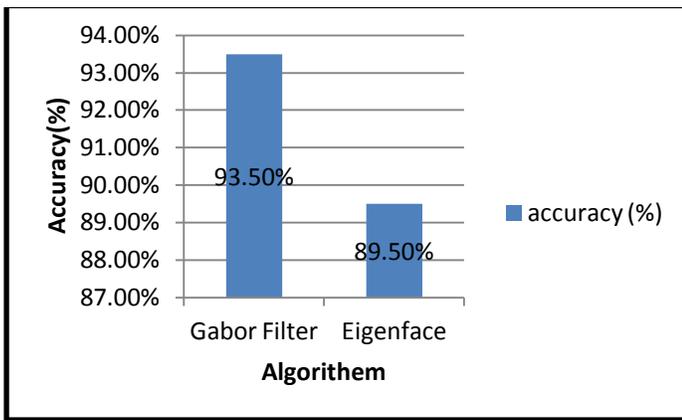
The results showed a 93.5% match in an average 70.8 Second for Gabor system. The same test set of images was used by the Eigenface system and resulted in 179 images were matched and 21 were not matched in an average of 57.48 Second. Total results are shown in figure 11, 12.

In cases of Eigenface, the recognition rates for a given number of Eigenfaces are reached relatively quickly. This indicates that in any implementation of such a recognition system there does not exist a meaningful advantage to using more Eigenfaces than first provide the desired level of accuracy and speed.

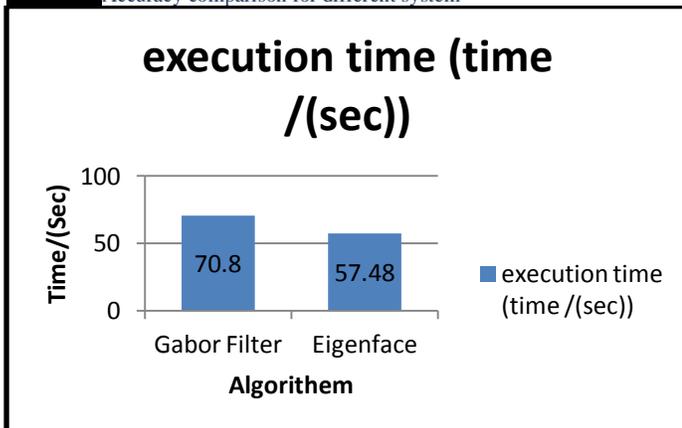
V. CONCLUSION

On line voting allow voter to vote 24 hour per day and 7 day per week also allow him to vote from anywhere in his state or out of state.

In this research, we proposed two FRD systems based on Gabor and Eigenface as authentication techniques in online voting. Both systems detect the face from an image captured using a webcam and recognize face from SCE database and check if the two images match. If a match accrues, then verify that the law and roles of voting are not violated then allow him to vote.



Accuracy comparison for different system



Execution time comparison for different system

Analysis of the Eigenface recognition technique using both averaging and removal methods gives evidence that the methods prove 89.5% accurate. But we achieve 93.5% accurate when we applied Gabor filter, but it take more time to recognize face from image.

In future work, we plan more extensive experimentation with a larger images database. We also plane on trying other good performance face detection and matching algorithms in aim to increase algorithm efficiency and improve execution time.

References

- [1] Face Recognition using Gabor Filters, Muhammad SHARIF, Adeel KHALID, Mudassar RAZA, Sajjad MOHSIN, Department of Computer Sciences, COMSATS Institute of Information Technology, Wah Cantt-Pakistan J.
- [2] Face Detection using Neural Network & Gabor Wavelet Transform, Avinash Kaushal¹, J P S Raina², IG CET, Greater Noida, U.P., India; 2BBSBEC, Fatehgarh Sahib, Punjab, India.
- [3] Heseltine, T., Pears, N., Austin, J.: Evaluation of image pre-processing techniques for eigenface based face recognition. In Proc. of the Second International Conference on Image and Graphics, SPIE vol. 4875, (2002) 677-685.
- [4] Oppliger, R., & Schwenk J., & Helbach, J. (2008). Protecting Code Voting Against Vote Selling. In A. Alkassar & J. H. Siekmann (Eds.), Sicherheit 2008; 128, 193–204.
- [5] Volkamer, M., & Alkassar, A., & Sadeghi, A., & Schultz, S. (2006). Enabling the Application of Open Systems like PCs for Online Voting. Proceedings of the Frontiers in Electronic Elections – FEE '06. Retrieved January 14, 2011, from http://fee.iavoss.org/2006/papers/fee-2006-avossEnabling_the_application_of_open_systems_like-PCs_for_Online_Voting.
- [6] <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>.
- [7] <http://www.clear.rice.edu/elec301/Projects99/faces/index.html>.
- [8] Novel Face Detection Method Based on Gabor Features, ie Chen¹, Shiguang Shan² 1 2, Peng Yang², Shengye Yan², Xilin Chen School of computer Science and Technology, Harbin Institute of Technology, 50001, China ICT-ISVISION JDL for AFR, Institute of Computing echnology, CAS, Beijing, 100080, China 1 and, Wen Gao^{1,2}
- [9] Highly Secure Online Voting System with Multi Security using Biometric and Steganography J. Cross Datson Dinesh Assoc. Prof. Dept of Computer Science and Engineering Rajalakshmi engineering College #2 Chennai, India
- [10] <http://aicat.inf.ed.ac.uk/category.php?id=12>
- [11] Verification And Validation Issue In Electronic Voting. Orhan cetinkaya¹, and deniz cetinkaya². ¹institute of applied mathematics, METU, Ankara, turkey. ² computer engineering , METU , ankara , turk

Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Social Networks

Roobaea AlRoobaea

Faculty of Computer Science and
Information Technology Taif
University, Saudi Arabia and School
of Computing Sciences UEA, Norwich, UK

Ali H. Al-Badi

Department of Information Systems,
Sultan Qaboos University, Oman

Pam J. Mayhew

School of Computing Sciences,
UEA, Norwich, UK

Abstract—The electronic information revolution and the use of computers as an essential part of everyday life are now more widespread than ever before, as the Internet is exploited for the speedy transfer of data and business. Social networking sites (SNSs), such as LinkedIn, Ecademy and Google+ are growing in use worldwide, and they present popular business channels on the Internet. However, they need to be continuously evaluated and monitored to measure their levels of efficiency, effectiveness and user satisfaction, ultimately to improve quality. Nearly all previous studies have used Heuristic Evaluation (HE) and User Testing (UT) methodologies, which have become the accepted methods for the usability evaluation of User Interface Design (UID); however, the former is general, and unlikely to encompass all usability attributes for all website domains. The latter is expensive, time consuming and misses consistency problems. To address this need, a new evaluation method is developed using traditional evaluations (HE and UT) in novel ways.

The lack of an adaptive methodological framework that can be used to generate a domain-specific evaluation method, which can then be used to improve the usability assessment process for a product in any chosen domain, represents a missing area in usability testing. This paper proposes an adaptive framework that is readily capable of adaptation to any domain, and then evaluates it by generating an evaluation method for assessing and improving the usability of products in a particular domain. The evaluation method is called Domain Specific Inspection (DSI), and it is empirically, analytically and statistically tested by applying it on three websites in the social networks domain. Our experiments show that the adaptive framework is able to build a formative and summative evaluation method that provides optimal results with regard to our newly identified set of comprehensive usability problem areas as well as relevant usability evaluation method (UEM) metrics, with minimum input in terms of the cost and time usually spent on employing traditional usability evaluation methods (UEMs).

Keywords—Heuristic Evaluation (HE); User Testing (UT); Domain Specific Inspection (DSI); adaptive framework; social networks domain.

I. INTRODUCTION

Online social gatherings are known by a variety of names, such as online community and social network websites (SNSs). SNSs are increasingly recognized as one of the most popular mediums of online communication, and they are increasingly attracting the attention of academic and industry researchers intrigued by their affordability and reach. They

have attracted millions of users around the world; many of them have integrated these sites to be part of their daily activities. Nowadays, SNSs play a vital role in many fields, such as e-government and business. They have gained in popularity not only because of their many interactive and innovative features, but also because their purpose is clearly established and the audience is targeted effectively [Pessagno, 2010]. Software development organizations are paying increased levels of attention to the usability of such social networking websites (SNSs); however, the majority of SNSs still have low levels of usability [Fu et al., 2008]. The motivation for this research is that SNSs are increasingly interesting as a topic of research in Information Systems, and so assessing and hence improving the usability of these SNSs is becoming crucial [Fox and Naidu, 2009]. Also, SNSs are becoming increasingly popular, and so it is important that the usability of individual sites be tested, assessed and improved in an objective manner as possible.

It is clear that Heuristic Evaluation (HE) and User Testing (UT) are the most important usability evaluation methods for ensuring system quality and usability [Chattrachart and Lindgaard, 2008; Chattrachart and Brodie, 2004]. Currently, the growth of a new breed of dynamic websites, complex computer systems, mobile devices and their applications have made usability evaluation methods even more important; however, usability differs from one website to another depending on website characteristics. It is clear that users have become the most important factor impacting on the success of a website; if a website is produced and is then deemed not useful by the end-users, it is a failed product (nobody can use it and the company cannot make money) [Nielsen, 2001]. Nayebe et al. (2012) asserted, “Companies are endeavoring to understand both user and product, by investigating the interactions between them”.

The traditional usability measures of effectiveness, efficiency and satisfaction are not adequate for the new contexts of use [Zaharias and Poylymenakou, 2009]. HE has been claimed to be too general and too vague for evaluating new products and domains with different goals; HE can produce a large number of false positives, and it is unlikely to encompass all the usability attributes of user experience and design in modern interactive systems [Hertzum and Jacobsen, 2001; Chattrachart and Lindgaard, 2008]. UT has been claimed to be costly, time consuming, prone to missing consistency problems and subject to environmental factors

[Oztekin et al., 2010]. Several studies have also emphasised the importance of developing UEMs as a matter of priority, in order to increase their effectiveness. To address these challenges, many frameworks and models have been published to update usability evaluation methods (UEMs) [Alias et al., 2013; Gutwin and Greenberg, 2000]; however, these frameworks and models are not applicable to all domains because they were developed to deal with certain aspects of usability in certain areas [Coursaris and Kim, 2011].

The adaptive methodological framework in this paper was originally constructed and then evaluated practically in the educational domain; in this, it delivered interesting results by discovering more real usability problems in specific usability areas than HE or UT [Roobaea et al., 2013a]; [Roobaea et al., 2013b]. The main objective of this paper is to address the challenges that were raised and to continue testing the validity of the adaptive framework by applying it on the social networks domain, through three case studies. Furthermore, it is to conduct a comprehensive comparison between UT, HE and our domain-specific Inspection (DSI) method, which is developed through the adaptive framework, in terms of number of real usability problems found and their severity in each of a number of usability problem areas, as well as in terms of certain UEM metrics and other measurements. The paper is organized in the following way. Section 2 starts with a brief literature review relating to this study; it includes a definition of usability problems, and describes the concept of severity rating. Section 3 details the construction of the adaptive framework. Section 4 is on the research methodology. Section 5 details the set of measurements and analysis metrics. Section 6 validates the adaptive framework by applying the new method (DSI), HE and UT in practice on three cases, and then provides an analysis and discussion of the results. Section 7 presents a discussion of the findings. Section 8 presents the conclusion and future work.

A. Research Hypotheses

This research hypothesizes that:

1) *There are significant differences between the results of HE and DSI, where the latter method outperforms the former in terms of achieving higher ratings from evaluators on the issues relating to the number of usability problems, the usability problem areas, the UEM performance metrics, and the evaluators' confidence, concluding that it is not essential to conduct HE in conjunction with DSI.*

2) *There are significant differences between results of UT and DSI, where the latter method outperforms the former in terms of achieving higher ratings on the issues relating to the number of usability problems, the usability problems areas, the UEM performance metrics, concluding that it is not essential to conduct UT in conjunction with DSI.*

II. LITERATURE REVIEW

SNSs have quickly become one of the most popular means of online communication in the last few years, and their users are dramatically growing in number by the day. SNSs can be defined as 'web-based services' that allow individuals to

construct a public or private profile within a bounded system, and to explore connections with others within the system. They can be used to seek out new friendships or to group together to chat with friends, share activities or interests and extend one's own social network [Ellison et al., 2007; Fox and Naidu, 2009]. Most of the existing social networks on the Internet offer a range of services to users, such as instant messaging, private messages and e-mail, video and file sharing, blogging and playing online games, but they are also used by businesses, advertisers and employers, and those who wish to follow companies in order to receive information, updates and RSS feeds [Ellison et al., 2007; Estes et al., 2009]. It is now apparent that SNSs have had a significant impact on how individuals and social groups communicate and exchange information. These networks involve a great many users at any one time, and they are divided into broad categories according to purpose; there are networks for making new friends, for study and for work, in addition to networks based on interest or activity. The most well-known SNSs are Facebook, MySpace and Twitter, but others are emerging as this is an evolving field. Consequently, this is a productive environment for informational conflict between these websites, who seek to increase their number of users by attracting new users, attracting users from competitors' websites and maintaining current users [Tufekci, 2008]. Essentially, the success of SNSs depends to a large extent on the degree of users' contributions and activities, and so they need to be highly usable; if websites are not usable, users will leave and find others that better cater to their needs.

Emanating from the development of Web 2.0, there is now a real need to study the usability of SNSs [Ali et al., 2013]. The reviewed literature shows that the techniques for measuring the quality of user experience have been classified under the headings of ergonomics and ease-of-use, but more lately under the heading of usability [Oztekin et al., 2010]. This aims to ensure that the user-interface is of sufficiently high quality. 'Usability' is one of the most significant aspects affecting the quality of a product and its user experience. A website is a product, and the quality of a product takes a significant amount of time and effort to develop. A high-quality website is one that provides all the main functions in a clear format, and that offers good accessibility and a simple layout to avoid users spending an inordinate length of time learning how to use it; these are the fundamentals of the usability of a website. However, poor website usability may have a negative impact on various aspects of the organization, and may not allow users to achieve their goals efficiently, effectively and with a sufficient degree of satisfaction [ISO, 1998]. Nielsen (1994b) stated, "usability is associated with learnability, efficiency, memorability, errors and satisfaction" [Nielsen, 1994b]. Usability is not a single 'one-dimensional' property of a user interface; there are many usability attributes that should be taken into account and measured. Shackel and Richardson (1991) proposed attributes covering four dimensions that influence the acceptance of a product, which are effectiveness, learnability, flexibility and attitude [Shackel and Richardson, 1991]. Nielsen (1994b) introduced five major attributes of usability based on a System Acceptability model [Nielsen, 1994b], and they are as follows; 1) Easy to learn: a system should be easy to learn for the first time; 2) Efficient to

use: the relationship between accuracy and time spent to perform a task; 3) Easy to remember: a user should be able to use the system after a period without spending time learning it again; 4) Few errors: the system should prevent users from making errors (this also addresses how easy it is to recover from errors); and 5) Subjectively pleasing: this addresses the user's feeling towards the system.

Usability evaluation methods (UEMs) are a set of techniques that are used to measure usability attributes. They can be divided into three categories: inspection, testing and inquiry. Heuristic Evaluation (HE) is one category of the inspection methods. It was developed by

[Molich and Nielsen, 1990], and is guided by a set of general usability principles or 'heuristics' as shown Table 1. It can be defined as a process that requires a specific number of experts to use the heuristics in order to find usability problems in an interface in a short time and with little effort [Magoulas et al., 2003]. It can be used early in the development process, and may be used throughout the development process [Nielsen and Molich, 1990]. However, it is a subjective assessment and depends on the evaluator's experience, and can produce a large number of false positives that are not usability problems at all or can miss some real problems [Holzinger, 2005; Nielsen and Loranger, 2006; Chatratchart and Lindgaard, 2008; Hertzum and Jacobsen, 2001].

TABLE I. HEURISTIC EVALUATION

Heuristic Evaluation
Visibility of system status
Match between system and the real world
User control and freedom
Consistency and standards
Error prevention
Recognition rather than recall
Flexibility and efficiency of use
Aesthetic and minimalist design
Helps users recognize, diagnose and recover from errors
Help and documentation

There are two kinds of expert evaluators in HE. One is a 'single' evaluator, who can be defined as a person with general usability experience. The second is a 'double' evaluator who can be defined as a person with a usability background in a specific application area. Molich and Nielsen (1990) recommended from previous work on heuristic evaluation that between three and five single expert evaluators are necessary to find a reasonably high proportion of the usability problems (between 74% and 87%). For the double expert evaluators, it is sufficient to use between two and three evaluators to find most problems (between 81% and 90%). There is no specific procedure for performing HE. However, Nielsen [Nielsen, 1994a] suggested a model procedure with four steps. Firstly, conducting a pre-evaluation coordination session (a.k.a training session) is very important. Before the expert evaluators evaluate the targeted website, they should take few minutes browsing the site to familiarize themselves with it. Also, they should take note of the actual time taken for

familiarisation. If the domain is not familiar to the evaluators, the training session provides a good opportunity to present the domain. Also, it is recommended that in the training session, the evaluators evaluate a website using the heuristics in order to make sure that the principles are appropriate [Chen and Macredie, 2005]. Secondly, conducting the actual evaluation, in which each evaluator is expected to take around 1 to 1.5 hours listing all the usability problems. However, the actual time taken for evaluation should always be noted. Next, there should be a debriefing session, which would be conducted primarily in a brainstorming mode and would focus on discussion of possible redesigns to address the major usability problems and general problematic aspects of the design. A debriefing is also a good opportunity for discussing the positive aspects of the design, as HE does not otherwise address this important issue. Finally, the results of the evaluations are collected into a series of evaluation tables, and then combined into a single table after removing any redundant data. After the problems are combined, the evaluators should agree on the severity of each individual problem [Nielsen, 1994a].

In the present context and in relation to HE, usability testing (UT; also known as user testing), is another important evaluation method for ensuring system quality, in particular for websites. It needs participants to perform a set of tasks, usually in a laboratory. These tasks are performed without information or clues as to how to complete them, and with no help provided to the user during the test session. Also, the completion of these tasks is monitored and assessed by an observer who records the usability problems encountered by the users. All the observed data, such as error numbers, time spent, success rate and user satisfaction, need to be recorded for analysis [Nielsen, 1994b]. Dumas and Redish (1991) stressed that a fruitful usability testing session needs careful planning and attention to detail. Accordingly, there is a general procedure for conducting user testing, thus: 1) Planning a usability test; 2) Selecting a representative sample and recruiting participants; 3) Preparing the test materials and actual test environment; 4) Conducting the usability test; 5) Debriefing the participants; 6) Analysing the data of the usability test; and 7) Reporting the results and making recommendations to improve the design and effectiveness of the system or product. The Think-Aloud technique (TA) is used with UT. There are three TA types, which are concurrent, retrospective and constructive interaction. The concurrent TA type is the most common; this involves participants verbalising their thoughts whilst performing tasks in order to evaluate an artefact. Retrospective TA is less frequently used; in this method, participants perform their tasks silently, and afterwards comment on their work on the basis of a recording of their performance. Constructive interaction is more commonly known as Co-Discovery Learning, where two participants work together in performing their tasks, verbalising their thoughts through interacting [Van den Haak et al., 2004].

One important factor in usability testing is setting the tasks. Many researchers are aware that task design is an important factor in the design of adequate usability tests. The tasks designed for web usability testing should be focused on

the main functions of the system. The tasks should cover the following aspects: 1) Product page; 2) Category page; 3) Display of records; 4) Searching features; 5) Interactivity and participation features; and 6) Sorting and refining features [Wilson, 2007]. Dumas and Redish (1999) suggested that the tasks could be selected from four different perspectives. These are: 1) Tasks that are expected to detect usability problems; 2) Tasks that are based on the developer's experience; 3) Tasks that are designed for specific criteria; and 4) Tasks that are normally performed on the system. They also recommended that the tasks be short and clear, in the users' language, and based on the system's goals [Dumas & Redish, 1999]. Alshamari and Mayhew (2008) found that task design can play a vital role in usability testing results, where it was shown that changing the design of the task can cause differences in the results [Alshamari & Mayhew, 2008].

The result of applying HE and UT is a list of usability problems [Nielsen, 1994a]. These problems are classified into different groups to which a numeric scale is used to measure the severity of each problem. Firstly, this issue is not a usability problem at all. Secondly, this is a cosmetic problem that does not need to be fixed unless extra time is available on the project. Next, this issue is a minor usability problem; fixing this should be given low priority. Then, this is a major usability problem; it is important to fix this, so it should be given high priority. Finally, this issue is a usability catastrophe; it is imperative to fix this before the product can be released.

In the early years of computing, HE was widely applied in measuring the usability of Web interfaces and systems because it was the only such tool available. These heuristics have been revised for universal and commercial websites as HOMERUN heuristics [Nielsen, 2000]. Furthermore, [Chattratchart and Brodie, 2002; Chattratchart and Lindgaard, 2008] proposed UEMs called HE-Plus and HE++, which are extensions to HE by adding what is called a "usability problem profile". However, some researchers have found that their tested websites failed in certain respects according to these extended or modified heuristics [Thompson and Kemp, 2009; Alrobai et al., 2012]. On the other hand, many researchers then sought to compare and contrast the efficiency of HE with other methods such as UT. They found that HE discovered approximately three times more problems than UT. However, they reported that more severe problems were discovered through UT, compared with HE [Liljegen and

Osvalder, 2004; Doubleday et al., 1997; Jeffries et al., 1991]. Lately, researchers' findings have been almost unanimous in one respect: HE is not readily applicable to many new domains with different goals and are too vague for evaluating new products such as web products because they were designed originally to evaluate screen-based products; they were also developed several years before the web was involved in user interface design [Ling and Salvendy, 2005; Hasan, 2009; Hart et al., 2008]. Nevertheless, each method seems to overcome the other method's limitations, and researchers now recommend conducting UT together with HE because each one is complementary to the other, and then combining the two methods to offer a better picture of a

targeted website's level of usability [Nielsen, 1992; Law and Hvannberg, 2002].

It can be seen from the above that there is need to an effective and appropriate methodology for evaluating the emerging domains/technology to measure their levels of efficiency, effectiveness and satisfaction, and ultimately to improve their quality. Also, there is need for a method that considers context of use and that includes expert and user perspectives. This finding and the criticality of website usability has encouraged researchers to formulate such a framework. This framework should be applicable across numerous domains. In other words, it should be readily capable of adapting in any domain and for any technology. This paper constructs such a framework, i.e. for generating a context-specific method for evaluating the chosen domain that can be applied without needing to conduct user testing. However, developing and testing a method is not quick and it should involve some key stages. The next section describes the stages employed in the adaptive framework; also, it describes the process used to test it.

III. CONSTRUCTION OF THE ADAPTIVE FRAMEWORK

The adaptive framework is developed according to established methodology in HCI research. It consists of two distinct phases: 1) Development phase that consists of four main stages for gathering together suitable ingredients to develop a context-specific Inspection(DSI) method for website evaluation; and 2) Validation phase for testing the developed DSI method practically (these are outlined in Figures 1 and 2). Below is an explanation of the four stages in the development phase:

Development Stage One (D1: Familiarization): This stage starts from the desire to develop a method that is context-specific, productive, useful, usable, reliable and valid, and that can be used to evaluate an interface design in the chosen domain. It entails reviewing all the published material in the area of UEMs but with a specific focus on knowledge of the chosen domain. Also, it seeks to identify an approach that would support developers and designers in thinking about their design from the intended end-users' perspective.

Development Stage Two (D2: User Input): This stage consists of mini-user testing (task scenarios, TA and a questionnaire). Users are asked to perform a set of tasks on a typical domain website, to 'think aloud' whilst so doing and then to fill out a questionnaire. The broad aim of *this* is to elicit feedback on a typical system from real users in order to appreciate the user perspective, to identify requirements and expectations and to learn from their errors. Understanding user needs has long been a key part of user design, and so this stage in the framework directly benefits from including the advantages of user testing.

Development Stage Three (D3: Expert Input): This stage aims to consider what resources are available for addressing the need. These resources, such as issues arising from the mini-user testing results and the literature review, require a discussion amongst experts (in the domain and/or usability) in order to obtain a broader understanding of the specifics of the prospective domain. Also, it entails garnering more

information through conversations with expert evaluators to identify the areas/classification schemes of the usability problems related to the selected domain from the overall results. These areas provide designers and developers with insight into how interfaces can be designed to be effective, efficient and satisfying; they also support more uniform problem description and they can guide expert evaluators in finding real usability problems, thereby facilitating the

evaluation process by judging each area and page in the target system.

Development Stage Four (D4: Draw Up DSI Method: data analysis): The aim of this stage is to analyse all the data gathered from the previous three. Then, the DSI method will be established (as guidelines or principles) in order to address each area of the selected domain.

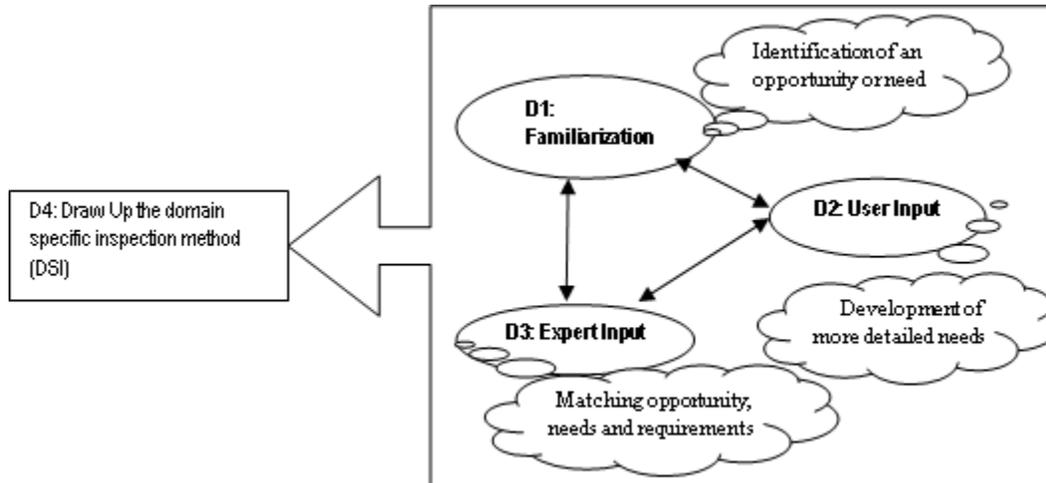


Fig.1. Development stages of the adaptive framework

After constructing the DSI framework, the researchers test it intensively through rigorous validation methods to verify the extent to which it achieves the identified goals, needs and requirements that the method was originally developed to address (this validation is outlined in Figure 2).

pilot experiment to make sure that everything is in place and ready for the actual evaluation.

2) *Heuristic Validation stage (Expert Evaluation (HE)):* The aim of this stage is the validation of the newly developed method by conducting a heuristic evaluation (HE). Expert evaluators need a familiarization session before the actual evaluation. The expert evaluation is then conducted using the newly developed DSI method alongside HE. The aim of this process is to collect data ready for analysis (analytically and statistically), as explained in stage 4.

3) *Testing Validation stage (User Evaluation (UT)):* The aim of this stage is to complement the results obtained from the expert evaluation, by carrying out usability lab testing on the same websites. [Nielsen, 1992] recommends conducting usability testing (UT) with HE because each one is complementary to the other. Then, the performance of HE is compared with the lab testing to identify which problems have been identified by UT and not identified by HE and/or DSI, and vice versa. The aim of this process is to collect data ready for analysis (empirically and statistically) in stage 4.

4) *Data Analysis stage:* This stage aims to analyse all the results and to answer all the questions raised from the above steps in a statistical manner. It is conducted in two steps; one focused on HE and the other on UT. The researchers extract the problems discovered by the experts from the checklists of both DSI and HE. Then, they conduct a debriefing session with the same expert evaluators to agree on the discovered problems and their severity, and to remove any duplicate problems, false positives or subjective problems. Then, the

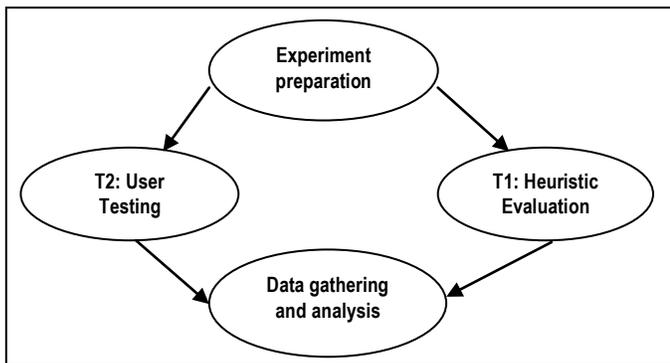


Fig.2. Testing stages of the adaptive framework

The validation phase of the adaptive framework again consists of four separate stages, as explained below:

1) *Experiment Preparation stage (for DSI, HE and UT):* Before the actual evaluation formally starts, the following initial preparative steps are needed: 1) Select a number of systems/websites that are typical of the chosen domain; 2) Recruit expert evaluators and users; 3) Plan the sequence of conducting the evaluations by each group in such a way as to avoid any bias; and 4) Prepare the experiment documents. This initial experiment preparation stage is concluded with a

problems approved upon are merged into a master problem list, and any problems upon which the evaluators disagree are removed. Ultimately, the researchers conduct a comparison on the results of both methods (DSI and HE) in terms of the number of problems discovered (unique and overlapping), their severity ratings, which problems are discovered by HE and not discovered by DSI and vice versa, the areas of the discovered problems, the UEM performance metrics, evaluator reliability and performance, and the relative costs entailed in employing the two methods.

In the second step, the researchers conduct a debriefing session with independent evaluators to rank the severity of the problems derived from the user testing and to remove any duplicate problems. Following this, they establish the list of usability problems for UT. Subsequently, a single unique master list of usability problems is consolidated from the three methods. A comparison of the results of the three methods is then conducted in terms of the number of problems discovered (unique and overlapping), their severity ratings, and the areas of the discovered problems; this is to identify which problems were discovered by HE and DSI and not discovered by UT, and vice versa. Also, the UEM performance metrics of each method are measured, together with other measures, such as their relative costs and reliability. Moreover, this final step seeks to prove or refute the efficacy of conducting UT and HE with DSI.

Having proposed the framework above, it was decided to evaluate its practicality by applying it to a real-life experiment. From the literature review, it was found that SNSs have recently been the subject of much study by researchers interested in areas such as privacy, identity, community dynamics and the behaviour of adolescents [Ellison et al., 2007]. However, it has not yet been fully explored, nor have any context-specific evaluation methods been generated for this domain (to overcome the shortcomings of HE and UT); this is an important area of research because these websites are now essential to many users and companies. A well-designed SNS (i.e. one that is aesthetically attractive and is easy to use) can positively affect the number of people who become members. If these

are considered, an SNS will gain members more quickly because it will allow users to carry out social tasks more easily [Pessagno, 2010].

IV. RESEARCH METHODOLOGY

The experimental approach was selected to address the research hypotheses outlined above. Essentially, this section describes the methodology employed in this comparative study. Before conducting this experiment, a set of procedures were followed by the researchers, as follows:

A. Design

This experiment employs the between-subject and within-subject designs. The independent variables are the three methods (HE, DSI and UT). The dependent variables are the UEM performance metrics, which are calculated from the

usability problems reported by the evaluators/users, and from the reliability and efficiency measurements.

B. Development; Evaluation of the Practicality of the Framework

In the first of the four stages within the development phase, the researchers conducted a literature review on the materials relating to usability and UEMs as well as on the requirements of the social networks domain. In stage two, a mini-user testing session was conducted through a brief questionnaire that entailed four tasks, which were sent to ten users who are regular SNS users. In stage three, a focus group discussion session was conducted with six experts in usability and the social networks domain (i.e. single and double experts). Cohen's kappa coefficient was used on the same group twice to enable a calculation of the reliability quotient for identifying usability problem areas. In stage four, the researchers analysed the results of the three stages and incorporated the findings. The intra-observer test-retest using Cohen's kappa yielded a reliability value of 0.9, representing satisfactory agreement between the two rounds. After that, the usability problems areas were identified to facilitate the process of evaluation and analysis, and to help designers and programmers to identify the areas in their website that need improvement. Then, the DSI method was established, closely focused on social networks as well as business networking websites, taking into an account what is called 'user experience'. The method was created and classified according to the usability problem areas detailed in Table 2 below.

C. Validation Stage 1; Preparation

a) Selection of the targeted websites

The first step in an initial preparation stage (of the validation phase) was selecting the websites. The researchers sought to ensure that the selected websites would support the research goals and objectives. The selection process was criteria-based; five aspects were determined and verified for each website, and these are: 1) Good interface design, 2) Rich functionality, 3) Good representatives of SNSs, 4) Not familiar to the users, and 5) No change will occur before and during the actual evaluation. In order to achieve a high level of quality in this research, the researchers chose three well-known websites in this domain, which are LinkedIn, Google+ and Ecademy. All of these have all the aspects mentioned above.

a) Experts and Users Recruitment

The selection of usability experts and users was the second important step in the initial preparation phase in this experiment. The researchers decided to recruit six expert evaluators, divided into two groups of three, who were carefully balanced in terms of experience. In each group, there are two double expert evaluators (usability specialists in SNSs) and one single expert evaluator (usability specialists in general). Selecting and recruiting users must be done carefully; the participants must reflect the real users of the targeted website because inappropriate users will lead to incorrect results, thereby invalidating the test.

TABLE II. FINAL VERSION OF DOMAIN SPECIFIC INSPECTION (DSI)

Usability problem areas/attributes	Domain Specific Inspection (DSI)
Layout and formatting (LF)	Design consistency
	Simple user interface
Content quality (CQ)	Correct, relevant, reliable, error-free & up to date
	Site upload time & less memory utilization
	Representation with familiar terminology & understandable content
	Appropriate & approachable content
Security and privacy (SP)	Awareness of security mechanisms/settings & protection
	Transparency of transactions
Business support (BS)	Advertise or sales pitches mechanism
	Trust & credibility of information sources and company advertising
	Easy to follow & share
	Forum/blog facilities and connectivity with different groups/businesses
	Syndication of Web content (such as RSS tools)
	Frequent posting & updating
User usability, sociability and management activities (USM)	Manageable personal profile & user-driven content
	Easy functionality, participation & user privileges, such as revoke & join friends/connection
	Supports user's skills & freedom, such as customize/modify user's content/messaging and notification.
	Offers informative feedback - action & reaction
	Appropriate multimedia with complete user control
	Help & support
Accessibility and compatibility (AC)	Accessibility and compatibility of hardware devices
	Accessible path-contact details & site map
	Easy access through universal design
Navigation system and search quality (NS)	Correct & reliable navigation/directions
	Easy identification of links and menus
	Search support & functionality

Appropriate users will deliver results that are more reliable; they will also be intrinsically motivated to conduct the experiment [Dumas & Redish, 1999]. There is no agreement on how many users should be involved in usability testing. Dumas and Redish (1999) suggested that 6 to 12 users are sufficient for testing, whereas other studies have recommended that 7, 15 and 20 users are the optimal numbers for evaluating small or large websites; particularly 20 users if benchmarking is needed [Nielsen & Loranger, 2006]. At this point, 30 users were engaged; they were chosen carefully to reflect the real users of the targeted websites and were divided into three groups for each website, i.e. a total of 10 users for each website. The majority of the users are students and employees, and they were mixed across the three users groups in terms of gender, age, education level and computer skills.

b) Task Sequencing

The third step was planning the sequence of the groups' evaluations. Each group employed two methods, namely DSI and HE, to evaluate the three different websites. The evaluations were carried out in a prescribed sequence, i.e. one group used DSI on Google+ and then HE on LinkedIn, and finally DSI on Ecademy, while the second group used HE on Google+ and Ecademy and then DSI on LinkedIn as shown in Figure 3 below. The researchers adopted this technique to avoid any bias in the results and also to avoid the risk of any

expert reproducing his/her results in the second session through over-familiarity with one set of heuristics, i.e. each evaluation was conducted with a fresh frame of mind.

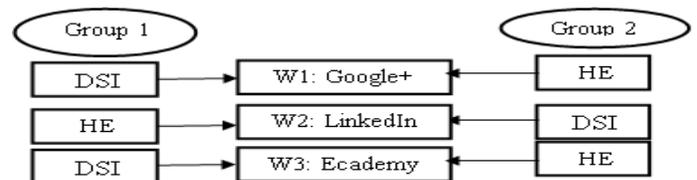


Fig.3. Usage of method sets by Group

c) Building the Experiment Documents

In the preparation phase, the fourth step entailed building a set of task-preparation documents for HE, DSI and UT, such as introduction sheets, HE and DSI checklists, tasks sheets, problem-ranking sheets, observer sheets, demographics, satisfaction and Likert questionnaires, sheets for collated problems and a master problem list. The introduction sheet contains the goals and objectives of the evaluation and the roles of users and experts. Before starting actual evaluation, users and experts completed a demographics questionnaire to obtain more information about them. Expert evaluators used the checklists that had been developed by the researchers to

facilitate the evaluation process for DSI and HE. Users used the task sheets that were designed according to the main functions that users would normally expect to perform on the three websites. A combination of task designs and TA approaches (as mentioned in the literature review) were used in this experiment. There were four sub-tasks in each task for all three task groups, they were kept to a reasonable time limit and they were interesting and engaging enough to hold the users' attention. As briefly mentioned above, usability testing requires an observer, and the researcher adopted this role in all the sessions, noting all the comments made by the users. The researcher used a stopwatch to record the time spent by each user on each task, and an observation sheet to write down the behaviour of each user and the number of problems encountered. The ranking sheet aims to help the expert evaluators and independent evaluators (for user testing) to rank the severity of usability problems by using Nielsen's scale as mentioned above. After the evaluators had finished their evaluations and had ranked HE and DSI problems, they were asked to complete a satisfaction questionnaire using the System Usability Scale (SUS) to complete the five-point scale (1 for strongly disagree and 5 for strongly agree) to rate their satisfaction on the evaluation method they had used (DSI or HE). It is made up of ten items in the form of scale questions ranging from 0 to 100 to measure the satisfaction of expert evaluators [Brooke, 1996]. Also, when the users finished their tasks, they were asked to rate their level of satisfaction in a questionnaire on a scale of one to seven, where one refers to 'highly unsatisfactory' and seven indicates 'highly satisfactory'. This scale has been suggested to truthfully measure the levels of satisfaction that are felt by users on a website interface following a test [Nielsen and Loranger, 2006]. Evaluators and users were asked to fill in an open-ended questionnaire by writing down their comments and feedback on the methods used and explaining any reaction that was observed during the test. Subsequently, the Likert scale was used by the evaluators for measuring either positive or negative responses to a statement in both the DSI and HE methods. Moreover, the researchers extracted the problems of three methods from the problems sheet and removed all false positive ('not real') problems, evaluators' 'subjective' problems and duplicated problems during the debriefing session. The problems agreed upon were merged into a unique master problem list and any problems upon which the evaluators disagreed were removed.

d) Piloting the Experiment

The final step in the initial preparation stage was a pilot experiment. It was conducted by two independent evaluators and fifteen users. All materials were checked to make sure that there were no spellings or grammatical errors and no ambiguous words or phrases. Furthermore, to assess the time needed for testing, the fifteen users were divided into three groups (five users in each). Each group performed its tasks. The users' behaviour was monitored, and all the usability measures were assessed as they would be in real testing. All of these steps resulted in useful corrections and adjustments for the real test. Also, the test environment was a quiet room. We attempted to identify the equipment that the users regularly use and set it up for them before the test, for example, using the same type of machine and browser.

D. Validation Stage 2: Heuristic Evaluation

The heuristic validation stage started with a training (familiarization) session for the eight expert evaluators. They were given a UEM training pack that contained exactly the same information for both groups. The researchers emphasized to each evaluator group that they should apply a lower threshold before reporting a problem in order to avoid misses in identifying real problems in the system. Then, the actual expert evaluation was conducted and the evaluators evaluated all websites consecutively, rating all the problems they found in a limited time (which was 90 minutes). After that, they were asked to submit their evaluation report, the SUS questionnaire and the Likert questionnaire and to give feedback on their own evaluation results.

E. Validation Stage 3: User Testing

The UT validation stage started with a training (familiarization) session for the 30 users; it involved a quick introduction on the task designs, the TA approach and the purpose of the study. The next step entailed explaining the environment and equipment, followed by a quick demonstration on how to 'think aloud' while performing the given tasks. Prior to the tests, the users were asked to read and sign the consent letter, and to fill out a demographic data form that included details such as level of computer skill. All the above steps took approximately ten minutes for each test session. The actual test started from this point, i.e. when the user was given the task scenario sheet and asked to read and then perform one task at a time. Once they had finished the session, they were asked to rate their satisfaction score relating to the tested website, to write down their comments and thoughts, and to explain any reaction that had been observed during the test, all in a feedback questionnaire. This was followed by a brief discussion session.

F. Validation Stage 4: Data Analysis and Measurements

To determine whether our adaptive framework has generated an evaluation method of sufficiently high quality, the results of the comparison process between the three methods needed a meta-analysis to be performed, as follows:

- 1) *Compare the average time spent by each group when using each method during the evaluation sessions.*
- 2) *Compare the results of the usability problems and their severity in order to assess the performance of each method in terms of identifying unique and overlapping problems and of identifying real usability problems in the usability problem areas.*
- 3) *Comparing the satisfaction scores and evaluators' attitude of HE and DSI by using System Usability Scale (SUS) and Likert Scale.*
- 4) *Reliability of HE and DSI: This can be measured from employing the 'evaluators' effect formula' (Any-Two-Agreement). It is used on each evaluator in order to measure their performance on an individual basis [Hertzum and Jacobsen, 2001].*
- 5) *Any-Two-Agreement = Average of $|P_i \cap P_j| / |P_i \cup P_j|$ over all / $n(n-1)$ pairs of evaluators, where P_i is the set of*

problem discovered by evaluator i and the other evaluator j , and n refers to the number of evaluators.

6) Evaluators' Performance: This can be measured by the performance of single and double expert evaluators in discovering usability problems by using HE and DSI in each group and website.

To make further comparisons between the performance of HE, DSI and UT in identifying usability problems, a set of UEM and other metrics were used for examining their performance; none of these metrics on their own addresses errors arising from false positive, subjective and missed problems. They are efficiency, thoroughness, validity, effectiveness, reliability and cost. Efficiency in UEMs is the "ratio between the numbers of usability problems detected to the total time spent on the inspection process" [Fernandez et al., 2011]. Thoroughness is perhaps the most attractive measure; it is defined as a measure indicating the proportion of real problems found when using a UEM to the total number of known real problems [Liljegren, 2006]. Validity is the extent to which a UEM accurately identifies usability problems [Sears, 1997]. Effectiveness is defined as the ability of a UEM to identify usability problems related to the user interface [Khajouei et al., 2011]. The reliability of user testing can be measured by the mean number of evaluators to the number of real problems identified [Chattratichart and Lindgaard, 2008]. The cost can be calculated by identifying the cost estimates. It can be done fairly simply by following Nielsen's equation who estimated the hourly loaded cost for professional staff at \$100 [Nielsen, 1994]. All of them are computed as follows:

$$1) \text{ Efficiency} = (\text{No. of problems}) / (\text{Average time spent})$$

$$2) \text{ Thoroughness} = (\text{No. of real usability problems found}) / (\text{Total no. of real usability problems present})$$

$$3) \text{ Validity} = (\text{No. of real usability problems found}) / (\text{No. of issues identified as a usability problem})$$

$$4) \text{ Effectiveness} = \text{Thoroughness} \times \text{Validity}$$

$$5) \text{ Reliability of UT} = (\text{Mean no. of evaluators}) / (\text{No. of real problems identified})$$

$$6) \text{ Cost} = (\text{No. of evaluation hours}) \times (\text{Estimate of the loaded hourly cost of participants})$$

To test the research hypotheses and choose the correct statistical test in SPSS, the normality of the data should be examined. The t-test, One-way ANOVA, Pearson's correlation, Mann-Whitney and Wilcoxon were chosen (at the 5% significance level) as our methods for statistical analysis, as the dependent variables in our data are independent of each other, improving the validity of using analysis of variance.

V. ANALYSIS AND DISCUSSIONS

This section describes the results obtained from using the three methods adopted in this experiment. It starts by detailing the results of the HE and DSI methods separately, including quantitative and qualitative analyses. This is followed by detailing the results of the UT method alone, including quantitative and qualitative analyses. Ultimately, all the results derived from the three methods were compared in terms of the numbers of problems and types, as well as the other usability

metrics as mentioned above.

A. Analysis for HE and DSI Results

1) Time spent: It is clear from Tables 3 and 4 below that the average time taken for doing the three experiments using HE was 56 minutes, whereas for DSI the average was 72 minutes. This difference in time spent between them is not significant ($F = 0.139$, $p = 0.714$) using the t-test. The group who used HE managed to evaluate the websites more quickly than the other group but discovered fewer usability problems, whereas, the group that used DSI spent slightly more time evaluating the websites, but discovered many more real usability problems. There was a statistically significant positive relationship between time spent and problems discovered through using Pearson's correlation test, where the 'Sig' value is 0.041 at the 0.05 level. This result reveals that the users who spent more time were able to discover more usability problems.

TABLE III. AVERAGE TIME TAKEN AND NUMBER OF PROBLEMS FOUND DURING THE EVALUATION BY GROUP 1

Website	Google+	LinkedIn	Ecademy
Evaluator 'G1'	Time	Time	Time
1	90	70	80
2	60	50	60
3	70	60	75
Method	DSI	HE	DSI
# of problems	55	13	33
Mean time taken	73	60	72

TABLE IV. AVERAGE TIME TAKEN AND NUMBER OF PROBLEMS FOUND DURING THE EVALUATION BY GROUP 2

Website	Google+	LinkedIn	Ecademy
Evaluator 'G1'	Time	Time	Time
1	60	80	60
2	50	70	50
3	40	65	60
Method	HE	DSI	HE
# of problems	22	47	12
Mean time taken	50	72	57

Explanations for the differences in time spent and number of problems located were gleaned from the evaluators' feedback. They said that HE was not particularly helpful, understandable or memorable for them. However, DSI helped them to develop their skills in discovering usability problems in this application area; also, this set was more understandable and memorable during their evaluations and covered most broad areas. To further analyse these factors of time spent and number of problems discovered, efficiency metrics were applied. DSI proved to be more efficient than HE in discovering usability problems (DSI = 0.6 vs. HE = 0.4) as Table 5 shows. The t-test reveals significant difference in terms of efficiency between HE and DSI ($t = -3.070$, $df =$

11.391, $p = 0.01$).

TABLE V. MEAN SCORE OF EFFICIENCY FOR THE TWO METHODS

Method	Google+	LinkedIn	Ecademy	Mean
	Efficiency	Efficiency	Efficiency	
HE	0.5	0.29	0.3	0.4
DSI	1.1	0.8	0.8	0.6

2) Number of usability problems: Table 5 shows that HE was able to uncover 26% of the total number of real usability problems. However, DSI was able to uncover 74% of the total number of real usability problems in the websites.

TABLE 5: SUMMARY OF USABILITY PROBLEMS (NUMBERS AND PERCENTAGES) UNCOVERED BY EACH WEBSITE, EACH GROUP, EACH EVALUATOR AND EACH METHOD

Website	Group	Expert and type	Method	# of problems found by each evaluator	Total # of problems with repetition	Total # of problems without repetition	Total # of problems in each site with repetition	% of problems found by each evaluator	% # of problems found by each group
Google+	G1	Ev. 1^	DSI	16	66	55	77	21%	71%
		Ev. 2+	DSI	33				43%	
		Ev. 3+	DSI	17				63%	
	G 2	Ev. 1+	HE	6	22	22		8%	29%
		Ev. 2^	HE	5				7%	
		Ev. 3+	HE	11				14%	
LinkedIn	G1	Ev. 1^	HE	2	16	13	60	3%	22%
		Ev. 2+	HE	8				13%	
		Ev. 3+	HE	6				10%	
	G 2	Ev. 1+	DSI	24	59	47		40%	78%
		Ev. 2^	DSI	8				13%	
		Ev. 3+	DSI	27				45%	
Ecademy	G 1	Ev. 1^	DSI	6	57	33	45	14%	73%
		Ev. 2+	DSI	28				67%	
		Ev. 3+	DSI	23				55%	
	G 2	Ev. 1+	HE	5	12	12		11%	27%
		Ev. 2^	HE	3				7%	
		Ev. 3+	HE	4				9%	
Total number of usability problems discovered by each method						Methods		Total number	%
						HE		47	26%
						DSI		135	74%

(+) Double Expert (^) Single Expert (Ev.) Evaluator

TABLE VI. RESULTS OF T-TEST BETWEEN GROUPS AND METHODS IN EACH WEBSITE

Website	Group	Method	t-value	df-value	p-value
Google+	Group 1	HE	-5.524	5.448	0.045
	Group 2	DSI			
LinkedIn	Group 1	HE	-5.429	5.455	0.040
	Group 2	DSI			
Ecademy	Group 1	HE	-5.922	5.973	0.001
	Group 2	DSI			

In terms of the performance of each method in discovering unique and overlapping problems, Table 5 illustrated that the total number of real problems discovered was 182 in all three

websites, out of which 47 were identified using HE and 135 using DSI. When the problems from the three evaluation groups were consolidated, there were 24 duplicates; we thus identified a total of 158 problems in all websites. The total for uniquely identified real problems in all websites was 128 problems.

The heuristic evaluation using DSI identified 96 real problems (61% of the 158 problems) that were not identified by HE, and there were 32 real problems (20% out of 158) identified by HE that were not identified by DSI. 30 real problems (19%) out of 158 were discovered by both methods (as depicted in Figure 4).

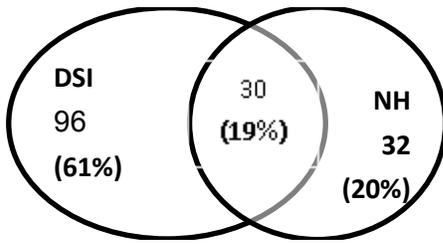


Fig.4. Overlap between both methods

In order to further compare the two methods, the severity ratings of the problems discovered (cosmetic, minor, major and catastrophic) were assessed, as Table 7 shows. Overall, a great many real usability problems were discovered across the rating scale. The most important results were obtained from using DSI, while HE found fewer (or no) usability problems. The Wilcoxon test revealed that there is a significant difference between the two methods in terms of problem severity (Cosmetic $z = -1.997$, $p = 0.04$; Minor $z = -2.207$, $p = 0.027$; Major $z = -2.003$, $p = 0.045$; Catastrophic $z = -2.100$, $p = 0.03$). This quantitative assessment between the two methods also entailed a comparison in terms of the usability problem areas in which the problems were found. These classifications/areas assisted in identifying how each method performed in each usability problem area or sub- heuristic. The six expert evaluators discussed, agreed and decided on the categories to which the problems should belong in both methods, as Tables 8 and 9 illustrate. The overall results from both tables show that the two groups (and the three websites) revealed more usability problems by using DSI than HE, particularly in User usability, sociability and management activities, business support (from assessing the DSI heuristics). The results show some level of agreement between the methods, regardless of the number of problems discovered in each usability area; both of them found the most problems

in the area entitled , Navigation system and layout and formatting (DSI) or User control and freedom (HE). These were followed in DSI by Content quality, Navigation system and sear quality, Layout and formatting and Security and privacy; these areas were not discovered efficiently or sufficiently by the equivalent areas in HE. This suggests that HE is rather general and unlikely to encompass all the usability attributes of user experience and design.

TABLE VII. TOTAL NUMBER OF USABILITY PROBLEMS BY SEVERITY RATING FOR BOTH METHODS

Website	Severity of Problems	Method			
		DSI		HE	
Google+	Cosmetic	Group 1	16	Group 2	6
	Minor		28		13
	Major		11		3
	Catastrophic		0		0
	Severity (average)		1.9		1.9
LinkedIn	Cosmetic	Group 2	11	Group 1	0
	Minor		19		8
	Major		11		5
	Catastrophic		6		0
	Severity (average)		2.3		2.3
Ecademy	Cosmetic	Group 1	16	Group 2	4
	Minor		11		8
	Major		6		0
	Catastrophic		0		0
	Severity (average)		1.7		1.7
Overall Severity (average)			2		2
No. of discovered problems			135		47

TABLE VIII. USABILITY PROBLEMS FOUND BY EACH HEURISTIC IN HE

Nielsen's Heuristics	Google+	LinkedIn	Ecademy
Visibility of system status	2	1	5
Match between the system and the real world	4	1	1
User control and freedom	5	3	1
Consistency and standards	3	2	0
Error prevention	0	1	0
Recognition rather than recall	2	3	1
Flexibility and efficiency of use	2	0	0
Aesthetic and minimalist design	1	0	0
Helps users recognize, diagnose and recover from errors	1	1	1
Help and documentation	2	1	3
Total problems	22	13	12

TABLE IX. USABILITY PROBLEMS FOUND BY CATEGORY THROUGH DSI

Usability problem area	Google+	LinkedIn	Ecademy
Layout and formatting	9	2	4
Content quality	12	11	5
Security and privacy	2	4	2
Business support	0	1	3
User usability, sociability and management activities	21	20	14
Accessibility and compatibility	1	1	3
Navigation system and search quality	10	8	2
Total problems	55	47	33

One striking result is that the number of problems identified by each evaluator who used HE was always fewer than the number of problems identified by any evaluator using DSI for the same website. An explanation of this was found in the evaluators' answers in the questionnaire. They said that the HE set was difficult to use and did not remind them of aspects they might have forgotten about, and they did not believe that this set encouraged them to be thorough in their evaluation. On the other hand, they said that the DSI set was easy to use; it did indeed help them to remember all the functions that needed to be tested, it is specific and was designed to cover all the aspects needed for social networking websites.

3) UEM Performance Metrics:

After employing the above formulae and as depicted in Table 10, the Mann-Whitney test was used to investigate the statistical differences between the two methods in terms of the UEMs and reliability. The thoroughness of DSI in identifying

the number of real problems was higher than for HE (0.3 vs. 0.1); also, Mann-Whitney revealed a significant difference between them ($z = -2.235, p = 0.025$). Further,

the validity of DSI was higher in accurately identifying real usability problems than HE (0.2 vs. 0.04); there was significant difference between them ($z = -2.600, p = 0.009$). The effectiveness of DSI was higher than that for HE (0.1 vs. 0.01); there was significant difference between them ($z = -2.230, p = 0.025$). Finally, the reliability values for DSI were slightly higher than for HE (0.76 vs. 0.64), and the results reveal that the difference between the two methods is significant ($z = -3.202, p = 0.001$). It can now be concluded that there is general agreement amongst the evaluators on the usability problems ($z = -3.202, p = 0.001$). Finally, the average results in terms of the cost of employing the two methods show that there is a slight difference in this research (Table 11); DSI = \$863.33 vs. HE = \$706.66.

TABLE X. MEAN SCORE OF UEM FOR TWO METHODS

Method	Google+		LinkedIn		Ecademy		Mean overall	
	HE	DSI	HE	DSI	HE	DSI	HE	DSI
Thoroughness	0.3	0.47	0.1	0.23	0	0.14	0.1	0.3
Validity	0.09	0.15	0.04	0.13	0	0.16	0.04	0.2
Effectiveness	0.03	0.07	0.004	0.03	0	0.03	0.01	0.1
Reliability	0.5	0.9	0.64	0.6	0.8	0.8	0.64	0.76

TABLE XI. COST OF EMPLOYING BOTH METHODS IN THIS RESEARCH

Mean cost	Ecademy	LinkedIn	Google+	Evaluation Method
\$706.66	\$710 Time spent by 3 evaluators (2.8 hours) + 1 hour collecting data from the evaluation sessions + 3.3 hours analysing data.	\$730 Time spent by 3 evaluators (3 hours) + 1 hour collecting data from the evaluation sessions + 3.3 hours analysing data.	\$680 Time spent by 3 evaluators (2.5 hours) + 1 hour collecting data from the evaluation sessions + 3.3 hours analysing data.	Heuristic evaluation (HE)
\$863.33	\$860 Time spent by 3 evaluators (3.5 hours) + 1.3 hours collecting data from the evaluation sessions + 3.8 hours analysing data.	\$860 Time spent by 3 evaluators (3.5 hours) + 1.3 hours collecting data from the evaluation sessions + 3.8 hours analysing data.	\$870 Time spent by 3 evaluators (3.6 hours) + 1.3 hours collecting data from the evaluation sessions + 3.8 hours analysing data.	Domain Specific Inspection (DSI)

4) *Post- test questionnaire*

- Satisfaction Score: The researchers used the System Usability Scale (SUS), and the results reveal that HE delivered a lower overall score, at 51, whereas DSI delivered slightly higher score, at 76. The evaluators gave this result because the process of the evaluation was smoother when using DSI and it was generated to cover all social network aspects.

- **Opinions and Attitudinal Questions (Likert scale)**
The Likert scores were collated for each statement in order

to obtain overall results concerning the opinions of the expert evaluators with respect to DSI and HE. A Likert score of 1-2 was regarded as a negative response, 4-5 a positive response, and 3 a neutral one. Cronbach's Alpha test was used to measure the reliability of responses and the result was 0.89. The Likert scores reveal that the evaluators were satisfied overall with DSI, and the results reveal significant differences between DSI and HE (using Mann-Whitney), as Table 12 shows.

TABLE XII. RESULTS OF MANN-WHITNEY FOR BOTH METHODS

Methods	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Mann-Whitney U	1.500	6.000	5.000	6.500	4.500	3.000	0.000

B. *Quantitative Analysis for Usability Testing Result*

1) *Time spent*

Table 13 shows the time spent by each user on performing the experiment. The Google+ groups spent the longest time, more than the LinkedIn and Ecademy groups, with 429, 377 and 372 minutes, respectively. This was probably due to problems in navigation, structure and function in the three websites, which caused the users to spend more time in accomplishing their tasks. This was particularly so in the Google+ website, as some tasks were abandoned because the users had doubts about how to accomplish them. Also, in the

LinkedIn website, the group spent time thinking about how to perform some tasks, such as the 'find group' task and the 'upload CV' task. The average time spent by each user in all three groups was more than 3.72 minutes. The efficiency formula used for UT for all the experiments, in terms of number of usability problems discovered over time spent, delivered a mean score 0.47 (Google+ = 0.64, LinkedIn = 0.46, Ecademy = 0.30). One-way ANOVA was used to determine whether there was a significant difference in terms of time spent, and the result reveals that indeed there was ($F = 15.033$, $p < 0.001$). Moreover, there was a statistical difference in terms of efficiency ($F = 24.694$, $p < 0.001$).

TABLE XIII. TIME TAKEN ON CONDUCTING THE EVALUATION

Usability measure	Google+	LinkedIn	Ecademy
Total time spent by all users (In minutes)	429	377	372
Average time per user per task (in minutes)	4.29	3.77	3.72
Average time per user over ten tasks	42.9	37.7	37.2

2) *User Satisfaction*

It can be seen clearly that Ecademy delivered the highest overall score, at 6.5, whereas LinkedIn delivered the second highest score, at 4.9, and Google+ delivered the lowest score among the three websites, at 4.2. This indicates that there were certain factors that influenced the users, which then affected the satisfaction rating for the tested website, as evidenced by the critical user comments on the design features of each website.

3) *Number of usability problems discovered*

Table 14 explains the total usability problems found by user testing and their severity rating. All the redundant problems are removed.

The usability problems detected in Google+ were 34, higher than in the LinkedIn and Ecademy websites (26 vs. 19). The One-way ANOVA test was used, and it delivered statistical differences amongst the number of

problems ($F = 15.033$, $p < 0.001$). Pearson's correlation was used and the result reveals a positive relationship between time spent and problems discovered, with a 'Sig' value of 0.02. This result reveals that the users who spent more time were able to discover more usability problems.

4) *UEM Performance Metrics*

By applying the UEM and reliability formulae, Table 15 explains that the thoroughness of UT in identifying real usability problems was 0.23. The validity of UT in finding the known usability problems was 0.04. The effectiveness of UT in identifying usability problems related to the user interface was 0.03. The One-way ANOVA test was used to find significant differences between the websites (as a dependent factor). The results reveal that there are no significant differences ($p > 0.05$). The results for the cost of employing UT on each website were a little different with an average of \$1,404, as shown Table 16.

TABLE XIV. NUMBER OF USABILITY PROBLEMS DISCOVERED

Problem type	Google+	LinkedIn	Ecademy	Total problems in all websites
	Total usability problems	Total usability problems	Total usability problems	
Catastrophic	4	2	0	6
Major	9	5	3	17
Minor	11	8	6	25
Cosmetic	10	11	11	32
No. of problems	34	26	19	79

TABLE XV. THE MEAN RESULT OF UEM METRICS

Metric	Google+	LinkedIn	Ecademy	Mean Total
Thoroughness	0.23	0.21	0.24	0.23
Validity	0.11	0.14	0.15	0.04
Effectiveness	0.02	0.03	0.032	0.03
Reliability	0.4	0.28	0.23	0.3

TABLE XVI. COST OF EMPLOYING UT IN THIS RESEARCH

Mean cost	Ecademy	LinkedIn	Google+	Evaluation Method
\$1,404	\$1,370 Time spent by 10 users (6.2 hours) + 5 hours collecting data from the evaluation sessions + 2.5 hours analysing data	\$1,378 Time spent by 10 users (6.28 hours) + 5 hours collecting data from the evaluation sessions + 2.5 hours analysing data.	\$1,465 Time spent by 10 users (7.15 hours) + 5 hours collecting data from the evaluation sessions + 2.5 hours analysing data.	User Testing (UT)

VI. COMPARATIVE ANALYSIS TO EVALUATE THE ADAPTIVE FRAMEWORK

This section represents comparative and comprehensive analysis between the three methods.

A. Types of problems found by UT in relation to DSI and HE

Two independent expert evaluators were involved in discussing, agreeing and deciding where the UT problems should be in HE, and to which category they should belong in DSI, as Tables 17 and 18 illustrate. The overall results from both tables show that all the UT problems were successfully classified into DSI. However, 30 problems out of 34 in Google+, and 12 problems out of 19 in Ecademy were successfully classified into HE. This proves that HE is rather general, and is unlikely to encompass all user problems, such as usability problems in the 'User usability, sociability and

management activities', 'Business support', and 'Security and privacy' areas. Thus, this proves that DSI was able to discover user problems, and the unique problems that were discovered by UT and did not discovered by HE and DSI; were classed as missed problems for DSI and HE. The tasks given to the users during the usability testing seem to have 'walked them through' the activities, which could have increased the opportunity to discover problems. Furthermore, the findings confirm that 'Visibility of system status', 'Match between the system and the real world', 'Aesthetic and minimalist design' in HE, as well as the seven areas in DSI are a common weakness in dynamic websites (particular for SNSs). All three websites found nearly equal numbers of usability problems related to navigation and visibility. In conclusion, UT worked better than HE because 11 problems were not classified in it. However, all the users' problems were classified in the DSI.

TABLE XVII. USABILITY PROBLEMS FOUND COMPARED TO THE HE

Nielsen's Heuristics	Google+	LinkedIn	Ecademy
Visibility of system status	4	2	4
Match between the system and the real world	5	3	2
User control and freedom	3	2	0
Consistency and standards	1	1	2
Error prevention	2	3	0
Recognition rather than recall	2	1	1
Flexibility and efficiency of use	0	2	0
Aesthetic and minimalist design	6	1	2
Helps users recognize, diagnose and recover from errors	4	2	1
Help and documentation	3	1	0
Total problems	30	19	12

TABLE XVIII. USABILITY PROBLEMS FOUND COMPARED TO THE DSI

Usability problem area	Google+	LinkedIn	Ecademy
Layout and formatting	3	4	3
Content quality	7	6	2
Security and privacy	3	1	0
Business support	5	3	0
User usability, sociability and management activities	8	5	6
Accessibility and compatibility	2	0	0
Navigation system and search quality	6	7	8
Total problems	34	26	19

B. Performance of the Three Methods

Generally, Tables 19, 20 and 21 show how UT, HE and DSI revealed different types and numbers of usability problems. One-way ANOVA reveals that there is significant difference between three methods in terms of discovering usability problems on the whole ($F = 13.32$, $p < 0.001$). UT, HE and DSI revealed 47%, 31% and 75% of the usability problems found in Google+, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant mean difference amongst the methods in finding usability problems in Google+ between HE and UT, where $p < 0.03$ and the mean difference = -14.667, as well as between DSI and HE, where $p < 0.003$ and mean difference = -16.767. In LinkedIn, UT, HE and DSI revealed 46%, 23% and 84% of the found usability problems, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant difference amongst the methods in finding usability problems in LinkedIn, particular between

HE and DSI ($p < 0.046$ and mean difference = -14.333) and between HE and UT ($p < 0.009$ and mean difference = -15.367). Finally, UT, HE and DSI revealed 50%, 32% and 87% of the found usability problems in Ecademy, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is significant difference amongst the methods in finding usability problems in Ecademy between HE and DSI, where $p = 0.012$ and mean difference = -15.000. The performance of HE in discovering usability problems during the experiment ranged from 23% to 31%. UT discovered usability problems ranging from 40% to 47%, while DSI discovered usability problems ranging from 69% to 84%. Also, UT and HE performed better in discovering major, minor and cosmetic real usability problems, but DSI was the best in discovering more catastrophic, major, minor and cosmetic real usability problems. Thus, it can be seen that DSI was the best in discovering real problems; this was followed by UT, and then finally HE.

TABLE XIX. FINDINGS IN GOOGLE+

Method \ Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicates)
Catastrophic	4 (100%)	0 (0%)	0 (0%)	4
Major	9 (82%)	3 (27%)	11 (100%)	11
Minor	11 (37%)	13 (43%)	28 (93%)	30
Cosmetic	10 (37%)	6 (22%)	16 (59%)	27
No. of problems	34 (47%)	22 (31%)	55 (75%)	72

TABLE XX. FINDINGS IN LINKEDIN

Method Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicates)
Catastrophic	2 (33%)	0 (0%)	6 (100%)	6
Major	5 (39%)	5 (39%)	11 (85%)	13
Minor	8 (32%)	8 (32%)	19 (76%)	25
Cosmetic	11 (92%)	0 (0%)	11 (92%)	12
No. of problems	26 (46%)	13 (23%)	47 (84%)	56

TABLE XXI. FINDING IN ECADEMY

Method Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicates)
Catastrophic	0 (0%)	0 (0%)	0 (0%)	0
Major	3 (50%)	0 (0%)	6 (100%)	6
Minor	6 (50%)	8 (67%)	11 (92%)	12
Cosmetic	11 (37%)	4 (13%)	16 (53%)	30
No. of problems	19 (40%)	12 (25%)	33 (69%)	48

C. Overlapping and Unique Problems

Many researchers recommend conducting UT together with HE because they have found that each method discovers unique problems [Nielsen, 1992], so when they are conducted together, they can reveal and present all the problems in the targeted website. Again, this experiment may confirm or deny this recommendation, depending on the following results.

Table 22 shows the performance of the three methods on a unique performance basis for the three websites, illustrating the number of problems revealed by the UT but not identified by the HE and DSI and vice versa. DSI was able to discover 6 catastrophic, 24 major, 41 minor and 25 cosmetic problems that were not revealed by the other methods. HE was not able to identify any catastrophic problems alone; however, it was able to identify 4 major, 19 minor and 9 cosmetic problems. UT was able to discover 6 catastrophic, 17 major, 25 minor and 32 cosmetic problems that were not revealed by the other methods.

In fact, each method revealed different types of problem (both unique and overlapping). However, DSI revealed the majority of real usability problems, indicating those with high severity ratings, and it also appeared to work fruitfully for the expert evaluators, who then revealed more real problems, both unique and overlapping.

For example, DSI found 41% uniquely of the total number of real usability problems (n = 73 out of 176). HE found 14% uniquely of the total number of real usability problems (n = 24 out of 176), and UT identified 32% uniquely of the total number of real usability problems (n = 56 out of 176). 23 (13%) real problems out of 176 were found to be 'overlapping' by the three methods. The clear superiority of DSI was due to involving user inputs in designing the method (as it is included in one stage of the adaptive framework), and due DSI being appropriate for the particular characteristics of the SNS domain.

TABLE XXII. SEVERITY PROBLEMS OF EACH METHOD'S PERFORMANCE, UNIQUELY AND WORKING IN PAIRS

Problem Types	HE (unique)	DSI (unique)	UT (unique)	HE & UT (overlapping)	DSI & UT (overlapping)	DSI & HE (overlapping)	Total number of problems in three websites (unique)
Catastrophic	0	6	6	0	8	0	10
Major	4	24	17	3	11	4	30
Minor	19	41	25	4	18	17	67
Cosmetic	9	25	32	5	21	9	69
Total	32	96	79	12	58	30	176

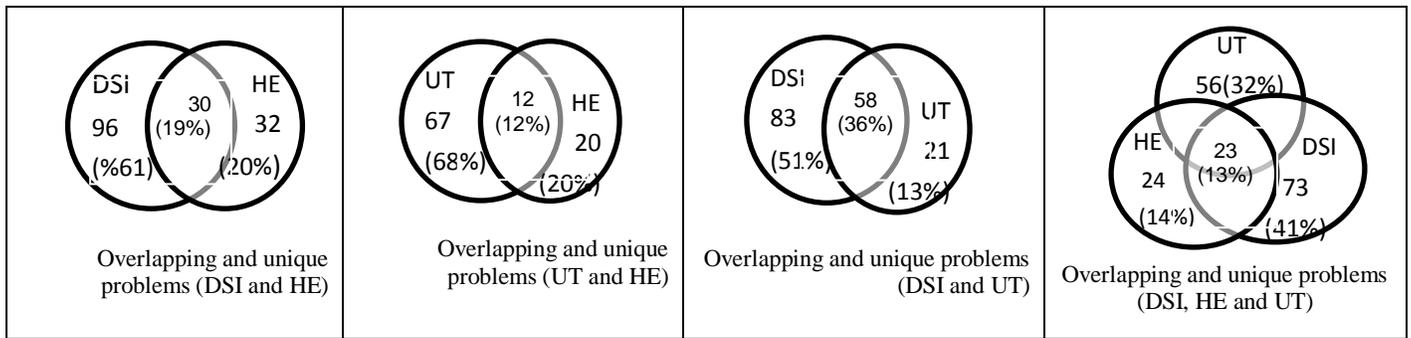


Fig.5. Overlapping and unique problems among the methods

It can also be seen that combining the results of DSI with either HE or UT offers better performance in terms of catastrophic, major, minor and cosmetic problems, whereas combining HE with either DSI or UT offers quite good results in terms of cosmetic problems. Combining UT with either DSI or HE offers better results in terms of minor and cosmetic problems. In summary, the result of comparison between UT and HE confirms conducting UT with HE in order to overcome the shortcomings of each, because each one is complementary to the other. On the other hand, DSI (as created by the adaptive framework) refutes this recommendation.

D. Usability Problem Areas

It can be seen in Table 23 that DSI helped to identify large numbers of real usability problems in all usability areas on the three websites (135). However, HE overall worked slightly better in discovering 47 real usability problems related to just four usability problem areas, although it failed to expose any usability problems in three main usability problems areas, which are 'Security and privacy', 'Business support', and 'Accessibility and compatibility'. Furthermore, UT worked better in discovering usability problems (76) in three usability areas, but it failed to identify a sufficient number of usability problems in the 'Accessibility and compatibility' area.

TABLE XXIII. NUMBER OF USABILITY PROBLEM AREAS IDENTIFIED BY THREE METHODS

Usability Problem Areas	UT	DSI	HE
Layout and formatting	10	15	9
Content quality	12	28	4
Security and privacy	4	8	-
Business support	8	4	-
User usability, sociability and management activities	19	55	19
Accessibility and compatibility	2	5	-
Navigation system and search quality	21	20	15
Total Problems	76	135	47

E. Comparison between the Three Methods in UEM performance

It can be seen from Table 24 that DSI are more efficient, thorough and effective in terms of identifying the total number of real problems against total time spent, and in its relative ability to identify usability problems related to the user

interface than the other methods. UT is the second best method, and HE is the last method. However, HE is the cheapest in terms of employment, and DSI is slightly more expensive than HE; both are cheaper than UT. One-way ANOVA reveals that there is significant difference among the methods used in terms of the UEM metric results, as shown in Table 25.

TABLE XXIV. COMPARING THE METRICS BETWEEN THE THREE METHODS

Metrics \ Methods	Efficiency	Thoroughness	Validity	Effectiveness	Reliability	Cost
HE	0.4	0.1	0.04	0.01	0.6	\$706,66
DSI	0.6	0.3	0.2	0.1	0.8	\$863,33
UT	0.5	0.023	0.04	0.03	0.3	\$1,404

TABLE XXV. ONE-WAY ANOVA RESULTS FOR THE THREE METHODS

Metrics	F	Sig. (p-value)
Efficiency	19.809	P< 0.001
Thoroughness	8.902	0.001
Validity	3.210	0.049
Effectiveness	3.367	0.48
Reliability	3.344	0.44

F. Advantages and Disadvantages of the Three Methods

We now assess the relative advantages and disadvantages of the three methods in evaluating user interfaces (see Table 26). Overall, DSI, as applied here, produced the best results; it found the most real problems, including more of the most serious ones, than did HE and UT, and at only a slightly

higher cost. HE missed a large number of the most severe problems, but it was quite good in identifying cosmetic and minor problems. UT is the most expensive method and it missed some severe problems; however, it helps in discovering general problems and it assists, as does DSI, in defining the users' goals.

TABLE XXVI. SUMMARY OF THE STUDY'S FINDINGS

Method	Advantages	Disadvantages
Usability Testing (UT)	<ul style="list-style-type: none"> * Helps define and achieve users' goals * Identifies the users' real problems * Identifies recurrent and general real problems 	<ul style="list-style-type: none"> * Misses some severe real problems * High cost * takes more time * Conducting under lab condition
Heuristics Evaluation (HE)	<ul style="list-style-type: none"> * Identifies little real problems * Low cost 	<ul style="list-style-type: none"> * Misses some severe problems * Too general * Not readily applicable to many new domains
Domain Specific Inspection (DSI)	<ul style="list-style-type: none"> * Identifies many more real problems * Identifies more serious, major, minor and cosmetic real problems * Improves the evaluator's performance * Identifies the real users' problems and helps define and achieve users' goals 	<ul style="list-style-type: none"> * A little higher in cost than HE and cheaper than UT. * Slightly higher in time than HE

VII. DISCUSSION AND FINDINGS

This section explores the results of this experiment and highlights the main findings. It thendraws out the lessons learned from the research. The main objective of this experiment was to evaluate the adaptive framework through its ability to generate a new method, specifically an inspection method designed for the social networks domain, by comparing its results with usability testing (UT) and Heuristic Evaluation (HE). It has been clearly shown that the hypotheses were accepted, and that Domain Specific Inspection (DSI) was able to find all the real problems that were discovered by UT and HE and more, but with greater

efficiency, thoroughness and effectiveness. Also, DSI was better at discovering catastrophic, major, minor and cosmetic real problems. It seemed to guide the evaluators' thoughts in judging the usability of the website through clear guidelines that included all aspects of the quality of the selected websites, which were represented in seven usability areas. As a result, it is unsurprising that the DSI method revealed a number of problems not discovered by the other two methods. HE method did not perform as well as either DSI or UT, based on the number of usability problems discovered during this experiment. The experts that used HE seemed to undermine their confidence whilst performing the evaluation, for example, when they performed the evaluation, they found no

readily applicable heuristic within HE for performing some of the main functions in these websites. Consequently, HE performed poorly in discovering problems. The UT method performed modestly against DSI, and well against HE, based on the number of problems identified. Thus, the findings indicate that it is not essential to conduct UT in conjunction with HE, in order to address the shortcomings of these methods; rather, to avoid wasting money, an alternative that is well-developed, context-specific and capable, such as the one generated here for SNSs (or in another research on the

educational domain [Roobaea et al., 2013c], should be employed. Furthermore, the adaptive framework provided optimal results regarding the identification of comprehensive 'usability problem areas' on the SNSs, with minimal input in terms of cost and time spent in comparison with the employment of usability evaluation methods. The framework was used here to generate DSI, which helped to guide the evaluation process as well as reducing the time that it would have taken to identify these usability issues through current evaluation methods. In terms of the definition of missed problems given by [Cockton and Woolrych, 2002], we can consider the problems found by any one method and not found by the others as missed problems. From this standpoint, DSI missed discovering 80 real usability problems. However, HE and UT missed 129 and 97 real usability problems, respectively.

The above findings facilitate decision-making with regard to which of these methods to employ, either on its own or in combination with another, in order to identify usability problems on websites. The selection of the method or methods will depend on the types of problem best identified by each of them.

VIII. CONCLUSIONS AND FUTURE WORK

Contrary to most of the efforts to construct and test enhanced usability methods, our work here has made explicit the process for so doing. The adaptive framework includes the views of users and usability experts to help generate a context-specific method for evaluating any chosen domain. The work presented here illustrates and evaluates this process for the generation of the DSI method to assess and improve the usability of social network websites. DSI outperformed both HE and UT, even when taken together. This clearly represents a step in the right direction. Further validation of the use of our adaptive framework will indicate whether it is indeed applicable across domains. In order to consolidate and confirm the findings, future research could include testing the adaptive framework by developing DSI for different fields such as e-commerce and healthcare systems.

In conclusion, this research contributes to the advancement of knowledge in the field. Its first contribution is the building of an adaptive framework for generating a context-specific method for the evaluation of whichever system in any domain (Figure 1). The second contribution is the introduction of DSI, which is specific for evaluating social network websites (Table 2). The third contribution is the identification of usability problem areas in the social network domain (seven areas in Table 2).

ACKNOWLEDGEMENTS

We thank the expert evaluators and users in the School of Computing Sciences at the University of East Anglia (UEA) and the Aviva company for their participation in the comparative study and the mini-usability testing.

References

- [1] Ali H. Al-Badi, Michelle, O. Okam, Roobaea Al Roobaea and Pam J. Mayhew (2013), "Improving Usability of Social Networking Systems: A Case Study of LinkedIn," *Journal of Internet Social Networking & Virtual Communities*, Vol. 2013 (2013), Article ID 889433, DOI: 10.5171/2013.889433.
- [2] Alias, N., Siraj, S., DeWitt, D., Attaran, M. & Nordin, A. B. (2013), Evaluation on the Usability of Physics Module in a Secondary School in Malaysia: Students' Retrospective. *The Malaysian Online Journal of Educational Technology*, 44.
- [3] Alrobai, A. AlRoobaea, R. Al-Badi, A., Mayhew, P. (2012). Investigating the usability of e- catalogue systems: modified heuristics vs. user testing, *Journal of Technology Research*.
- [4] Alshamari, M. and Mayhew, P. (2008). Task design: Its impact on usability testing. In *Internet and Web Applications and Services, 2008, ICIW'08. Third International Conference on*, pages 583-589. IEEE.
- [5] Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*, pages 189-194.
- [6] Chatrtrichart, J. & Brodie, J. (2004). Applying user testing data to UEM performance metrics. In *CHI'04 extended abstracts on Human factors in computing systems* (pp. 1119- 1122). ACM.
- [7] Chatrtrichart, J. and Brodie, J. (2002). Extending the heuristic evaluation method through contextualisation. *Proc. HFES2002, HFES (2002)*, 641-645.
- [8] Chatrtrichart, J. and Lindgaard, G. (2008). A comparative evaluation of heuristic-based usability inspection methods, In the proceeding of *CHI'08 extended abstracts on Human factors in computing systems*, 2213-2220.
- [9] Chen, S. Y. and Macredie, R. D., (2005), The assessment of usability of electronic shopping: A heuristic evaluation, *International Journal of Information Management*, vol. 25 (6), pp. 516-532.
- [10] Cockton, G. and Woolrych, A. (2002). Sale must end: should discount methods be cleared off HCI's shelves? *interactions*, 9(5):13-18. ACM.
- [11] Coursaris, C. K. & Kim, D. J. (2011), A meta-analytical review of empirical mobile usability studies. *Journal of usability studies*, 6(3), 117-171.
- [12] Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 101-110). ACM.
- [13] Dumas, J. and Redish, J. (1999). *A practical guide to usability testing*. Intellect Ltd.
- [14] Ellison, N. et al. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1): 210-230.
- [15] Estes, J., Schade, A. and Nielsen, J. (2009), 109 *User Experience Guidelines for Improving Notifications, Messages, and Alerts Sent Through Social Networks and RSS*, Accessed on 5/8/2012, Available at: [<http://www.nngroup.com/reports/streams/>].
- [16] Fernandez, A., Insfran, E. and Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study, *Information and Software Technology*.
- [17] Fox, D. and Naidu, S. (2009). Usability evaluation of three social networking sites. *Usability News*, 11(1): 1-11.
- [18] Fu, F., Liu, L., and Wang, L. (2008). Empirical analysis of online social networks in the age of web 2.0. *Physica A: Statistical Mechanics and its Applications*, 387(2-3):675-684.
- [19] Gutwin, C. & Greenberg, S. (2000), The mechanics of collaboration: Developing low cost usability evaluation methods for shared workspaces. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2000. (WET ICE 2000). Proceedings. IEEE 9th International Workshops on* (pp. 98-103). IEEE.

- [20] Hart, J., Ridley, C., Taher, F., Sas, C. and Dix, A. (2008), Exploring the Facebook experience: a new approach to usability. In Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges, pages 471-474. ACM.
- [21] Hasan, L. (2009), Usability evaluation framework for e-commerce websites in developing countries.
- [22] Hertzum, M. and Jacobsen, N. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4): 421-443.
- [23] Holzinger, A. (2005), Usability engineering methods for software developers *Communications of the ACM*, vol. 48 (1), pp. 71-74.
- [24] ISO (1998), ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability.
- [25] J. Nielsen. (2001), "Did poor usability kill e-commerce", in www.useit.com.
- [26] Jeffries, R., Miller, J.R., Wharton, C. & Uyeda, K.M. (1991). User interface evaluation in the real world: A comparison of four techniques. *Proceedings of ACMCHI'91*, pp. 119-124. New York: ACM Press.
- [27] Khajouei, R., Hasman, A. and Jaspers, M. (2011), Determination of the effectiveness of two methods for usability evaluation using a CPOE medication ordering system, *International Journal of Medical Informatics*, vol. 80 (5), pp. 341-350.
- [28] Latchman, H., Salzmann, C., Gillet, D. and Bouzekri, H. (1999), Information technology enhanced learning in distance and conventional education, *Education, IEEE Transactions on*, vol. 42 (4), pp. 247-254.
- [29] Law, L. and Hvannberg, E. (2002). Complementarily and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. In *Proceedings of the second Nordic conference on human-computer interaction*, pages 71-80, ACM.
- [30] Liljgren, E. (2006), Usability in a medical technology context assessment of methods for usability evaluation of medical equipment, *International Journal of Industrial Ergonomics*, vol. 36 (4), pp. 345-352.
- [31] Liljgren, E., & Osvalder, A. L. (2004). Cognitive engineering methods as usability evaluation tools for medical equipment. *International Journal of Industrial Ergonomics*, 34(1), 49-62.
- [32] Ling, C. and Salvendy, G. (2005), Extension of heuristic evaluation method: a review and reappraisal, *Ergonomia IJE & HF*, vol. 27 (3), pp. 179-197.
- [33] Mack, R. and Nielsen, J. (1994). *Usability inspection methods*. edited book, John Wiley & Sons, Inc., ISBN 0-471-01877-5.
- [34] Magoulas, G. D., Chen, S. Y. and Papanikolaou, K. A. (2003), Integrating layered and heuristic evaluation for adaptive learning environments. In the proceeding of UM2001, 5-14.
- [35] Mankoff, J., Dey, A., Hsieh, G., Kientz, J., Lederer, S. and Ames, M. (2003). Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 169-176. ACM.
- [36] Masip, L., Granollers, T. and Oliva, M. (2011). A heuristic evaluation experiment to validate the new set of usability heuristics. In *Proceedings of the 2011 Eighth International Conference on Information Technology: New Generations*, pages 429-434. IEEE Computer Society.
- [37] McCarthy, J. and Wright, P. (2004). Technology as experience. *Interactions*, 11(5):42-43.
- [38] Nayebi, F., Desharnais, J. M. & Abran, A. (2012). The state of the art of mobile application usability evaluation. In *Electrical & Computer Engineering (CCECE)*, 2012 25th IEEE Canadian Conference on (pp. 1-4). IEEE.
- [39] Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings ACM CHI'92 Conference (Monterey, CA, May 3-7)*, pages 373-380. ACM.
- [40] Nielsen, J. (1994), *Heuristic evaluation, Usability Inspection Methods*, vol. 24, pp. 413.
- [41] Nielsen, J. (1994a), *Usability engineering*, Morgan Kaufmann.
- [42] Nielsen, J. (2000), *HOMERUN Heuristics for Commercial Websites*, in www.useit.com.
- [43] Nielsen, J. and Molich, R. (1990), Heuristic evaluation of user interfaces, *Proc. ACM CHI'90 (Seattle, WA, 1-5 April 1990)*, 249-256.
- [44] Oztekin, A., Kong, Z. J. and Uysal, O. (2010), UseLearn: A novel checklist and usability evaluation method for eLearning systems by criticality metric analysis, *International Journal of Industrial Ergonomics*, vol. 40 (4), pp. 455-469.
- [45] Pessagno, R. (2010), *Design and usability of social networking web sites*, BSc dissertation at California Polytechnic State University - San Luis Obispo.
- [46] Pinelle, D., Wong, N. and Stach, T. (2008). Heuristic evaluation for games: usability principles for video game design. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1453-1462. ACM.
- [47] Preece, J. and Maloney-Krichmar, D. (2003). Online communities: focusing on sociability and usability. *Handbook of Human-computer Interaction*, pages 596-620.
- [48] Rabiee, F., (2004), Focus-group interview and data analysis, *Proceedings of the Nutrition Society*, vol. 63 (4), pp.655.
- [49] Roobaea AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew. (2013a). A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites, 2nd International conference on Human Computer Interaction & Learning Technology (ICHCILT 2013), MARCH 05-06, 2013, Abu Dhabi, United Arab Emirates (UAE).
- [50] Roobaea AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew.(2013b) .A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites- Further Validation, 2nd International conference on Human Computer Interaction & Learning Technology (ICHCILT 2013), MARCH 05-06, 2013, Abu Dhabi, United Arab Emirates (UAE).
- [51] Roto, V., Rantavuo, H. and Väänänen-Vainio-Mattila, K. (2009), Evaluating user experience of early product concepts, In the proceeding of DPPI, 2009, 199-208.
- [52] Sears, A. (1997), Heuristic walkthroughs: Finding the problems without the noise, *International Journal of Human-Computer Interaction*, vol. 9 (3), pp. 213-234. Shackel, B. and Richardson, S. J. (1991), *Human factors for informatics usability*, Cambridge University Press.
- [53] Smith-Atakan, S. (2006), *Human-computer interaction*. Thomson Learning Emea.
- [54] Sutcliffe, A. and Gault, B. (2004), Heuristic evaluation of virtual reality applications. *Interacting with Computers*, 16(4): 831-849.
- [55] Tan, W., Liu, D. and Bishu, R. (2009), Web evaluation: Heuristic evaluation vs. user testing, *International Journal of Industrial Ergonomics*, vol. 39 (4), pp. 621-627.
- [56] Thompson, A. and Kemp, E. (2009), Web 2.0: extending the framework for heuristic evaluation. In *Proceedings of the 10th International Conference NZ Chapter of the ACM's*
- [57] Special Interest Group on Human-Computer Interaction, pages 29-36. ACM.
- [58] Tsui, K. M., Abu-Zahra, K., Casipe, R., M Sadoques, J. and Drury, J. L. (2009), A Process for Developing Specialized Heuristics: Case Study in Assistive Robotics, University of Massachusetts Lowell, Tech. Rep, vol. 11, pp. 2009.
- [59] Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1), 20-36.
- [60] Van den Haak, M., de Jong, M. and Schellens, P. (2004), Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with computers*, 16(6): 1153-1170.
- [61] Wilson, C. (2007). Taking usability practitioners to task. *Interactions*, 14(1): 48-49.
- [62] Zaharias, P. & Poylymenakou, A. (2009), Developing a usability evaluation method for e-learning applications: Beyond functional usability. *Intl. Journal of Human-Computer Interaction*, 25(1), 75-98.
- [63] Roobaea AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew., (2013c). Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Educational Websites, *International Journal of Human Computer Interaction (IJHCI) Volume 4, Issue 2*.

Proposed Multi-Modal Palm Veins-Face Biometric Authentication

S.F.Bahgat

College of computers and IT
Taif University
Taif, KSA

S. Ghoniemy

College of computers and IT
Taif University
Taif, KSA

M. Alotaibi

College of computers and IT
Taif University
Taif, KSA

Abstract—Biometric authentication technology identifies people by their unique biological information. An account holder's body characteristics or behaviors are registered in a database and then compared with others who may try to access that account to see if the attempt is legitimate. Since veins are internal to the human body, its information is hard to duplicate. Compared with a finger or the back of a hand, a palm has a broader and more complicated vascular pattern and thus contains a wealth of differentiating features for personal identification. However, a single biometric is not sufficient to meet the variety of requirements, including matching performance imposed by several large-scale authentication systems. Multi-modal biometric systems seek to alleviate some of the drawbacks encountered by uni-modal biometric systems by consolidating the evidence presented by multiple biometric traits/sources. This paper proposes a multi-modal authentication technique based on Palm Veins as a personal identifying factor, augmented by face features to increase the accuracy of security recognition. The obtained results point at an increased authentication accuracy.

Keywords—Biometric authentication; Face Recognition; Feature Fusion; Palm veins; Statistical features.

I. INTRODUCTION

Biometrics is automated methods of recognizing a person based on a physiological or behavioral characteristic. Among the features measured are: face, fingerprints, hand geometry, iris, retinal, veins, handwriting, gait, and voice. Biometric systems are superior because they provide a nontransferable means of identifying people, not just cards or badges. The key point about an identification method that is "nontransferable" means it cannot be given or lent to another individual, so nobody can get around the system - they personally have to go through the control point.

A key advantage of biometric authentication is that biometric data is based on physical characteristics that stay constant throughout one's lifetime and are difficult to fake or change. Fingerprints, palm vein, and iris scans can produce absolutely unique data sets when done properly. It is not easy to determine which method of biometric data gathering and reading does the "best" job of ensuring secure authentication. Each of the different methods has inherent advantages and disadvantages [1].

Palm vein authentication uses an infrared beam to penetrate the users hand as it is held over the sensor; the veins within the palm of the user are returned as gray lines. As each Biometrics technology has its merits and shortcomings, it is

difficult to make direct comparisons, but because vein authentication relies on biological information on the interior of the body, it is more effective than the others at reducing the possibility of falsification. Also, vein pattern recognition requires just a scan of the palm, thus making it the easiest and most natural to use among the various biometric technologies [2].

Moreover, to confirm the accuracy of personal authentication to an even greater degree, vein recognition can be combined with face recognition systems to enable "multi-modal authentication" that guarantees accuracy through multiple layers of security. In addition to enhanced security, vein authentication used in conjunction with face recognition systems would also keep a log of facial information should it be necessary to be used as evidence [3].

This paper proposes a bi-modal biometric authentication system that fuses the features of the palm veins with that of the face for increasing authentication accuracy. At first, 4 types of statistical features are tested for the best recognition accuracy for palm veins and face independently. The feature sets best performing for each biometric are fed to a feature fusion strategy for increased authentication accuracy.

The rest of the paper is organized as follows. Section 2 scans the previous work in the related areas. Section 3 presents the structure of the proposed system. Section 4 analyzes the obtained results. The paper is terminated by a conclusion summarizing the obtained results and specifying problems for future work.

II. PREVIOUS WORKS

Ishani Sarkar et al [4] presented a review on the palm vein authentication device that uses blood vessel patterns as a personal identifying factor. As biometric technology matures, there will be an increasing interaction among the market, technology, and the applications. This interaction will be influenced by the added value of the technology, user acceptance, and the credibility of the service provider. It is too early to predict where and how biometric technology would evolve and get embedded in which applications. But it is certain that biometric-based recognition will have a profound influence on the way we conduct our daily business.

Masaki Watanabe et al [2], have shown a biometric authentication using contactless palm vein authentication device that uses blood vessel patterns as a personal identifying factor. Implementation of these contactless identification systems enables applications in public places or in

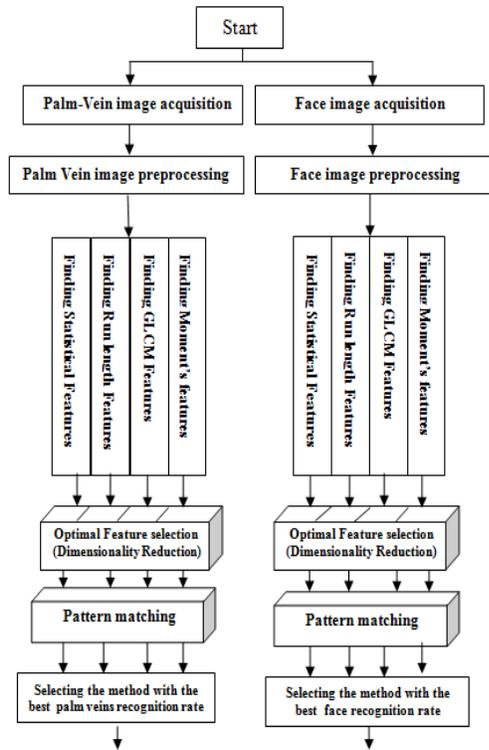


Fig.1 Flow of processes of the preparation phase

environments where hygiene standards are required, such as in medical applications. In addition, sufficient consideration was given to individuals who are reluctant to come into direct contact with publicly used devices.

Yingbo Zhou and Ajay Kumar [5] presented two palm vein representations, using Hessian phase information from the enhanced vascular patterns in the normalized images, and secondly from the orientation encoding of palm vein line-like patterns using localized Radon transform. The experimental results suggest that the proposed representation using localized Radon transform achieves better or similar performance than other alternatives while offering significant computational advantage for online applications. The proposed approach achieves the best equal error rate of 0.28%. Finally, they proposed a score level combination strategy to combine the multiple palm vein representations, thus achieved consistent improvement in the performance, both from the authentication and recognition experiments, which illustrates the robustness of the proposed schemes.

Yi-Bo Zhang et al [6] proposed a scheme of personal authentication using palm vein. The proposed system includes: 1) Infrared palm images capture; 2) Detection of Region of Interest; 3) Palm vein extraction by multiscale filtering; 4) Matching. The experimental results demonstrate that the recognition rate of their system is fine but not good enough to be a real system. The capture device is very sensitive to the outside lights. The outside lights can affect the inside infrared light source so that some images have very poor quality. If the capture device can be improved, the system performance should be better. Further, the database is too small to be

convincible. More data are required to be collected for the evaluation of the system.

Alaa ELEYAN, and Hasan DEM'IREL [7], introduced a new face recognition technique based on the gray-level co-occurrence matrix (GLCM). GLCM represents the distributions of the intensities and the information about relative positions of neighboring pixels of an image. They proposed two methods to extract feature vectors using GLCM for face classification. The first method extracts the well-known Haralick features from the GLCM, and the second method directly uses GLCM by converting the matrix into a vector that can be used in the classification process. The results demonstrate that the second method, which uses GLCM directly, is superior to the first method that uses the feature vector containing the statistical Haralick features in both nearest neighbor and neural networks classifiers. The proposed GLCM based face recognition system not only outperforms well-known techniques such as principal component analysis and linear discriminant analysis, but also has comparable performance with local binary patterns and Gabor wavelets. It is obvious from the results that the GLCM is a robust method for face recognition with competitive performance.

Muhammad Imran Razzak et al [3], presented multimodal face and finger veins recognition systems in which multilevel score level fusion was performed. Since there is no database for finger veins and face, thus they test the CAIRO employer and students. The imposter and genuine score are combined using Fuzzy fusion to increase the face recognition system. Simulation results shows that proposed multimodal recognition system is very efficient in reducing the FAR 0.05 and increasing GAR 91.4. The GAR and FAR can further be optimized by applying class to client approach on finger veins.

III. PROPOSED SYSTEM

Due to the increase in security requirements, biometric systems have been commonly utilized in many recognition applications. Multimodal systems have great demands to overcome the issue involved in single trait systems and this has become one of the most important research areas of pattern recognition. This paper presents a multimodal palm veins and face biometric verification system to improve the performance that fuses palm veins and face features for better authentication accuracy. The proposed system proceeds as follows (for both biometric systems (palm veins and face):

A. Preparation Phase:

In this phase, we estimate the performance of four statistical feature extraction methods and adopt the one with the best recognition rate. The process proceeds as shown in Fig. 1.

1) Image Acquisition:

The palm vein images were captured using "M2-PalmVein™ Reader" [8], for 18 persons.

The face images were captured usingfor the same 18 persons; examples of the captured images are shown in Figure 2.



Figure 2 Examples of the captured palm veins and face images

2) Preprocessing:

As the quality of the palm veins images were very bad, several preprocessing techniques were used to enhance the image quality. On the other hand, the facial images were noise-cleaned and contrast enhanced. An example of the enhanced images is shown in Figure 3.



Fig. 3 Examples of palm veins images after preprocessing

3) Feature selection

Four statistical approaches are applied for feature extraction; namely: Gray-level Co-occurrence Matrix (GLCM), Run-Length Matrix (RLM), Statistical Features (SF), and Moment Invariants (MIs). The paper aims at selecting the best performing approach for both biometrics. For that, we distorted the original images with three different types of noise; each with different noise levels. Salt and pepper, impulse and Gaussian noise types with intensity of 5%, 10%, and 15% are used throughout the study. The used feature selection techniques are as follows:

1) *Statistical features* The following set of features is used [9]:

“Mean, Variance, Smoothness, Skewness, Kurtosis, Uniformity, and Entropy”

2) *Gray-level-Concurrence Matrix (GLCM)* The following set of features is used [10]:

“Homogeneity (Angular Second Moment (ASM)), Energy, Entropy, Means, Variance, and Correlation”

3) *Run-length matrix (RLM)* The following set of features is used [11]:

“Short run emphasis (SRE), Long run emphasis (LRE), High gray-level run emphasis (HGRE), Low gray-level run emphasis (LGRE), Pair-wise combinations of the length and gray level emphasis (SRLGE, SRHGE, LRLGE, LRHGE), Run-length non-uniformity (RLNU), Grey-level non-uniformity (GLNU), and Run percentage (RPC)”

4) *Moment Invariants (MIs)* [12,13]:

Most frequently used are the Hu set of invariant moments which are invariant under translation, changes in scale, and also rotation.

5) *Dimensionality reduction* [14]:

To avoid the curse of dimensionality, PCA is applied by projecting the data onto a lower-dimensional space. This technique is applied for the feature vectors of the four feature extraction approaches.

6) Selecting the best-performing approach:

The recognition rate of each approach is determined using ten data sets; the original (clean) data set, and nine corrupted data sets with different types of noise with different noise levels. The average recognition rate is calculated and the best performing feature set is adopted for post processing.

7) The best approach for each biometric is selected.

B. Training Phase

The training phase algorithm is depicted in Fig. 4. In this phase:

- The palm veins and face images are captured in the same way as in the preparation phase
- The captured images pass through the same preprocessing stages.
- The feature vectors are extracted using the technique selected in the preparation phase.
- PCA is applied on both feature vectors for dimensionality reduction.
- The sets of features of both biometrics are fused.
- The fused feature vector is stored in database for future comparisons.

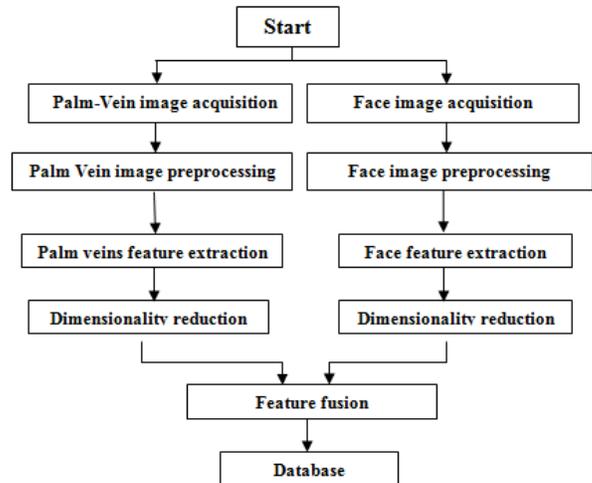


Fig. 4 Flow of processes of the training phase

C. Testing Phase

- The images of the person under test are acquired and preprocessed typically as in the preparation phase.
- The feature vector of the adopted approach is calculated and reduced using PCA.
- The resulting feature vector is compared with those stored in database and the person is recognized.

The testing phase algorithm is depicted in Fig. 5

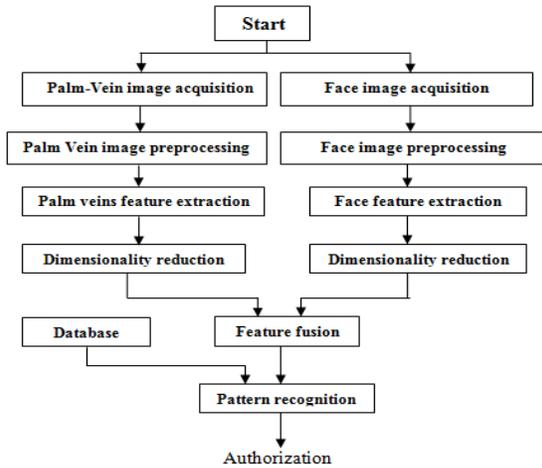


Fig.5 Flow of processes of the testing phase

IV. RESULTS AND DISCUSSION

For palm veins, and according to Fig. 6, it is clear that the Moment Invariants method gives the best recognition rate for all types of noise. However, for face; and according to Fig. 7, both GLCM and Moment Invariants give comparable recognition rates. As the GLCM requires more computations, we adopted Moment Invariants feature vectors for both biometrics.

Table 1 Average Palm-veins Recognition rate for different noise types

	salt & pepper	impulse noise	Gaussian
Statistical	29.16666667	31.94444444	29.16666667
RLM	29.16666667	29.16666667	29.16666667
GLCM	36.11111111	36.11111111	29.16666667
Moment	72.22222222	77.77777778	72.22222222

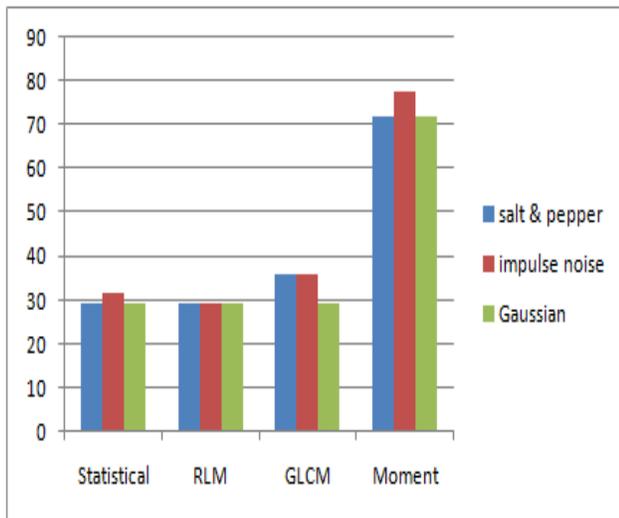


Fig. 6 Average Palm-veins Recognition rate for different noise types

Table 2 Average face Recognition rate for different noise types

	salt & pepper	impulse noise	Gaussian
RLM	63.88889	58.33333	29.16667
GLCM	70.83333	69.44444	29.16667
Moment	66.66667	70.83333	30.55556
Statistical	40.27777778	36.11111111	29.16667

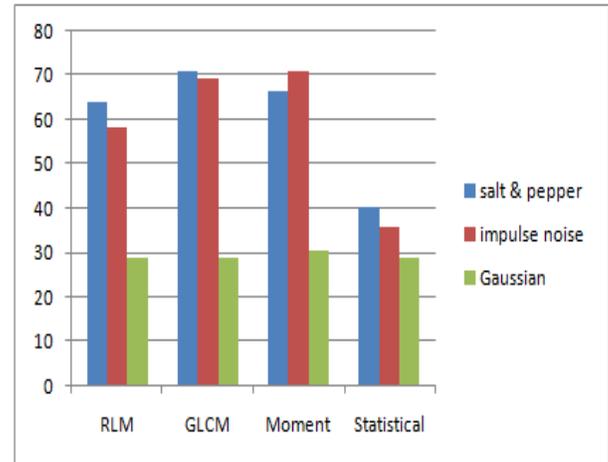


Fig. 7 Average Face Recognition rate for different noise types

For evaluating the performance of the proposed system, we calculated the recognition rate for palm veins images corrupted by salt & pepper noise with intensities of 0, 15, 20, and 25% respectively. The obtained results are shown in Table 3. Simultaneously, we calculated the recognition rate using the fused feature vector for the same type of noise with the same intensity levels. The results are shown in Table 4. It is clear that an average of 30% enhancement in the recognition rate is obtained, Fig. 8..

Table 3 Recognition rate of images corrupted by salt & pepper noise with different intensities using palm veins and fused feature vectors respectively

Noise Intensity	Recognition rate using palm veins feature vector	Recognition rate using fused feature vector
0%	100 %	100 %
15%	84.444 %	100%
20%	66.66 %	94.44 %
25%	44.44 %	90.44 %
average	73.882 %	96.22 %

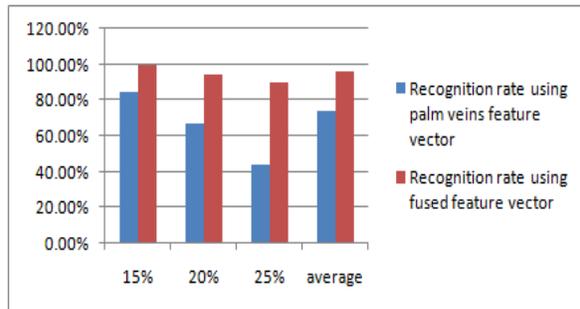


Fig. 8 Recognition rate enhancement using feature fusion

V. CONCLUSIONS AND FUTURE WORK

This paper studies the contribution of four different statistical approaches in recognizing persons through their palm veins and face images. These approaches are Gray-level Co-occurrence Matrix (GLCM), Run-Length Matrix (RLM), Statistical Features (SF), and Moment Invariants (MIs). The images under consideration are corrupted by different types of noise with different noise intensities, and the recognition rate is evaluated in each case. It was found that Moment Invariant (MIs) feature vector guarantees the best recognition rate in all cases. The MIs are then fused and applied for estimating the resulting recognition rate. It was found that this fusion enhances the recognition rate by more than 30%. Further analysis is required for more enhancements.

References

[1] Bhudev Sharma, "Palm Vein Technology", Technical Report, Electronics Engineering Department, National Institute of Technology, India, 2010.

[2] Masaki Watanabe, Toshio Endoh, Morito Shiohara, and Shigeru Sasaki, "Palm vein authentication technology and its applications", The Biometric Consortium Conference, September 19-21, 2005, USA.

[3] Muhammad Imran Razzak, Rubiyah Yusof and Marzuki Khalid, "Multimodal face and finger veins biometric authentication", Scientific Research and Essays Vol. 5(17), pp. 2529-2534, ISSN 1992-2248 ©2010 Academic Journals.4 September, 2010.

[4] Ishani Sarkar, Farkhod Alisherov, Tai-hoon Kim, and Debnath Bhattacharyya, "Palm Vein Authentication System: A Review", International Journal of Control and Automation, Vol. 3, No. 1, March, 2010.

[5] Yingbo Zhou, Ajay Kumar, "Contactless Palm Vein Identification using Multiple Representations", Department of Computing, The Hong Kong Polytechnic University.

[6] Yi-Bo Zhang, Qin Li, Jane You, and Prabir Bhattacharya "Palm Vein Extraction and Matching for Personal Authentication", Biometrics Research Centre, Department of Computing, Concordia University,

Quebec, Canada, Spring – Verlag Berlin Heidelberg 2007. <http://www.academicjournals.org/SRE>

[7] Alaa ELEYAN, Hasan DEM'IREL, "Co-occurrence matrix and its statistical features as a new approach for face recognition", Turk J Elec Eng & Comp Sci, Vol.19, No.1, 2011.

[8] <http://www.m2sys.com/pdf/M2-PalmVein.pdf>

[9] Fazal Malik and Baharum Baharudin, "The Statistical Quantized Histogram Texture Features Analysis for Image Retrieval Based on Median and Laplacian Filters in the DCT Domain", IAJIT First Online Publication, 2012.

[10] Fritz Albrechtsen, "Statistical Texture Measures Computed from Gray Level Cooccurrence Matrices," International Journal of Computer Applications, November 5, 2008.

[11] Xiaoou Tang, "Texture Information in Run-Length Matrices", IEEE Transactions on Image Processing, Vol. 7, No. 11, November 1998.

[12] Z. Song, B. Zhao, Z. Zhu, E. Mao "Research on Traffic Number Recognition Based on Neural Network and Invariant Moments", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.

[13] Zhihu Huang, Jinsong Leng, "Analysis of Hu's Moment Invariants on Image Scaling and Rotation", 2010 2nd International Conference on Computer Engineering and Technology, vol 7 pp 476-480, 2010.

[14] Aamir Khan and Hasan Farooq, "Principal Component Analysis -Linear Discriminant Analysis Feature Extractor for Pattern Recognition", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2, November 2011, ISSN (Online): 1694-0814.

Micro Sourcing Strategic Framework for Low Income Group

Noor Habibah Arshad

Department of Information Systems
Faculty of Computer and Mathematical Sciences,
University Teknologi MARA,
Shah Alam, Malaysia

Siti Salwa Salleh, Syaripah Ruzaini Syed Aris,

Norjansalika Janom, Norazam Mastuki
Faculty of Computer and Mathematical Sciences,
Faculty of Accountancy University Teknologi MARA,
Shah Alam, Malaysia

Abstract—The role of ICTs among poor people and communities has increased tremendously. One of the ICT industries – the micro sourcing industry – has been identified as one of a potential industry to help increase income for the poor in Malaysia. Micro sourcing is an effective way to accomplish tedious tasks at a faster rate. It involves large projects that are broken down into micro tasks. These micro tasks are well-defined and then distributed to a group of workers. The objective of this study is to develop the strategic framework of micro sourcing to generate income for the low income group. Four methods were used to gather information for this study. The methods used were documentation and literature reviews, focus group meetings, workshops and interviews. Based on the analysis of the current scenario of local micro sourcing industry, strategic framework was developed based on the five Strategic Thrusts identified. The Strategic Thrusts are harnessing demand side (job providers) of domestic and international market; platform capacity and capability building; leverage and utilise existing infrastructure; uplift and enhance capability of the supply side (micro workers); and instruments to expedite growth of local micro sourcing industry. The Strategic Framework is intended to provide strategic direction at national level to all stakeholders; to highlight key areas that need to be addressed in order to grow a sustainable micro sourcing industry in the country; and to serve as a guideline in the implementation of programs and plans related to micro sourcing industry development

Keywords—Capability building; expedite growth; harnessing demand; platform capacity; strategic thrusts

I. INTRODUCTION

The role of developing countries and poor people and communities as consumers and producers of ICTs is evolving. There is growing interest in developing countries as potential growth markets for ICT goods and services. In response, ICT producers are adjusting their products as well as business models to target low-income consumers. More resources are allocated to find ways to reach the “bottom of the pyramid” [1]. Improving mobile access – partly as a result of cheaper imports of technology – at increasingly affordable rates, and new service models are facilitating access for people without large or predictable incomes. This development has allowed for greater involvement of enterprises from developing countries in ICT-related innovation processes [2].

ICTs can also strengthen internal information systems for those (predominantly growth-oriented) enterprises that own PCs and are able to make effective use of computer-based applications. There is further evidence that ICTs can provide other benefits involving the strengthening of social and human capital such as enhancement of skills, increased self-confidence, increased participations of women, empowerment, and security against income loss. ICT use helps cement or even accentuate existing power relations and inequalities. ICT can reinforce the market position and power of existing trading intermediaries, whose actions may not impact positively on the livelihoods of the poor. Finally, the role of ICTs might be more limited in local value chain systems (particularly of subsistence based enterprises) that rely heavily on pre-existing, informal and culturally rooted communication where the exchange of valued information is by means of personal contact.

The extent to which technologies are available and used by the poor varies a great deal, with mobile phones and radio appearing as the most widely diffused and Internet-connected PCs (and especially with a broadband connection) the least. Beyond availability, the uptake of certain technologies by the poor also depends on the needs and capabilities of potential users. What matters is whether people have the access to what they want and need, not that they have access to technologies which are identical to other people with different needs.

The Malaysian Government realising its responsibility to upgrade the quality of life of the poor, has seriously list down its commitment for the RMK-10 period (2011-2015), in which key strategies to provide equitable opportunity to participate in the economy as well as work towards greater socioeconomic inclusiveness among all Malaysians [3]. Poverty in Malaysia is conceptualised and defined as income poverty and measured using a poverty line income to demarcate poor and non-poor households. The poverty line is determined in both absolute and relative terms. Absolute poverty line is calculated based on the income required to purchase a minimum food basket and other basic necessities. The relative concept of poverty stresses income inequality as its fundamental manifestation and is reflected in the definitions of poverty in the lower quintiles of the population, the welfare ratio and the index of poverty. Relative poverty in Malaysia is defined as per capita household income level of less than RM2300 per month with average

This study is conducted by researchers from Universiti Teknologi MARA, Shah Alam, Malaysia in collaboration with Malaysia Development Corporation (MDeC) and fully funded by Malaysia Ministry of Finance.

salary of RM1400 that cuts off the bottom 40 percent of the population, also known as the B40 group [3].

As Malaysia enters the last stretch in achieving Vision 2020, the strategic thrust has focused on ICT as an Industry, ICT as an Enabler and ICT for Society [4]. Therefore, to materialise these National Agenda, ICT together with micro sourcing can be used as one of the mechanism to uplift the low household income group. Through micro sourcing, workers have the flexibility of doing work during hours, locations and duration of their own choices. It also provides an additional income to complement existing income. In light of this initiative, the objective of this paper is to propose a micro sourcing development framework for Malaysia

II. MICRO SOURCING

Debates about IT outsourcing have lead to the discovery of micro sourcing implementation. As a result of high dependencies in IT project, micro sourcing activities could be emphasized as one of the outsourcing alternatives. If before this, an organization will outsource their IT project to a vendor, but now it can be done through a new mechanism known as micro sourcing.

A. Outsourcing vs. Micro sourcing

Outsourcing could be defined as an act of delegating or transferring some or all of the information technology related decision making rights, business processes, internal activities, and services to external providers, who develop, manage, and administer these activities in accordance with agreed upon deliverables, performance standards and outputs, as set forth in the contractual agreement [5]. According to [6] outsourcing related to a new process to create new framework involving the relationships of employee/employer. Payroll processing, email, web services and hosting, programming, call centres, and storage area network are examples services that are widely being outsourced [7], [8], [9]. This statement also supported by [10], [11] which has revealed that outsourcing brings competitive advantage, profits and customer satisfaction. It can be concluded, the willingness of an organizations to embark in IT outsourcing is being driven by several factors such as lower costs, faster development cycle, performance assurance and quality, professional and geographically dispersed service and creative and structured leases [12], [13], [14].

Due to the high demand of IT outsourcing, micro sourcing has been taken into consideration for some of the organizations in order to reduce the operational cost, increase the revenue and decrease the employee pressure. Micro sourcing is an effective way to accomplish tedious tasks at a faster rate. Task can be done either online or offline.

Normally, micro sourcing involves large projects that are broken down into micro tasks. These micro tasks are well-defined and then distributed to a group of workers [15]. There were several models proposed for micro sourcing such as FORT framework [16], Structuration Theory [17] and Entrepreneurial model [18]. Research conducted by [18] has proposed to utilize the Structuration Theory in micro sourcing by integrating other ideas related to opportunity research which lead to the discovery of micro sourcing research.

Structuration theory could be defined as entrepreneurial action that is enabled and constrained by conscious selection, imitation, and modification of business “scripts” that occur within social and business structures [18]. Scripts could be used in an environment which needs quick respond. Reference [18] has classified scripts as “legitimate,” “powerful,” and “competent”. Those scripts could be utilized from the organization level to the individual level, which contributes towards the practice of micro sourcing.

Based on the models stated above, it is clearly noted that micro sourcing occurs at individual level and it also facing several challenges that need to be addressed. Hence, emphasize should be given on how does IT supports the micro sourcing activities [19]. Furthermore, like any other IT services, micro sourcing also has to face several key issues that need to be addressed such as data security, agreement between the outsourcer and outsourcees, culture, and communication systems. According to [19], security and intellectual property security, risk and trust between micro sourced employees and contract are the main concern in micro sourcing activities.

B. Global Micro sourcing Industry

According to Crowdsourcing Industry Report [20], demand in the global micro sourcing industry is driven by start-up and small companies. Collectively they account for over 60% of the market revenues. Start-up companies drive majority of the revenues in the industry, contributing 39% of the total revenues. Furthermore, the industry’s total revenue grew approximately US\$140.8 million in 2009 and accelerated further to US\$375.7 million in 2011 as depicted in Figure 1.

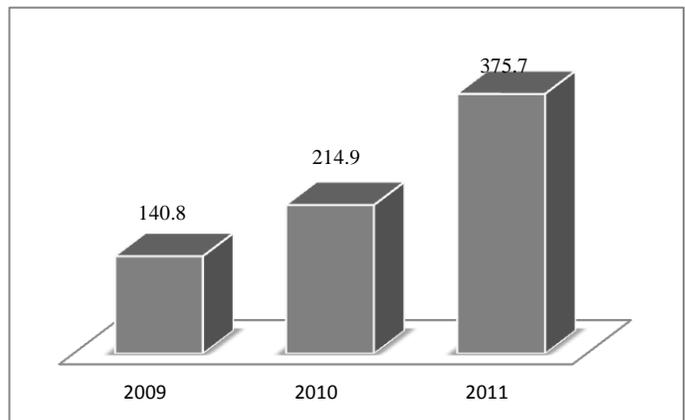


Fig. 1. Global micro sourcing industry’s revenues (US\$ million) [20]

Demand in the global micro sourcing industry as shown in Figure 2 is driven by start-up and small companies. Collectively they account for over 60 % of the market revenues. Start-up companies drive majority of the revenues in the industry, contributing 39 % of the total revenues. Large enterprises with revenue of more than US\$1 billion represented only 8% of total job providers but contributed 21% of total revenues due to huge transaction volume. In the global micro sourcing industry, North America and Europe are the largest job providers which contributed to 91% of jobs collectively. Meanwhile, 41% of micro workers from North America

followed by Asia Pacific which is 35% as depicted in Figure 3. Income for a micro worker vary significantly, depending on type of tasks, number of hours and efficiency in performing the tasks. For example in Figure 4, the top five workers handling micro tasks earn on average US\$500 per month while the top five workers handling expert and software-based tasks earn on average US\$8,300 per month. On average, the top five workers from each task category earn US\$4105 per month.

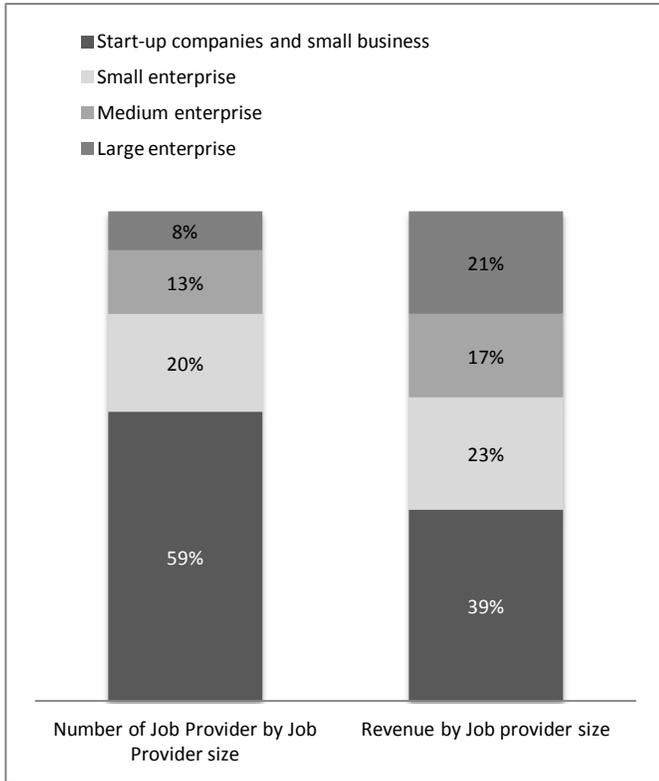


Fig. 2. Number of job provider and revenue by job provider's size in 2011 [20]

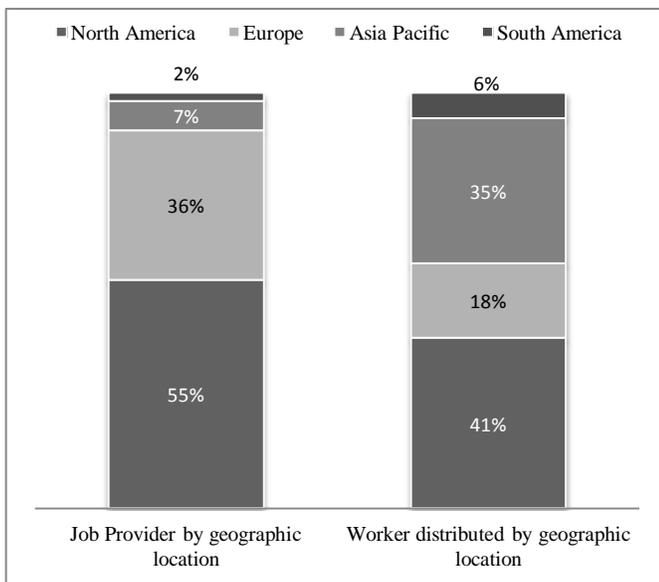


Fig. 3. Job provider and worker by geographic location in 2011 [20]

Platforms are needed to run the software and processes for micro works for use with internal or external micro workers. In the international market, number of platforms has been growing in the past few years. Among the platforms well-known worldwide are AmazonMechanicalTurk, CrowdFlower, SamaSource, Ushahidi, Micro sourcing and ODesk. AmazonMechanicalTurk was launched in November 2005 and which the requestors are restricted to US-based entities, however the workers can be sourced globally [21]. CrowdFlower was founded in 2007 and uses the MTurk platform to distribute work, but provides its own interface on which work is completed. It also has sophisticated APIs to create and manage works [22]. SamaSource was founded in 2008 and claims to have a dedicated team of remote workers but does not post jobs on a public portal like Mturk [23]. In 2012, Samasource has employed over 3,000 workers from low-income backgrounds in places as diverse as rural Haiti, informal settlements in East Africa, and peri-urban parts in India, paying out over US\$2 million in wages. At the current growth rate, Samasource expects to pay and train about 20,000 workers by 2017. Ushahidi was founded in 2008 and provides a platform for information collection, visualization and interactive mapping, especially for crises [24]. Micro sourcing is Philippines based company providing traditional outsourcing solutions. Micro sourcing started operations in 2004 and as of July 2011, they employ more than 1500 people working for more than 60 clients from all over the world [25]. Meanwhile, ODesk [26] was founded in 2003 and focus more on long-term work through remote staffing than real micro sourcing.

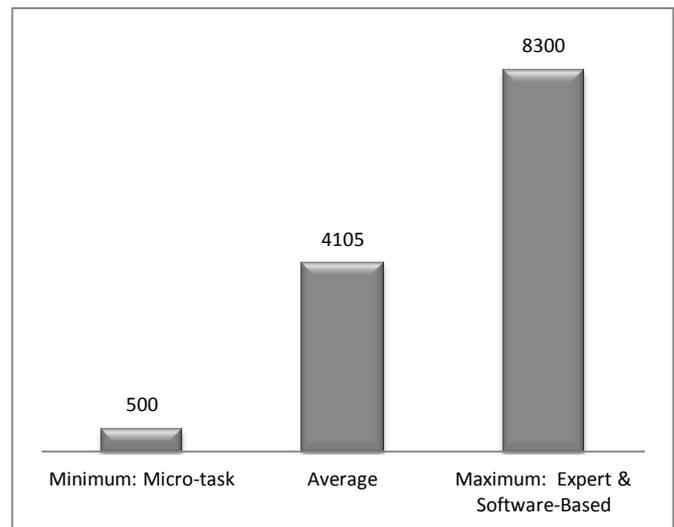


Fig. 4. Monthly earnings(\$USD) of top five workers by task category [20]

C. Micro sourcing Scenario in Malaysia

Micro sourcing project is not a green-field project but instead will tap on existing initiatives/ projects under the National ICT Policies. By tapping on existing information and infrastructure, replication and redundancy can be avoided while the impact of these existing initiatives/ projects can be enhanced. Some the existing initiatives/ projects are e-Government, myGovernment, National Broadband initiative

and Telecentre Development Program. Opportunities for micro sourcing industry are available in these initiatives. For example, some tasks under e-Government initiatives can be done through micro sourcing. Government agencies manage huge amounts of data under one system for internal and external use. Managing these data requires huge resources that can be minimised through micro sourcing by dividing the tasks required into micro tasks and outsourced them to micro workers. Some of the suitable tasks for micro sourcing are data entry, maintenance of databases, data back-up/ recovery and data protection.

Digitisation service is another suitable task for micro sourcing. Data or document digitisation is a task where data or information are extracted from hard copies (newspapers, books, paper documents, business cards, and periodicals) and soft copies (image/ video/ audio files and analogue signal) which then converted into digital formats. Other Government initiatives such as National Broadband Initiative and Telecentre Development Program meanwhile, could become enablers for the local micro sourcing industry.

Reference [27], the micro sourcing project is targeted to generate RM2.2 billion of income in the micro sourcing industry by 2020. About 70% of the income will be generated by the micro workers (employees) as payments for tasks completed. The remaining of the income will be generated by micro sourcing platform operators via profits and salaries. This project is also expected to create 1,425 full time jobs and provide income for 337,000 task workers by 2020. Another positive impact from this Project will be the value-added benefits to the low income group. Other than earning additional income to complement their existing income, they will also acquire new skills from the training provided.

The existing micro sourcing business model in Malaysia involved three groups namely Job Providers, Platforms and Micro workers [15], [28]. Job providers will outsource their jobs to micro sourcing platforms. Number of job providers is also quite limited and concentrated within the private sector. In the local market the number of platforms is relatively small and their roles are very limited; as mediator between job providers and micro workers. They advertise the tasks source from job providers and once tasks completed by micro workers; they verify those tasks before submitting them to job providers. Their roles in the local micro sourcing industry are not as wide as international platforms. When a task is advertised on a platform, micro workers will pull the task based on their interests. After the task is completed, the micro worker will submit it together with proof of the completed task to the platform. Once the task is verified as completed by the platform, the macro worker will get paid directly by the job provider. However, the tasks available in the market are not targeted to any specific micro workers and these workers are not given proper training to perform the task. The existing scenario could make the industry unsustainable in the long run.

Study on micro sourcing in Malaysia is very limited since the micro sourcing industry in Malaysia is still at its infancy stage and not properly structured. After analysing the current scenario of local micro sourcing industry, we are proposing development of a "Strategic Framework" to develop micro

sourcing industry in Malaysia with focus on the participation of the low income group as micro workers in the industry.

III. METHODOLOGY

This study was conducted through documentation review, observations, meetings, workshops and discussions held with relevant stakeholders. Literature review and information gathering were from Digital Malaysia lab report, academic and industry literature, RMK10, EPP annual reports, and market studies.

Focus group meetings and discussion were held starting with a focus group meeting and discussion with Micro sourcing for B40 Implementation Committee (MSIC-B40). MSIC-B40 is responsible to lead and oversee the development of this Project and is also responsible to formulate plans for local micro sourcing growth and identify micro sourcing opportunities for B40. Members of MSIC-B40 come from various non-governmental organisations (NGOs), government agencies, private organisations, universities and distinguished individuals. From these focus group meetings, initial direction of the study was identified and agreed. The research group also participates in the Crowd Business Model Summit (Oct 2012) to identify benchmark and best practices.

Two workshops were held to get the stakeholders engagement and to gather information. The first workshop was held with the objectives to gather information on micro sourcing potentials, opportunities, potential growth and critical enablers from the perspectives of stakeholders. The second workshop was held to discuss the findings and recommendations on the micro sourcing proposed framework and its strategic directions. Participants of both workshops were from MSIC-B40, government ministries and agencies, NGOs and private sectors.

To enrich the data and information gathered, follow up interviews were done with some related government agencies and private sectors. Interviews were also done with local, international platforms and potential job providers from Small Medium Enterprises (SME), Multi National Corporations (MNC) and Public Listed companies. The interviews help to uncover current practices, requirements and challenges in existing Malaysian micro sourcing environment. Interviews with related organisations also help to identify potential partnerships for micro sourcing strategic implementation. The information gathered from all these sources was then compiled and analysed as inputs for this study.

IV. MICRO SOURCING STRATEGIC FRAMEWORK

This study defines micro sourcing as the distribution of well-defined tiny tasks (also referred as micro tasks) to a large group of networked users (micro workers) through the internet. These tasks can be completed under flexible circumstances – own time and locations – using basic internet connected devices such as mobile phones, smart phones, tablets, notebooks and personal computers (PCs).

Based on the analysis of the current scenario of local micro sourcing industry, we are proposing a strategic framework for the development of local micro sourcing industry based on five Strategic Thrusts (Figure 5). The Strategic Thrusts are:

- 1) **Strategic Thrust 1: Harnessing Demand Side (Job Providers) of Domestic and International Market;**
- 2) **Strategic Thrust 2: Platform Capacity and Capability Building;**
- 3) **Strategic Thrust 3: Leverage and Utilise Existing Infrastructure;**
- 4) **Strategic Thrust 4: Uplift and Enhance Capability of the Supply Side (Micro Workers); and**
- 5) **Strategic Thrust 5: Instruments to Expedite Growth of Local Micro sourcing Industry.**

eliminating the gaps identified [15], [28]. Success of the Ecosystem relies on effective roles played by the relevant stakeholders. Thus, this Strategic Framework is intended to be as follows:

- 1) *To provide strategic direction at national level to all stakeholders;*
- 2) *To highlight key areas that need to be addressed in order to grow a sustainable micro sourcing industry in the country; and*
- 3) *To serve as a guideline in the implementation of programs and plans related to micro sourcing industry development.*

Five Strategic Thrusts have been identified as the foundation for the Strategic Framework of the micro sourcing industry development in Malaysia. These Thrusts are based on the SWOT and gap analysis involving the four groups of role players in the proposed micro sourcing industry ecosystem.

B. Strategic Thrust 1: Harnessing Demand Side (Job Providers) of Domestic and International Market

Strategy on the demand side (job providers) should focus on capturing local and international demand. The job providers should be from local public and private sectors as well as international firms and organizations.

1) Rationale

Key to the micro sourcing ecosystem is the volume of micro tasks available in the market. Without enough volume, there will not be enough jobs available for micro workers to sustain the industry ecosystem. There is a wide range of firms/public agencies, local and international that can be tapped as micro sourcing job providers for local market. The market size for global micro sourcing is estimated to grow to US\$20 billion in 2015 with 780,000 workers. The Malaysian outsourcing industry meanwhile, is predicted to be worth US\$1.9 billion by 2013.

In general, there is still lack of awareness about micro sourcing among local firms and public agencies. There is a need to promote micro sourcing as a way to lower operation costs especially for Small Medium Enterprises (SMEs) and public sector. On the international front, Malaysia with its diverse population with multicultural background should be promoted as a regional micro sourcing hub to attract international firms outsourcing their micro tasks in Malaysia.

2) Components

- Public Sector

One of the main sources on the demand side is Federal and State public sectors. Micro sourcing firms could anchored their businesses to public sector's operations, providing them the opportunity to build a track record and scale up their operations before handling wider range of business processes in the private sector. There are large pools of Governments' operations that can be utilised as micro sourcing tasks such as archive digitisation and e-Government platforms.

- Private Sector

There are many local businesses especially mobile carriers and banks that can provide tasks to local micro sourcing firms.

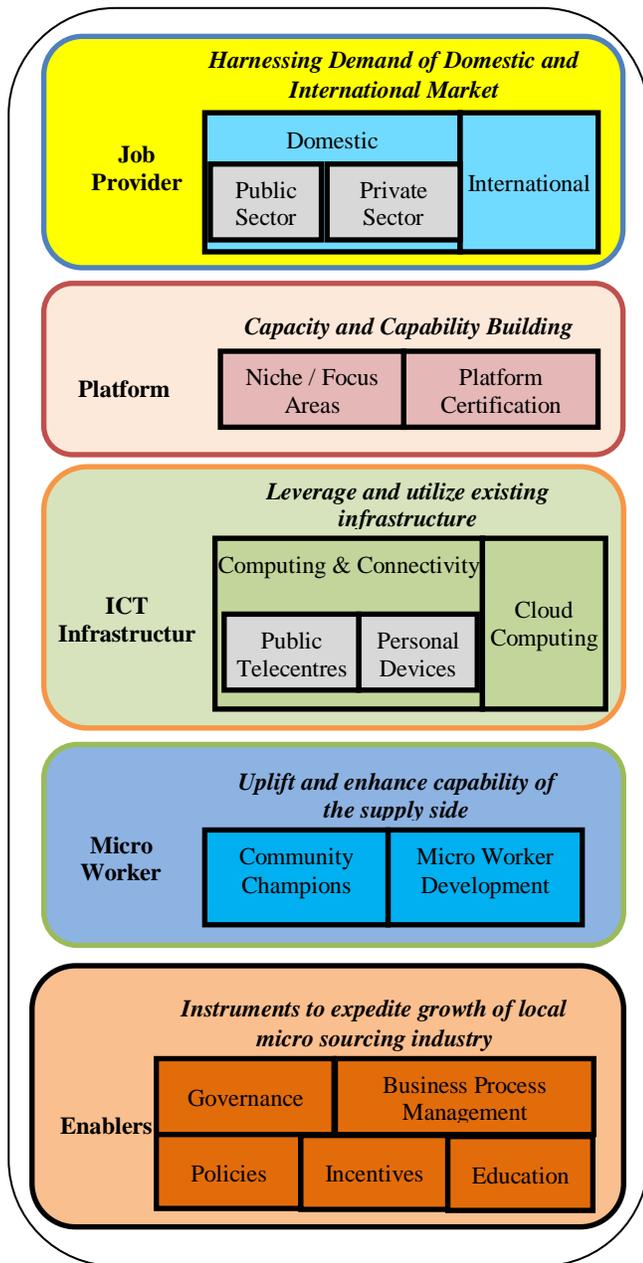


Fig. 5. Proposed micro sourcing strategic framework

The rationale behind the Strategic Framework is to create the proposed Micro sourcing Ecosystem by narrowing or

For example, there is increasing demand for digitisation of non-digitised records across the private sector. This includes one-time drive to digitised old records and continuous need to digitised new data/ information. Recipients of Government's contracts and concessions can also sub-contract parts of those contracts and concessions to local micro sourcing firms. SMEs are another private sector group that has huge potential as micro sourcing job providers. SMEs can utilise micro sourcing firms' to reduce their operating costs.

Micro tasks are not limited to 'profitable' business operations. Non-profitable operations can also be micro tasked such as activities by non-profit associations and societies. Large corporate also involve in non-profitable tasks under their Corporate Social Responsibility (CSR) activities. There are increasing number of "Triple" bottom line companies that focus on social, environment and financial results, which can utilise micro tasks for their CSR activities.

- International

Local micro sourcing firms should be encouraged to collaborate with large international business process outsourcing (BPO) firms to tap into jobs from international firms that can be utilised for local micro sourcing market. North American and European firms are the largest job providers for micro sourcing, collectively offering 90% of job market. Today, the impact sourcing market generates an estimated US\$4.5 billion in revenues globally, representing 3.8% of the entire US\$119 billion BPO industry, and directly employs about 144,000 people across all segments. US\$1.2 billion of these revenues is estimated to be incomes for impact sourcing workers. Share of impact sourcing in total BPO is expected to increase to approximately US\$20 billion in 2015, just over 11% of total BPO market, and directly accounting for 780,000 jobs. More than US\$10 billion is expected to be income for impact sourcing workers [20].

C. Strategic Thrust 2: Platform Capacity and Capability Building

Platform is important in the micro sourcing ecosystem to link job providers and micro workers. There should be enough platforms with diverse niche and focus areas that suit different key target groups of micro workers. These platforms should also be accredited to make them competitive.

1) Rationale

As there is no boundary in digital world, local micro sourcing platforms should be at the highest level of capacity and capability to compete with international platforms. Competitive platforms in local micro sourcing ecosystem require skilled and experienced talents that can design reliable, safe and diverse platforms.

Currently, there is no established national minimum standard and certification program for micro sourcing platforms. Without this standard and certification, we will not be able to promote Malaysia as a micro sourcing industry hub. Standard and certification will not only ensure platforms adhere to services of the highest quality but also ensure there are enough platforms suitable for the key target groups of micro workers.

Targeted micro workers have different competent levels and skills. Therefore, available platforms should be as diverse as possible to cater for these differences.

2) Components

- Niche/ Focus Areas

Each platform is specialised based on the type of task completed. These types of tasks can be categorised into ICT enabler or ICT related categories. ICT enabler refers to the use of ICT as an enabler to advertise a job or obtaining a task. ICT enabler does not require ICT to complete the tasks, but instead used to communicate the task to the workers. On the other hand, ICT related tasks refer to types of work that require ICT to complete the tasks. ICT related tasks include ideation-based tasks, knowledge-based task, expertise-based tasks and micro task.

Diverse platforms should be made available with different niche/ focus areas to cater for different types of targeted micro workers. Targeted micro workers have different competency levels and skills. Diverse platforms will allow wider pool of workers to participate in the micro sourcing industry.

- Platform Certification

Platform certification is important to ensure platforms' credibility. It provides guidance to better integrate the application process into micro sourcing ecosystem and ensures the quality of functional design incorporated into the micro sourcing platform. This guidance includes information on the technical and business processes, implementation of platform requirements, and certification requirements for a micro sourcing platform. A micro sourcing platform that has gone through certification processes means that it has met or exceeded the standards required. The certification processes should cover a platform's operations, security, performance, and resources to ensure the platform's capabilities and services

D. Strategic Thrust 3: Leverage and Utilise Existing Infrastructure

Existing infrastructure should be leveraged and utilises to reach out targeted groups of micro workers and reduce cost of platforms. Existing public telecentres can be used as micro tasks centres for targeted micro workers. In addition, targeted micro workers can also be assisted to have personal devices for micro tasks usage. Existing cloud computing infrastructure meanwhile, can provide flexibility, lower cost and improved scalability for platform providers

1) Rationale

Government has invested in various programs in rural areas, as well as selected urban areas to increase people's access to the Internet. These programs, such as setting up public telecentres and building telecommunication towers, are intended to reduce the digital divide between rural and urban areas, and also increase access to Internet for the under-served sections of the society. Currently there are 2,477 telecentres for different target groups that can be used as micro tasks centres to train micro workers as well as job locations for micro workers. However, proper coordination is needed to ensure that these telecentres are coordinated and managed properly to

ensure that they are properly utilised for the benefits of targeted micro workers.

Other than telecentres, personal devices such as mobile phones, tablets, Personal Digital Assistants (PDAs) and smart phones are among essential connectivity infrastructures to be available among B40 micro workers. However, since most of B40 cannot afford to have smart devices with Internet connection, there is a need to leverage on existing public telecentres.

Cloud computing is another existing IT infrastructure that can be utilised, especially with the increasing usage of personal mobile devices by micro workers. Cloud computing allows transfer of information and functionality between users and systems via mobile devices. Platform operators can optimise their IT infrastructure through cloud computing by leveraging on various enterprise-level applications provided by many MSC status companies. Utilising existing infrastructure will lower operating cost of platform operators and thus, make local platforms more competitive.

2) Components

- Computing and Connectivity (through Public Telecentres and Personal Devices)

In order to reach the targeted micro workers, computing infrastructure and connectivity are important. There are two models that can be utilized:-

- usage of personal computing devices for individuals to access to Internet and complete their tasks; and
- usage of public telecentres as distribution centres for micro workers to complete their tasks.
- Cloud Computing

Cloud computing infrastructure allows platform operators to reallocate their IT operational costs from infrastructure-related costs to other important costs such as platform designing. In addition, with applications hosted centrally, updates can be released without the need for platform operators to install new software. In a business model using cloud computing, platform operators are provided with access to application software and databases hosted at a centralised server. Cloud computing allows platform operators to get their applications up and running faster, with improved manageability and less maintenance, and also enables them to rapidly adjust their resources to meet fluctuating and unpredictable business demand. Cloud-based applications will also make it easier for micro workers to access platforms via web browser from a desktop or mobile application.

Using cloud computing, platform providers will have the opportunity to leverage enterprise-level applications and development without the associated upfront capital expenditure pinch or complex IT roll-out. Cloud computing may also be more flexible in developing new micro sourcing applications because the cloud includes middleware, and users are not limited to one server or one data centre. Cloud computing also has been proven to be very scalable, and can overcome technical fault or maintenance related downtime.

E. Strategic Thrust 4: Uplift and Enhance Capability of the Supply Side (Micro Workers)

Supply side in a micro sourcing industry ecosystem is not only about the numbers of micro workers available but also the capability of these micro workers and the quality of the micro tasks that they completed. Investment in training for micro workers' development is important to ensure high quality of labour supplies in the ecosystem. Community Champions should be groomed to engage these targeted micro workers.

1) Rationale

Skills training are necessary to ensure sustainability of any industry. This is particularly pertinent when considering the employability gap that is being seen today, where graduates can be deemed as non-employable due to skills gaps. As the proposed micro sourcing industry ecosystem is targeted at certain groups of micro workers especially the B40 group, skills training for these groups is a must since 52.3% of people in the B40 group have no education certificate and most of them have low to moderate competency in computer and Internet skills, as well as English proficiency [15].

Programs and activities to enhance capability of these targeted micro workers should involve organisations/ agencies that have direct contact with these groups. NGOs such as Yayasan Basmi Kemiskinan (YBK) and Yayasan Pembangunan Islam Malaysia (YAPEIM), and Government agencies such as Regional Development or Kementerian Pembangunan Luar Bandar dan Wilayah (KPLBW), Jabatan Kemajuan Masyarakat (KEMAS) and Pusat Zakat Selangor (PZS) should be roped in as they are well-verse with these targeted groups' environment. These organisations/agencies should be groom as Community Champions to engage these targeted groups into the micro sourcing industry ecosystem.

Funding for these programs and activities should come from the Government or sourced from private sectors' contributions as most these targeted micro workers cannot afford to pay for them.

2) Components

- Community Champions

Community Champions should be created within the micro sourcing ecosystem to engage participation of targeted micro workers from specific community. Platform owner should identify the Community Champion for a distribution centre and work together closely in engaging these micro workers. Candidates for Community Champions most preferably are organisations/ societies/ individuals that have established relations with the community. Platform firms will be the one to aggregate the tasks while the Community Champions will be responsible to maintain the infrastructure at the distribution centres.

- Micro Worker Development.

To support the sustainability of the Micro sourcing Ecosystem, targeted micro workers especially the B40 group need to be continuously trained. Among crucial training required are ICT-related courses, communication skills and personality development courses. Training, skills and

knowledge development must be on-going and therefore require considerable investment.

F. Strategic Thrust 5: Instruments to expediate Growth of Local Micro Sourcing Industry

An effective Micro sourcing Ecosystem requires proper governance. As with any established industry, proper monitoring or regulatory bodies are required to monitor fair play, enforce equitable standards, and prevent exploitation of stakeholders, especially micro workers. Government's policies and incentives are also important to kick start the micro sourcing industry development.

1) Rationale

The micro sourcing industry in Malaysia is still at its infancy stage. Therefore, lots of questions arise from potential role players from the four groups in the ecosystem – Job Providers, Micro Workers, Platforms Providers and Enablers – on the sustainability and security of the industry. Proper regulatory and monitoring frameworks to govern the industry will provide confidence to public and private sectors, organisations and individuals to participate in the industry. Proper regulatory and monitoring frameworks will also create a healthy ecosystem with proper marketing, enforcement, security and documented procedure.

The Government has important role to play to expedite industry's growth. Government's incentives are crucial as there will be lack of private investment at the initial stage. Potential local job providers need to be attracted, potential micro workers need to be trained, and potential platform developers need to be assisted to compete with international platforms. Certain existing incentives need to be reshaped to suit micro sourcing industry's requirement. Government's policies also need to be designed to support the industry's growth.

The micro sourcing industry is still lacking skilled/experienced talents across the ecosystem. There is a need to develop skilled and experienced talents as well as project management and business development teams to support micro sourcing platforms. Therefore, there is need to incentivise training and skilling of talents across the industry to ensure industry sustainability.

Engagement with professional organisations, particularly those in the field of micro sourcing and outsourcing advisory services is important to spread awareness and understanding of what micro sourcing industry can offers. These organisations can provide leadership and promote sharing of solution and best practices, as well as spur new business process management and innovation.

2) Components

• Governance

The micro sourcing industry ecosystem involves many stakeholders with their own agenda and objectives. As with any established industry, proper monitoring or regulatory bodies are required to supervise fair play, employ equitable standards, and prevent the exploitation of its players, particularly the employees.

A proper governance structure is required to monitor payment, conditions, and competition, as well as to uphold the objective of micro sourcing industry as a mean to alleviate disadvantaged from poverty.

• Policies

Competing effectively in the micro sourcing market place requires Government's role to create and introduce business friendly legal systems, policies and industry support mechanisms. For example, the Government could impose uniform restrictions on hiring foreign workers in an effort for private sectors to find ways in hiring local workers, including via micro sourcing. The Government could also provide leadership in the industry by outsourcing its non-core tasks and introducing micro sourcing-related tasks in its procurement policy. Other enabling policies could also be developed such as policies on micro sourcing self-regulation, telecentres opening hours and payment mechanism.

• Incentives

The Government can also promote micro sourcing industry through incentives. Domestic outsourcing can be encouraged by incentivising organisations that provide outsourced services. Government can foster investment in the industry by providing special incentives to firms who invest in programs and activities that benefit marginalised and disadvantaged households such as those to engage participation of the B40 group in the micro sourcing industry. Guideline on homeworking could be strengthened to encourage job providers to hire more micro worker especially the B40.

• Business Process Management.

New business process management that supports micro sourcing initiatives should be promoted. All relevant stakeholders such as Government agencies, private sectors, SMEs and NGOs need to be engaged and educated on micro sourcing and how to formulate a micro sourcing strategy. Micro sourcing pilot projects have to be monitored to ensure they have the necessary support and backing from all relevant stakeholders within the ecosystem.

• Education

All relevant stakeholders in the industry require continuous education and trainings. In addition, awareness programs are required from time to time to educate the general public about the industry. The industry requires high skilled workers to manage micro sourcing platforms and therefore, training/educational programs should be encouraged. Linkages between industry and institutes of higher learning (IHLs) are essential to ensure continuous supply of skilled talents into the industry.

For micro sourcing industry to grow in Malaysia, all relevant stakeholders will need to be aware of the micro sourcing business processes. The role of workplace is changing with technology development, combined with an increasingly cross-generational and distributed workforce, challenging traditional concepts of the workplace. Thus, the implementation of the Strategic Framework require strong governance and leadership, mandated actions where

appropriate, enforcement via spend controls, and monitoring and reporting.

V. CONCLUSION

As highlighted earlier, jobs for local micro sourcing industry would be sourced from local and international job providers. Contrary, awareness level of micro sourcing is still low among potential local job providers. In addition, potential job providers from the public sector have concerns on the local micro sourcing mechanism such as data confidentiality and payment mechanisms. These issues need to be addressed to get their confidence in the industry. As a pre-requisite to the local micro sourcing industry development, potential job providers in public and private sectors should have a thorough understanding, readiness, and awareness of the industry. Other than local job providers, efforts must also be made to capture jobs from international market through international platforms as well as international BPO service providers.

As in any new industry, support from the Government is very important in developing local micro sourcing industry. Each Strategic Thrust involves Government's participation in different capacity. Supports from other ministries, agencies and other divisions in public sector are nevertheless important in making this project and the micro sourcing industry in general a success. This project cuts across multiple sectors and thus, requires different types of Government support from different ministries and agencies. Other stakeholders also have to perform their roles in to ensure the effectiveness of the proposed Strategic Framework. With full support from these stakeholders, this industry would be able to achieve its true economic potential and this project would be able to help improve the socioeconomic status of the low income group in Malaysia.

References

- [1] Prahalad. C.K., "The fortune at the bottom of the pyramid: eradicating poverty through profit". Upper Saddle River, NJ: Wharton School Publishing, 2004.
- [2] Heeks, R., "Where next for ICTs and international development?. In ICTs for Development. Paris". Organisation for Economic Co-operation and Development (OECD) 2009, pp. 29-74.
- [3] EPU, "Economic Planning Unit, Tenth Malaysia Plan 2011-2015". Prime Minister's Department, Putrajaya, 2010.
- [4] Ling, R., Chan, H., Choon, L.S., Singh, D and Lim, Victor, "Special study: MSC Malaysia 2.0 state ICT blueprint: Negeri Sembilan", 2010
- [5] Dhar.S and Balakrishnan, B., "Risks, benefits and challenges in global it outsourcing: perspectives and practices", Journal of Global Information Management, Vol. 14, Issue 3, pp. 59-89., September 2006.
- [6] Kishore, R., H. R. Rao, K. Nam, S. Rajagopalan et al., " A relationship perspective on it outsourcing, "Communications of the ACM (46) 12, pp. 87-92. 2003
- [7] Dhar.S, "From outsourcing to cloud computing: evolution of it services, IEEE Int'l Technology Management Conference, Vol. 11, 2011, pp. 434-438"
- [8] Arshad, N.H., Yap,M.L., Mohamed, A., and Affandi, S., "Inherent risks in ICT outsourcing projects". Proceedings of the 8th WSEAS International Conference on Mathematics and Computers in Business and Economics, Vancouver, Canada, June, 2007, pp.141-146.
- [9] Arshad, N.H., Yap,M.L., and Mohamed, A., "ICT outsourcing: inherent risks, issues and challenges". WSEAS Transactions on Business and Economics, Iss 8, Vol. 4, 2007, pp.117-124.
- [10] Ang, S. and Straub, D.W., "Production and Transaction Economies and IS Outsourcing: A Study of the U.S. Banking Industry" MIS Quarterly, Volume 22, No. 4. 1998.
- [11] Dibbern, J., Goles, T., Hirschheim, R., & Jayatilaka., "Information systems outsourcing: A survey and analysis of the literature", The Data Base for Advances in Information Systems, 35(4), 2004, pp. 6-102.
- [12] Arshad, N.H., Hanapi, H., and Buniyamin, N., "IT outsourcing and knowledge transfer in Malaysia". Proceeding of 2nd International Congress on Engineering Education, Kuala Lumpur, December, 2010, pp.19-21.
- [13] Bahli, B. and Rivard S., "The information technology outsourcing risk: a transaction cost and agency theory-based perspective". Journal of Information Technology, vol. 18, no. 3, 2003, pp. 211-221.
- [14] King, W.R. & Malhotra, Y., Developing a framework for analyzing IS sourcing. Information & Management, 37, 2000, pp. 323-334.
- [15] Arshad, N. H., Salleh, S.S., Aris, S.R.S, Janom, N., Mastuki, N., "Micro sourcing: The SWOT analysis on the demand, supply and platform". Science and Information Conference (SAI), October, London, 2013. "in-press".
- [16] Kishore, R., H. R. Rao, K. Nam, S. Rajagopalan et al., " A relationship perspective on it outsourcing, "Communications of the ACM (46) 12, 2003, pp. 87-92.
- [17] Giddens, A., "The Constitution of Society: Outline of the Thoery of Structuration", University of California Press, Berkely and Los Angeles, 1984, pp.1-28.
- [18] Chiasson, M. and Saunders, C., "Reconciling diverse approaches to opportunity research using the structuration theory," Journal of Business Venturing (20), 2005, pp. 747-767.
- [19] Obal, L., "Microsourcing – "Using Information Technology to Create Unexpected Work Relationships and Entrepreneurial Opportunities". Communications of the Association for Information Systems. Vol (24), No 1., 2009.
- [20] Crowdsourcing Industry Report. "Enterprise Crowdsourcing: Market, Provider and Worker Trends", February 2012, Massolution. URL (accessed 20 September 2012), <http://www.massolution.com>
- [21] AmazonMechanicalTurk, (accessed 29 September 2012), <https://www.mturk.com/mturk/>.
- [22] CrowdFlower, (accessed 29 September 2012), <http://crowdfower.com/>.
- [23] SamaSource, (accessed 29 September 2012), <http://samasource.org/>.
- [24] Ushahidi, (accessed 29 September 2012), <http://www.ushahidi.com/>.
- [25] Micro sourcing, (accessed 29 September 2012), <http://www.microsourcing.com/>.
- [26] ODesk, (accessed 29 September 2012), <https://www.odesk.com/>.
- [27] Multimedia Development Corporation Sdn. Bhd. (MDeC), 2012, Digital Malaysia Lab Report.
- [28] Salleh, S.S., Arshad, N. H., Aris, S.R.S, Janom, N., Mastuki, N., "Formulating Cohesive Digital Ecosystem of Micro Sourcing Business Process in Malaysia". Science and Information Conference (SAI), October, London, 2013. "in-press".

A New Algorithm to Represent Texture Images

Silvia María Ojeda

Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina

Grisel Maribel Britos

Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina

Abstract—In recent times the spatial autoregressive models have been extensively used to represent images. In this paper we propose an algorithm to represent and reproduce texture images based on the estimation of spatial autoregressive processes. The image intensity is locally modeled by a first spatial autoregressive model with support in a strongly causal prediction region on the plane. A basic criteria to quantify similarity between two images is used to locally select this region among four different possibilities, corresponding to the four strongly causal regions on the plane. Two global image similarity measures are used to evaluate the performance of our proposal.

Keywords—Autoregressive Models; Texture Images; Similarity Measures.

I. INTRODUCTION

The goal of this work is to introduce a new algorithm to represent and reproduce texture images that uses and improves other recent proposals concerning this topic.

Most of the images of interest, for example, the images of cultivated fields and concentration of population are naturally rich in texture, level of gray, etc. The same thing happens to the images of geographical regions that allow the making of maps and, in general, almost all the images of the earth. During the past decades, image representation and image texture recovery have been two important and challenging topics. In this sense the spatial autoregressive model (AR-2D model) has been extensively used to represent images ([3], [14]) due to its two main properties. First, simulation experiments have shown that this model is adequate to represent a diversity of real scenarios ([4]). Second, the AR-2D model does not require a large number of parameters to represent different real scenarios (parsimony) ([4]). In particular, the first-order AR-2D model is able to represent a wide range of texture images, as is shown in Figure 1; the image (a) have been generated by a first-order AR-2D model with three parameters, while images (b), (c), (d) and (e) have been generated by a model of the same type with two parameters. Theoretical properties of the first-order AR-2D model were studied by Basu and Reinsel ([2]). They derived the correlation structure of the model and the maximum likelihood estimators of the parameters. Also, the spatial autoregressive models have benefited other topics in image processing like image segmentation. An approach to perform image segmentation based on the estimation of AR-2D processes has been recently suggested by Ojeda et al. ([15]).

First an image is modeled using a spatial autoregressive model for the image intensity. Then the residual autoregressive image is computed. This resulting image possesses interesting texture features. The borders and edges are highlighted, suggesting that the algorithm can be used for border detection. Later, a new scheme was proposed to enhance the segmentation yielded by the previous algorithm (Vallejos et al., 2012, [18]). It is based on the identifying of the best prediction window, and generalizes the previous algorithm to different prediction windows associated with unilateral processes on the plane. An analysis of the association between the original and fitted images shows how the algorithm works in practice. In all experiments carried out in [18], the first step of the segmentation algorithm was implemented using the same strongly causal prediction window. Consequently the support of the the local autoregressive models used to represent the images was always the same. Our proposal is a methodology that allows to identify locally the strongly causal prediction window (and consequently the support of the first-order AR-2D processes) associated with the better local representation of the image. To analyze the performance of our method, we quantified the similarity between the original and fitted images by two image measures. The study shows that the new method is capable to enhance the capacity of the AR-2D models to reproduce and represent images.

The rest of the paper is organized as follows: Section II presents a brief description of most used schemes of neighborhoods in \mathbb{Z}^2 . Section III provides an overview of the spatial ARMA models. Section IV presents the recent algorithm developed by Ojeda et al. ([15]) and the improvement due to Vallejos et al. ([18]) in the context of image segmentation. In Section V we explain our proposal to reproduce and represent texture images based on the estimation of spatial autoregressive processes. Section VI shows the results of our study and provides an analysis of the performance of our methodology using two similarity image measures. Conclusions and future scopes will appear in sections VII and VIII respectively.

II. NEIGHBORHOODS IN \mathbb{Z}^2

In time series, there is a natural neighbor structure induced by the existing total order of \mathbb{Z} (the set of all past values of $t \in \mathbb{Z}$; is the set of all integers that are less than t). However, for points on the plane, for instance $(m, n) \in \mathbb{Z}^2$, there are several different notions of neighborhood.

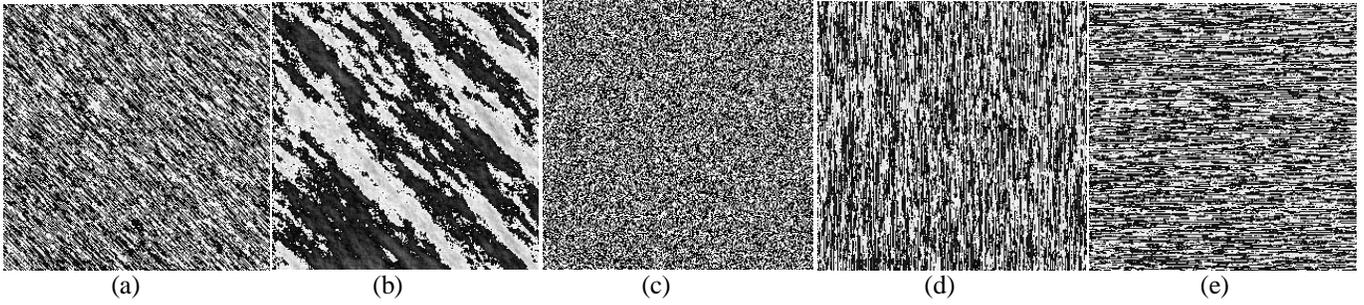


Fig. 1 Texture images generated by AR-2D process: (a) $\phi_1 = 0.0100, \phi_2 = 0.0100, \phi_3 = 0.8000$; (b) $\phi_1 = 0.5000, \phi_2 = 0.4999$; (c) $\phi_1 = 0.0100, \phi_2 = 0.0200$; (d) $\phi_1 = 0.0100, \phi_2 = 0.9000$; (e) $\phi_1 = 0.9000, \phi_2 = -0.0200$

In general, definitions of neighborhood of a point (m, n) on the plane are motivated by the physical acquisition system of the data like in the case of images that have been captured by satellites. A description of the most commonly used neighbor structures in statistical image processing can be found in [13]. In this paper we define the structure of neighborhood based on strongly causal regions.

For all $(m, n) \in \mathbb{Z}^2$ we distinguish the following strongly causal regions:

$$\begin{aligned} S_1(m, n) &= \{(k, l) \in \mathbb{Z}^2: k \leq m, l \leq n\} - \{(m, n)\} \\ S_2(m, n) &= \{(k, l) \in \mathbb{Z}^2: k \geq m, l \leq n\} - \{(m, n)\} \\ S_3(m, n) &= \{(k, l) \in \mathbb{Z}^2: k \geq m, l \geq n\} - \{(m, n)\} \\ S_4(m, n) &= \{(k, l) \in \mathbb{Z}^2: k \leq m, l \geq n\} - \{(m, n)\} \end{aligned}$$

Considering the region $S_1(m, n)$, a strongly causal prediction window containing two elements is

$$W_1 = \{(k, l) \in S_1(m, n): m-1 \leq k \leq m, n-1 \leq l \leq n\} - \{(m-1, n-1)\}$$

W_1 is shown in Figure 2 (a). Similarly, the strongly causal prediction windows W_2, W_3 , and W_4 , with two elements each, can be defined considering $S_2(m, n), S_3(m, n)$ and $S_4(m, n)$ respectively (Figure 2, (b)-(d)).

III. SPATIAL ARMA MODELS

Spatial ARMA processes have also been studied in the context of random fields indexed over $\mathbb{Z}^d, d \geq 2$, where \mathbb{Z}^d is endowed with the usual partial order; that is, for $s = (s_1, s_2, \dots, s_d), u = (u_1, u_2, \dots, u_d)$ in $\mathbb{Z}^d, s \leq u$ iff for $i = 1, 2, \dots, d, s_i \leq u_i$. Let $S[a, b] = \{x \in \mathbb{Z}^d: a \leq x \leq b\}$

and $S(a, b) = S[a, b] / \{a\}$, where $a, b \in \mathbb{Z}^d; a \leq b$ and $a \neq b$. A random field $(X_s)_{s \in \mathbb{Z}^d}$ is said to be a spatial ARMA (p, q) with parameters $p, q \in \mathbb{Z}^d$ if it is weakly stationary and satisfies the equation

$$X_s = \sum_{j \in S(0, p]} \phi_j X_{s-j} + \sum_{k \in S(0, q]} \vartheta_k \varepsilon_{s-k} + \varepsilon_s \quad (1)$$

where $(\phi_j)_{j \in S(0, p]}$ and $(\vartheta_k)_{k \in S(0, q]}$ denotes, respectively the autoregressive and moving average parameters with

$\phi_0 = \vartheta_0 = 1$; and $(\varepsilon_s)_{s \in \mathbb{Z}^d}$ denotes a sequence of independent and identically distributed centered random variables with variance σ^2 . Notice that if $p = 0$, the sum over $S(0, p]$ is supposed to be zero and the process is called spatial moving average MA(q) random field,

$$X_s = \sum_{k \in S(0, q]} \vartheta_k \varepsilon_{s-k} + \varepsilon_s \quad (2)$$

Similarly if $q = 0$ the process is called spatial autoregressive AR(p) random field, and it is defined as:

$$X_s = \sum_{j \in S(0, p]} \phi_j X_{s-j} + \varepsilon_s \quad (3)$$

The ARMA random field is called causal if it has the following unilateral representation

$$X_s = \sum_{j \in S[0, \infty]} \psi_j \varepsilon_{s-j} \quad (4)$$

with $\sum_j |\psi_j| < \infty$. Similarly to the time series case, there are conditions for the (AR or MA) polynomials to have stationarity and invertibility respectively. Let $\Phi(z) = 1 - \sum_{j \in S(0, p]} \phi_j z^j$ and $\Theta(z) = 1 - \sum_{j \in S(0, q]} \vartheta_j z^j$, where $z = (z_1, z_2, \dots, z_d)$ and $z^j = z_1^{j_1} z_2^{j_2} \dots z_d^{j_d}$. A sufficient condition for the random field to be causal is that the AR polynomial $\Phi(z)$ has no zeros in the closure of the open disc D^d in \mathbb{C}^d .

Applications of spatial ARMA processes and the study of spatial unilateral first order ARMA model have been developed in [10], [7], [2]. Other extensions of the theory developed for time series to spatial ARMA models can be found in [11], [6], [1], [17], [9], [5], [8]. As an example, consider a particular case of model (3) when $d = 2$ and $p = (1, 1)$. This model is called a first-order autoregressive process. In this case, $S((0, 0), (1, 1)) = \{(0, 1), (1, 1), (1, 0)\}$ and the model is of the form:

$$X_{(i, j)} = \phi_{(1, 0)} X_{(i-1, j)} + \phi_{(1, 1)} X_{(i-1, j-1)} + \phi_{(0, 1)} X_{(i, j-1)} + \varepsilon_{(i, j)}$$

Note that $\phi_{(1, 1)} = 0$, implies that for all $(i, j) \in S_1(i, j)$, $W_1(i, j)$ is the strongly causal prediction window of the intensity $X_{(i, j)}$.

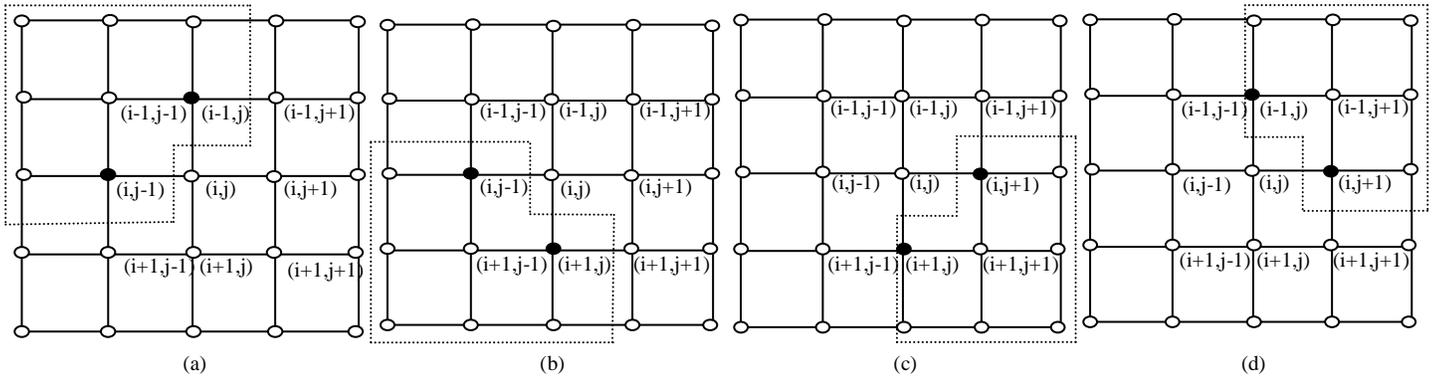


Fig. 2 Prediction windows for the first order spatial AR model with two parameters: (a) W_1 , (b) W_2 , (c) W_3 and (d) W_4 .

IV. APPROXIMATION OF IMAGES BY USING 2D UNILATERAL AR PROCESSES

In this section, we describe an algorithm to approximate an image by first-order AR-2D processes with two parameters using blocks. This algorithm was originally defined by [14] and later adapted by [15]. Following, we briefly describe this methodology.

Suppose that a real image is available. This fitted image is constructed by dividing the original image into squared sub-images (e.g., 5×5) and then fitting a first-order AR 2D model to each sub-image (i.e., block). Then, a sub-image is generated from each local fitted model and the final fitted image is yielded by putting together all generated sub-images. Let $Z = Z_{(m,n)}$, $1 \leq m \leq M$, $1 \leq n \leq N$ be the original image, and let $X = X_{(m,n)}$, $1 \leq m \leq M$, $1 \leq n \leq N$ where, for all $1 \leq m \leq M$, $1 \leq n \leq N$, $X_{(m,n)} = Z_{(m,n)} - \bar{Z}$; and \bar{Z} is the mean of Z . Let $4 \leq k \leq \min(M, N)$ (e.g. $k = 5$). For simplicity we shall consider from now on that the images to be processed (Z and X) are arranged in such a way that the number of columns and the number of rows are multiples of k ; that is,

$$Z = (Z_{(m,n)}); 1 \leq m \leq M'; 1 \leq n \leq N';$$

$$X = (X_{(m,n)}); 1 \leq m \leq M'; 1 \leq n \leq N';$$

Where $M' = \lceil \frac{M}{k} \rceil k$ and $N' = \lceil \frac{N}{k} \rceil k$. Considering the values

$M'' = \lceil \frac{M'-1}{k-1} \rceil$ and $N'' = \lceil \frac{N'-1}{k-1} \rceil$. For all $i_b = 1, \dots, M''$ and $j_b = 1, \dots, N''$, define the $k \times k$ block (i_b, j_b) of the image X by

$$B_X(i_b, j_b) = (X_{(r,s)}),$$

Where r and s are sub-index that satisfy:

$$(k-1)(i_b-1) + 1 \leq r \leq (k-1)i_b + 1,$$

$$(k-1)(j_b-1) + 1 \leq s \leq (k-1)j_b + 1.$$

The $M'' \times N''$ approximated image \hat{X} of X is provided by the following algorithm.

Algorithm 1:

1) For each block $B_X(i_b, j_b)$ compute estimators of least squared $\hat{\phi}_1^1$, $\hat{\phi}_2^1$ of ϕ_1 and ϕ_2 corresponding to the block $B_X(i_b, j_b)$ in the strongly causal prediction region S_1 , using the prediction windows W_1 .

2) Let $\hat{X}^1(i_b, j_b)$ be defined in the block $B_X(i_b, j_b)$ by

$$\hat{X}^1(i_b, j_b)_{(r+1,s+1)} = \hat{\phi}_1^1 B_X(i_b, j_b)_{(r+1,s)} + \hat{\phi}_2^1 B_X(i_b, j_b)_{(r,s+1)}$$

when $r = 1, \dots, (k-1)$ and $s = 1, \dots, (k-1)$.

3) Let \hat{X}^1 be defined in the block $B_X(i_b, j_b)$ by

$$\hat{X}^1_{((i_b-1)(k-1)+r, (j_b-1)(k-1)+s)} = \hat{X}^1(i_b, j_b)_{(r,s)}$$

with $r = 1, \dots, k$ and $s = 1, \dots, k$.

4) Define the approximated image \hat{Z}^1 of the original image Z as:

$$\hat{Z}^1_{(m,n)} = \hat{X}^1_{(m,n)} + \bar{Z}$$

with $1 \leq m \leq M'$ and $1 \leq n \leq N'$.

In order to propose a more efficient algorithm, Vallejos et al. ([18]) suggested new variants of this algorithm specially to address the problem of determining the most convenient (in terms of the quality of the segmentation) prediction window of unilateral AR-2D processes. In effect, they generalize the Algorithm 1 to different prediction windows associated with unilateral processes on the plane. Three variants of the Algorithm 1 (called **Algorithm 2**, **Algorithm 3** and **Algorithm 4**) were implemented in the strongly causal prediction regions S_2 , S_3 and S_4 , using the prediction windows W_2 , W_3 and W_4 , respectively. In each block $B_X(i_b, j_b)$, and for $t = 1, 2, 3, 4$, we denote the output corresponding to step 2-Algorithm t , as $\hat{X}^t(i_b, j_b)$. Similarly, \hat{X}^t denote the output corresponding to step 3 - Algorithm t . The computation of the distance between each filtered image and the original was done by using Q and CQ image quality measures ([18], [19], [16]). We described briefly these measures:

Let two weakly stationary processes, $(X_s)_{s \in D}$ and $(Y_s)_{s \in D}$, $D \subset \mathbb{Z}^d$, the index Q ([19]) is

$$Q = \frac{4S_{XY}\bar{X}\bar{Y}}{(S_X^2 + S_Y^2)[\bar{X}^2 + \bar{Y}^2]} = \frac{S_{XY}}{S_X S_Y} \frac{2\bar{X}\bar{Y}}{[\bar{X}^2 + \bar{Y}^2]} \frac{2S_X S_Y}{(S_X^2 + S_Y^2)}$$

$$= CMV$$

where X is the mean of $(X_s)_{s \in D}$, S_X is the standard deviation of $(X_s)_{s \in D}$, and S_{XY} is the covariance between $(X_s)_{s \in D}$ and $(Y_s)_{s \in D}$ (and similarly for Y and S_Y). The quantity $C = \frac{S_{XY}}{S_X S_Y}$ models the linear correlation between $(X_s)_{s \in D}$ and $(Y_s)_{s \in D}$, $M = \frac{2\bar{X}\bar{Y}}{[\bar{X}^2 + \bar{Y}^2]}$ measures the similarity between the sample means (luminance) of $(X_s)_{s \in D}$ and $(Y_s)_{s \in D}$, and $V = \frac{2S_X S_Y}{(S_X^2 + S_Y^2)}$ measures the similarity related to the contrast

between the images. Coefficient Q is defined as a function of the correlation coefficient; hence, it is able to capture only the linear association between $(X_s)_{s \in D}$ and $(Y_s)_{s \in D}$. The CQ index was suggested by [16] and it is defined as:

$$CQ(h) = \hat{\rho}(h)MV$$

where $\hat{\rho}(h)$ is the sample codispersion coefficient in the direction h . This index, can quantify the similarity between images that are generated using a local approximation of AR-2D processes with different window sizes. Moreover, CQ captures different levels of spatial similarity between two images by considering different directions in two-dimensional space. Note that values of Q and CQ close to 0 point out a low similarity level between two images; instead, values of $|Q|$ or $|CQ|$ close to 1, indicate a high similarity level (direct or inverse, respectively).

V. A NEW ALGORITHM TO REPRESENT TEXTURE IMAGES

In order to represent and reproduce texture images, we suggest a methodology that we have called Algorithm 5. As Algorithms 1-4, our proposal is also based on the idea that is advantageous approximate an image by first-order AR-2D processes using blocks, just that we introduce the possibility to select in each block the causal prediction window of the model. The goal is to get \hat{Z}^* , a new and better representation of the original image Z , that improves the results achieved by the algorithms 1-4. Essentially in each block, our approach selects one among the algorithms 1, 2, 3 or 4, and uses it to approximate the image in the block. Previously to defined \hat{Z}^* , a $M' \times N'$ approximated image \hat{X}^* of X is provided by the algorithm. This selection is based on the mean square error (MSE). This basic similarity measure for images is computed from the difference of intensity, of pixel to pixel, between two images. More formally, if $(X_s)_{s \in D}$ and $(Y_s)_{s \in D}$, with $D \subset \mathbb{Z}^2$, are two weakly stationary processes,

$$MSE(X, Y) = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (X_{(i,j)} - Y_{(i,j)})^2$$

More details about the MSE can be consult in [19]. In the following we describe the new algorithm.

Algorithm 5:

1) In each block $B_X(i_b, j_b)$, and for $t = 1, 2, 3, 4$, compute the least squares estimators $\hat{\phi}_1^t$; $\hat{\phi}_2^t$ of the ϕ_1 and ϕ_2 corresponding to the block $B_X(i_b, j_b)$, in the strongly causal prediction region S_t , using the prediction windows W_t (Apply step 1-Algorithm t , with $t = 1, 2, 3, 4$).

2) In each block $B_X(i_b, j_b)$, and for $t = 1, 2, 3, 4$, compute $\hat{X}^t(i_b, j_b)$ (Apply step 2-Algorithm t , with $t = 1, 2, 3, 4$).

3) In each block $B_X(i_b, j_b)$, and for $t = 1, 2, 3, 4$, compute $MSE(\hat{X}^t(i_b, j_b), B_X(i_b, j_b))$. Choose the approximated image

generated by the lowest mean square error. Denote it as $X^*(i_b, j_b)$.

4) Let \hat{X}^* be defined in the block $B_X(i_b, j_b)$ by

$$\hat{X}_{((i_b-1)(k-1)+r, (j_b-1)(k-1)+s)}^* = X^*(i_b, j_b)_{(r,s)}$$

with $r = 1, \dots, k$ and $s = 1, \dots, k$.

5) The approximated image \hat{Z}^* of the original image Z is: $\hat{Z}_{(m,n)}^* = \hat{X}_{(m,n)}^* + \bar{Z}$ with $m = 1, \dots, M' = \lfloor \frac{M}{k} \rfloor k$ and $n = 1, \dots, N' = \lfloor \frac{N}{k} \rfloor k$.

Note how Algorithm 5 (step 3), allows to locally identify (based on the MSE) the strongly causal prediction window associated with the more adequate local representation of the original image.

VI. EXPERIMENTS AND RESULTS

In this section we developed some examples in different scenarios to explore the performance of Algorithm 5. Six real images were used, Elaine, Lenna, Peppers, Threads, Rind and Aerial (Figure 3 (a)- (f)), all taken from the USC-SIPI image database [12]. We applied the Algorithms 1 to 5, for each original image, considering a block of size 5×5 , and we obtained five representations of the original image. Figure 4 shows the image representations obtained by the five methods, from the original image Elaine.

To gain insight on the quality of each image representation, the images produced by the five algorithms were compared with the original image using the similarity index Q ([19]) and CQ ([16]), described in section IV. The results are shown in TABLE I. In all cases the highest values of the image quality measures were obtained for the image representation produced by the Algorithm 5. In practice, this means that the residual image (difference between the original and the approximated) is more compatible with a null image, when the Algorithm 5 is used.

Visually, in the residual images produced from the Algorithm 5, it is more difficult to detect the patterns of the respective original images, in comparison with the residual images generated from the others methods. As an example, Figure 5 shows the residual images obtained by applying the five algorithms to the original image Elaine, and Figure 6 shows the histograms of these images. Note that the residual image produced from Algorithm 5 (Figure 5 (e)) does not highlight the original borders and boundaries because the original image and the image representation are too similar. The histogram of the image (Figure 6 (e)) confirms this fact.

The results presented in this section are not restrictive to the images treated in this paper. There is a large set of images for which the experiments developed in this article can be replicated.

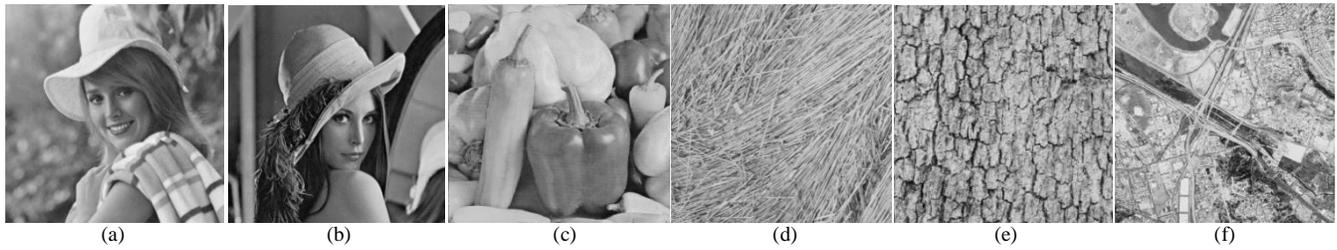


Fig. 3 Images of USC-SIPI image database. (a) Elaine, (b) Lenna, (c) Peppers, (d) Threads, (e) Rind and (f) Aerial.

TABLE I. PERFORMANCE OF THE ALGORITHMS 1, 2, 3, 4, AND 5 BY Q AND CQ INDEX.

Met.	(a) Elaine		(b) Lenna		(c) Peppers		(d) Threads		(e) Rind		(f) Aerial	
	Q	CQ	Q	CQ	Q	CQ	Q	CQ	Q	CQ	Q	CQ
1	0.9769	0.5115	0.9345	0.6372	0.9895	0.8071	0.8456	0.7321	0.9155	0.6885	0.9186	0.6247
2	0.9784	0.5298	0.9529	0.6874	0.9880	0.8220	0.8804	0.7905	0.9196	0.7425	0.9154	0.6737
3	0.9780	0.5126	0.9628	0.6556	0.9877	0.8066	0.8419	0.7284	0.9193	0.6908	0.9110	0.6214
4	0.9778	0.5305	0.9472	0.6827	0.9898	0.8223	0.8782	0.7879	0.9219	0.7451	0.9059	0.6670
5	0.9878	0.7520	0.9735	0.8119	0.9947	0.8871	0.9343	0.8844	0.9579	0.8379	0.9589	0.8169



Fig. 4 (a)-(e), Image representations generated by Algorithms 1-5 respectively. Original image: Elaine.

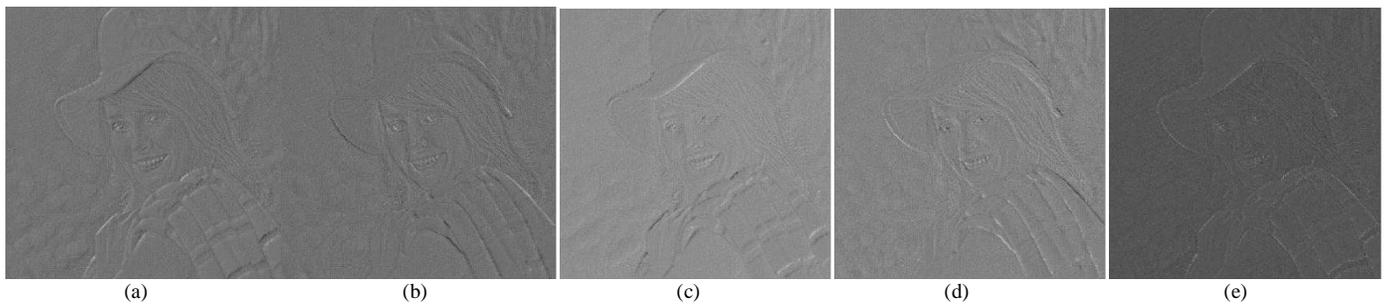


Fig. 5 (a)-(e), Residual images (difference between the original and image representations) generated from Algorithms 1-5 respectively. Original image: Elaine.

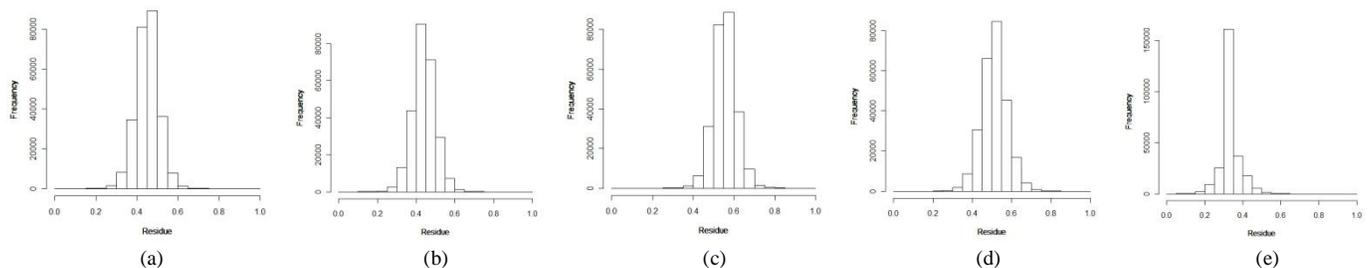


Fig. 6 a)-(e) Image residual histograms produced from Algorithms 1-5 respectively. Original image: Elaine.

VII. CONCLUSION

This paper proposes a new algorithm to represent and reproduce texture images based on the estimation of spatial autoregressive processes. Our proposal, as other methods, suggests approximating an image by first-order AR-2D processes using blocks, just incorporating the option to select in each block the causal prediction window of the model. This selection is based on the mean square error, a quantitative performance metric in the field of signal processing ([20]). The method, called Algorithm 5, can generate a wide range of textures using four different local approximations of AR-2D processes, corresponding to the four strongly causal regions on the plane. Using two image quality index that are extensively used in image similarity analysis, we carried out experiments that support our algorithm ([19], [20], [16]). Specifically, a set of images belonging to the image database ([12]) were processed and provided satisfactory results in order to reproduce and represent image textures. In addition the paper performs a review about the main characteristics and applications of the spatial autoregressive and moving average models. In the light of the examples presented in this article, we suggest in practice to use the Algorithm 5 as a replacement of other similar methods based on only one strongly causal region on the plane, to produce a more adequate representation of the texture images.

VIII. FUTURE RESEARCH

The performance of the algorithm under different kinds of contamination in images is an interesting open problem to be addressed in the future. The effect of considering noncausal and semi-causal prediction windows in the spatial AR model, with different window sizes, is also an open issue to be addressed in future research.

ACKNOWLEDGMENT

S. Ojeda and G. Britos thank SeCyT-UNC and CIEM-FAMAF for financial assistance.

References

[1] S. Baran, G. Pap, and M. C. A. Zuijlen, "Asymptotic inference for a nearly unstable sequence of stationary spatial AR models", *Statistics & Probability Letters*, Vol. 69(1), pp. 53-61, 2004.

[2] S. Basu and G. Reinsel, "Properties of the spatial unilateral first-order ARMA model", *Advances in Applied Probability*, Vol. 25(3), pp. 631-648, 1993.

[3] J. Bennet and A. Khotanzad, "Maximum likelihood estimation methods for multispectral random field image models" *IEEE Transaction Pattern Analysis and Machine Intelligence* Vol. 21, pp. 537-543, 1999.

[4] O. Bustos, S. Ojeda and R. Vallejos, "Spatial ARMA models and its applications to image filtering", *Brazilian Journal of Probability and Statistics*, Vol. 23 (2), pp. 141-165, 2009.

[5] O. Bustos, S. Ojeda, M. Ruiz, R. Vallejos and A. Frery, "Asymptotic Behavior of RA-estimates in Autoregressive 2D Gaussian Processes", *Journal of Statistical Planning and Inference*, Vol. 139(10), pp. 3649-3664, 2009.

[6] B. Choi, "On the asymptotic distribution of mean, autocovariance, autocorrelation, crosscovariance and impulse response estimators of a stationary multidimensional random field", *Communications in Statistics- Theory and Methods*, Vol. 29(8), pp. 1703-1724, 2000.

[7] B. R. Cullis and A. C. Glesson, "Spatial analysis of field experiments an extension to two dimensions", *Biometrics*, Vol. 47(4), pp. 1449-1460, 1991.

[8] C. Gaetan, and X. Guyon, "Spatial Statistics and Modelling", Springer, New York, 2010.

[9] M. G. Genton and H. L. Koul, "Minimum distance inference in unilateral autoregressive lattice processes" *Statistica Sinica*, Vol. 18, pp. 617-631, 2008.

[10] M. R. Grondona, J. Crossa, P. N. Fox and W. H. Pfeiffer, "Analysis of variety yield trials using two-dimensional separable ARIMA processes", *Biometrics*, Vol. 52(2), pp. 763-770, 1996.

[11] J. Guo and L. Billard, "Some inference results for causal autoregressive processes on a plane", *Journal of Time Series Analysis*, Vol. 19(6), pp. 681-691, 1998.

[12] Image database, Signal and Image Processing Institute, University of Southern California, <http://sipi.usc.edu/database/>

[13] A. K. Jain, "Fundamentals of Digital Image Processing", Prentice Hall.

[14] R. Kashyap and K. Eom, "Robust images techniques with an image restoration application", *IEEE Trans. Acoust. Speech Signal Process*, Vol. 36(8), pp. 1313-1325, 1988.

[15] S. M. Ojeda, R. Vallejos and O. Bustos, "A New Image Segmentation Algorithm with Applications to Image Inpainting", *Computational Statistics & Data Analysis*, Vol. 54(9), pp. 2082-2093, 2010.

[16] S. M. Ojeda, R. Vallejos and W. P. Lamberti, "Measure of Similarity Between Images Based on the Codispersion Coefficient", *Journal of Electronic Imaging*, Vol. 21, 2012.

[17] R. Vallejos and G. Garcia-Donato, "Bayesian analysis of contaminated quarter plane moving average models", *Journal of Statistical Computation and Simulation*, Vol. 76 (2), pp. 131-147, 2006.

[18] Ronny Vallejos and Silvia Ojeda, Chapter in book: "Segmentation of Images and Time Series Based on Spatial ARMA Processes", *Advances in Image Segmentation*, October 24, 2012, ISBN 980-953-307-581-0.

[19] Z. Wang and A. Bovik, "A universal image quality index", *IEEE Signal Processing Letters*, Vol. 9(3), pp. 81-84, 2002.

[20] Z. Wang and A. Bovik, "Modern Image Quality Assessment", Morgan & Claypool Publishers, United States of America, 2006.

Image and Video based double watermark extraction spread spectrum watermarking in low variance region

Mriganka Gogoi

Electronics and Communication
Assam Don Bosco University
Guwahati,India

H.M.Khalid Raihan Bhuyan

Electronics and Communication
Assam Don Bosco University
Guwahati,India

Koushik Mahanta

Electronics and Communication
Assam Don Bosco University
Guwahati,India

Dibya Jyoti Das

Electronics and Communication
Assam Don Bosco University
Guwahati,India

Ankita Dutta

Electronics and Communication
Assam Don Bosco University
Guwahati,India

Abstract— Digital watermarking plays a very important role in copyright protection. It is one of the techniques which are used for safeguarding the origins of the image, audio and video by protecting it against Piracy. This paper proposes a low variance based spread spectrum watermarking for image and video in which the watermark is obtained twice in the receiver. The watermark to be added is a binary image of comparatively smaller size than the Cover Image. Cover Image is divided into number of 8x8 blocks and transform into frequency domain using Discrete Cosine Transform. A gold sequence is added as well as subtracted in each block for each watermark bit. In most cases, researchers has generally used algorithms for extracting single watermark and also it is seen that finding the location of the distorted bit of the watermark due to some attacks is one of the most challenging task. However, in this paper the same watermark is embedded as well as extracted twice with gold code without much distortion of the image and comparing these two watermarks will help in finding the distorted bit. Another feature is that as this algorithm is based on embedding of watermark in low variance region, therefore proper extraction of the watermark is obtained at a lesser modulating factor. The proposed algorithm is very much useful in applications like real-time broad casting, image and video authentication and secure camera system. The experimental results show that the watermarking technique is robust against various attacks.

Keywords—Watermark;Gold Code; Variance; Correlation.

I. INTRODUCTION

Digital watermarking plays a very important role in multimedia transmission. Consequently, digital watermark technique needs to be incorporated in digital rights to address different aspects of the content supervision. Due to the availability of digital equipments and rapid development of internet, access to digital information has become very easier. As a result protection of copyright and intellectual property of

the media has become very much essential. Digital watermarking came as an efficient solution of the above mention. The watermarking embeds information into the host data in some invisible as well visible way that is supposed to identify the owner .Watermarking schemes can be classified into various ways: There are basically two types of watermarking scheme, spatial domain and frequency domain according to working field. Watermarking in transform domain provides more robustness to the watermarking process. In transform domain, the effect of noise and distortion of original Image during the watermarking process is less. The frequency domain schemes are generally considered more robust than the spatial domain schemes and are based on DCT (discrete cosine transform) and DWT (discrete wavelet transform), Fast Fourier Transform (FFT) in general. In most of transform domain watermarking techniques in which watermark is embedded once and the same is extracted in the receiver. However, if the watermarked image or video undergoes any attacks in the channel then a distorted watermark is obtained in the receiver. But, finding the error of a particular bit in the watermark is quite challenging when the original watermark is not present in the receiver. In this proposed method a single watermark is embedded twice with less distortion in the low variance region of the image. Also, the same watermark is obtained twice, comparing both the watermarks gives the error occurred.

II. RELATED WORK

Previously, some researchers have proposed various papers and implemented different techniques for the watermark algorithm. Ingemar J. Cox et al[1] proposed an algorithm to insert a watermark into the spectral components of the data using techniques analogous to spread spectrum communications, hiding a narrow band signal in a wideband channel that is the data. Mercy George et al[2] proposed the

direct sequence spread spectrum technique of watermarking for video and images which can be used for both special and spectral domain watermarking. T. Kohda et al [3] proposed a color image communication system through code division multiple access (CDMA) channels with spreading sequences of variable-period to transmit YIQ signals. In this paper, they proposed a digital watermarking system where YIQ signal channels and embedded watermark image channel are separately used. R. Bangaleea et al [4] proposed a spatial domain-watermarking scheme for data hiding and copyright protection of still images using the attack characterisation approach. Here they have added two watermarks and used two keys for the watermarking system. The use of the key introduce uncertainty about the location of the watermark bits in spatial domain. Fan Zhang et al[5] proposed Digital Image Watermarking algorithm Based on CDMA Spread Spectrum where orthogonal gold code is used to spread spectrum of the copyright messages. The copyright messages can be extracted without the original image as it is a blind watermarking algorithm. Bo Chen et al [6] proposed a new robust-fragile double image watermarking algorithm. They have used the improved pixel-wise masking model and pseudo-random sequence to embed robust watermark and fragile watermark into the insensitive (robust) part and sensitive part of the wavelet coefficients of the host image. This makes the two watermarks non-interfering and increases the watermarking capacity of the host image without reducing watermark robustness and also the PSNR value is found to be high compared to the single watermarking scheme while considering the same robustness. Xiu-mei Wen et al[7] proposed a digital image watermarking algorithm based on Fourier domain. The fourier transform can embed watermark not only in the amplitude of transform domain but also in the phase degree of transform domain if it is satisfied with the real numbers inverse transform. In this paper, the algorithm embeds the watermark in the amplitude of transform domain.

HO ATS et al[8] has proposed a robust digital image in image watermarking using Fast Hadamard Transform and he pointed out some of the advantages in terms of shorter processing time. Kutter et al[9] embedded the watermark by modifying a selected set of pixels in the blue channel because human eyes are less sensitive to changes in this color channel. The blue channel, however, is the frailest to JPEG compression among the three color channels, so the hidden watermark information is easy to be lost.

III. WATERMARKING ALGORITHM

The watermark algorithm is viewed as a sequence of stages. The block diagram for transmitter and receiver is shown in the transmitter and in the receiver part.

A. Transmitter part

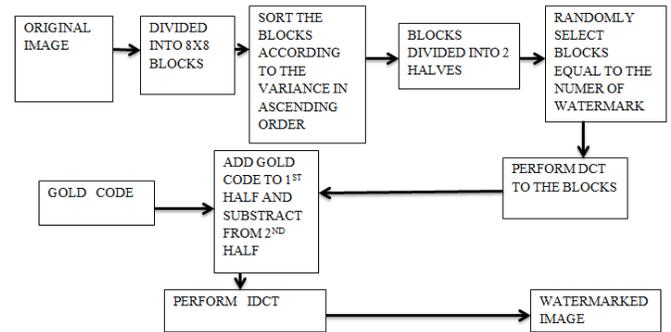


Fig. 1 Watermark embedding block

In the transmitter part, the image is taken from pc and the image is divided into (8x8) blocks. Each block will contain 64 pixel values. The blocks are sorted according to the variance in ascending order. The low variance blocks are then taken because they can create less distortion in the image and so the watermark can easily be extracted by using the algorithm known to sender and the receiver. The equation of variance is given by—

$$\text{Var}(x) = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{\sum x_i^2}{N} \quad (1)$$

Where ‘x’ is a 1D matrix, ‘x̄’ is the mean and ‘N’ is the length of the matrix.

After sorting the blocks according to the variance, the blocks are divided into two halves and half of them are considered having low variance and from those blocks some blocks are randomly selected and with them orthogonal gold codes are added whose number of bits is equal to the randomly selected blocks. Now DCT is performed to the selected blocks. Again the randomly selected blocks are divided into two halves by considering each having 31 AC co-efficients by ignoring the DC co-efficient and the last AC co-efficient. Now by performing XOR operation between two p-n sequences we get a gold code which is added to one half and subtracted from the other half. Now inverse discrete cosine transform of the blocks gives the watermarked image. The equation for watermarking is-

CASE 1-

$$X_w = X_i + m_k g_m \text{ if watermark bit is '1'}. \quad (2)$$

$$X_w = X_i - m_k g_m \text{ if watermark bit is '0'}.$$

CASE 2-

$$X_w = X_i - m_k g_m \text{ if watermark bit is '1'}. \quad (3)$$

$$X_w = X_i + m_k g_m \text{ if watermark bit is '0'}.$$

Where X_w is watermarked image, X_i is the original image, m_k is the modulation factor and g_m is the orthogonal gold code.

B. Receiver Part

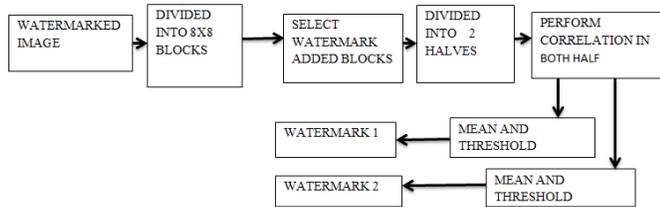


Fig. 2 Watermark extraction block.

In the receiver section, from the watermarked image, blocks in which watermark is added are considered and DCT is performed. Like in the transmitter each block is again divided into two halves eliminating the DC component and last AC co-efficient. Now correlation is obtained between each half and the corresponding gold code added in the transmitter. Two means are obtained from the two sets of correlation values. For first set If the co-relation value is greater than the mean ,watermark bit is ‘1’ else it is ‘0’ and for the second set If the co-relation value is greater than the mean, watermark bit is ‘0’ else it is ‘1’. The equation for correlation is given as-

$$C_r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (4)$$

Finally, two similar watermarks are obtained in the receiver. The advantage of getting two watermarks is that we can get the information hidden from any one of the watermark even if due to various attacks one watermark is lost. Again, in most of the cases, the researchers generally considered all the blocks of the image in the transform domain. But in this paper DCT is performed only to the blocks where watermark is to be added.

C. DCT Watermarking

A DCT expresses finite number of data points in terms of sum of cosine functions oscillating at different frequencies. (DCT)-based digital image watermarking which is used for image authentication and copyright protection is the transpose method. In this method the 2 Dimensional DCT is obtained by taking two 1- dimensional DCTs in series. The image pixel value is first divided into 8x8 blocks and the row-wise 1D DCT of each block is taken. The transpose of the blocks is then determined and a column-wise 1D DCT is ascertained which gives the 2D DCT of the data. Two dimensional discrete cosine transform (2D-DCT) is defined as—

$$F(j, k) = a(j)a(k) \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} f(m, n) \cos\left[\frac{(2m+1)j\pi}{2N}\right] \cos\left[\frac{(2n+1)k\pi}{2N}\right] \quad (5)$$

$$f(m, n) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} a(j)a(k) F(j, k) \cos\left[\frac{(2m+1)j\pi}{2N}\right] \cos\left[\frac{(2n+1)k\pi}{2N}\right] \quad (6)$$

IV. EXPERIMENTS AND RESULTS

In this section, some experimental results for a low variance DCT based watermarking algorithm are shown. The algorithm is implemented in Matlab version 7.8.0.347(R2009a) with the standard [256x256] cameraman.tif image and rhinos.avi video.

The watermark is a [8x8] binary image for Image watermarking and three [8x8] binary image for video watermarking. Fig.3 and Fig.4 Represents the original image and the watermarked image. Fig.5 represents the added watermark in different portion of the image.

The various attacks performed are Gaussian filter, Salt and pepper noise, Median filter, Histogram equalization and the results are shown in Fig.8, Fig.9, Fig.10, Fig.11.



Fig. 3 Original image



Fig. 4 Watermarked image



Fig. 5 Added watermark

The PSNR value of the watermarked image with different values of modulation factor (m_k) is shown in TABLE I .

TABLE I.

Modulation factor(m _k)	PSNR value
0.5	62.5257
0.7	66.5024

Fig. 6 PSNR values for different modulation factor.

Also, PSNR values of the two extracted watermarks after different attacks are shown in TABLE II.

TABLE II.

Watermarking Attacks	WATERMARKS AND PSNR VALUES			
	Watermark image(1)	Watermark image(2)	PSNR1	PSNR2
Gaussian filter			24.0824	26.5812
Salt & pepper noise			30.1030	36.1236
Median filter			30.1236	36.1258
Histogram equalization			20.5821	26.3504

Fig. 7 Extracted watermarks and PSNR values.

The experimental results shows that the PSNR values of the extracted watermarks are very much acceptable. The various attacks performed are Gaussian filter, Salt and pepper noise, Median filter, Histogram equalization and the results are shown in Fig.8, Fig.9, Fig.10, Fig.11.

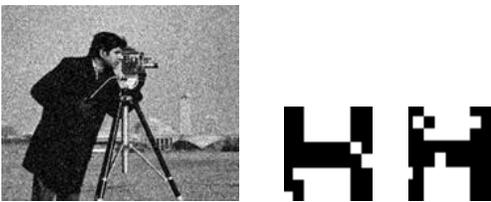


Fig. 8 Gaussian noise and extracted watermark

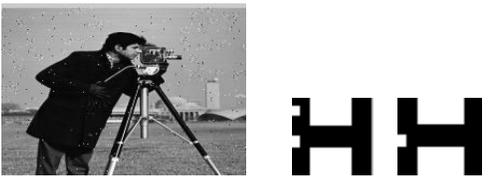


Fig. 9 Salt and Pepper noise and extracted watermark



Fig. 10 Median filter and extracted watermark



Fig. 11 Histogram equalization and extracted watermark

The algorithm is also performed with [256 x 256] colored image as well as with various frames of a video. The original image, watermarked image and extracted watermark are shown in Fig. 12(a), Fig. 12(b) and Fig. 12(c) for images and Fig. 13 is for some of original frames of a video, Fig. 14 shows the watermarks, Fig. 15 shows watermarked frames along with PSNR values and Fig. 16 represents extracted watermarks of 10 frames.



Fig. 12 a)Original image, (b) Watermarked image, modulation factor(m_k)=0.5, (c) Added watermark, (d) Extracted watermarks, (e) Watermarked image, modulation factor(m_k)= 0.7.

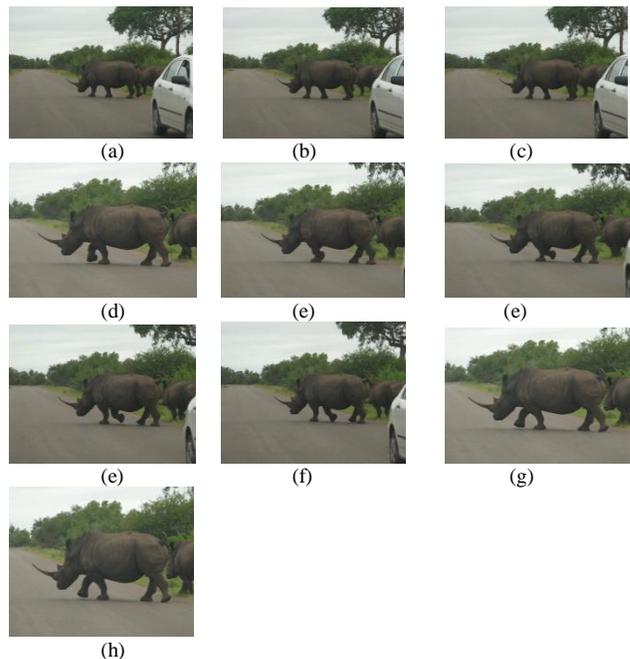


Fig. 13 Original Frames.

E C E

Fig. 14 Watermarks.

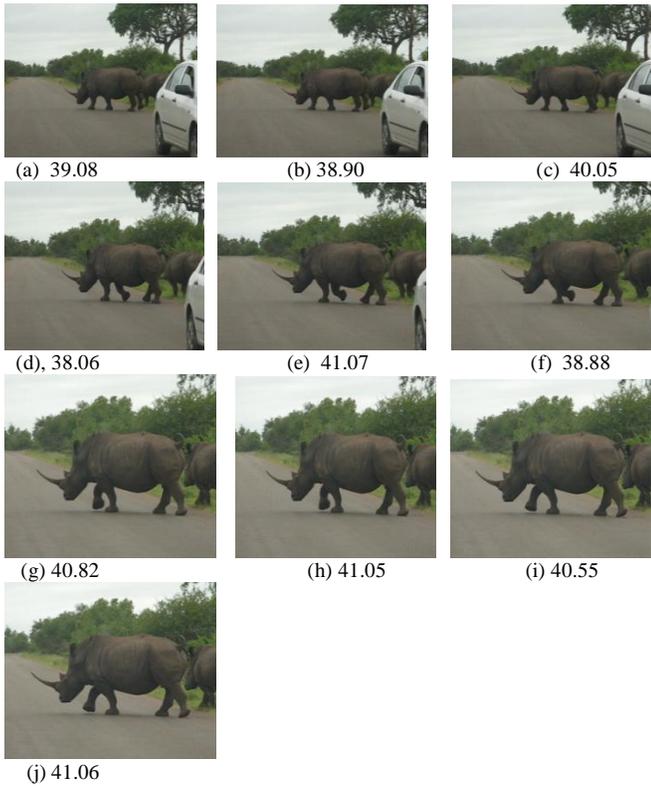


Fig. 15 Watermarked Frames, PSNR values.

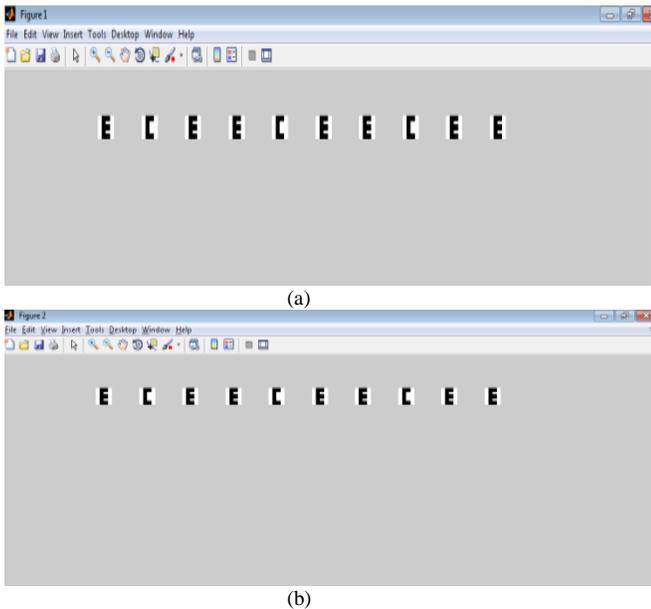


Fig. 16 Double Extracted Watermarks.

V. CONCLUSION

In this paper, the watermark algorithm is implemented based on low variance of various blocks of image and video frames. For the same watermark, it is embedded twice with gold codes, within the same randomly selected blocks equivalent to the number of bits in the watermark. Here there is no need of converting all the blocks to transform domain as a result time required for transformation of the blocks is reduced.

Also the watermark algorithm extracts two watermarks in the receiver; therefore even if the information is lost in one watermark, we can retrieve the information from the other one. As compared to other watermarking techniques it helps in finding the location of the error bit of the watermark without the help of original watermark. But, the size of the watermark must be comparatively smaller than the original image.

References

- [1] Cox IJ, Kilian J, Leighton T, Shamoon T. Secure spread spectrum watermarking for multimedia. *IEEE Trans Image Process* 1997;6(12):1673–87.
- [2] George, M., Chouinard, J.V., Georganas, N., 1999. Digital watermarking of images and video using direct sequence spread spectrum techniques. *IEEE Can. Conf. Electrical Comput. Eng.* 1 (9–12), 116–121.
- [3] T. Kohda, Y. Ookubo, and K. Shinokura, “Digital watermarking through CDMA channels using spread spectrum techniques,” in *Proc. IEEE International Symposium on Spread-Spectrum Techniques and Applications*, vol. 2,
- [4] [4] R. Bangaleea, H.C.S. Rughooputh, Performance improvement of spread spectrum spatial-domain watermarking scheme through diversity and attack characterization, *IEEE 6th Africon Conference*, Vol. 1, 2002, pp. 293-298.
- [5] [5] F. Zhang, B. Xu, and X.Zhang, “Digital Image Watermarking Algorithm based on CDMA spread spectrum,” in *12th International Multi-Media Modelling Conference Proceedings*, 2006, pp. 405-408.
- [6] [6] C. Bo and S. Hong, "A new robust-fragile double image watermarking algorithm," in *Multimedia and Ubiquitous Engineering*, 2009. MUE '09. Third International Conference on, 2009, pp. 153-157..
- [7] [7] Xiu-mei Wen, Wei Zhao, Fan-xing Meng “Research of a Digital Image Watermarking Algorithm Resisting Geometrical Attacks in Fourier Domain” pp: 265-268, *IEEE* 2009.
- [8] [8] Ho ATS, Shen J, Chow AKK, Woon J. Robust digital image-in-image watermarking using the fast Hadamard transform. In: *Proceedings of the international symposium on circuits and systems (ISCAS 2003)*; 2003. p. 826-9.
- [9] [9] Kutter M, Jordan F D, Bossen F, Digital signature of color images using amplitude modulation . *Storage and Retrieval for Image and Videos Dabbles* Vol. 3022, San Jose: SPIE, 1997: 518-526.

A Framework for Creating a Distributed Rendering Environment on the Compute Clusters

Ali Sheharyar

IT Research Computing
Texas A&M University at Qatar
Doha, Qatar

Othmane Bouhali

Science Program and IT Research Computing
Texas A&M University at Qatar
Doha, Qatar

Abstract—This paper discusses the deployment of existing render farm manager in a typical compute cluster environment such as a university. Usually, both a render farm and a compute cluster use different queue managers and assume total control over the physical resources. But, taking out the physical resources from an existing compute cluster in a university-like environment whose primary use of the cluster is to run numerical simulations may not be possible. It can potentially reduce the overall resource utilization in a situation where compute tasks are more than rendering tasks. Moreover, it can increase the system administration cost. In this paper, a framework has been proposed that creates a dynamic distributed rendering environment on top of the compute clusters using existing render farm managers without requiring the physical separation of the resources.

Keywords—distributed; rendering; animation; render farm; cluster

I. INTRODUCTION

A. Background

Rendering is a process of generating one or more digital image(s) from a model or a collection of models, characterized as a virtual scene. A virtual scene is described in a scene file that contains the information such as geometry, textures, lights, etc. It is modelled in a 3D modelling application. Most commonly used modelling applications are Blender [1], Autodesk 3D Studio Max [2] and Autodesk Maya [3]. All modelling applications have a user interface with a drawing area where users can create a variety of geometrical objects, manipulate them in various ways, apply textures, and even animate etc. Fig. 1 shows the user interface of Blender 3D modelling application [1]. A virtual scene is then given to the renderer that generates a set of high quality images later to be used to produce the final animation. Some of the most popular renderers are mental ray [4], V-Ray [5] and Blender [1].

Rendering is a compute-intensive and time-consuming process. Rendering time for an individual frame may vary from a few seconds to several hours. The rendering time depends on scene complexity, degree of realism (shadows, lights etc.) and output resolution. For example, a short animation project may be two-minutes in length, but at 30 frames per second (fps), it contains 3,600 frames. An average rendering time for a fairly simple frame can be approximately 2 minutes, resulting in a total of 120 hours. Fortunately, rendering is a highly parallelizable task as rendering of individual frames does not depend on any other frame. In order to reduce the total

rendering time, rendering of individual frames can be distributed to a group of computers on the network. An animation studio, a company dedicated to production of animated films, typically has a cluster of computers dedicated to render the virtual scenes. This cluster of computers is called a render farm.

B. Objectives

In a university environment, it can be complicated to do the rendering because many researchers do not have access to a dedicated machine for rendering [6]. They do not have access to a rendering machine for a long time as it may become unavailable for other research use. Moreover, they can lose their data due to hardware failure. By creating a distributed rendering environment, these problems can be addressed. However, some universities have one or more compute clusters that are normally used to perform high performance computing tasks. Distributed rendering on a compute cluster is possible, but it requires a lot of manual interaction with a cluster's job scheduler and is cumbersome.

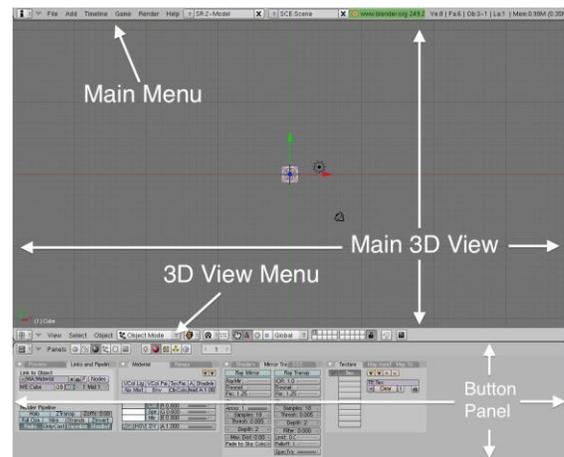


Fig.1. Screen shot of the Blender user interface

In this paper, a framework has been presented to create a render farm-like environment on a compute cluster by using existing render farm software. This way, researchers and students will be presented with an interface similar to what typically animation studios have. Moreover, it does not require manual interaction with the cluster's job scheduler and makes the rendering workflow smoother.

C. Related Work

There has been some work on doing the image rendering on the cluster and on grids [7]. In this section, related work is briefly presented.

Huajin and Bing [8] discuss the design and implementation of a render farm manager based on OpenPBS. OpenPBS is a resource managers used for compute clusters. They have proposed to extend the OpenPBS functionality in order to facilitate the render job management. They maintain a state table to hold the render jobs status. They implement a new command "qsubr" to submit the job and another command "qstatr" to monitor the render jobs. They also provide a web-based interface in order to facilitate the job submission and monitoring.

Gooding et al. [6] talk about implementation of distributed rendering on diverse platforms rather than a single cluster. They consider utilizing all available resources such as recycled computers, community clusters and even TeraGrid [9]. One benefit of this approach is that it gives access to diverse computing resources, but on the other hand it requires significant changes in the infrastructure. They require adding a couple of new servers to host the software for job submission and distribution to render machines. They also need a new central storage system because it is essential for the network distribution of the render job's resources (textures etc.) so that all render nodes could access it and save the output back. It is obvious that it requires a change in networking infrastructure. They offer only command line interface for job submission and support, only RenderMan rendering engine [10].

Anthony et al. [11] propose a framework of distributed rendering over Grid by following two different approaches. One approach is to setup the portal through which a user can submit a rendering job on-demand and download the rendered images from the remote server at a later time. Another approach is to submit the job to the Grid through a GUI plug-in within the 3D modelling software where every rendered image will be sent back individually to the client concurrently. Both of these approaches require significant effort for implementation. They also talk about compression methods that are beyond the scope of this paper.

All of these approaches focus on implementing the render farm manager (or job manager). They provide users a way to interface to submit and control the render jobs in the form of command line using SSH, online web portal and/or plug-in within 3D modelling software. In summary, they need a significant amount of time and resources to implement all the nitty-gritties of various software components. In the next section, a new approach will be proposed.

This paper is divided in to several sections. Section II and III give an overview of render farms and compute clusters respectively. Section IV describes the current approaches and proposed approach to render the computer animation in distributed computing environments. Section V presents the experimental results and, finally, section VI and VII presents the conclusion and future work respectively.

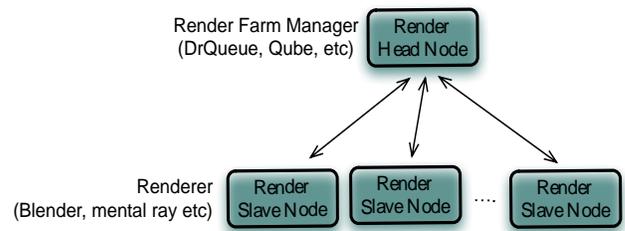


Fig.2. Render farm architecture

II. RENDER FARM

A render farm is a cluster of computers connected to a shared storage dedicated to render the compute-generated imagery. Usually, a render farm has one master (or head) machine (or node) and multiple slave machines. The head node runs the job scheduler that manages the allocation of resources on slave machines to jobs submitted by users (artists). Fig. 2 shows the client/server architecture of a render farm. In the diagram, Render Head Node runs the server software of render farm manager, whereas, Render Slave Node runs the client software.

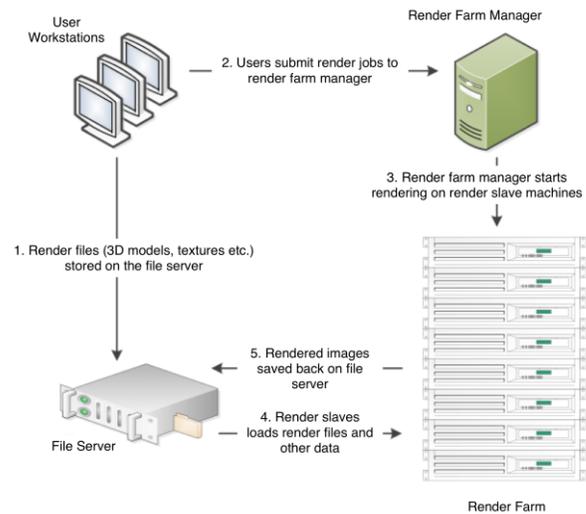


Fig.3. Rendering workflow

A typical rendering workflow (see Fig. 3) can be described in the following steps:

- 1) Artists create the virtual scenes on their workstations.
- 2) Artists store their virtual scene files, textures etc. on the shared storage.
- 3) Artists submit rendering jobs to the queue manager, a software package running on the head node.
- 4) Queue manager divides the job into independent tasks and distributes them to slave machines. A task could be rendering of one full image, a few images, or even a subsection (tile) of an image. A job may have to wait in the queue depending on its priority and load on the render farm.
- 5) Slave machines read the virtual scene and associated data from the shared storage.
- 6) Slaves render the virtual scene and save the resulting image(s) back on the file server.

7) *User is notified of job completion or errors, if any.*

A render queue manager (also known as render farm manager or job scheduler), typically a client-server package facilitates the communication between the slaves and master. The head node runs the server component whereas all slave nodes run the client component of the render queue manager; although some queue managers have no central manager. Some common features of queue managers are prioritization of queue, management of software licenses, and algorithms to optimize the throughput in the farm. Software licensing handled by the queue manager might involve dynamic allocation of available CPUs or even cores within CPUs.

III. COMPUTE CLUSTER

A compute cluster is a group of computers linked with each other through a fast local area network. Clusters are used mainly for computational purposes rather than handling the IO-oriented operations such as databases or web services. For instance, a cluster might support weather forecast simulation or flow dynamics simulation of a plane wing.

A typical compute cluster comprises one head node and multiple compute (or slave) nodes. All clusters run a resource management software package that accepts jobs from users. They preserve them until they are run, run the jobs, and deliver

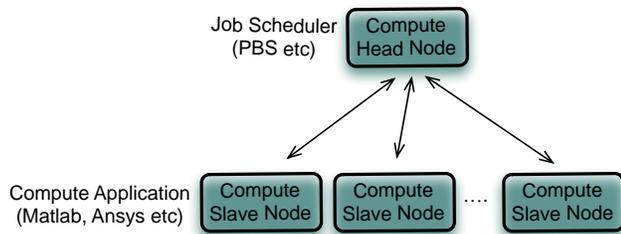


Fig.4. Compute cluster architecture

the output back to user. Fig. 4 shows the architecture of a compute cluster. Compute Head Node runs the server software of the resource manager, whereas, Compute Slave Node runs client software.

A typical workflow to execute a job on a cluster is described below:

- 1) *User prepares the job file that contains some parameters and path to the executable or script to run. For instance, the amount of memory and number of CPU cores required are specified.*
- 2) *User submits the job file to the job scheduler. Submission is done usually over SSH terminal but some job schedulers also offer the web interface for job submission.*
- 3) *Scheduler puts the job in to appropriate queue.*
- 4) *When the job's turn comes, scheduler allocates the resources and starts the execution of executable/script specified in job description on the allocated slave node.*
- 5) *When job finishes its execution, the output and error log is saved to disk. The scheduler can terminate a job if its execution time exceeds a predefined amount of time.*
- 6) *User is notified of job completion.*

IV. RENDERING ON A COMPUTE CLUSTER

It is clear that both render farms and compute clusters have similar architecture. Both of them contain one master machine (or head node) and one or more slave machines. Both run a resource manager software package and both have similar workflows. A render farm can be considered as a special kind of compute cluster that uses resource manager and other software (renderers) specific to rendering the computer-generated imagery.

This section focuses on running the render-farm ecosystem over a compute cluster. First, current approaches to solve this problem and their limitations have been described. Then, a new approach has been proposed that not only can present existing and familiar interfaces to users but also requires less implementation effort.

A. Proposed Cluster-based Rendering Framework

As mentioned above, all of existing work [6][8][11] have focused on developing all components of rendering job management and scheduling from scratch. However, this paper proposes a new approach using existing open-source or commercial render farm managers, meeting the requirements mentioned later, and using the compute cluster's resource with minimal or no change in existing setup.

Recall from earlier sections that a render farm manager has client/server architecture. The server software (r-server) runs on the head node whereas the client component (r-client) runs on slave nodes. The key difference between existing approaches and the proposed approach is that current approaches schedule the render jobs submitted by users directly to the cluster or grids and manage their state and execution process themselves. However, the new approach proposes to schedule the client-component of the render farm manager rather than the render jobs directly. The server component can either run on the cluster's head node (compute head node) along with cluster's existing job manager or a new server machine (render head node) that can be added to the existing environment. The render head node also runs a software module called Rendering on Cluster Meta Scheduler (RCMS) (see Fig. 5).

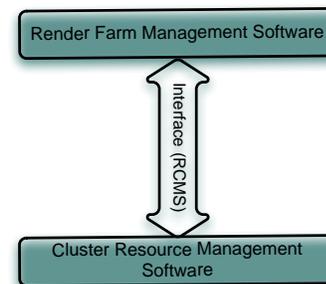


Fig.5. Interface between render farm manager and cluster resource manager

The RCMS scheduling module queries the render farm manager and dynamically adjusts the number of active r-client jobs. Each r-client job appears to the r-server as a render slave node. The number of active r-client jobs depends on the current

load of the compute cluster and the number of render jobs in queue. If there are pending render jobs, RCMS can submit new r-client jobs to the compute cluster. It also kills the active r-client jobs if there is no render job in queue and releases the resources to make them available for compute jobs. It maintains a state table to keep track of r-client jobs submitted to the cluster (see Fig. 6).

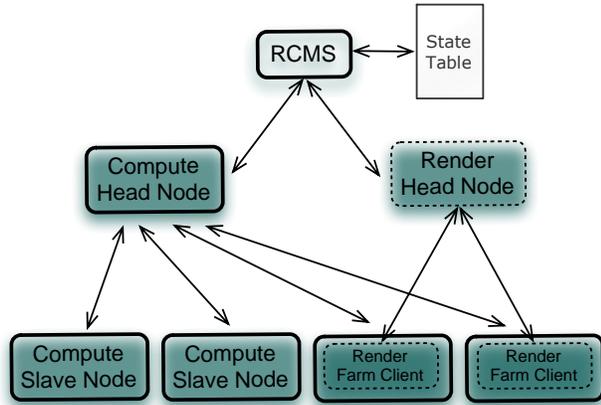


Fig.6. Render farm integration with the compute cluster

The RCMS module should be able to talk to render farm manager and cluster resource manager in order to keep the rendering system running effectively. Interfacing with cluster managers is trivial because most of cluster managers support at least command line interface for job submission etc. On the other hand, not all render farm managers are compatible with this framework. There is a set of requirements that a render farm has to meet in order to be compatible with the proposed framework. These requirements are described in the following sub-section.

B. Render Farm Manager Compatibility Requirements

There are several open-source and commercial render farm managers available. But, in order to be compatible with proposed distributed rendering on cluster framework, it should meet the following requirements:

1) *Job control API:* Render farm manager should support a set of calls to query the information about the render jobs submitted by users. This interface can be either command-line programs or API of any programming language such as Python, Ruby, C/C++ etc.

2) *Failsafe rendering:* The render farm manager should be able to detect failed or incomplete rendered images. As an r-slave job can run only for a fixed amount of time. After that, the cluster job manager will terminate it. It is important that the render farm manager should detect the incomplete rendered images and reschedule them.

3) *Automatic client recognition:* The server should automatically detect active clients on slave nodes. Clients should send so-called heartbeats to the server, so that the server will automatically know their existence. It is required as clients are expected to be active dynamically over a set of compute nodes in the cluster.

4) *Supervisor required:* Some render farm managers do not need the server component to manage the resources. However, the proposed approach requires that render farm management software has a supervisor to centrally control the jobs and resources.

Table 1 shows some of the popular render farm managers along with some features. As it is clear that Smedge and Spider are not compatible with the proposed framework because they do not support either supervisor and/or job control API.

Cluster Resource Manager Compatibility Requirements

All major cluster resource managers like PBS and LSF support at least SSH over command-line interface for user interaction. Some resource managers also support online web-interface.

Proposed distributed rendering framework requires that the cluster resource manager should have support for the following operations via command-line:

1) *Job submission:* A cluster manager should support job submission via command line. RCMS will prepare a job file that will specify the desired resources like number of cores, memory and execution time.

2) *Query jobs:* It should support querying the currently active jobs by their names. RCMS will use its own naming scheme in order to identify the r-client jobs.

3) *Job deletion:* As r-client jobs on cluster will be dynamically deleted in order to release the cluster resources to be used for other computation tasks, it is necessary that cluster resource manager support this feature.

TABLE I. RENDER QUEUE MANAGEMENT SOFTWARE

Name	Supported 3D Applications	Supervisor Required?	Job Control API
DrQueue	Blender, Maya, mental ray, Pixie, command-line tools	Yes	Yes
Qube!	Maya, mental ray, SoftImage, RenderMan, Shake, command-line tools	Yes	Yes
Smedge	3ds max, After Effects, Maya, mental ray, SoftImage	No	Yes
Spider	Maya	No	No
RenderPal	3ds max, Blender, Cinema 4D, Houdini, Maya, mental ray, SoftImage	Yes	Yes
ButterflyNetRender (BNR)	All major applications and command-line tools	Yes	Yes

TABLE II. HARDWARE AND SOFTWARE SPECIFICATION OF TEST PLATFORMS

Machine	Dell 690	Suqoor (single node)	HP Z800
CPU	2x Intel Clovertown X5355 @ 2.66 Ghz	2x Intel Harpertown E5462 @ 2.80 GHz	2x Intel Westmere-EP X5650 @ 2.66 GHz
Cores (per CPU)	4	4	6
Threads (per CPU)	4	4	12
L2 Cache (per CPU)	8 MB	12 MB	12 MB
CPU Launch Date	Q4'06	Q4'07	Q1'10
Memory	16 GB	32 GB	16 GB
Operating System(s)	Win XP 64-bit/Cent OS 6 (64-bit)	SuSE Linux Enterprise Server 10 (64-bit)	Red Hat Enterprise Linux 5 (64-bit)
GPU	Quadro FX 4600	None	Quadro Plex 6000



Fig. 7. A 3D virtual scene modeled in Maya

V. EXPERIMENTAL RESULTS

As a proof of concept, a prototype of proposed distributed rendering framework is implemented and benchmarked. In this section, the benchmark results are presented. The prototype uses PipelineFX Qube [12] as render farm manager and PBS [13] for cluster resource management. It is implemented in Python language. The compute cluster (named *Suqoor*) at Texas A&M University at Qatar [14] has been used as a test environment. Out of ten available licenses for Qube, one is consumed by Qube Supervisor that manages all render jobs and remaining nine licenses are used by Qube workers. Each worker requires one license. A virtual scene (see Fig. 7) that comes with Autodesk Maya [3], a 3D modelling application, is used for the benchmarking. This scene has a 30-second long animation that comprises of 720 frames at 24 frames per second. For performance comparison, the same animation has been rendered (software rendering) on *Suqoor* and three other workstations as well. Software rendering refers to a rendering process that is unassisted by any specialized graphics hardware (such as graphics processing units or GPUs). Hardware rendering, utilizing the graphics hardware for rendering, cannot be performed on *Suqoor* due to lack of graphics hardware. However, performance of hardware rendering on workstations has also been compared to software rendering on *Suqoor*.

Table 2 summarizes the hardware specification of the workstations and single compute node of the compute cluster (*Suqoor*). Note that two of the workstations have the same hardware specification (Dell 690) but have different operating systems. One has Windows XP x64 and other has Cent OS 6.

Fig. 8 shows the average rendering time per frame on a single node (8 cores) of *Suqoor* and other workstations. Note that rendering time on a single compute node of cluster having 8 cores (25.22s) and HP Z800 workstation having 12 cores (25.86s) differs just by a fraction of a second. It has also been observed that rendering on Windows XP is almost 2.76 times slower than CentOS Linux on the same hardware configuration.

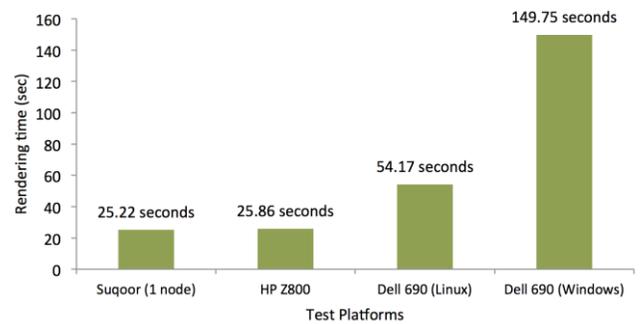


Fig. 8. Average rendering time per frame (software rendering)

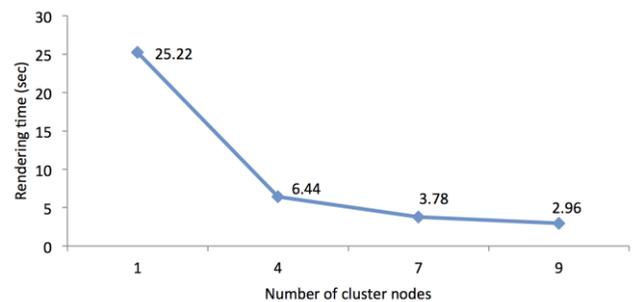


Fig. 9. Average rendering time per frame versus number of cluster nodes (software rendering)

Fig. 9 shows the average rendering time per frame by using 1, 4, 7 and 9 compute nodes, respectively. Due to the limited number of available Qube licenses (10), the experiment could not be performed with more than 9 compute nodes.

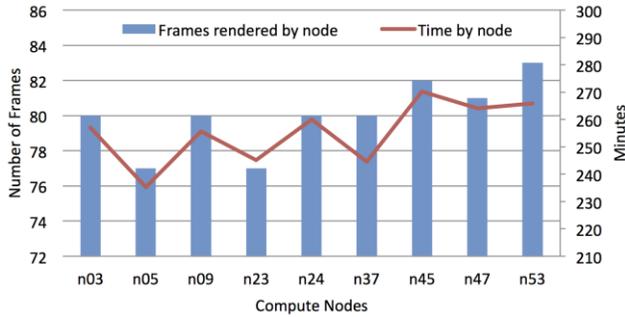


Fig. 10. Number of frames rendered and accumulated time spent

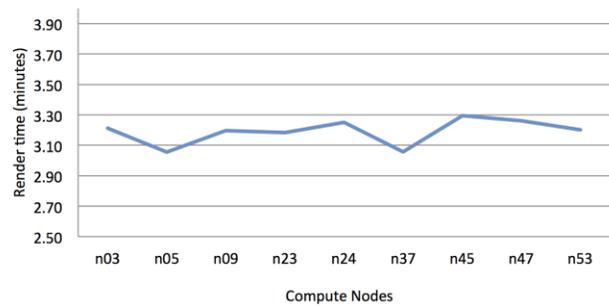


Fig. 11. Average rendering times per frame with respect to compute nodes

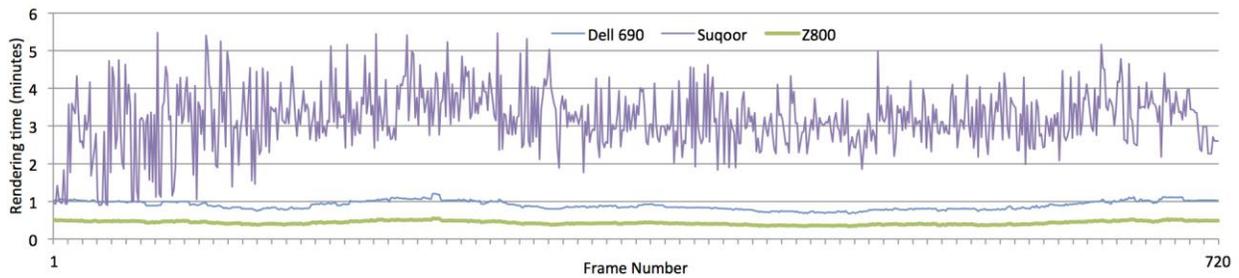


Fig. 12. Rendering time of individual frames (software rendering)

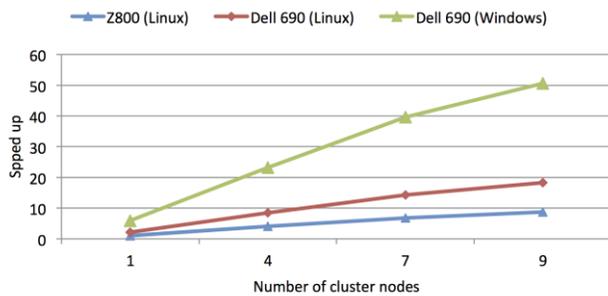


Fig. 13. Speed up with cluster (software rendering)

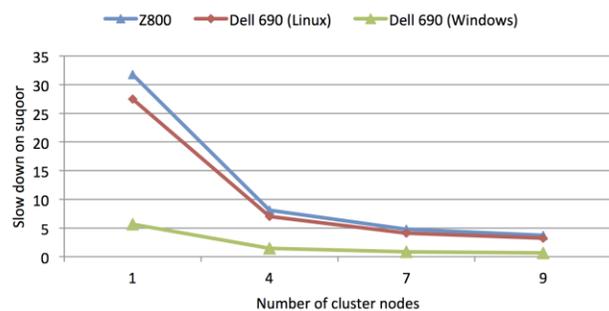


Fig. 14. Software rendering on cluster versus hardware rendering on Windows and Linux workstations

Fig. 10 shows the number of rendered frames and aggregated rendering time spent by each compute node. It shows how the Qube supervisor has distributed the render jobs across compute nodes. With an average distribution of 80 frames per node, it is apparent that work distribution is almost equal.

Fig. 11 shows the average rendering times per frame with respect to each compute node. The variation in the result can be characterized to the variation in the 3D model complexity from different view angles.

Fig. 12 shows the rendering time of all frames. Remember that there are 720 frames in the animation. It is observed that workstations render the frames one after the other by utilizing multiple CPU cores for single-frame rendering. On the other hand, *Suqoor* renders multiple frames on individual compute nodes. The numbers of active frames depend on the number of available cores on the compute nodes by assigning each core to

an individual frame. Due to this, rendering time of individual frames is higher on *Suqoor* than other workstations.

Fig. 13 shows the performance speed-up with respect to other platforms. For instance, a cluster with nine compute nodes performs nearly 51 times better than the Dell 690 workstation with Windows, nearly 18 times better than the Dell 690 workstation with Linux, and nearly 9 times better than the HP Z800 workstation.

Fig. 14 compares the performance of software rendering on *Suqoor* to hardware rendering on other workstations. This plot shows how fast hardware rendering is on other workstations with respect to *Suqoor*. Remember that *Suqoor* does not support hardware rendering. It is observed that hardware rendering, especially on Linux workstations, is remarkably faster than software. The performance gap is drastically reduced by using more nodes on the cluster.

VI. CONCLUSION

This paper has presented a framework that creates a distributed rendering environment on a general-purpose compute cluster by using an existing render farm management application. It can be used to create the rendering environment similar to that of an animation studio in a university environment where users do not have exclusive access to the computers to perform time-consuming image renderings. The prototype of the proposed framework, using Qube! for render farm management and PBS for compute cluster management, has been implemented. The experimental results show that the compute cluster reduces the rendering time significantly in case of software rendering. Moreover, by using the existing render farm manager, the overall rendering workflow becomes efficient.

VII. FUTURE WORK

One thing, where compute cluster lacks behind is the hardware rendering that is due to the absence of GPUs in the compute nodes. Texas A&M University at Qatar is soon expected to acquire a larger cluster that will also have GPUs in several compute nodes. For the future work, the same experiment will be repeated on the new cluster and performance of the hardware rendering will be analyzed. The new cluster is expected to outperform the workstation by a large margin.

References

- [1] Blender, <http://www.blender.org>
- [2] Autodesk 3D Studio Max, <http://www.autodesk.com>
- [3] Autodesk Maya, <http://www.autodesk.com>
- [4] Mental ray, <http://www.mentalimages.com/products/mental-ray.html>
- [5] V-Ray, <http://chaosgroup.com/en/2/index.html>
- [6] Gooding, S. Lee, Laura Arns, Preston Smith, and Jenett Tillotson. "Implementation of a distributed rendering environment for the TeraGrid." In Challenges of Large Applications in Distributed Environments, 2006 IEEE, pp. 13-21. IEEE, 2006.
- [7] Grid Computing, http://en.wikipedia.org/wiki/Grid_computing
- [8] Jing, Huajun, and Bin Gong. "The design and implementation of Render Farm Manager based on OpenPBS." In Computer-Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference on, pp. 1056-1059. IEEE, 2008.
- [9] TeraGrid, <http://www.teragrid.org>
- [10] RenderMan, <http://renderman.pixar.com>
- [11] Chong, Anthony, Alexei Sourin, and Konstantin Levinski. "Grid-based computer animation rendering." In Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia, pp. 39-47. ACM, 2006.
- [12] Pipeline FX Qube!, <http://www.pipelinefx.com>
- [13] PBS Guide, <http://hpc.sissa.it/pbs>
- [14] High Performance Computing at Texas A&M University at Qatar, <http://technology.qatar.tamu.edu/rc/2000.aspx>

Integrating Social Network Services with Vehicle Tracking Technologies

Ahmed ElShafee
Assistant Professor
Ahram Canadian University
Egypt

Mahmoud ElMenshawi
B.Sc. Computer Science
Ahram Canadian University
Egypt

Mena Saeed
B.Sc. Computer Science
Ahram Canadian University
Egypt

Abstract—This paper gives design, and implementation of a newly proposed vehicle tracking system, that uses the popular social network as a value added service for traditional tracking system. The proposed tracking system make use of Google maps service to trace the vehicle, each vehicle has an account that contains a posts of Google maps that display the vehicle location on real time mode. A hardware module is inside the vehicle that uses Global Positioning System (GPS) – to detect vehicle location- and Global system for mobile communication (GSM) – to update vehicle location in vehicle account on social network -. System uses the well-known Arduino microcontroller to control GSM-GPS Modem. The proposed system can be used for a broad range of applications such as traffic management and vehicle tracking/anti theft system, and finally traffic routing and navigation. it can be applied in many business cases, like public transportation, so passengers can track their buses, trains, by following the vehicle account on social network. It also can be used in private business sector as an easy and simple fleet tracking and management system , or can be used by anyone who wants to track his car, or to find his way in case he get lost.

Keywords—Vehicle Tracking; GSM; GPS; Microcontrollers; Twitter; Google maps.

I. INTRODUCTION

A. Overview

Vehicle tracking systems [1], this term covers a range of products which, uses communications technology, or a combination of technologies, identify vehicle land report and, its real-time location and present this information to vehicle trackers on a remote server through internet. Vehicle tracking systems are commonly used by fleet operators for fleet management functions such as routing, dispatch, on-board information and security. Other applications include monitoring driving behavior, such as an employer of an employee, or a parent with a teen driver. Vehicle tracking systems are also popular in consumer vehicles as a theft prevention and retrieval device. Police can simply follow the signal emitted by the tracking system and locate the stolen vehicle.

social network sites (SNSs) [2] such as Twitter, Facebook, have attracted millions of users, many of whom have integrated these sites into their daily practices. Till now there are alot of SNSs, with various technological affordances, supporting a wide range of interests and practices. The great benefit behind SNSs that they help strangers connect based on shared interests, political views, or activities. They attract people

based on common language or shared racial, sexual, religious, or nationality-based identities. Most SNSs incorporate new information and communication tools, such as mobile connectivity, blogging, and photo/video-sharing.

B. Background and Related Works

Vehicle Tracking service [3] [4] is a GPS based solution that provides instant location information to the vehicle owner/authorized person through web/SMS with other flexibilities. The basic feature of popular vehicle tracking systems is locating real-time position of the vehicle. There are many other value added features and services, like applying rules on the vehicles (e.g.: speed limit, No Go Area, etc.), securing the vehicles by adding security features like remote immobilization, panic alarm. Some of the other features are: Speed Violation Alert/Report (speed of all vehicles can be controlled, monitored, and hence when violated, immediate contact can be made to reduce such violation). Area Alarm (an area can be assigned to vehicles, the owner/authorized person will be notified, if the rule is violated); 'No-Go' area (A 'No-Go' area can be created for the vehicles, the owner/authorized person will be notified, if the rule is violated); in addition, the car owner can find out whether the ignition is on or off; if on, then whether the vehicle is moving or stationary, thus vehicles can be monitored even if they are switched off.

Most social networking websites are supported by paid advertisements that appear on member pages. Because of this, most social networking websites do not carry membership charges and offer free services to all users. The main goal of social networking websites are designed to allow members to connect and communicate with one another - so features of these sites foster interaction, activity, and of course, community.

II. SYSTEM ANALYSIS

A. Problem definition

The proposed vehicle tracking system is an open system that uses a free and open source software and is composed of commodity hardware that is easy-to-find. Our system is composed of four components, a GPS/GSM Tracking Device, a web server with database, social network, and finally Map. The GPS tracking device is an embedded system that transmits location information to the server through GPRS networks. The server is a personal computer that receives the information and put it in the database. Twitter is selected as value added service to common vehicle tracking system, vehicle presented on the

social network as virtual profile that users can simply follow vehicle account. Vehicle send tweets in regular bases, that contains a link to a map showing the current location of the vehicle.

B. Proposed system feature

The proposed vehicle tracking system provides the following features;

- Vehicle position Tracking system
- Intelligent Transportation System (ITS)
- Fleet Management System
- Vehicle anti-theft system.

C. Proposed system objectives and scope

- Exploring GPS based tracking systems
- Developing Automatic Vehicle Location system using GPS for positioning information and GSM/GPRS for information transmission with following features:
 - Acquisition of vehicle's location information (latitude and longitude) after specified time interval.
 - Transmission of vehicle's location and other information (including ignition status, door open/close status) to the monitoring station/Tracking server after specified interval of time.
 - Server is capable to place the latitude and longitude on Google maps, preparing a simple URL for a map containing the current location of the tracked vehicle
 - Server posts the vehicle status to vehicle account on the social network
 - Now users who follow vehicle profile will find vehicle feeds attached with a Google map showing vehicle current location.

III. SYSTEM DESIGN

A. Proposed System layout

- Overall system is partitioned into four major units.
- In-Vehicle unit
- Server
- Social network
- User interact with the system using web browser through vehicle account on the social network, Figure 1 shows system layout

B. In-Vehicle Unit functions and components

This is major part of the system and it will be installed into the vehicle. It is responsible for capturing the following information for the vehicle.

- Current location of vehicle
- Speed of vehicle
- Door open/close status

- Ignition on/off status

In-vehicle unit is also responsible for transmitting this information to Tracking Server through the internet. To achieve all these functionalities In-Vehicle unit uses following modules.



Fig. 1. Proposed system layout

1) GPS module

GPS [5] module is responsible of capturing the current location and speed. Location and speed data provided by GSM/GPS module need some fragmentation to be compatible with Google maps format. CPU is required to process this raw data. SiRF Star III single-chip GPS receiver is used which comes integrated with GM862-GPS. GPS receiver can also provide information of altitude, time of last reported location, status of GPS last reported location, number of satellite used to compute current location information along with location and speed. System truncate the vehicle coordinates and time. Other data provided by GPS receiver is used to determine the validity of location information, and will be ignored

2) Central Processing Unit

CPU captures raw data from GSM/GPS receiver to extract the required vehicle location and speed information. CPU is also responsible for monitoring vehicle door open/close, engine status on/off and controlling the vehicle ignition on/off status.

CPU is also responsible of establishing connections between GSM/GPS module and remote server, through internet over the GSM network, CPU sends detected vehicle location, speed, door status, and engine status. On other hand, CPU process commands being sent from the remote server to control the vehicle like ignition on/off.

The microcontroller selected to serve as CPU for In-vehicle unit is Atmel's ATmega328. a popular microcontroller based called Arduino[6] uses the Atmega328. It has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, a 16 MHz ceramic resonator, a USB connection, and power connector. It contains everything needed to support the microcontroller; it can simply programmed through USB connection, and can be directly interfaced to GSM/GPS module, door, generator, and ignition

3) Data Transceiver

For real time tracking of vehicle, a reliable wireless network is required to transmit data to remote server. Existing GSM network is selected because of its broad coverage, and its cost effective rather than to deploy own network for transmission of vehicle information.

For data transmission over GSM network, GSM modem is required as a data transceiver module. GM862-GPS [7] GSM/GPRS modem is selected to transmit data over GSM network because of its features and capabilities. GM862-GPS provides AT commands interface i.e. all functions can be accessed using AT commands. AT commands can be sent to it using UART serial interface

C. Design of In-Vehicle Unit

In-Vehicle unit consists of two main modules, the Telite GM862-GPS GSM/GPRS modem, and Arduino microcontroller board. Figure 2, shows the block diagram of in-vehicle unit.

GPS antenna should be directed toward sky in a correct computation of the GPS satellites location to be able to receive GPS satellites signals. Arduino microcontroller communicates with modem through a Simple UART serial interface. Arduino process raw data received from GPS/GSM modem, then transmits this information to remote server using GSM/GPRS modem over the internet through GSM network. Microcontroller controls the operation of GSM/GPRS modem through serial interface using AT commands. To assure a reliable transmission and receiving of data an external GSM antenna is required by the GSM/GPRS modem.

Arduino microcontroller receives commands and information passed from remote server through GPS/GSM modem, then passes this information to which analyses received information and performs action accordingly (i.e. turns on/off ignition of vehicle, transmits current location, etc).

Arduino microcontroller is directly connected to vehicle different components, as a normal digital sensors or actuators, like ignition on/off circuitry and door status output of vehicle. Information packet sent to server also contains status information of these I/O ports



Fig. 2. block diagram of in-vehicle unit.

Arduino microcontroller receives commands and information passed from remote server through GPS/GSM modem, then passes this information to which analyses received information and performs action accordingly (i.e. turns on/off ignition of vehicle, transmits current location, etc).

Arduino microcontroller is directly connected to vehicle different components, as a normal digital sensors or actuators, like ignition on/off circuitry and door status output of vehicle.

Information packet sent to server also contains status information of these I/O ports.

D. GM862-GPS Arduino shield card

As mentioned before, Arduino microcontroller communicates with GM862-GPS through UART serial interface [8]. The first step is to design and implement extended PCB with GM862-GPS modem that can be installed over the Arduino microcontroller, called GM862-GPS Arduino shield. Which acts as an upgrade module that can easily interface the GM862-GPS module to Arduino microcontroller. The following figure shows the designed GM862-GPS shield

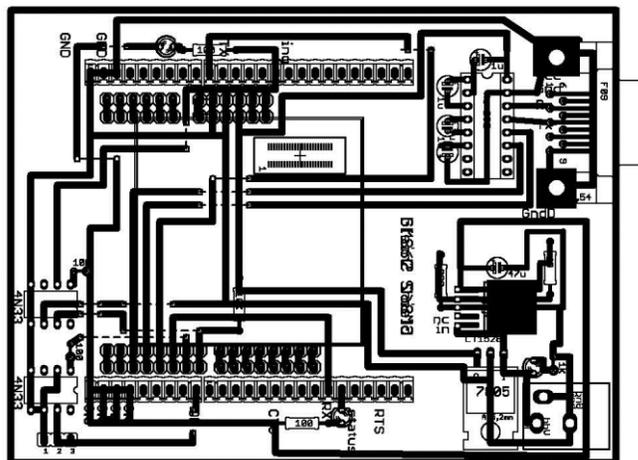


Fig. 3. GM862-GPS shield

E. In-Vehicle Unit Software Design

Arduino Microcontroller is the Central Processing Unit for In-Vehicle unit, it controls all its operations.

A software program written in C language, that is compiled then burnt into the microcontroller's flash memory. Complete software consists of small modules as shown in the Figure 4.

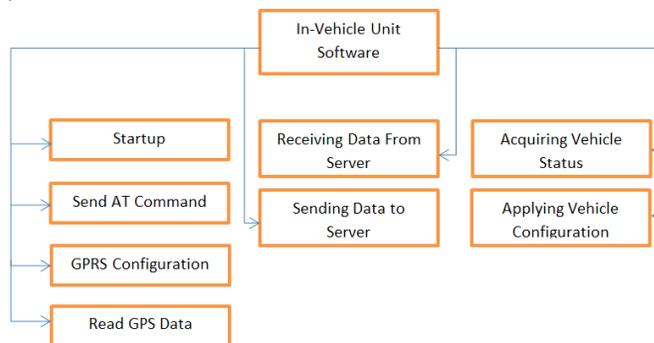


Fig. 4. block diagram of in-vehicle unit.

All these modules are implemented as subroutines in the software. Each subroutine performs series of its designated tasks. Flow chart of each subroutine is described below

1) Startup- subroutine

Startup routine is executed only when device is powered on and reset. It initializes all hardware of the In-Vehicle unit and configures GM862-GPS. It powers up the GM682-GPS modem then performs various tests to ensure that modem is working

properly and is ready to use. If microcontroller failed to communicate with modem, alarm led set on, and sends error description to its serial interface.

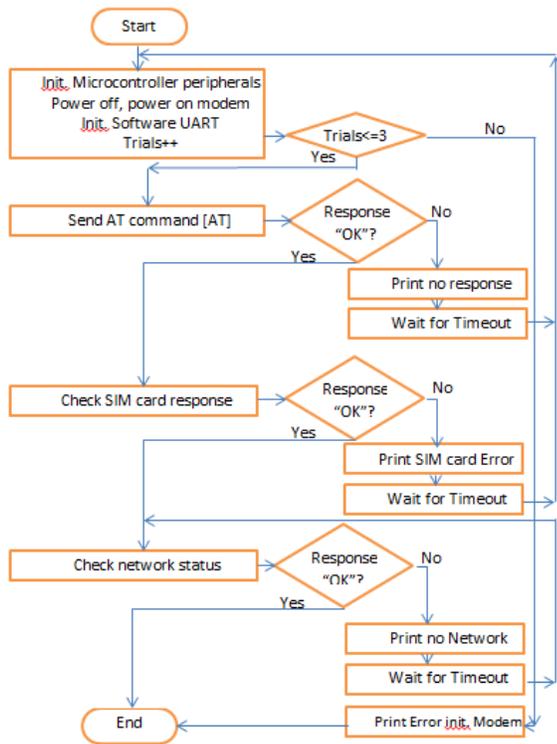


Fig. 5. Startup routine flow chart.

Figure 5 shows the startup subroutine flowchart. The flow chart shows that the subroutine starts with initializing peripherals of the microcontroller.

All peripherals in use need to be initialized in this step. After initializations Microcontroller power up the GM862-GPS modem then it starts modem checking process. Microcontroller sends “AT” command to the modem using “Send AT Command” subroutine.

All AT commands sent to the modem are sent using this subroutine. If the device responds with “OK”, it means microcontroller can communicate with module. If device doesn’t respond after expiration of timeout modem is restarted, and the microcontroller sends “no response” message to its serial port (device console port) . If the modem doesn’t respond to microcontroller for three successive trials, and problem then persists definitely something in hardware is damaged. Then microcontroller sends “error initializing modem” message, then go to halt mode.

After receiving “OK” response from the modem, which refers to modem status that is connected and ready. Then various parameters of modem need to be initialized, and checked, SIM presence is checked by sending command “AT+CPIN?” If device responds with “+CPIN: READY” message, SIM is ready to be used. Microcontroller will consider any other response message as an error. Microcontroller will send “SIM card error” message to

console port, and module will be restarted after expiration of timeout.

When SIM card check is OK,, microtroller is going to make sure that modem is connected to network or not. Network status can be tested with command “AT+CREG?” If module responds with “+CREG: 0, 1” module is connected to network and ready to communicate over the network. If any other response is received module keeps on checking for network status until it is connected. Once it made sure that module is connected to network, subroutine is ended

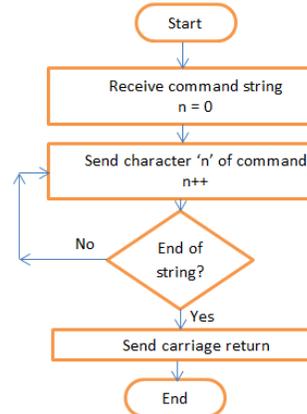


Fig. 6. Send AT command routine flow chart.

2) Send AT commands - Subroutine

This subroutine is the basic routine that handles all the communications with GM82-GPS modem. This routine accepts AT command as input string argument or parameter then sends it (character by character) to the modem followed by carriage return (‘r’) as a command terminating character. Figure 6 shows routine flowchart During the startup routine, a soft serial port (UART) is initialized on two pins which is connected to modem serial port, as shown in figure 3. The communication baud-rate (between microcontroller and modem) is specified during the initialization process too. Transmit buffer is a software register of UART. As soon as a 8-bit data is written into the transmit buffer. It will be transmitted through UART at the specified baud-rate. Each character of command string will be sent in same way. After sending the command characters, microcontroller terminates the command by sending carriage return to the modem. Response received from the modem will be handled in another subroutine.

3) GPRS configuration - Subroutine

The in-vehicle unit sends vehicle information for the server through internet using GPRS service [9]. The first step is to configure modem.

Figure 6 shows the required steps to configure the GM682/GPS module for GPRS data transmission. First step is to define GPRS context, which means identify the internet entry point interface of your network provider. Hence microcontroller sends the following command “AT+CGDCONT” with some parameters to identify network entry point interface in order to gain access to the internet and define the value of IP address of the module as follows;

AT+CGDCONT=1, “IP”, “wap.vodafone.com.eg”, “0.0.0.0”, 0, 0.

First parameter is context id, it is possible to define up to 5 contexts. Next parameter is communication protocol, third parameter is APN assigned by network server provider.

The next step is to set the parameters for Quality of service. Commands used are ;

“AT+CGQMIN= 1,0,0,0,0” and
“AT+CGREQ=1,0,0,3,0,0”.

These parameters are recommended by manufacturer of the GM862-GPS module.

Network service provider provides a user name and a password to authenticate the network connection, so the next step is to set user name and password for current GPRS context. Commands used are;

“AT#USERID=WAP” and
“AT#PASSW=WAP”.

Next step is to configure the TCP/IP stack, which mainly sets the minimum packet size, data sending timeout and socket inactivity timeout. Command used for configuring TCP/IP stack is:

“AT#SCFG=1,1,140,30,300,100”.

The first parameter of command is connection identifier; 2nd parameter is the context identifier for which stack is being configured. the 5th parameter (300) is the minimum number of bytes that will be sent in one packet. The last parameter (100) is the inactivity timeout, connection timeout, and data sending timeout.

Next step of the subroutine is to configure the firewall settings, which allows certain computers to connect to the module. In this case server IP address will be provided to firewall so that Tracking server can connect to In-Vehicle unit. Command used for firewall settings is;

AT#FRWL=1,”server ip”, “subnet mask”

Server IP address is to the Tracking server address on the internet and subnet mask.

Last step is activate current GPRS context through the following command;

“AT#SGACT=1, 1.”

First parameter is context id to be activated and next parameter is status i.e. 1 for activation and 0 for deactivation.

4) Read GPS data- Subroutine

Microcontroller requests current location from GM862-GPS by sending the following command “AT\$GPSACP”. Microcontroller waits for modem reply as follows;

“\$GPSACP:080220,4542.82691N,01344.26820E,259.07,3,2.1,0.1,0.0,0.0,270705,09”

Microcontroller extract the latitude and longitude from the modem message, which are the 2nd and 3rd parameters “4542.82691N,01344.26820E” in the following format “ddmm.mmmm N/S” for latitude and “dddmm.mmmm E/W”

for longitude. dd: degree (0-90 for latitude) and (0-180 for longitude), mm.mmmm: minutes (0- 59.9999 minutes), N/S:

North / South, and E/W: East / West. Microcontroller converts [10] these values first to decimal values instead of degree value in order to send these coordinates to server. Figure 7 shows the subroutine flow chart.

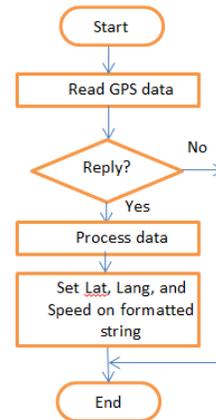


Fig. 7. Read GPS data subroutine

5) Receiving data from server- Subroutine

Microcontroller is always listening to GM-862-GPS module data stack, checking for its buffer, if any data is available, it starts processing data saved in data stack buffer. Microcontroller decide what is server asking for then call the appropriate subroutine, like acquiring vehicle status, or acquiring vehicle location, or applying a command to vehicle. Figure 8 shows subroutine flowchart.

6) Sending data to server- Subroutine

In order to send data over the internet (IP network), application needs an interface to medium access and physical layers, which named as socket. The socket is an interface contains three main entries, the transport protocol type (TCP/UDP) the server IP address, and the port number .This subroutine starts with opening socket for currently configured TCP/IP stack [11]. Command used to open socket for configured embedded TCP/IP stack is “AT#SD=1, 0, 80,”server address””.

The first parameter is connection identifier of TCP/IP stack, 2nd is protocol i.e. 0 for TCP and 1 for UDP, the 3rd parameter is the port number, and the last parameter is the IP address/host name of Tracking server. If command returns the response CONNECT; connection is accepted. Now the connection to the server is established and ready to send data to the server broadcast and activated on the module. When GPRS connection is alive, module can't accept AT commands and GPS data can't be read from module. To return to At command mode socket is suspended using escape sequence +++.

If In-Vehicle unit is configured for continuous transmission of vehicle information after regular intervals then microcontroller will automatically update server with vehicle status including vehicle location (latitude, and longitude), and vehicle sensors status like door, and ignition status. Figure 9 shows the flowchart for this subroutine.

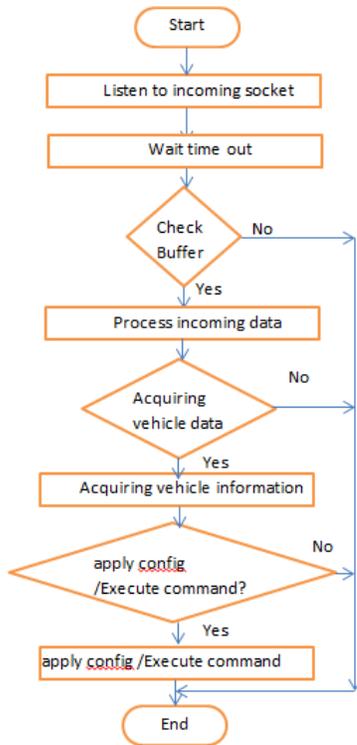


Fig. 8. Receiving data from server- Subroutine flow chart.

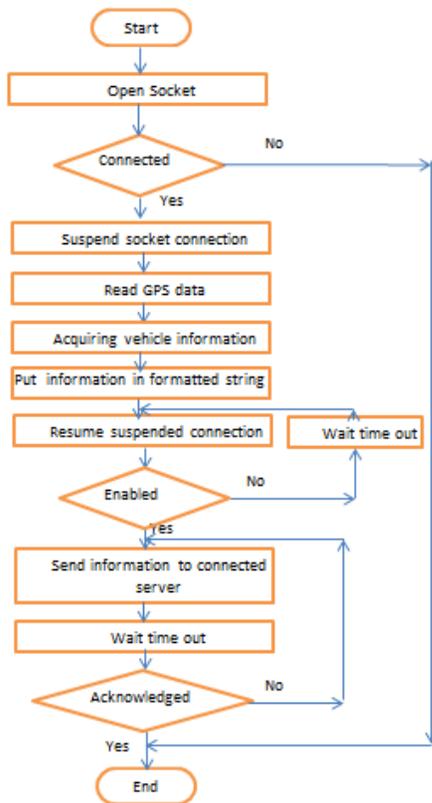


Fig. 9. Sending data to server- Subroutine

7) *Acquiring vehicle status - Subroutine*

In this subroutine the microcontroller collects the vehicle sensors status like the door, and ignition status. The second part

of this subroutine is to get the vehicle location and speed, so microcontroller should call the “Read GPS data” subroutine. Finally the information string received from Read GPS data subroutine is appended with status of I/O ports then sent back to tracking server after resuming socket connection. All above steps are repeated otherwise module waits for incoming requests from Tracking server. Figure 10 shows subroutine flow chart.

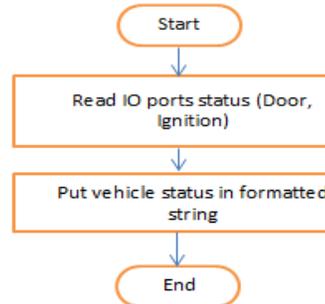


Fig. 10. Acquiring vehicle status – Subroutine flowchart

8) *Applying vehicle configuration/ Execute vehicle Command - Subroutine*

Server can send request for vehicle shutdown, restart the In-Vehicle unit. Microcontroller process the received data from server, then accordingly execute the proper action. Figure 11 shows subroutine flow chart

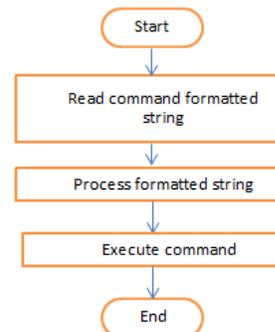


Fig. 11. Applying vehicle configuration/ Execute vehicle Command – Subroutine flow chart

9) *Main Routine of In-Vehicle Unit*

The main routine just calls all above subroutines. When in-vehicle unit starts up, the microcontroller calls startup subroutine once, if the GM862-GPS modem successfully started, microcontroller calls the GPRS configuration subroutine, finally microcontroller enters an endless loop.

In this loop, microcontroller listens to data being received from tracking server by calling “receiving data from server” subroutine, if any data is received microcontroller starts processing requests then take necessary action. If the in-vehicle unit is configured to continuously update tracking server with vehicle status, microcontroller will read GPS data, acquire vehicle status, then finally send these data to tracking server in regular bases, based on unit configuration. All subroutines are implemented in C language. Compiler used Arduino IDE software . Figure 12 shows main subroutine flow chart.

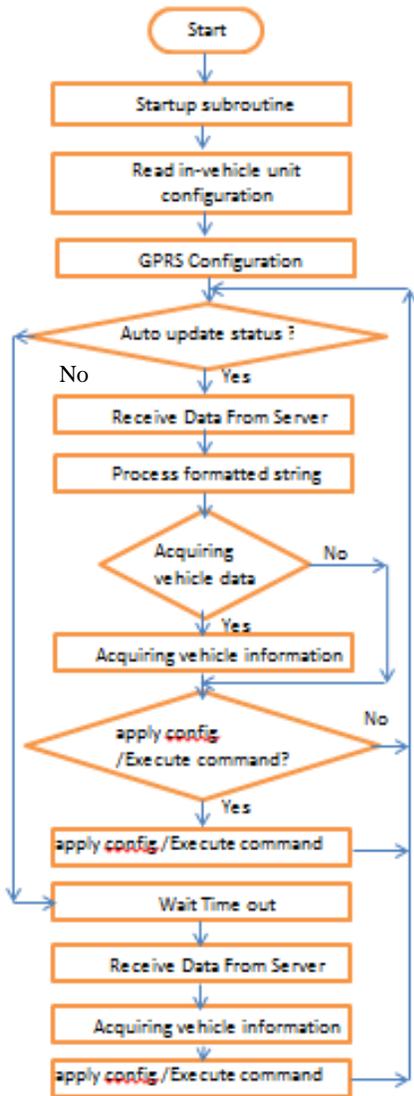


Fig. 12. In-Vehicle unit main program

F. Tracking Server

1) Tracking Server software design

Tracking server maintains all received information from all In-Vehicle units installed in different vehicles into a central database. This database is accessible from internet through user friendly interface to authorized users through a web application. Here all vehicle updates are available, like vehicle location, door status, ignition status, and authorized user can send commands to in-vehicle unit like shut down the vehicle or restart the in-vehicle unit. Authorized users control vehicle accounts on twitter social network, and can automatically make server posts vehicle updates (vehicle location) to vehicle's account on twitter. Vehicle location is automatically placed on Google maps, which make it easier for tracking the vehicle by vehicle trackers. Tracking Server consists of four major parts.

- (i) Communication Software with GM862-GPS
- (ii) Communication Software with Twitter social network

(iii) Database

(iv) Web Interface

2) Web Interface software module

As described in previous section Tracking Server maintains all information in a database. To display this information to authorized End users, front end software is required. The Authorized end users are the persons who have installed the In-Vehicle unit in their vehicle and also the system administrators who are managing Vehicle Tracking System. Server is designed to handle many In-Vehicle units at once; each unit presents a car in the tracking server. Each in-Vehicle unit has a unique identifier that identifies the vehicle to server and their authorized users. Whenever In-Vehicle unit is installed, information about that vehicle is stored in the database. Web interface supports this functionality. As the tracking server will be available on the internet, access to the vehicle information should be restricted to the authorized users.

3) Database module

Database is designed to store all received vehicle information (vehicle updates), information about In-Vehicle units and users of the system. Figure 13 show the ER diagram of tracking server database

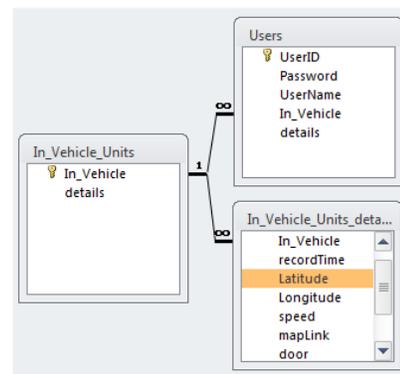


Fig. 13. tracking server DB ER diagram

4) Communication Software with GM862-GPS software module

GM862-GPS is GSM/GPRS modem that was used in the In- Vehicle unit. From tracking server point of view it's a seamless TCP/IP communication protocol. Server simply listens to pre-defined socket port, after receiving information from in-vehicle system, it extract the vehicle location and vehicle status, and save it in database. If server is configured to posts vehicle location to vehicle account on twitter, server also forwards a Google map traking the vehicle location on it. Authorized end users who are authorized to access vehicle account on server can monitor vehicle sensors status, and send commands to in-vehicle unit by forwarding these information to in-vehicle unit as a reply to the same socket that in-vehicle unit opened before to communicate with tracking server.

5) Communication Software with twitter social network software module

As mentioned before tracking server can be configured to automatically post vehicle location on vehicle's twitter account.

To do so user needs first to create an application via a vehicle twitter account with Read/Write permission, and of course the twitter's account password is known to tracking server, and saved in its database. To gain access to vehicle twitter's account we used The OAuth 2.0 authorization framework, which enables a third-party application to obtain limited access to an HTTP service. If user is storing protected data on original user's behalf, they shouldn't be spreading their passwords around the web to get access to it. OAuth is used to give users access to data while protecting original users accounts' credentials. Finally authors use PHP Library to support OAuth for Twitter's REST API. Figure 14 shows the flow chart of that software module

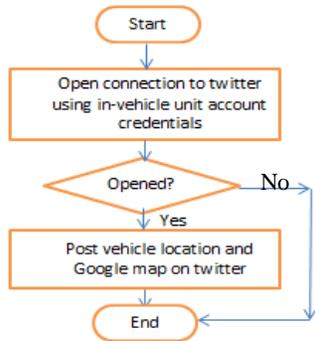


Fig. 14. tracking server and Tracking Server software design

G. System Data Flow diagram

Figure 14 shows the flowchart of main program. Main program listens to incoming TCP/IP connections from in-vehicle units, and creates separate thread for each incoming connection, which allows any number of In-Vehicle units to connect to server at once

IV. SYSTEM TESTING AND RESULTS

The proposed system is verified by testing it after integration of all components of the system. The following section contains a few testing information of each module.

A. Testing In-Vehicle Unit

GM862-GPS interface board was connected to microcontroller board through a serial cable. When In-Vehicle unit is powered on, it executes Startup routine. At first it reads and displays the existing configuration of the system. Next step, the microcontroller configures the GM862-GPS. It tests the communication interface by sending "AT" command. GM862-GPS responded with "OK" message which shows that interface is working. +CPIN: READY response shows that SIM card is ready and +CREG:0, 1 response shows that module is connected to network. Figures 15 shows in-vehicle unit.

B. Testing Tracking Server

In order to test server, an i7 PC was configured to act as a server. Apache server was installed to make it act like server. MySQL database server was installed.

C. Web Interface Testing

After logging to the website it displays vehicle location placed on Google maps, and the status of vehicle sensors like door, and ignitions status. User can send commands to in-vehicle unit to restart it or to shut down the vehicle. Figure 16 shows web interface.

D. Twitter social network integration

Each in-vehicle unit has its account on twitter social network. in-vehicle unit update vehicle information on tracking server in frequent basis, and server return update vehicle location on twitter's vehicle profile. The following figure 17, shows vehicle account on twitter social network



Fig. 15. In-Vehicle Unit

V. FUTURE WORK

System can get vehicle speed from some sensor installed in the vehicle, then posts remaining time for the next stop in social network. That would be a good value added for public/private transportation services. System can also analyze the time between stops, then report the traffic flow status on social network.

VI. CONCLUSION

This paper propose a new vehicle tracking and security system, that make use of social network as a value added service for traditional tracking systems. For vehicle tracking in real time, in-vehicle unit and a tracking server is used. The information is transferred to Tracking server using GSM/GPRS module on GSM network by direct TCP/IP connection with Tracking server through GPRS. Vehicle information is recorded in tracking server database. This information like vehicle location (on google maps), and vehicle status (door, and ignition) is only available to authorized users of the system via web interface over the internet. User can send different commands to in-vehicle unit (restart, shut down) to remotely controls his vehicle, which can be used as vehicle security and tracking system. Tracking server posts vehicle location placed on a Google maps to vehicle's twitter account, which make the vehicle followers easily find targeted vehicle, which can be applied to public transportation tracking. Currently In-Vehicle unit was implemented with Arduino. Microcontroller board which is connected to GM862-GPS through extension board named GM862-GPS shield

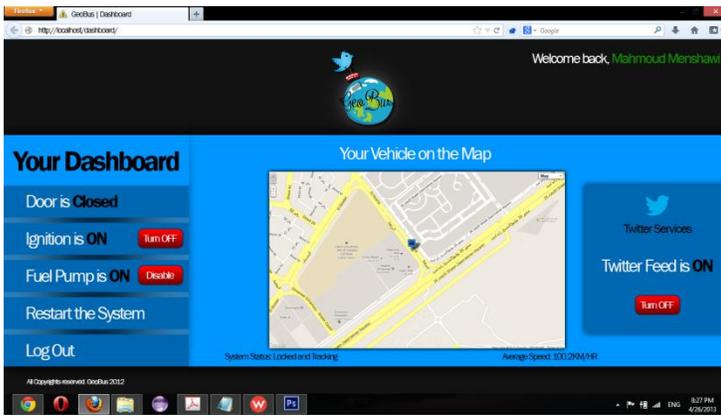


Fig. 16. proposed web interface



Fig. 17. Twitter posts of vehicle location “#GeoBus”

ACKNOWLEDGMENT

This project won the 1st prize of ACU “Make it Tweet” competition Fall 2011 which was organized by Ahram Canadian university. Then won the 1st prize of “Microsoft Geeks” competition 2012, organized by Microsoft Egypt, and hosted in Ahram Canadian University, Spring 2012.

The following video stream show development and operating process of prototype.

<http://youtu.be/mVTgX3bpx9Y>

<http://youtu.be/cxjfSAub1iI>

REFERENCES

- [1] Muruganandham, P.R.Mukes, “Real Time Web based Vehicle Tracking using GPS”, World Academy of Science, Engineering and Technology 37 2010
- [2] boyd, d. m., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), article 1.
- [3] "Transportation District's Automatic Vehicle Location System" [online:] http://www.itdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/13589.html
- [4] "Vehicle Tracking Systems Overview" [Online:] <http://www.roseindia.net/technology/vehicltracking/VehicleTrackingSystems.shtml>
- [5] J.B. TSUI, "Fundamentals of Global Positioning System Receivers", 1st Edition. John Willey & Sons Inc., 2000.
- [6] "Arduino microcontroller", [online:] <http://Arduino.cc>
- [7] "Telit Wireless Solutions", GM862-GPS Modem, 2008.
- [8] "Telit Wireless Solutions" "GM862-GPS Hardware user guide". 1vv0300728 Rev. 8 - 20/09/07
- [9] "GPRS (General Packet Radio Service), HSCSD & EDGE" [online:] <http://www.mobile-phones-uk.org.uk/gprs.htm>
- [10] R. Parsad, M. Ruggieri, "Applied Satellite Navigation Using GPS, GALILEO, and Augmentation Systems", London, ARTECH HOUSE, 2005.
- [11] T. Halonen et al, "GSM, GPRS and EDGE Performance", 2nd Edition, Chichester, John Willey & Sons Ltd., 2003.

An Efficient Approach for Image Filtering by Using Neighbors pixels

Smrity Prasad¹

Research Scholar, Department of Computer Science
Christ University
Bangalore, Karnataka, India

N.Ganesan²

Director (MCA)
RICM, Padmanabav Nagar
Bangalore, Karnataka, India

Abstract—Image Processing refers to the use of algorithm to perform processing on digital image. Microscopic images like some microorganism images contain different type of noises which reduce the quality of the images. Removing noise is a difficult task. Noise removal is an issue of image processing. Images containing noise degrade the quality of the images. Removing noise is an important processing task. After removing noise from the images, the visual effect will not be proper. This paper presents an approach to de-noise based on averaging of pixels in 5X5 window is proposed.

Keywords—Salt & Pepper Noise; Filter; PSNR; MSE

I. INTRODUCTION

Images of microorganism are extensively used in the area of medicine and biotechnology. Microorganism image analysis is having very important role in modern diseases diagnosis. The study of microorganism needs identification of different type of microorganism. For that qualitative analysis is required. By the term qualitative analysis mean the differentiation of different type of microorganism that are present in industrial sludge. In microscopic image capturing, impulse noise is caused due to environmental conditions, system noise, and motion of the object and so on, there will be difference between the original image and the resulting image. Impulse Noise must be removed for its improvement so that real information about image will be obtained for special purpose. There are two types of impulse noise (i) salt and pepper noise (ii) random valued noise. Salt and Pepper Noise can have values either 0 or 255 but random valued impulse noise can have any value from 0 to 255[2]. There are number of algorithms for noise removal [1]-[5].

In this paper, a simple method of removal of impulse noise for gray scale image is presented. The proposed method includes two steps 1) Detection of noisy pixels and noise free pixels 2) Filtering of noisy pixels. Here noisy pixel and noise free pixels are separated based on averaging of neighborhood pixels along each direction. After that noisy pixels are removed and replaced by the pixel using adaptive median. Here optical microscope (400X) image of Cyanobacteria with a size of 583 X 345 has been taken for analysis.

The rest of the paper is organized as follows:-

In the second section the impulse noise is described. In the third section detection algorithm and reduction algorithm is described and in fourth section assessment parameter is

discussed. Experimental result and discussion is presented in section 5. Section 6 contains the conclusion.

II. IMAGE IMPULSE NOISE

The Image impulse noise is a very common noise in communication [7, 8]. Let $x_{i,j}$ be the grey level of noisy image x at (i, j) and can be described as follows:-

$$x_{i,j} = \begin{cases} b_{i,j} & \text{with probability } p \\ f_{i,j} & \text{with probability } 1-p \end{cases} \quad (1)$$

Where $b_{i,j} \in [W_{\min}, W_{\max}]$ is the noisy pixel at location (i,j) with probability P . where W_{\min} and W_{\max} be the maximum and minimum intensity value. $f_{i,j}$ is the noise free pixel with probability $(1-P)$.

Impulse noise alters at random the value of some pixels. In Binary image some white pixel become black and some black pixel become white [4]. In binary image this means that some black pixels become white and white pixels become black. This is also called salt and pepper noise.

III. PROPOSED ALGORITHM

A. Detection Algorithm

In this paper, algorithm based on averaging of pixels in 5x5 windows is proposed. There will be four main directions that will include 7 pixels as shown in the figure 1. An edge aligned with each direction is considered separately. Pixels aligned with each direction will be considered to find average. There are four steps in detection algorithm and is followed.

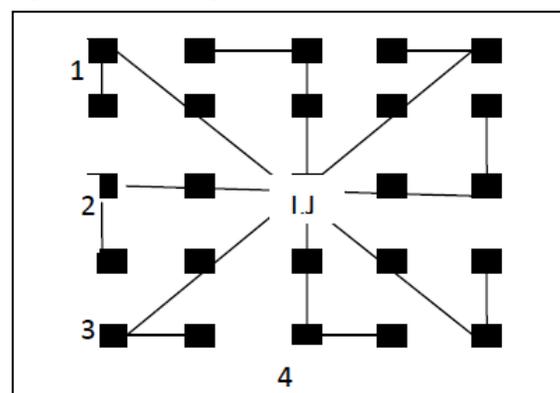


Fig. 1 Four Directional Pixels in the 5x5 window

1) Let R_k ($k=1$ to 4) denotes a set of seven pixels in k th direction, origin at (i, j) i.e.

$$\begin{aligned} R_1 &= \{(i-1, j-2)(i-2, j-2)(i-1, j-1)(i, j)(i+1, j+1)(i+2, j+2)(i+1, j+2)\} \\ R_2 &= \{(i+1, j-2)(i, j-2)(i, j-1)(i, j)(i, j+1)(i, j+2)(i-1, j+2)\} \\ R_3 &= \{(i+2, j-1)(i+2, j-2)(i+1, j-1)(i, j)(i-1, j+1)(i-2, j+2)(i-2, j+1)\} \\ R_4 &= \{(i+2, j+1)(i+2, j)(i+1, j)(i, j)(i-1, j)(i-2, j)(i-2, j-1)\} \end{aligned}$$

2) Detection of pixels as noise candidates or noise free is done by temple window of size 5×5 centered at (i, j) . The center pixel $X_{i,j}$ is considered as noisy by comparing the maximum and minimum intensity value in the 5×5 temple window. The algorithm first gets the minimum and maximum intensity value in the temple window 5×5 of the central pixel. If the test pixel lies within the range of its neighbor it is considered as non impulsive otherwise it is considered as noisy pixel. Let S be the set of noise free pixel and NP is the set of noisy pixels. W_{min} and W_{max} be the maximum and minimum intensity value.

$$X_{i,j} \in \begin{cases} NP & W_{min} \geq X_{i,j} \geq W_{max} \\ S & W_{min} < X_{i,j} < W_{max} \end{cases} \quad (2)$$

Once the noise free candidates are identified, they are separated and noisy pixels are separated.

For NP , algorithm goes second level detection.

3) For all noisy candidates, in each direction shown in the figure 1, average of the absolute difference between two closest pixels from the center pixel is denoted by A_{mcl} . Average of absolute difference between two far pixels from the center pixel is denoted by A_{mfr} . Average of absolute difference between two corner pixels from the center pixel is denoted by A_{mcr} .

$$A_{mcl} = \frac{1}{2} \sum_{k=1}^2 W_{kclm} \quad (3)$$

where $1 \leq m \leq 4$

And

$$\begin{aligned} W_{1cl_1} &= |X_{i,j} - X_{i-1,j-1}|, W_{2cl_1} = |X_{i,j} - X_{i+1,j+1}| \\ W_{1cl_2} &= |X_{i,j} - X_{i,j-1}|, W_{2cl_2} = |X_{i,j} - X_{i,j+1}| \\ W_{1cl_3} &= |X_{i,j} - X_{i+1,j-1}|, W_{2cl_3} = |X_{i,j} - X_{i-1,j+1}| \\ W_{1cl_4} &= |X_{i,j} - X_{i+1,j}|, W_{2cl_4} = |X_{i,j} - X_{i-1,j}| \\ A_{mfr} &= \frac{1}{2} \sum_{k=1}^2 W_{kfr_m} \end{aligned} \quad (4)$$

where $1 \leq m \leq 4$

And

$$W_{1fr_1} = |X_{i,j} - X_{i-2,j-2}|, W_{2fr_1} = |X_{i,j} - X_{i+2,j+2}|$$

$$\begin{aligned} W_{2fr_1} &= |X_{i,j} - X_{i,j-2}|, W_{2fr_2} = |X_{i,j} - X_{i,j+2}| \\ W_{1fr_3} &= |X_{i,j} - X_{i+2,j-2}|, W_{2fr_3} = |X_{i,j} - X_{i-2,j+2}| \\ W_{1fr_4} &= |X_{i,j} - X_{i+2,j}|, W_{2fr_4} = |X_{i,j} - X_{i-2,j}| \\ A_{mcr} &= \frac{1}{2} \sum_{k=1}^2 W_{kcr_m} \end{aligned} \quad (5)$$

Where $1 \leq m \leq 4$

And

$$\begin{aligned} W_{1cr_1} &= |X_{i,j} - X_{i-2,j-2}|, W_{2cr_1} = |X_{i,j} - X_{i+1,j+2}| \\ W_{1cr_2} &= |X_{i,j} - X_{i+1,j-2}|, W_{2cr_2} = |X_{i,j} - X_{i-1,j+2}| \\ W_{1cr_3} &= |X_{i,j} - X_{i+2,j-1}|, W_{2cr_3} = |X_{i,j} - X_{i-2,j+1}| \\ W_{1cr_4} &= |X_{i,j} - X_{i+2,j+1}|, W_{2cr_4} = |X_{i,j} - X_{i-2,j-1}| \end{aligned}$$

$$4) \quad r_{i,j} = \text{mean} \{A_{mcl}, A_{mfr}, A_{mcr}\} \quad (6)$$

where $0 \leq r_{i,j} \leq 255$

For an image the pixels in the set NP are considered as noisy pixels based on the value $r_{i,j}$. For an image with grey label in the interval $(0, 255)$, the pixel will be noisy if $r_{i,j}$ is in between 230 and 255. When $r_{i,j}$ is less than 230, the pixel is not noisy. In the case of an image with grey label $(0, 1)$, $r_{i,j}$ should be less than 0.90 for noiseless pixel. So complete detection rule as

$$X_{i,j} \in \begin{cases} NP & \text{if } 230 \leq r_{i,j} \leq 255 \\ S & \text{Otherwise} \end{cases} \quad (7)$$

B. Reduction Algorithm

The signal pixels are kept same and only noisy pixels are corrected. There are number of filtering methods which can be adopted. When the noisy pixels are identified, they should be filtered. In this paper filtering is done as follows. Here adaptive median filter is used to remove noise.

If the processing pixel is noisy, it should be replaced by median of $N \times N$ window. But it may be possible that median itself will be noise i.e. maximum or minimum point and if this is the case then window size should be increased by 2 and median is calculated. This process will go on until the maximum window size is reached. So filtering process will be as follows

$$y_{i,j} = \begin{cases} \text{adpmed} & \text{if } x_{i,j} \in NP \\ x_{i,j} & \text{if } x_{i,j} \in S \end{cases} \quad (8)$$

Where adpmed is adaptive median filter.

IV. ASSESSMENT PPARAMETER FOR ANALYZING THE OUTPUT OF THE ALGORITHM

There are number of parameters such as Noise Standard Deviation (NSD), Mean Square Error (MSE), Equivalent Numbers of Looks (ENL), and Peak Signal to Noise the algorithm.

A. Mean Square Error(MSE)

The Mean Square Error is used to find the total amount of difference between two images. It indicates average difference average difference of the pixels of throughout the image where K is the de noised image and I is the original image with noise. A lower MSE indicates that there is small difference between the original image with noise and de noised image. The formula is

$$MSE = 1/mn \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i, j) - K(i, j))^2 \quad (9)$$

B. Peak Signal to Noise Ratio

To assess the performance of the noise removal method, PSNR is used. The formula is

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

V. RESULT AND DISCUSSION

The microscopic image of Cyanobacteria with a size of 583 X 345 has been corrupted by salt and pepper noise at different density. In this section result are presented to illustrate the performance of proposed algorithm. An original noise free image shown in figure 2 is given as reference. A quantitative comparison is performed between different techniques in terms of PSNR. Figure 3 shows the result of Cyanobacteria corrupted by noise at different density. Noise of different densities ranging from 30% to 90%.The proposed method has been compared with simple median, progressive median and 3X3 median filter. Progressive median and 3x3 median filter is giving better result compare to simple median filter. Noisy image is filtered using proposed algorithm and result is shown in the figure 3, 4,5,6,7. Figure 3 is the image of Cyanobacteria which is corrupted by salt and pepper noise of different density. Figure 4 is filtered image of Cyanobacteria on which simple median filter is implemented. .Figure 5 is filtered image of Cyanobacteria by progressive median filter. .Figure 6 is filtered image of Cyanobacteria by 3x3 median algorithm. Figure 7 is filtered image of Cyanobacteria by proposed algorithm. It can be seen that result using the proposed method are significantly better than other three methods when noise density is more than 30%.The results are measured quantitatively using PSNR.Table 1 shows the comparison table of PSNR of different techniques.

Figure 8 show the comparison graph of PSNR of different techniques for Cyanobacteria.

VI. CONCLUSION

Here an efficient approach for impulse noise removal is proposed. The algorithm goes in two stages. Stage one identifies noisy and noise free pixels. This stage separates those two sets of pixels.

Again in these stage noisy pixels is considered as undetected pixels and goes for second level detection. Second stage does filtering to restore the image. The noisy pixels are replaced by adaptive median which is calculated recursively by increasing the size of the window up to limited size of window. It shows that the method proposed in the paper is effective for microbiologist in digital image processing. With experimental result it is seen that proposed algorithm gives good result for noise removal, edge preservation and image detail preservation. The peak signal to noise ratio also shows improvement as compared to other methods.

TABLE I. Comparison of PSNR of Different Techniques for Cyanobacteria

Noise Density	Simple Median Filter	Progressive Median Filter	Algorithm With 3X3 window	Proposed Algorithm
30	29.3076	32.5432	32.5632	32.6886
50	19.7264	24.3708	24.3708	24.3809
60	14.0519	22.9781	23.002	23.7285
80	10.6808	18.7064	19.0809	19.8350
90	8.7102	16.4250	17.5643	19.2911
95	6.4048	15.0521	16.0008	18.6506

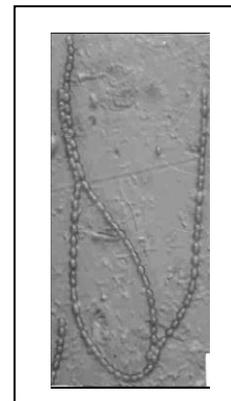


Fig. 2 Original microscopic image of Cyanobacteria.

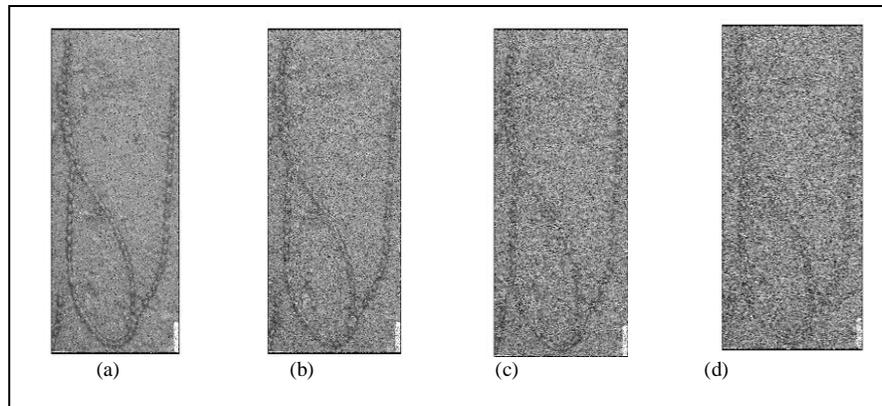


Fig. 3 Image Cyanobacteria corrupted by salt & pepper noise. (a) Noise Density 30%, (b) Noise Density 60%, (c) Noise Density 80%, (d) Noise Density 90%

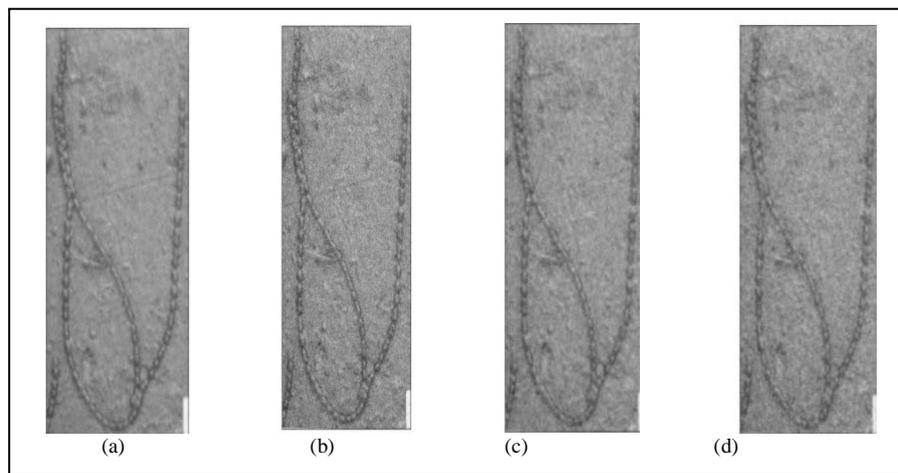


Fig. 4 De-noising by Simple Median filter (a) De-noising image of figure 3(a), (b) De-noising image of figure 3(b), (c) De-noising image of figure 3(c), (d) De-noising image of figure 3(d)

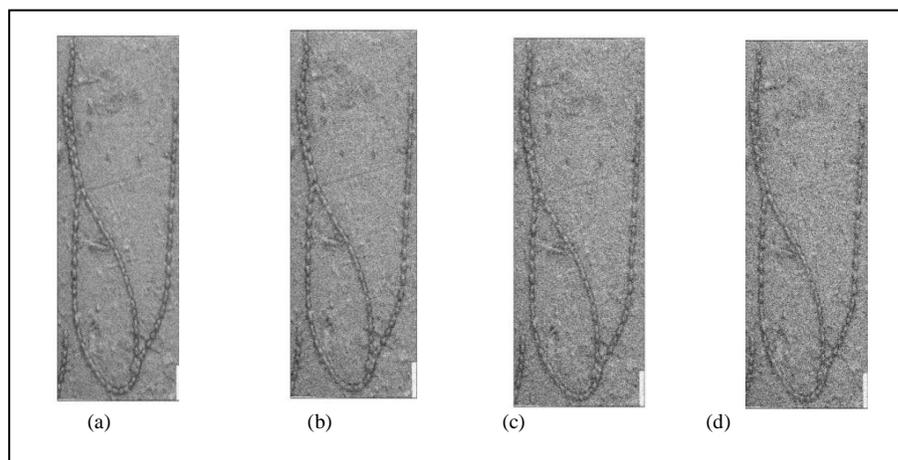


Fig. 5 De-noising by Progressive median (a) De-noising image of figure 3(a), (b) De-noising image of figure 3(b), (c) De-noising image of figure 3(c), (d) De-noising image of figure 3(d)

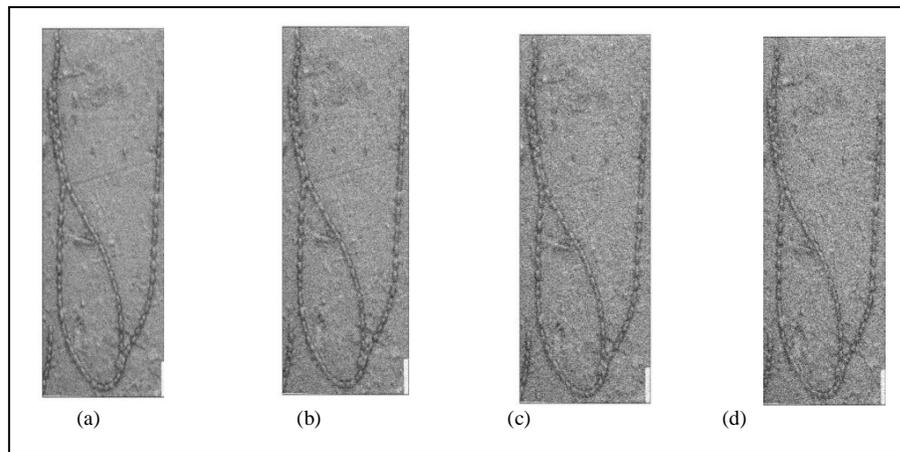


Fig. 6 De-noising by 3X3 median (a) De-noising image of figure 3(a) ,(b) De-noising image of figure 3(b) ,(c) De-noising image of figure 3(c) , (d) De-noising image of figure 3(d)

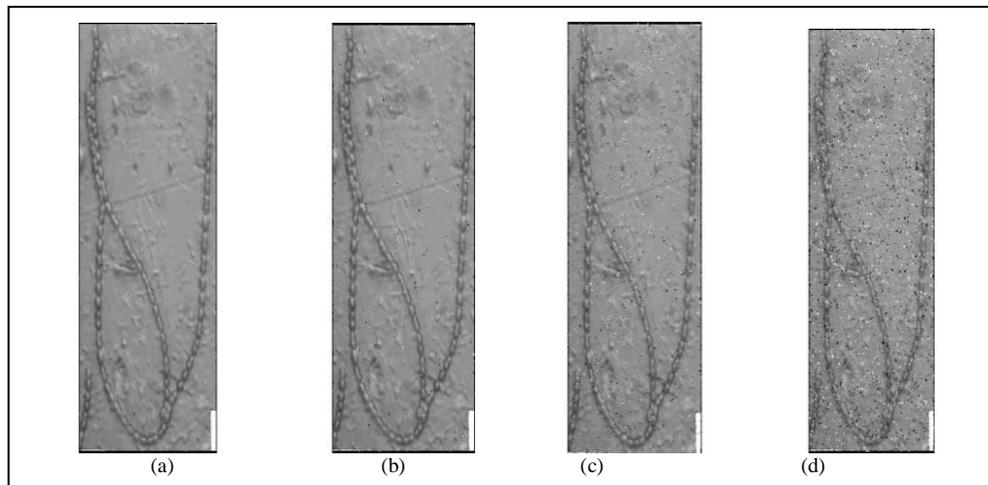


Fig. 7 De-noising by Proposed Algorithm (a) De-noising image of figure 3(a) ,(b) De-noising image of figure 3(b) ,(c) De-noising image of figure 3(c) , (d) De-noising image of figure 3(d)

REFERENCES

- [1] Zhou Wang and David Zhang, "Progressive Switching Median Filter for the Removal of Impulse Noise from Highly Corrupted Images" IEEE Transaction on Circuits and Systems, vol.46, pp.78-80 January, 1999.
- [2] Raymond H. Chan, Chung-wa Ho and Mila Nikolova, "Salt and pepper noise removal by median-type noise detectors and detail preserving regularization," IEEE Transactions on image Processing, vol.14, no 10, pp. 1479-1485, Oct 2005.
- [3] Eduardo Abreu, Michael Lightstone, Sanjit K. Mitra and Kaoru Arakawa, "A New Efficient Approach for the Removal of Impulse Noise from Highly corrupted Images," IEEE Transactions Image Processing, vol. 5 ,no.6, pp. 1012-1025, June 1996.
- [4] RCGonzalez and R.E.Woods, "Digital Image Processing" Prentice Hall, 2002.
- [5] T. Chen and H.R. Wu, "Application of partition based median type filters for suppressing noise in images," IEEE Transactions Image Processing, vol.10, no.6, pp.829-836, June.2001.
- [6] J. Astola and P.Kousmanen, Fundamentals of Nonlinear Digital Filtering. CRC Press, 1997.
- [7] L. Ilizzo and L. Paura, "Error probability for fading CPSK signals in gaussian and impulsive atmospheric noise environments," IEEE Transactions on Aerospace and Electronic Systems, vol. 17, no. 5, pp.719-722, Sep. 1981.
- [8] G. A. Tsihrintzis and C. L. Nikias, "Performance of optimum and suboptimum receivers in the presence of impulsiveness modeled as an alpha-stable process," IEEE Transactions on communications, vol. 43, no. 234, pp.904-914, Feb./Mar./Apr. 1995
- [9] Sun and Y. Neuvo, "Digital-preserving median based filters in image processing." Pattern Recognit. Lett., vol. 15, pp. 341-347, Apr 1994.
- [10] F. Russo and G. Ramponi, "A fuzzy filter for images corrupted by impulse noise," IEEE Signal Processing Lett., vol. 3, pp. 168-170, June 1996.
- [11] H. Kong and L. Guan, "A noise-exclusive adaptive filtering framework for removing impulse noise in digital images," IEEE Trans. Circuits Syst. II, vol. 45, pp. 422-428, Mar. 1998.

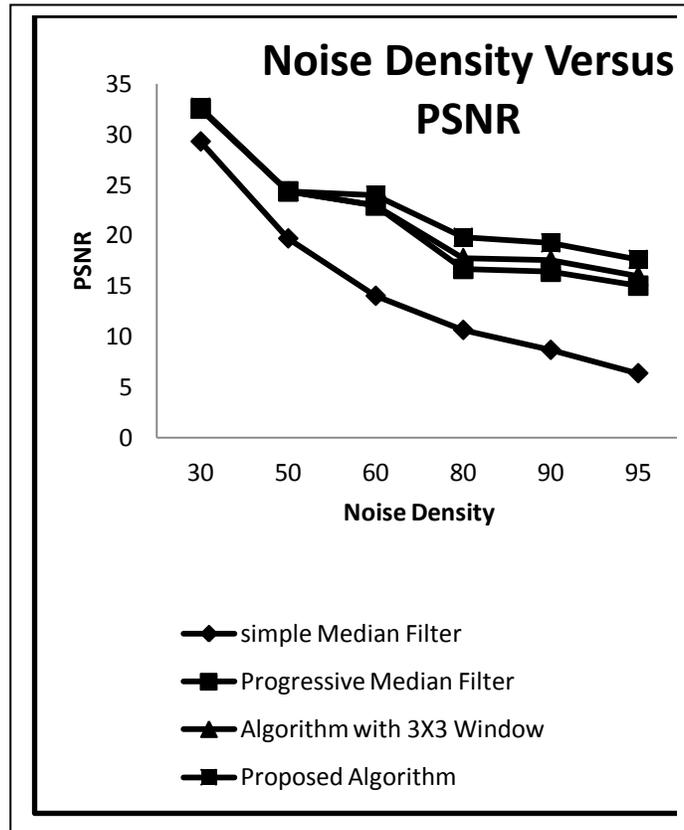


Fig. 8 Comparison graph of PSNR at different noise density for different techniques

- [12] T. S. Huang, G. Y. Tang, "Fast two dimensional median filtering algorithm," IEEE Transactions on Acoustic, speech, and Signal Processing, 1(1979), pp. 13-18.
- [13] Y. Dong, R. H. Chan, and S. Xu, "A detection statistics for random valued impulse noise," IEEE Trans. Image Process., vol. 16, no. 4, pp. 1112-1120, April 2007.
- [14] F. Cai, R. H. Chan, and M. Nikolova, "Fast two phase image deblurring under impulse noise," J. Math. Imag. Vis., vol. 36, no. 1, pp. 46-53, 2010.
- [15] Changhong Wang, Taoyi Chen and Zhenshen Qu "A Novel Improved Median Filter for Salt and Pepper Noise from Highly corrupted Images" in proc. System and Control in Aeronautics and Astronautics (ISSCAA), 2010 p.718-722.
- [16] GAI Qiang. Research and application on the theory of local wave time frequency analysis method [D]. Dalian: Dalian University of Technology, 2001.
- [17] U. Ranjith, P. Caroline, H. Martial. Toward Objective Evaluation of Image Segmentation Algorithms. IEEE Trans P.A.M.I., vol. 29, no. 6, pp. 929-944, 2007.
- [18] A. Mike Burton, Rob Jenkins, Robust representations for face recognition: The power of averages, Cognitive Psychology, vol. 51, no. 3, pp. 256-284, 2005.
- [19] Forouzan, A.R. Araabi, B.N., "Iterative median filtering for restoration of images with impulsive noise", Electronics, Circuits and Systems, vol. 1, pp. 232-235, 2003.
- [20] D.R.K Brownrigg, "The weighted median filter", Communications of the ACM, vol. 27, no. 8, pp. 807-818, August 1984.
- [21] J. K. Mandal and Somnath Mukhopadhyay, "A Novel Technique for Removal of Random Valued Impulse Noise using All Neighbor Directional Weighted Pixels (ANDWP)", International Conference on Parallel, Distributed Computing Technologies and Applications, PDCTA 2011, Communications in Computer and Information Science, Springer, vol. 203, pp. 102-111, September 2011.
- [22] S. Balasubramanian, S. Kalishwaran, R. Muthuraj, D. Ebenezer, V. Jayaraj, "An Efficient Non-linear Cascade Filtering Algorithm for Removal of High Density Salt and Pepper Noise in Image and Video sequence", In Intl. Conf. on Control, Automation, Communication and Energy Conservation, 2009.
- [23] M. Mahmoudi, G. Sapiro, "Fast Image and Video denoising via Nonlocal Means of Similar Neighborhoods", IEEE Signal Processing Letters, Vol. 12, No. 12, 2005, pp. 839-842.
- [24] S.Q. Yuan, Y.H. Tan, "Difference-Type noise detector for adaptive median filter", IEEE Electronic Letters, Vol. 42, No. 8 2006.
- [25] A. Sawant, H. Zeman, D. Muratore, S. Samant, and F. DiBianka, "An adaptive median filter algorithm to remove impulse noise in X-ray and CT images and speckle in ultrasound images," Proc. SPIE, vol. 3661, pp. 1263-1274, Feb. 1999.
- [26] J.-H. Wang, "Prescanned minmax center-weighted filters for image restoration," Proc. Inst. Elect. Eng., vol. 146, no. 2, pp. 101-107, 1999.

A Comparative Study of Three TDMA Digital Cellular Mobile Systems (GSM, IS-136 NA-TDMA and PDC) Based On Radio Aspect

Laishram Prabhakar

Manipur Institute of Management Studies
Manipur University
Canchipur, Manipur, India

Abstract—As mobile and personal communication services and networks involve providing seamless global roaming and improve quality of service to its users, the role of such network for numbering and identification and quality of service will become increasingly important, and well defined. All these will enhance performance for the present as well as future mobile and personal communication network, provide national management function in mobile communication network and provide national and international roaming. Moreover, these require standardized subscriber and identities. To meet these demands, mobile computing would use standard networks. Thus, in this study the researcher attempts to highlight a comparative picture of the three standard digital cellular mobile communication systems: (i) Global System for Mobile (GSM) -- The European Time Division Multiple Access (TDMA) Digital Cellular Standard, (ii) Interim Standard-136 (IS-136) -- The North American TDMA Digital Cellular Standard (D-AMPS), and (iii) Personal Digital Cellular (PDC) -- The Japanese TDMA Digital Cellular Standard.

Keywords—Comparative Study; GSM; IS-136 TDMA; PDC; Radio Aspect.

I. INTRODUCTION

In the last two centuries, mobility has been redefined such that both physical and virtual object are now mobile. The foundation of mobility of information was laid by Joseph Henry (1797-1878) who invented the techniques for distant communication. Later Samuel F.B Morse used the same property of electricity to invent the Telegraph. In 1876, Alexander Graham Bell laid the foundation by making the first voice call over wire i.e. "Mr. Watson, come here, I want to see you."

After the launch of Sputnik by USSR in 1957, US formed Advance Research Project Agency (ARPA) and laid the foundation of packet switched data networks. With the evolution of computer and the packet switched network movements, now byte has move to a new state of maturity. The convergence of telecommunication and information technology in 1965 leads to Information Communication Technology (ICT) that addresses the need to accommodate data formation and knowledge from anywhere and anytime.

Due to the achievement and advancement of the ICT, mobile computing has become a very important part. Now it can be defined as a computing environment over physical mobility. In this environment, a user has the capacity to

perform anywhere, using a computing device, in the public, corporate and personal information spaces. While on the move, the preferred device could be a mobile device, and back home or in the office, a desktop computer could be preferred. Nevertheless computing should be through wired and wireless media -- be it for the mobile workforce, holidaymakers, enterprises or rural population. The access to information and virtual objects through mobile computing are absolutely necessary for optimal use of resource and increased productivity. Thus, mobile computing is used in different contexts such as virtual home environment and nomadic computing.

II. MOBILE COMMUNICATION SYSTEMS

It has been known for centuries that knowledge is power but in this information age, communication is becoming the real power. In this present age, mobile communication takes a great role. For example, a modern aircraft with 800 seats already offer limited Internet access. However, aircraft of the next generation would offer easy Internet access. In this scenario, a mobile network moving at high speed above the ground with a wireless link will be the only means of transporting data to and from passengers. Again the underlying vision for the emerging mobile and personal communication service and the system is to enable communication any time, any place and in any form. Thus, for seamless communication, personal communication cover terminal mobility provided by wireless access, personal mobility based on personal number and service portability through use of intelligent network capabilities.

Terminal mobility systems are characterized by their ability to locate and identify a mobile terminal or it moves and allows the mobile terminal to access telecommunication services from any location and even while on the move. In this scenario, communication is always between the network and the static terminal. So call delivery and billing are always based on terminal identity and mobile station number.

In personal mobility, the relation between the terminal and the user is dynamic and the call delivery and billing can be based on a personal identity (personal no) assign to the user. This is characterized by the ability to identify end user on the move and allows end users to originate, receive calls, and access subscribed telecommunication services, in any location. It is applicable to both the wired and the wireless network.

In service portability, it refers to the capability of a network provided subscribed services at the terminal or location designated by the user. It is accomplished through the use of IN concept whereby the user service profile can be maintained in a suitable data base and the user can access, interrogate and modify to manage and control subscribed services.

III. TIME DIVISION MULTIPLE ACCESS DIGITAL CELLULAR MOBILE SYSTEM

A. Global System for Mobile

Development of GSM started in 1982 within European Conference of Postal and Telecommunications Administrations (CEPT) for a future pan-European Cellular system. This was designed to replace the incompatible analog systems. The development was transferred to European Telecommunications Standards Institute (ETSI) in 1989 and the phase 1 standards were frozen in 1990. The first commercial GSM service was launched in 1992 and the first GSM-1800 is also called DCS (Digital Cellular Service) DCS1800 was launched in September 1993. GSM standardization continues with Phase 2 standards completed in 1995. The enhancement of GSM services from the original concept of a pan-European standard shows that GSM was an attractive option to operators around the world including USA, and has become the number one digital cellular standard.

B. Interim Standard-136: North American Time Division Multiple Access

North American TDMA, often referred as TDMA, was developed in response to the need to increase cellular capacity. Unlike Europe and Japan where additional spectrum was made available for second generation digital systems, US operators were constrained to re-use the same spectrum used for Advance Mobile Phone System (AMPS). As a result the TDMA standard was developed to be compatible with the analogue AMPS system. Again the pressure on capacity forced the Telecommunications Industry Association (TIA) to consider a rapid development of a digital standard. As a result two TDMA standards were developed. Interim Standard – 54 (IS-54) often referred to as Digital-AMPS (D-AMPS), was the first of these. It shares the same 21 analogue call set-up channels with AMPS so that the call processing is the same between the two systems and handsets can support dual AMPS/ D-AMPS. The second phase standard is IS-136 which implements digital call set-up channels to enable stand-alone TDMA handsets. IS-136 has effectively replaced IS-54.

C. Personal Digital Cellular

Personal Digital Cellular (PDC) is a second-generation technology used in digital cellular telephone communication in Japan. It uses a variation of TDMA which divides each cellular channel into individual time slots in order to increase the amount of data that can be carried. PDC is currently used only in Japan, with the first systems introduced by Nippon Telegraph and Telephone (NTT) DoCoMo in 1991 as a replacement for the earlier analog networks. It operates in the 800MHz and 1,500MHz bands, making very efficient use of the available bandwidth. With bandwidth demand so high in Japan, the system can operate in two modes: full rate and half rate. Half-rate channels have reduced speech quality and data

transmission rates, but allow more channels to occupy the same bandwidth. Subscriber numbers are so high in Japan that, although PDC is only operational in this one country, it accounted for 12% of global digital subscriptions in December 1999.

Along with the other mobile communication standards, PDC can be developed along a gradual evolutionary path to the global International Mobile Telecommunications -2000 (IMT-2000) standards. Indeed, one of the IMT-2000 technologies, Wideband Code Division Multiple Access (WCDMA), is going through initial testing in Japan.

IV. NEED OF THE STUDY

As mobile and personal communication services and networks involve providing seamless global roaming and improve quality of service to its users, the role of such network aspect as numbering and identification and quality of service will become increasingly important and well defined. To provide national management function in mobile communication network and provide national and international roaming, well defined standardized subscriber and identifies are required. To meet these demands, mobile computing will use standard networks. Some of the standard digital cellular mobile communication systems are (i) Global System for Mobile (GSM) -- The European Time Division Multiple Access (TDMA) Digital Cellular Standard, (ii) IS-136 -- The North American TDMA Digital Cellular Standard (iii) Personal Digital Cellular (PDC) -- The Japanese TDMA Digital Cellular Standard and (iv) IS-95 -- The North American Code Division Multiple Access (CDMA) Digital Cellular Standard. The cellular industry continues to experience massive growth. While there remains a large subscriber base for analog systems, most of the recent growth has been on digital systems. So this is an attempt in trying to enhance knowledge of the networks by comparing the networks based on radio aspects.

V. OBJECTIVE OF THE STUDY

The main objective of this paper is to study the three TDMA based Cellular mobile system namely (i) GSM: The European TDMA Digital Cellular Standard, (ii) IS-136: The North American TDMA Digital Cellular Standard (D-AMPS) (iii) PDC: The Japanese TDMA Digital Cellular Standard and to prepare a comparative analysis based on the Radio Aspects.

VI. COMPARATIVE ANALYSIS

The International Telecommunication Union (ITU), which manages the international allocation of radio spectrum, allocated the bands 890-915 MHz for the uplink and 935-960 MHz for the downlink for mobile in Europe. Since this range was already used in the early 1980s by the analog systems of the day, the CEPT had the foresight to reserve the top 10MHz of each band for the GSM network that was still being developed. Eventually, GSM will be allocated to the entire 2 x 25 MHz bandwidth.

On the other hand the radio technology used in the IS-136 system provides a channel for advance services and improved system efficiency through the use of voice digitization, speech compression (coding), efficient radio modulation, enhanced radio frequency (RF) power control, and a flexible approach to

spectrum usage. D-AMPS will utilize the currently allocated spectrum for analog AMPS that is a total of 50 MHz (uplink) and 869-894 MHz (downlink), with each frequency channel to 30 KHz spacing. Each frequency channel then is time-multiplexed with a frame duration of 40ms, which is partitioned into six slots of 6.67 ms duration.

PDC is the most spectrally efficient of TDMA technologies, with six half-rate or three full-rates channels possible in a 25 kHz frequency space, compared to three channels in 30 kHz in IS-136 and eight channels in 200 kHz for GSM. It even compares favorably to Code Division Multiple Access (CDMA), using spread-spectrum technology to allow up to 131 channels in a 1,250 kHz spectrum band.

Full-rate speech normally requires a digital data transfer rate of 9.6kbps (kilobits per second), as is used in GSM, and TDMA IS-136 networks. PDC offers two alternative rates: 9.6kbps in full-rate channels or 5.6kbps in the half-rate channel. The quality of speech along a 5.6kbps connection is significantly lower than the standard 9.6kbps connection, but is a useful trade-off with the number of channels available.

About the advantages of PDC, the newly developed Linearized Saturated Amplifier with Bidirectional Control (LSA-BC) improves efficiency. Although it is a saturated amplifier, the voltage controlled power supply results in linear operation. Coherent detection with Adaptive Carrier Tracking (ACT) has been developed for digital systems. ACT gives excellent performance under fast Rayleigh fading because of fast carrier tracking ability. ACT is made by all digital circuits so that adjustment are not necessary and power dissipation is very low (1/40 that of conventional differential demodulator). A direct modulator made of only one differential amplifier is advantageous in terms of compactness and power consumption (1/8 that of previously developed modulator IC's).

The comparison of three cellular digital systems are done through a number of sub parameters like downlink frequency, uplink frequency, symbol rate frequency channel, modulation, access etc. and shown in the table no.1.

It is seen that the number of channels/carriers of GSM is greater than IS-136 and PDC. It is also observed that PDC employs diversity reception in the mobile station which obviates the need for equalizers, which are an essential component of GSM. PDC uses a much lower transmission bit rate (42kb/sec vs. 270.83 kb/sec in GSM/48.6kb/sec in IS-136) which leads to better spectrum utilization, higher capacity and lower cost. PDC has significant similarities with the IS-136 D-AMPS system.

The call sequence of GSM is similar to that of IS-136. However, it uses gateway Mobile Switching Centre (MSC) or International Mobile Subscriber Identity (IMSI). The call sequence of IS-136 is similar to that of GSM but recently it does not use gateway MSC and IMSI. It uses mobile identification number (MIN). However, PDC uses gateway mobile control centre. The access signaling protocols in PDC are simpler and require fewer procedures and are simpler to use and lower in cost.

TABLE I. RADIO PARAMETER AND CHARACTERISTICS FOR GSM, IS-136 AND PDC

Parameter	GSM	IS-136	PDC
Downlink frequency (MHz)	935-960 1805-1880 1930-1990	869-894 1930-1990	810-826 1429-1453
Uplink frequency (MHz)	890-915 1710-1785 1850-1910	824-849 1850-1910	940-956 1477-1501
Symbol rate	271 ksymbols (271 kbit/s)	24.3 ksymbols (48.6 kbits/s)	21 ksymbols (42 kbits/s)
Frequency channels	200 KHz	30 KHz	25 KHz
Modulation	GMSK* (BT=0.3)	$\pi/4$ -DQPSK** ($\alpha=0.35$)	$\pi/4$ - DQPSK ($\alpha=0.5$)
Access	FH/TDMA	TDMA	TDMA
Channels/Carrier	8 Full Rate 16 Half Rate	3 Full Rate 6 Half Rate	3 Full Rate 6 Half Rate
Voice Codec	RPE-LTP*** 22.8kb/s full, 13kb/s Source, VSELP 11.4kb/s half	VSELP**** 13kb/s full, 7.95kb/s source, 6.5kb/s half	VSELP 11.2kb/s full, 6.7kb/s Source, 3.45kb/s half source
Max delay time	16 μ s	50 μ s	10 μ s

*GMSK: Gaussian Minimum Shift Keying

**DQPSK: Differential Quaternary Phase Shift Keying

***RPE-LTP: Regular Pulse Excitation with Long-Term Predictor

****VSELP: Vector Sum Excited Linear Prediction

VII. FINDING

From the comparative analysis of the three standards the following are the highlights of the findings:

A. GSM

1) From the original concept of a pan-European standard, it soon became clear that GSM was an attractive option to operators around the world including USA, and has become the number one digital cellular standard.

2) Low terminal and service cost is an attractive feature for the users.

3) It has also an emergency service, where the nearest emergency service provider is notified by dialing three digits similar to 911.

4) Interestingly the key drivers for GSM are pan-European roaming to offer compatibility throughout Europe and interaction with Integrated Services Digital Network (ISDN) and bill to home.

5) It has SIM (Subscriber Identity Module) card which contains the IMSI used to identify the subscriber to the system, a secret key for authentication, and other information. The International Mobile Equipment Identity (IMEI) and the International Mobile Subscriber Identity (IMSI) are independent, thereby allowing personal mobility. The SIM card may be protected against unauthorized use by a password or personal identity number.

6) In GSM, Regular pulse excitation – long term prediction (RPE-LTP) scheme is employed in order to reduce the amount of data sent between the mobile station and base transceiver station. In essence, when a voltage level of a particular speech sample is quantified, the mobile station's internal logic predicts the voltage level for the next sample. When the next sample is quantified, the packet sent by the Mobile Station (MS) to the Base Transceiver Station (BTS) contains only the error (the signed difference between the actual and predicted level of the sample).

B. IS-136

1) It share the same 21 analog call set up channels with AMPS so that the call processing is the same between the two systems and handsets can support dual AMPS/D-AMPS.

2) Its modulation is $\pi/4$ – DQPSK with ($\alpha=0.35$) so it provides lower envelopes fluctuation than standard DQPSK, allows non-coherent detection and performs well in a multipath environment.

3) The IS-136 system adds new power class of mobile phone to allow reduces the minimum cell site radius.

4) It supports a new function Mobile Assisted Channel Assignment (MACA) similar to Mobile Assisted Hand Over (MAHO). MACA is a process in which signal strength reporting takes place while mobile phone is monitoring a Digital Control Channel (DCCH) camping.

5) Each cell site in a cellular system has its own unique Digital verification color code (DVCC). A unique DVCC for each cell site ensures that the correct mobile phone is communicating with the proper station since frequencies are reused in most cellular systems.

6) It allows several types of phone identities. They are (a) TMSI – Temporary mobile station identity, (b) IMSI – International mobile subscriber number and (d) ESN – Electronic serial number

7) Vector sum excited linear prediction (VSELP) is a speech coding method used in IS-136 (D-AMPS). It was used in the first version of RealAudio for audio over the Internet. D-AMPS (IS-54 and IS-136) VSELP specifies an encoding of each 20 ms of speech into 159-bit frames, thus achieving a raw data rate of 7.95 kbit/s. In an actual TDMA cell phone, the vocoder output is packaged with error correction and signaling information, resulting in an over-the-air data rate of 16.2 kbit/s. For internet audio, each 159-bit frame is stored in 20 bytes, leaving 1 bit unused. The resulting file thus has a data rate of exactly 8 kbit/s.

C. PDC

1) It is found from the study that PDC is one of the most spectrally efficient of TDMA technologies. It has six half-rate or three full-rates channels possible in a 25 kHz frequency space, compared to three channels in 30 kHz in IS-136 and eight channels in 200 kHz for GSM.

2) PDC compares favorably to Code Division Multiple Access (CDMA), using spread-spectrum technology to allow up to 131 channels in a 1,250 kHz spectrum band.

3) PDC offers two alternative rates: 9.6kbps in full-rate channels or 5.6kbps in the half-rate channel. The quality of speech along a 5.6kbps connection is significantly lower than the standard 9.6kbps connection, but is a useful trade-off with the number of channels available.

4) The PDC network supports many advanced features in-line with the other second-generation technologies, such as text messaging and caller identification.

5) Utilizing its Intelligent Network (IN) capabilities, PDC also supports pre-paid calling, personal numbers, Universal Access Numbers, advanced charging schemes and wireless virtual private networks (VPNs).

6) As already mentioned PDC is a TDMA system and it operates by splitting each channel into several time slots and thereby allowing several users to use the same frequency channel. For each channel it is possible to support three users under normal circumstances. However when traffic levels are high it is possible to use half data rate speech. Although this reduces the speech quality, it enables six calls to be supported by each channel. This compares very favourably to GSM that manages eight within each 200 kHz channel.

7) Speech encoding is an important factor. PDC uses a different encoder to that used on IS54/IS136. The standard rate is 9.6 kbps along with similar technologies such as GSM, but when half rate encoding is used this falls to 5.6 kbps. Although this gives a significant reduction in voice quality, it is still adequate to maintain intelligibility and enables the network capacity to be increased to accommodate further calls.

8) It is also found that PDC has similar speech coding method like IS-136.

VIII. CONCLUSION

In this paper the researcher gives an overview of mobile communication and the standards of digital cellular especially for GSM: The European TDMA digital cellular standard, IS-136: TDMA based digital cellular system in United States and PDC: The Japanese TDMA based digital cellular system. Nowadays industry speeding up the development of mobile communication system where both voice and data services can be delivered regardless of location, network or terminal. The study clears that the three digital cellular systems have their own special features that satisfy diverse needs of mobile communication system.

REFERENCE

- [1] A Sokeke, K Talukda, "A text Book on Mobile Computing" Tata McGraw Hill, 2005
- [2] J. Schiller, "Mobile Communications" Pearson Education Limited, 3rd Edition, 2006
- [3] M. Mouly, M.B. Pautet, "The GSM System for Mobile Communications", published by M.Mouly et Marie-B. Pautet, Palaiseu, France, 1992
- [4] N. Spencer, "An Overview of Digital Telephony Standards", 1998 The Institution of Electrical Engineers . 1.1 Printed and Published by IEE, Savoy Place, London WCPR OBL, UK.
- [5] N. Nakajima, M Kuramoto and K. Kinoshita, "Development of a Digital Cellular Systems Using TDMA Technique", NIT Radio Communication Systems Laboratories, IT Mobile Communications Division, 1-2356 Take. Yokosukashi, 238-03 Japan

- [6] R. Ganesh and K. Pahlavan, "Wireless Network Deployments", Kluwer Academic Publishers, 2000
- [7] R. Pandya, "Mobile and Personal Communication Systems and Services", Prentice Hall of India, 4th Indian Reprint 2003
- [8] S. Sampei, "Application of Digital Wireless Technologies to Global Wireless Communication", Prentice Hall PTR, 1997
- [9] William C.Y. Lee, "Mobile Cellular Telecommunication – Analog and Digital Systems", MCGraw Hill International Edition, 1995.
- [10] Yi_Bing Lin, Imrich Chlamtac, "Wireless and Mobile Network Architecture", Wiley, 2002

Format SPARQL Query Results into HTML Report

Dr Sunitha Abburu,

Professor & Director, Dept of Computer Applications,
Adhiyamaan College of Engineering
Hosur, Tamilnadu, India

G.Suresh Babu

JRF, Dept of M.C.A
Adhiyamaan College of Engineering
Hosur, Tamilnadu, India

Abstract—SPARQL is one of the powerful query languages for querying semantic data. It is recognized by the W3C as a query language for RDF. As an efficient query language for RDF, it has defined several query result formats such as CSV, TSV and XML etc. These formats are not attractive, understandable and readable. The results need to be converted in an appropriate format so that user can easily understand. The above formats require additional transformations or tool support to represent the query result in user readable format. The main aim of this paper is to propose a method to build HTML report dynamically for SPARQL query results. This enables SPARQL query result display, in HTML report format easily, in an attractive understandable format without the support of any additional or external tools or transformation.

Keywords—SPARQL query; Oracle database 11g semantic store; Jena adapter; HTML report.

I. INTRODUCTION

The goal of semantic web [1] is to extend the current web standards and technology so that machine understands the web content. The knowledge representation technology used in the semantic web is ontology. An ontology is a common, shared and formal description of important concepts in specific domain [2]. Researchers have developed several ontology languages such as RDF, RDFS, and OWL etc. There are several query languages based on the ontology data format [3]. For example XPath and XQuery query languages for XML format, RDQL [4] and SPARQL [5] for RDF format and OWL-QL for OWL format of ontologies. Among the ontology query languages, SPARQL is one of the most efficient query languages for the Semantic Web [6] and it is recommend by W3C.

The SPARQL query language for RDF has several query result forms such as CSV, TSV [7] and XML etc. These formats are not clear to the user to realize and analyze query results. There is a need to represent SPARQL query result in an attractive format so that user can easily understand the SPARQL query results. The above SPARQL query result formats require additional conversions or tool support to represent query results in user readable format. This paper proposed a method to represent SPARQL query results in more attractive and user graspable format.

The rest of the paper is organized as follows. Section II describes survey on SPARQL query result formats. Section III defines the problem statement and over view of the proposed system architecture. Section IV describes the implementation

details of the proposed system. Section V shows results of the proposed method. Finally section VI concludes.

II. RELATED RESEARCH WORK

Most RDF stores use one of the common RDF query languages like RDQL [4] or SPARQL [5]. Among them SPARQL is efficient query language for semantic web and it has W3C recommendation. Basically the SPARQL language for RDF defines several query result forms such as CSV, TSV [7], and XML etc. The SPARQL binds the variables of query results into XML notation. The conversion process of SPARQL query results into XML format is described by RDF data access working group (DAWG). DAWG has described four implementations. Among them two implementations produce SPARQL result in XML format and other two implementations, consumes the query results that are in XML format. The producer implementations are Joseki [8] and AllegroGraph [9]. The producer implementations produce SPARQL query results in XML format. The two consumers are python [10] and XSL Transform (XSLT). Python parses the format into an internal graph to check for correctness. XSLT consumes SPARQL query results XML format and generates an HTML document [11].

SPARQL 1.1 query results CSV (comma separated values) and TSV (tab separated values) formats provide simple and easy to process formats for the transmission of tabular data. These formats are supported as input to many tools like spreadsheets. SPARQL 1.1 query results in JSON format [12] is designed to represent the query results in an array object. The results of a SELECT query are serialized as an array, where each array element is one "row" of the query results. BIRT [13], an open source Eclipse-based reporting system that can be used to generate charts and other reports from input data. BIRT takes input from relational databases or spreadsheets. TopBraid Composer [14] integrates with BIRT to generate reports for SPARQL query results [15]. TopBraid Composer's Maestro Edition provides an interface between any OWL/RDF data source and BIRT.

The SPARQL query result formats CSV, TSV, JSON and XML require additional transformations or tool support to represent query result in an appropriate format. BIRT is not developed specifically for SPARQL query results. To use BIRT SPARQL query result needs to be converted into any BIRT specific input format. Our approach enables the user to build HTML document dynamically for variable binding SPARQL query results and browse the constructed HTML document automatically to view the report.

III. PROPOSED SYSTEM ARCHITECTURE

From the literature study, the conclusion is that, there is no direct method to represent SPARQL query result in user understandable format. All the existing formats require

results and browse the constructed HTML document automatically to view the query result.

The architecture describes generation of HTML page for variable binding SPARQL query results. Fig1 shows the

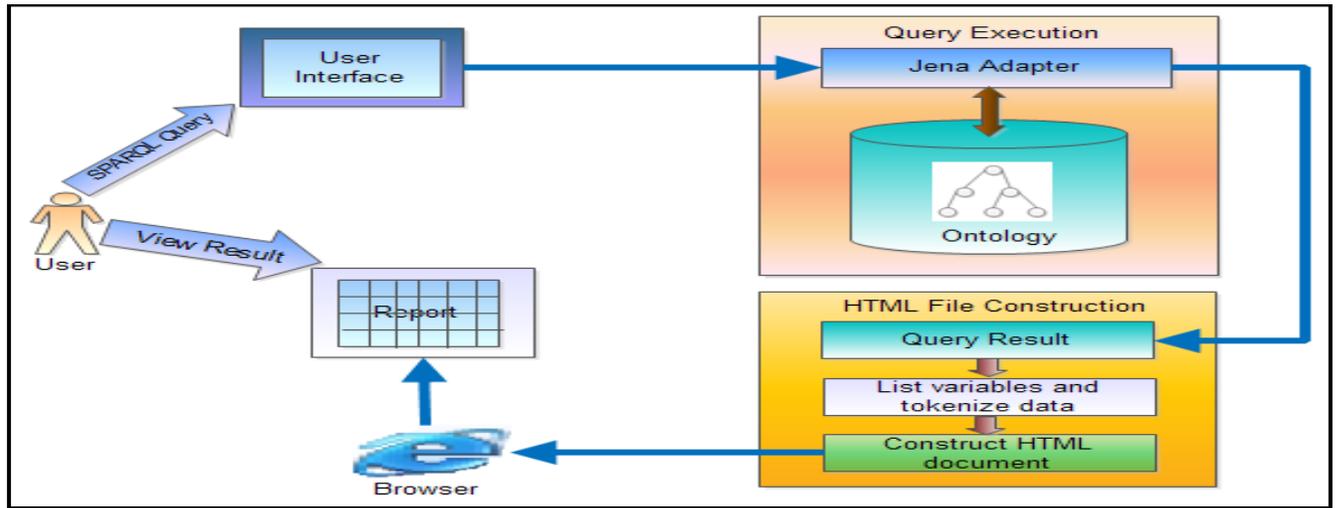


Fig. 1. Proposed System Architecture

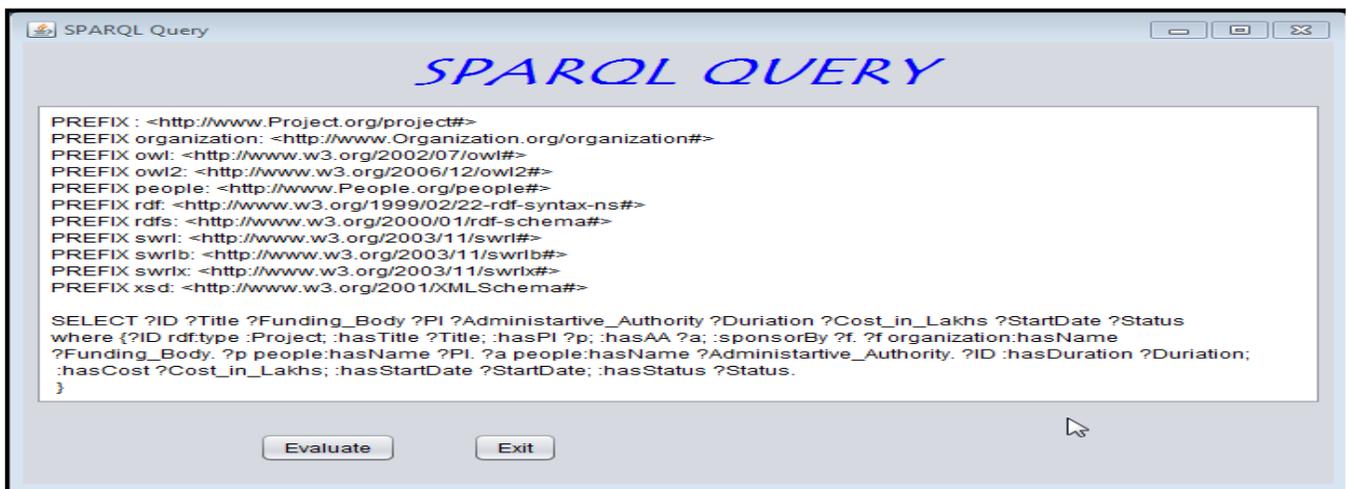


Fig.2. User interface to write SPARQL query

additional conversions or tool support to represent query results in user understandable format. The problem is to design and implement a method to represent results of a SPARQL SELECT query that is executed on semantic data which is stored in the oracle database 11g semantic store using Jena adapter [16].

The method should enable the user to build HTML document dynamically for variable binding SPARQL query

system architecture. The user interface allows user to write SPARQL SELECT query. The query is executed on semantic data stored in the oracle database using the oracle Jena adapter. The HTML file construction section lists the variables involved in the query and extracts the variable binding values. A HTML Document is constructed with the variables and query results. Finally the constructed HTML document is displayed to view the report using any web browser like internet explorer, etc.

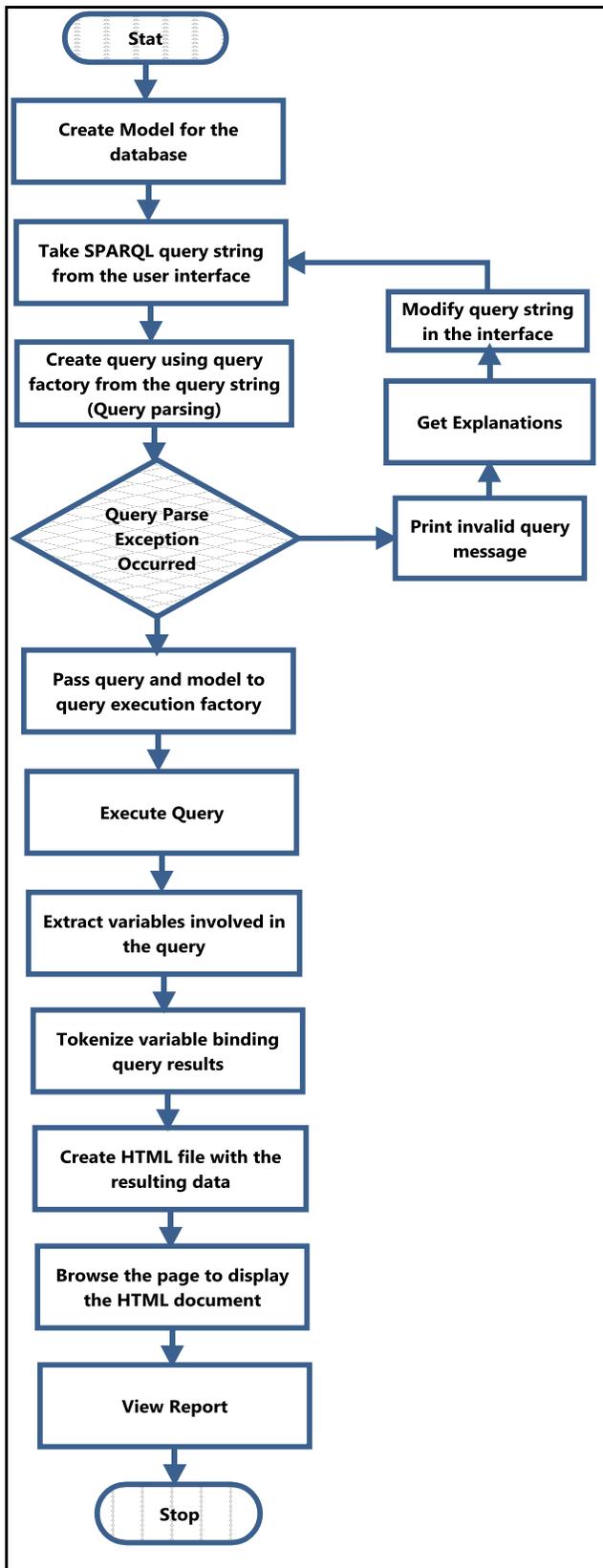


Fig.3. Execution process of SPARQL query and construction of HTML

IV. IMPLEMENTATION

The proposed architecture is implemented using Java NetBeans IDE and Jena API. A simple user interface is designed for writing SPARQL query using Java NetBeans IDE. Fig2 shows the interface for writing SPARQL query. Implementation of the proposed method has two phases. First phase consist query execution process using the Oracle Jena adapter and the second phase describes generation of HTML document for SPARQL query results. The entire process is described in fig3

Phase1: SPARQL query execution

Executing SPARQL query using Jena adapter has the following steps.

- Create model for the ontology store
- Model
`model=ModelOracleSem.createOracleSemModel(oracle,"RDF Model");`
- Take SPARQL query string from the interface
- `String q=txtQuery.getText();`
- Create query using QueryFactory.
- Query
`query=QueryFactory.create(q, Syntax.syntaxARQ);`
- Pass query and model to the QueryExecutionFactory and execute the query
`QueryExecution exec=QueryExecutionFactory.create(query, model);`

Phase2: HTML document construction for SPARQL query results

Constructing HTML document from the variable binding SPARQL query result has the following steps.

- List out the variables involved in the query
`ResultSet rs=exec.execSelect();`
`List l=rs.getResultVars();`
- Extract variable binding values (it is an iterative process)
`QuerySolution qs=rs.nextSolution();`
`String val=qs.get(l.get(i).toString()).toString();`
- Constructing HTML document with the listed variables and binding values. The following code shown in fig4 constructs HTML document dynamically.
- Pass the constructed HTML document to a web browser to display the result
`Desktop dt=Desktop.getDesktop();`
`dt.browse(new URI("result.html"));`

V. RESULTS

```

PrintWriter pw=new PrintWriter(new File("result.html"));
pw.print("<html><body bgcolor=#EAE6F5>");
pw.print("<h2 align=center><font color=#FF00FF>SPARQL  
RESULT</font></h2>");
pw.print("<table border=1 align=center>");
pw.print("<tr>");
for(int i=0;i<l.size();i++)
    pw.print("<th bgcolor=#FFA500><font  
size=6>"+l.get(i)+"</font></th>");
pw.write("</tr>");
pw.print("<tbody bgcolor=#C0C0C0>");
while(rs.hasNext())
{
    QuerySolution qs=rs.nextSolution();
    pw.print("<tr>");
    for(int i=0;i<l.size();i++)
    {
        val=qs.get(l.get(i).toString()).toString();
        pw.print("<td>"+val+"</td>");
    }
    pw.print("</tr>");
}
pw.print("</tbody></table>");
pw.print("</body></html>");
pw.close();

```

Fig.4. Code to generate HTML document for SPARQL query result

To evaluate the proposed method, OWL ontology for R&D projects is constructed using NeOn toolkit [17] and loaded into the oracle database 11g semantic store. This section describes the difference between actual SPARQL query result format and the proposed method output.

To show the difference, a sample SPARQL SELECT query is taken to print all the principal investigators involved in various research projects.

SPARQL SELECT Query: To print various R&D project details

```

SELECT ?ID ?Title ?Funding_Body ?PI
?Administrative_Authority ?Duration ?Cost_in_Lakhs
?StartDate ?Status where {?ID rdf:type :Project; :hasTitle
?Title; :hasPI ?p; :hasAA ?a; :sponsorBy ?f. ?f
organization:hasName ?Funding_Body. ?p people:hasName
?PI. ?a people:hasName ?Administrative_Authority. ?ID
:hasDuration ?Duration; :hasCost ?Cost_in_Lakhs;
:hasStartDate ?StartDate; :hasStatus ?Status. }

```

The query is executed and fig5 shows actual format of SPARQL query result. The format has unnecessary data and query results are combined with variables. It is not effective, difficult to read and understand.

Fig6 shows output of the above SPARQL query using proposed method. The resulting format is effective understandable and readable.

VI. CONCLUSION

This paper presents a method to represent results of SPARQL query executed on semantic data stored in the oracle 11g database. This method uses the Oracle Jena adapter to execute SPARQL query. The proposed method is implemented using Java code and HTML elements to render the query results in presentable format. Compared to other approaches, this approach does not require intermediate

```

(?Funding_Body = "DRDO") (?Cost_in_Lakhs = "4.85600014E1"^^xsd:float) (?Title = "Rooting System") (?Status = "On going") (?ID =
<http://www.Project.org/project#P0002>) (?StartDate = "2011-05-12T00:00:00"^^xsd:dateTime) (?Administrative_Authority = "Prof. K. Lakshmi") (?PI =
"Prof. M. R. Naidu") (?Duration = "3 Years")
(?Funding_Body = "DRDO") (?Cost_in_Lakhs = "4.05600014E1"^^xsd:float) (?Title = "Knowledge Representation System") (?Status = "Completed") (?ID =
<http://www.Project.org/project#P0004>) (?StartDate = "2006-05-26T00:00:00"^^xsd:dateTime) (?Administrative_Authority = "Dr. Gurdeep") (?PI =
"Prof. Mohanlal") (?Duration = "2 Years")
(?Funding_Body = "ISRO") (?Cost_in_Lakhs = "7.52600021E1"^^xsd:float) (?Title = "Resource Management System") (?Status = "Completed") (?ID =
<http://www.Project.org/project#P0005>) (?StartDate = "2007-01-24T00:00:00"^^xsd:dateTime) (?Administrative_Authority = "Dr. M. Jawahar") (?PI =
"Dr. A. Usha Rani") (?Duration = "3 Years")
(?Funding_Body = "MOES") (?Cost_in_Lakhs = "1.05E1"^^xsd:float) (?Title = "Managing Cloud in the Internet") (?Status = "Completed") (?ID =
<http://www.Project.org/project#P0001>) (?StartDate = "2009-03-24T00:00:00"^^xsd:dateTime) (?Administrative_Authority = "Dr. Ajaipal") (?PI = "Dr. L.
Kanhaiya Lal") (?Duration = "2 Years")
(?Funding_Body = "MOES") (?Cost_in_Lakhs = "3.03999996E1"^^xsd:float) (?Title = "Knowledge Management System") (?Status = "On going") (?ID =
<http://www.Project.org/project#P0003>) (?StartDate = "2011-06-25T00:00:00"^^xsd:dateTime) (?Administrative_Authority = "Prof. M K Naidu") (?PI =
"Prof. Dimpu Rani") (?Duration = "3 Years")

```

Fig. 5. SPARQL query result actual format

SPARQL RESULT								
ID	Title	Funding_Body	PI	Administrative_Authority	Duration	Cost_in_Lakhs	StartDate	Status
P0002	Rooting System	DRDO	Prof. M. R. Naidu	Prof. K. Lakshmi	3 Years	4.85600014E1	2011-05-12T00:00:00	On going
P0004	Knowledge Representation System	DRDO	Prof. Mohanlal	Dr. Gurdeep	2 Years	4.05600014E1	2006-05-26T00:00:00	Completed
P0005	Resource Management System	ISRO	Dr. A. Usha Rani	Dr. M. Jawahar	3 Years	7.52600021E1	2007-01-24T00:00:00	Completed
P0001	Managing Cloud in the Internet	MOES	Dr. L. Kanhaiya Lal	Dr. Ajaipal	2 Years	1.05E1	2009-03-24T00:00:00	Completed
P0003	Knowledge Management System	MOES	Prof. Dimpu Rani	Prof. M K Naidu	3 Years	3.03999996E1	2011-06-25T00:00:00	On going

Fig. 6. Variable binding SPARQL query result using proposed method

transformations to present query results in readable format and user need not to take much effort to view the result. The proposed method is examined with R&D project OWL ontology loaded in the oracle 11g database. As shown in the results, the proposed method constructs an HTML report dynamically for SPARQL query results and displays automatically using a web browser to view the result.

ACKNOWLEDGMENT

The work presented in this paper is done as part of a sponsored project funded by government of India, Ministry of Defence, DRDO (ER&IPR), and done in the labs of Adhiyamaan College of Engineering where the author is working as a Professor & Director in the department of Master of Computer Applications. The author would like to express her sincere thanks to DRDO for providing the support.

REFERENCES

- [1] N. Shadbolt, W. Hall and T. Berners-Lee, "The Semantic Web Revisited", *Intelligent Systems*, IEEE, 2006, vol. 21, no. 3, pp. 96-101.
- [2] O. Lassila, F. van Harmelen, I. Horrocks, J. Hendler, D.L. McGuinness, "The Semantic Web and its Languages", *Intelligent Systems and their Applications*, IEEE, 2000, Vol. 15, Issue 6, pp. 67-73.
- [3] J. Bailey, F. Bry, T. Furche, S. Schaffert, "Semantic Web Query Languages", *Encyclopedia of Database Systems*, Springer, 2009, pp. 2583-2586.
- [4] Seaborne A, "RDQL - A query language for RDF W3C Member Submission". Available at <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109>.
- [5] E. Prud'hommeaux and A. Seaborne, "SPARQL query language for RDF", Technical report, W3C Recommendation, 2008. Available on <http://www.w3.org/TR/rdf-sparql-query/>
- [6] S. M. Patil and D. M. Jadhav, "Semantic Search using Ontology and RDBMS for Cricket", *International Journal of Computer Applications*, May 2012, Vol. 4, No.14, pp.26-31.
- [7] Andy Seaborne, "SPARQL 1.1 Query Results CSV and TSV Formats", 2012, available at <http://www.w3.org/TR/2012/WD-sparql11-results-csv-tsv-20120501/>
- [8] <http://www.joseki.org/>
- [9] <http://www.franz.com/agraph/allegrograph/>
- [10] <http://www.franz.com/agraph/support/documentation/current/python-tutorial/python-tutorial-40.html>.
- [11] Dave Beckett and Jeen Broekstra, "Format SPARQL Query Results XML Format into XHTML (XSLT)", 2013, available at <http://www.w3.org/TR/rdf-sparql-XMLres/>
- [12] Andy Seaborne, "SPARQL 1.1 Query Results JSON Format", 2013, The Apache Software Foundation, available at <http://www.w3.org/TR/2013/REC-sparql11-results-json-20130321/>
- [13] BIRT Tutorial, version 8.2, May 2011; available at <http://www.eclipse.org/birt/phoenix/tutorial/>
- [14] www.topbraidcomposer.org.
- [15] Moritz Weiten, "OntoSTUDIO as a Ontology Engineering Environment", *Semantic Knowledge Management*, Springer Book, Chapter 5, 2009, pp.51-60.
- [16] Chuck Murray, "Oracle Database Semantic Technologies Developer's Guide", 11g Release 2 (11.2), may 2012.
- [17] NeOn toolkit: http://neon-toolkit.org/wiki/Main_Page.

A Comprehensive Evaluation of Weight Growth and Weight Elimination Methods Using the Tangent Plane Algorithm

P May
K College,
Brook Street, Tonbridge,
Kent, UK

E Zhou
Engineering, Sports and Sciences
Academic Group,
University of Bolton, UK

C. W. Lee
Engineering, Sports and Sciences
Academic Group
University of Bolton, UK

Abstract—The tangent plane algorithm is a fast sequential learning method for multilayered feedforward neural networks that accepts almost zero initial conditions for the connection weights with the expectation that only the minimum number of weights will be activated. However, the inclusion of a tendency to move away from the origin in weight space can lead to large weights that are harmful to generalization. This paper evaluates two techniques used to limit the size of the weights, weight growing and weight elimination, in the tangent plane algorithm. Comparative tests were carried out using the Extreme Learning Machine which is a fast global minimiser giving good generalization. Experimental results show that the generalization performance of the tangent plane algorithm with weight elimination is at least as good as the ELM algorithm making it a suitable alternative for problems that involve time varying data such as EEG and ECG signals.

Keywords—neural networks; backpropagation; generalization; tangent plane; weight elimination; extreme learning machine

I. INTRODUCTION

In Lee [1] an algorithm was described for supervised training in multilayered feedforward neural networks giving faster convergence and improved generalization relative to the gradient descent backpropagation algorithm. This tangent plane algorithm starts the training with the connection weights set to values close to zero in the expectation that the minimum weights necessary will be activated.

The results based on two real world datasets indicated that the tangent plane algorithm gives improved generalization over a range of network sizes and that it is robust with respect to the choice of its internal parameters.

Despite the success of the tangent plane algorithm there is, however, strong evidence to suggest that growing the weights to assume large values can actually hurt generalization in different ways. Excessively large weights feeding into output units can cause wild outputs far beyond the range of the data if an output activation function is not included. To put it another way, large weights can cause excessively large variances in the output. According to Bartlett [2], the size of the weights is more important than the number of weights in determining good generalization. This poses the following question: can we modify this algorithm so that it discourages the formation

of weights with large values? Further, can the algorithm encourage weights with small values to decay rapidly to zero thus producing a network having the optimum size for good generalization?

Weight decay is a subset of regularization methods. The principal idea of weight decay is to penalize connection weights with small values so that the network removes the superfluous weights itself. The simplest method is to subtract a small proportion of a weight after it has been updated [3]. This is equivalent to adding a penalty term $\sum_j w_{ji}^2$ to the objective function and performing gradient descent on the resulting total error. Unfortunately this method penalizes more of the w_{ji} 's than necessary whilst keeping the relative importance of the weights unchanged. This can be cured by using a different penalty term, $\sum_j w_{ji}^2 / (1 + w_{ji}^2)$, so that the small w_{ji} 's decay faster than the larger ones [4]. Williams [5] proposed yet another type of penalty function which is proportional to the logarithm of the l_1 norm of the weights,

$\sum_j |w_{ji}|$. It was shown in [5] that using this penalty term is more appropriate for internal weights than weight decay. Hoyer [6] proposed a sparseness measure based on the l_1 norm and l_2 norm of weights. Experiments with Hoyer's method indicate that it performs well in comparison with weight decay and weight elimination. A further refinement involves using a mixed norm penalty term [7]. In this procedure the l_1 norm of the weight vector is minimised subject to the constraint that the l_2 norm equals unity.

II. OBJECTIVES

The principal objective of this paper is to describe an alternative strategy for improving generalization in neural networks trained using the tangent plane algorithm. In the newly developed algorithm, the training is started from arbitrary initial conditions and the inactive weights in the network encouraged decaying to zero by using the weight elimination procedure.

Unlike other implementations of weight elimination procedures [4, 6, 8 - 9], the method used here is built into the geometry used in the derivation of the algorithm. A secondary objective is to compare the newly developed algorithm with the extreme learning machine [10], which obtains the least squares solution with the minimum training error and minimum norm of weights.

III. DERIVATION OF THE ALGORITHM

In Lee [1] an algorithm is described that accepts almost zero starting conditions for the connection weights, and which moves away from the origin in a direction indicated by the training data with the expectation that only the minimum weights would be activated. This tangent plane algorithm uses the target values of the training data to define a $(n - 1)$ surface in weight space \mathbf{R}^n . The weights are adjusted by moving from the current position to a point near to the foot of the perpendicular to the tangent plane to this surface, but displaced somewhat in the direction away from the origin, on the expectation that the smaller the distance moved from the foot of the perpendicular the less disturbance there will be to the previous learning.

Two enhancements are made to the tangent plane algorithm to obtain the improved tangent plane algorithm referred to as iTPA. Firstly, a directional movement vector is introduced into the training process to push the movement in weight space towards the origin.

This movement vector simulates weight decay which is known to have a beneficial effect on generalization in backpropagation learning. Secondly, the directional vector is further modified to give a heavier weighting to weights with small weight values to avoid penalizing more of the weights than necessary; one large weight costs much more than many smaller ones. A high degree polynomial term is used to select the proportion of weights for pruning. This term can be adjusted so that a weight decay procedure is implemented or refined in a way that specific weights are removed by causing them to decay more rapidly to zero.

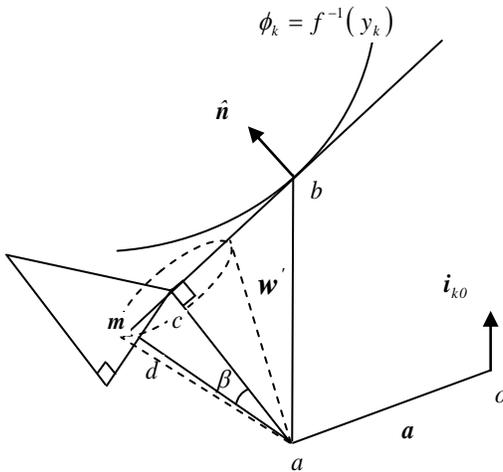


Fig. 1 Movement from the present position a to the point d inclined at an angle β to the perpendicular from a to the tangent plane to the constraint surface $\phi_k = f^{-1}(y_k)$ at point b in the weight space \mathbf{R}^n . The vector m represents the orthogonal projection of the weight elimination vector w' orthogonally onto the normal n to the constraint surface at point b

The method assumes a feed-forward neural network of units $\{u_j\}$, where the connection between u_i and u_j is mediated by w_{ji} . ϕ_j and θ_j denote the input and output of u_j , so that $\theta_j = f(\phi_j)$, and $\phi_j = \sum_i w_{ji} \theta_i$. The single output unit u_k is trained to mimic the target value y_k . Let n denote the number of weights

in the network. For a given set of inputs we can consider ϕ_k to be a function of the weights $\phi_k: \mathbf{R} \rightarrow \mathbf{R}^n$. The tangent plane algorithm adjusts the weights by moving along the line at an angle β to the perpendicular from the current position a to the tangent plane to the surface $\phi_k = f^{-1}(y_k)$, on the side of the perpendicular away from the origin (see Fig. 1).

Let $a = \sum_{ji} w'_{ji} i_{ji}$ be the current values of the weights, where i_{ji} is a unit vector in the direction of the w_{ji} axis. Use the equation $f^{-1}(y_k) = w_{k0} + \sum_{i \neq 0} w_{ki} \theta_i$ to find a value, w''_{k0} , for the bias weight w_{k0} from the values w_{ji} of the other weights, so that the surface $\phi_k = f^{-1}(y_k)$ contains the point

$b = w''_{k0} i_{k0} + \sum_{i, i \neq k, 0} w'_{ji} i_{ji}$. Now, if we use the equation $f^{-1}(y_k) = w''_{k0} + \sum_{i \neq 0} w'_{ki} \theta_i$ and $f^{-1}(\theta_k) = w'_{k0} + \sum_{i \neq 0} w'_{ki} \theta_i$, and note that b differs from a only in the value of w_{k0} , we get

$$\begin{aligned} b - a &= (w''_{k0} - w'_{k0}) i_{k0} \\ &= (f^{-1}(y_k) - f^{-1}(\theta_k)) i_{k0} \end{aligned} \quad (1)$$

Let \hat{n} be the unit normal to the surface at b , so $\hat{n} = \nabla \phi_k / \|\nabla \phi_k\|$. The length of the perpendicular from a to the tangent plane at b is $(b - a) \cdot \hat{n}$. If c is the foot of the perpendicular from a to the tangent plane at b ,

$$\begin{aligned} c - a &= (f^{-1}(y_k) - f^{-1}(\theta_k)) (i_{k0} \cdot \hat{n}) \hat{n} \\ &= \frac{f^{-1}(y_k) - f^{-1}(\theta_k)}{\|\nabla \phi_k\|} \frac{\nabla \phi_k}{\|\nabla \phi_k\|} \end{aligned} \quad (2)$$

And

$$\|c - a\| = \frac{f^{-1}(y_k) - f^{-1}(\theta_k)}{\|\nabla \phi_k\|} \quad (3)$$

The vector that is directed towards the origin and biased along the axes of the weights w_{ji} that have small weight values relative to some small positive constant w_a is

$w' = -\sum_{ji} (w_{ji}/w_a) i_{ji} / (1 + w_{ji}^2/w_a^2)$. The projection of w' onto the tangent plane is given by

$$\begin{aligned} m &= w' - (w' \cdot \hat{n}) \hat{n} \\ &= w' - \frac{1}{\|\nabla \phi_k\|} \sum_{l,m} \left(w'_{lm} \frac{\partial \phi_k}{\partial w_{lm}} \right) \frac{\nabla \phi_k}{\|\nabla \phi_k\|} \end{aligned} \quad (4)$$

Where

$$w' = -\sum_{j,i} \frac{(w_{ji}/w_a)}{1 + (w_{ji}^2/w_a^2)} i_{j,i} \quad (5)$$

Thus, if d is the point of intersection with the tangent plane of a line from a inclined at angle β to the perpendicular, then

$$\begin{aligned} d - a &= (d - c) + (c - a) \\ &= \|c - a\| \tan \beta \frac{m}{\|m\|} + (c - a) \end{aligned} \quad (6)$$

Let $\delta = f^{-1}(y_k) - f^{-1}(\theta_k)$ be the error in the input to final unit. Hence using equations (2), (3) and (4) in (5) yields

$$d - a = \frac{1}{\|\nabla \phi_k\|^2} \delta \nabla \phi_k + \frac{|\delta|}{\|\nabla \phi_k\|} \tan \beta \frac{1}{\|m\|} \quad (7)$$

$$\left(w' - \frac{1}{\|\nabla \phi_k\|} \times \sum_{lm} w'_{lm} \frac{\partial \phi_k}{\partial w_{lm}} \frac{\nabla \phi_k}{\|\nabla \phi_k\|} \right)$$

Thus, to adjust a given weight w_{ji}

$$\Delta w_{ji} = \frac{1}{\|\nabla \phi_k\|^2} \delta \frac{\partial \phi_k}{\partial w_{ji}} + \frac{|\delta|}{\|\nabla \phi_k\|} \quad (8)$$

$$\tan \beta \frac{1}{\|m\|} \left(w'_{ji} - \frac{1}{\|\nabla \phi_k\|^2} \times \sum_{lm} w'_{lm} \frac{\partial \phi_k}{\partial w_{lm}} \frac{\partial \phi_k}{\partial w_{ji}} \right)$$

$$\text{where } \|m\|^2 = \sum_{j,i} \left(w'_{ji} - \frac{1}{\|\nabla \phi_k\|^2} \sum_{l,m} \left(w'_{lm} \frac{\partial \phi_k}{\partial w_{lm}} \right) \frac{\partial \phi_k}{\partial w_{ji}} \right)^2 \quad (9)$$

The term $\partial \phi_k / \partial w_{ji}$ is the partial derivative of the net input to the output unit. The treatment of this term follows from Lee [1]

$$\frac{\partial \phi_k}{\partial w_{ji}} = \frac{\partial \phi_k}{\partial \phi_j} \theta_i$$

and

$$\frac{\partial \phi_k}{\partial \phi_j} = \begin{cases} 1, & \text{if } j = k \\ f'_j(\phi_j) \sum_{m \in M_j} \frac{\partial \phi_k}{\partial \phi_m} w_{mj}, & \text{if } j \neq k \end{cases}$$

Where M_j is the set of units to which u_j passes its output.

The new iTPA algorithm requires two parameters that need to be set manually. First parameter is the angle parameter $\tan \beta$, which gives the angle between the movement vector and the perpendicular from the current position to the tangent plane. Its value is usually chosen to be small, typically 0.05, so that movement is to a point nearby the foot of the perpendicular. Second parameter is the weight sensitivity parameter w_a , which gives the value an individual weight w_{ji} receives a large push towards the origin. w_a is preferred to be small, typically 0.5, so that weights with small values are selected for removal from the network. This will produce the required separation of active and inactive weights in the network.

An individual term $w'_{ji} = - (w_{ji}/w_a) / (1 + (w_{ji}/w_a)^2)$ in the directional vector w' varies according to (w_{ji}/w_a) in an anti-symmetric fashion. This permits the necessary sign changes in w'_{ji} so that the movement along a weight dimension is

always directed towards the origin. When $|w_{ji}| < w_a$ the directional term for that weight is approximately linear. On the other hand, when $|w_{ji}| > w_a$ the directional term approaches to zero. Thus a weight will receive a large push when w_{ji} equals w_a . A potential difficulty arises when both w_{ji} and w_a are less than 0.5. If $|w_{ji}| = w_a$, the resulting push may be large enough to overshoot the origin. This situation can be avoided through the appropriate choice of $\tan \beta$.

The approach of the new iTPA algorithm is reminiscent of the Newton-Raphson method of first degree [10] used to find the zero points of functions that depend on one value. An important difference is that it provides a whole R^n plane of suitable points to move towards. Any method that does a zero-point search of a function cannot get trapped in a local minimum unless it hits one by accident. The new iTPA algorithm uses the vector $\nabla \phi_k$ to do a linear extrapolation of the surface $\phi_k = f^{-1}(y_k)$ in order to gain a new weight vector that is hoped to be on, or at least close, to this surface.

A simple cost saving can be made by replacing the term $\|m\| = \|w' - (w' \cdot \hat{n}) \hat{n}\|$ in the algorithm with $w' \cdot \|w'\|$ is greater than or equal to $\|w' - (w' \cdot \hat{n}) \hat{n}\|$ with equality holding when w' is perpendicular to \hat{n} . Its use will result in a reduction in the size of m , but this term is scaled by $\tan \beta$ anyway. This reduction is greatest when w' is perpendicular to the tangent plane. According to equation (9) $\|m\|$ involves adding n products of the terms w'_{lm} and $\partial \phi_k / \partial w_{lm}$, and then using this result to scale n partial derivatives, $\partial \phi_k / \partial w_{ji}$. Thus the total computational saving is $2n$ operations per weight update.

The algorithm can be further improved by using a high degree polynomial in the denominator of each term w_{ji} in the directional vector w' , so that $w'_{ji} = -(w_{ji}/w_a) / (1 + (w_{ji}/w_a)^n)$ where n is a positive constant. The curves of an individual directional term w'_{ji} for three values of parameter n are shown in Fig. 2. Examination of the curves for $|w_{ji}| > w_a$ yield the following observation. When n is large (typically > 6), the directional term w'_{ji} for that weight decays rapidly to zero. Thus the proper choice of the parameter n will permit some weights in the network to assume values that are larger than with $n = 2$.

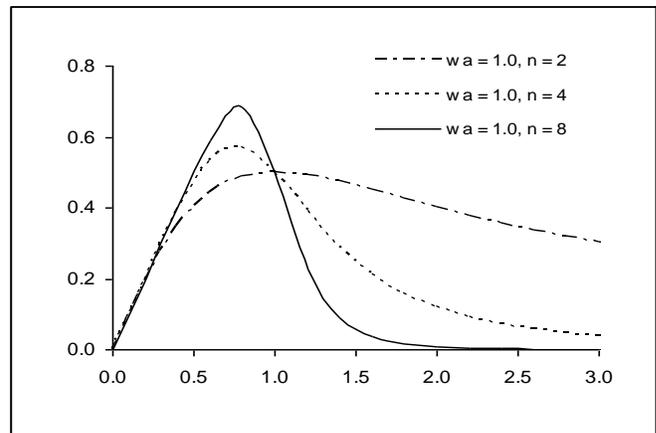


Fig. 2 Variation of an individual directional term w'_{ji} for different values of the parameter n

IV. ESTIMATING THE WEIGHT SENSITIVITY

The rationale of pruning is to reduce the number of free parameters in the network by removing dispensable ones. If applied properly, this approach often reduces overfitting and improves generalization. At the same time it produces a smaller network. The approach adopted in this paper is to automatically prune superfluous weights by using the method of weight elimination [4]. But how do we know whether the method of weight elimination actually produces the required separation of active and inactive weights? One approach might be to measure the significance or importance of each weight, as the magnitude of the weights is not the best measure of their contribution to the training process [11].

There are several methods suggested for calculating the importance of connection weights. Karnin [12] measures the sensitivity s_{ji} of each weight by monitoring the sum of all the changes to the weights during training. Thus the saliency of a weight is given as $s_{ji} = \sum_t \partial \mathcal{E}_k / \partial w_{ji} \Delta w_{ji}^{(t)} w_{ji}^f / (w_{ji}^f - w_{ji}^0)$, where t is the number of epochs trained, w_{ji}^f and w_{ji}^0 are the final and initial values of weight w_{ji} . LeCun et al [13] measure the saliency of a weight by estimating the second derivative of the error. They also reduce the network complexity by constraining certain weights to be equal. Low saliency means low importance of the weights. A more sophisticated approach avoids the drawbacks of approximating the second derivatives by computing them exactly [14].

The last two methods have the disadvantage of requiring training down to the error minimum. The autoprune method [11] avoids this problem. It uses a statistic t to allocate an importance coefficient to each weight based upon the assumption that a weight becomes zero during the training process

$$t(w_{ji}) = \log \left(\frac{\left| \sum_t w_{ji} - \Delta w_{ji}^{(t)} \right|}{\sum_t (\Delta w_{ji}^{(t)} - (\Delta \bar{w}_{ji}))^2} \right) \quad (10)$$

In the above formula, sums are over all training examples t of the training set, and the overline means arithmetic mean over all examples. A large value of t_{ji} indicates high importance of weight w_{ji} .

V. SIMULATIONS AND RESULTS

The convergence behaviour of the new iTPA algorithm was evaluated and compared with the gradient descent backpropagation algorithm. The dataset used was the two spiral problem [15 – 16]. Like most published work classifying the two spiral problem [17], a network with three hidden layers was used. 10 trials were performed with the classification error on the training and test sets, mean number of epochs to converge, and number of successful trials recorded. Network training was terminated when all the training patterns were learned correctly or 5,000 epochs or presentations of the entire dataset.

Next, the ability of the new iTPA and original tangent plane algorithms to generalise from a given set of training data was evaluated and compared with the Extreme Learning

Machine [18]. The Extreme Learning Machine (ELM) is a fast learning algorithm that obtains the least squares solution with the minimum training error and minimum norm of the weights. The benchmark datasets used were the Henon map [19] and the non-linear dynamic plant [20]. A standard feedforward neural network with two hidden layers was utilised with the number of hidden units determined by grid search. For each test, 10 trials were performed with the normalised mean square error on the training and test sets recorded together with the number of successful trials. Network training was terminated after 5,000 epochs or presentations of the entire training set.

Finally, the evolution and development of the weights in the new iTPA algorithm was evaluated and compared with the original tangent plane algorithm. The benchmark datasets used were the Henon map [19] and the non-linear dynamic plant [20]. The method used to estimate the sensitivity of the weights was autoprune [11]. Histograms of weight sensitivities were plotted after 100, 300 and 500 epochs.

A. Network initialization

The algorithms used in the study require manually set parameters. Preliminary tests showed that the best results were obtained with the parameters set as follows. First test is the new iTPA algorithm. For the two spiral problem, $\tan\beta = 0.01$. The weight sensitivity parameter w_a and n were varied according to a grid search. The input weights were set to random values in the range [-2, 2]. For the Henon map and non-linear dynamic plant, $\tan\beta = 0.01$, $w_a = 0.5$, and $n = 4.0$. The input weights were set to random values in the range [-0.5, 0.5]. Next test is the original tangent plane algorithm. The angle parameter $\tan\beta = 0.01$. The input weights were set to random values in the range [-0.01, 0.01]. Finally test is the standard back-propagation algorithm. For the two spiral problem, the learning rate $\eta = 0.01$, and momentum coefficient $\alpha = 0.3$. The input weights were set to random values in the range [-2, 2].

B. Simulation problems

The two spiral problem consists of two interlocking spirals, each made up of 97 data points. The network must learn to discriminate the two spirals. Traditionally this is known to be a very difficult problem for the back-propagation algorithm to solve. There are two inputs and one output. The inputs are the x and y co-ordinates, and the output notifies which spiral the point belongs to. For the points in the first spiral the output is set to +1, and for points on the other spiral the output is set to -1. The number of training samples is 194. A test set of 192 samples was generated by rotating the two spirals by a small angle.

The Henon map problem is a chaotic time-series prediction problem. The time series is computed by

$$x^{(t+1)} = 1 - c [x^{(t)}]^2 + b x^{(t-1)} \quad (11)$$

Where $x^{(t)}$ is the value at taken time t , and the parameters $b = 0.3$, and $c = 1.4$. Initial values for the time series are $x^{(1)} = x^{(0)} = 0.63133545$. This point is called the fixed point of the time series. In neural network simulations, four successive values of the time series are used in predicting the next value.

Thus, the number of inputs is four and the number of output is one. Data values were taken from the range [31,230] as given in Lahnajärvi et al [19]. The number of training samples is 100, and testing samples is 100.

The non-linear dynamic plant problem is a high order non-linear system introduced in Narendra and Parthasarathy [20]. It is modelled by the following discrete time equation

$$y^{(t)} = \frac{y^{(t-1)}y^{(t-2)}y^{(t-3)}u^{(t-1)}[y^{(t-3)} - 1] + u^{(t)}}{1 + [y^{(t-2)}]^2 + [y^{(t-3)}]^2} \quad (12)$$

Where $y^{(t)}$ is the model output at time t . Like Narendra and Parthasarathy [20], training data was generated using a random input signal uniformly distributed over the interval [-1, 1]. Five hundred data points were generated, the first three hundred used as training data whilst the remaining used as test data.

C. Error metrics used to determine convergence

The error metrics used in the simulations were CERR (Classification ERROR) for classification problems and NMSE (Normalized Mean Square Error) for regression problems. The CERR was calculated using the 40-20-40 criteria e.g. the actual output does not differ from the target output by more than 0.4 [15 – 16]. The NMSE was calculated by dividing the MSE by the variance of the target output.

D. Discussion of results

Two spiral problem. The first test is a difficult non-linearly separable problem where a set of co-ordinates (x,y) is classified as belonging to one of two interwoven spirals. A 2-100-100-100-1 network topology was chosen as given in [16]. For the new iTPA algorithm, 10 trails gave no failures and a mean number of steps to converge of 28 with standard deviation 9. The classification error on the test set was 1.3×10^{-2} (e.g. % test set learned = 98.7) with all of the points on the training set correctly classified. Using the standard backpropagation algorithm, there was one failure and a mean number of steps to converge of 736 with standard deviation 462. The classification error on the test set was 1.5×10^{-2} (e.g. % test set learned = 98.5). The results compare very favourably with those given in Linder et al [16] (Aprop: epochs = 67, % test set learned = 96.6; Rprop: epochs = 246, % test set learned = 65.6).

Table 1 demonstrates the effects of changing the weight sensitivity parameters w_a and n of the new iTPA algorithm. A 2-20-20-20-1 architecture was chosen to determine the degree to which the iTPA algorithm could generalize in a large network with many free parameters. It was found that the classification error improved slightly when large values were chosen for the weight sensitivity parameter w_a .

Further, increasing the value of the parameter n caused the classification error to dip to a clearly defined minimum. When w_a is large (e.g. typically > 0.5) the directional vector w' will push more of the network weights to small values close to zero thus implementing a weight decay procedure. This suggests that weight decay is a far more effective strategy for

improving generalization than weight elimination in large neural networks trained using the tangent plane algorithm.

TABLE I. CONVERGENCE SPEED AND CLASSIFICATION ACCURACY FOR DIFFERENT PARAMETERS IN THE iTPA ALGORITHM

Note: the columns in Table 1 refer to the classification error on the training set (Cerr) and test set (Cerr*), the mean

w_a	n	Cerr	Cerr*	Steps	Succ
0.10	4.0	0.46	1.35	497	10
0.20	4.0	0.36	1.56	634	10
0.50	4.0	0.41	0.78	563	10
1.00	4.0	0.36	0.78	460	10
0.50	2.0	0.31	0.89	435	10
0.50	4.0	0.41	0.78	563	10
0.50	6.0	0.52	0.99	532	10
0.50	8.0	0.41	0.89	559	10

number of steps to converge, and number of success trials for different values of the weight sensitivity parameter w_a and n .

Fig 3 and 4 show some typical test curves for both algorithms on the two spiral problems. Different sets of initial weights were used in each test. The test curves of the new iTPA algorithm show some variation (Fig 3). In many of the curves generated the test error was found to diminish slowly at the start of the training run with intermittent rises in the test error fairly typical (test 2). When the new algorithm was close to a solution, the convergence was usually rapid (test 1 and 3). The test curves of the standard back-propagation algorithm also show wide variation in the test error. Some curves exhibited turbulent behaviour similar to the new iTPA algorithm (test 3). Other curves got trapped in local minima of the error landscape resulting in very long runtimes (test 1). Generally speaking the new iTPA algorithm is prone to problems with stability. Introducing a 50% staged reduction in the step size resulted in faster convergence speeds.

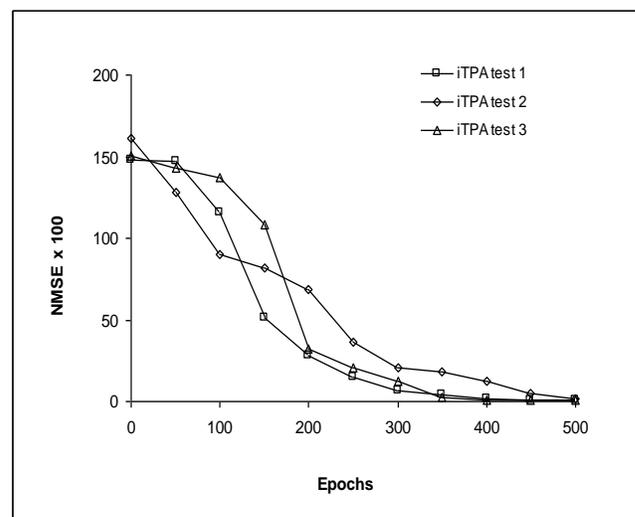


Fig. 3 Typical generalization behaviour of the new iTPA algorithm on the two spiral problem

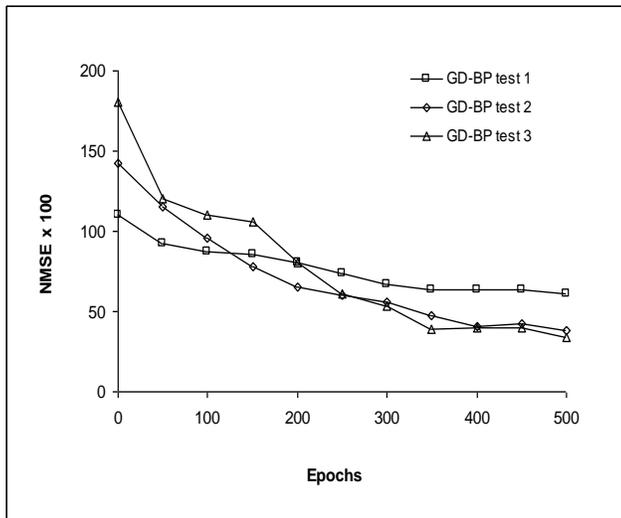


Fig. 4 Typical generalization behaviour of the gradient descent backpropagation algorithm on the two spiral problem

Henon map time series. The second test is a classical deterministic one-step-ahead prediction problem. Preliminary tests showed that the best results for both tangent plane algorithms were obtained using a 4-15-15-1 architecture. Network training was terminated after 5,000 epochs or presentations of the entire dataset. For the new iTPA algorithm, 10 trials gave a normalised mean square error on the training set and test set of 0.00007 and 0.00008 respectively. Using the original tangent plane algorithm, 10 trials gave (training set = 0.00005, test set = 0.00009). There was little evidence of overtraining. The performance of both tangent plane algorithms compare favourably with the Extreme Learning Machine. For ELM a single hidden layer feedforward neural net with 80 hidden units gave (training set = 0.00001, test set = 0.00011).

Fig 5 and 6 show histograms of the importance coefficients of the weights for both algorithms on the Henon map problem. The importance coefficients were recorded from the same trial at epochs 100, 300 and 500. The coefficient sizes were grouped in classes of width one and histograms plotted to show the distribution of the t_{ji} values at three different stages of training. The new iTPA algorithm gave average coefficient sizes at 100, 300, 500 epochs of 1.82, 1.95, and 1.95 respectively. The original algorithm gave 1.70, 1.96, and 2.25. Notice the right skewness of the histograms produced by the new algorithm (see Fig 5). After 500 epochs the histogram is dominated by a single high peak and long right tail. This suggests that many of the weights have taken on equally important roles in the network. Thus are likely to be fewer outlier weights with extreme values that are known to produce overtraining in neural networks.

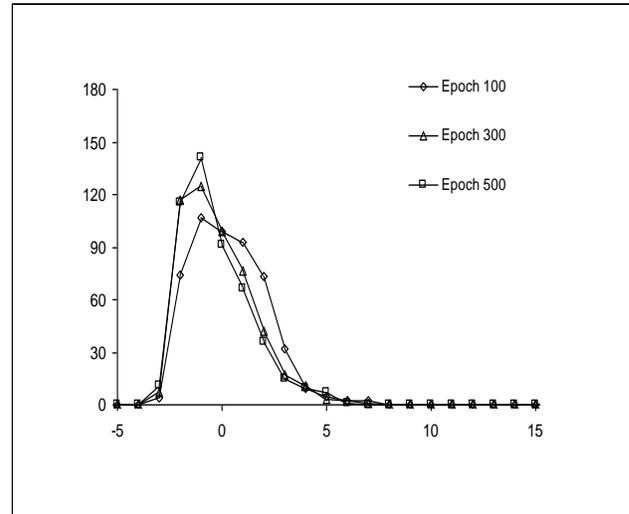


Fig. 5 Importance coefficient histograms for the new iTPA algorithm (Henon map problem). Horizontal axis: coefficient size grouped in classes of width 1. Vertical axis: absolute frequency of weights with this coefficient size.

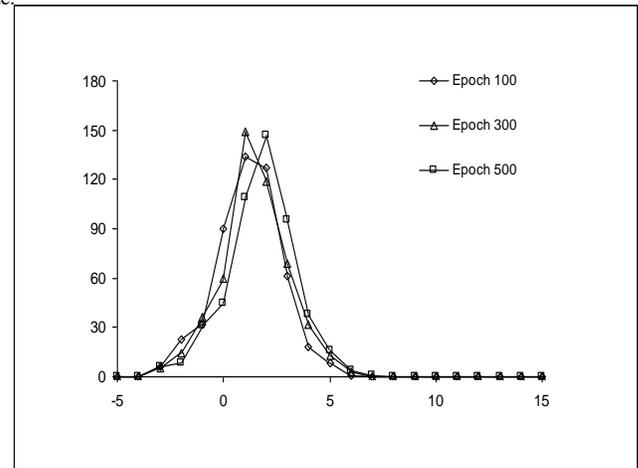


Fig. 6 Importance coefficient histograms for the original tangent plane algorithm (Henon map problem). Horizontal axis: coefficient size grouped in classes of width 1. Vertical axis: absolute frequency of weights with this coefficient size

Non-linear dynamic plant. The third test is a high order non-linear discrete time system. Preliminary tests showed that the best results for both tangent plane algorithms were obtained using a 2-10-10-1 architecture. Network training was terminated after 5,000 epochs or presentations of the entire dataset. For the new iTPA algorithm, 10 trials gave an average normalised mean square error on the training and test sets of 0.00045 and 0.00086 respectively. Using the original tangent plane algorithm, 10 trials gave (training set = 0.00142, test set = 0.00393).

Once again the performance of both tangent plane algorithms compare favourably with the Extreme Learning Machine. For ELM a single hidden layer neural net with 200 hidden units gave (training set = 0.00007, test set = 0.00568). The results on the training set suggest that the new iTPA algorithm is an effective global minimizer capable of reaching the smallest training error.

Fig 7 and 8 show histograms of the importance coefficients of the weights for both algorithms on the non-linear dynamic plant problem. The importance coefficients were recorded from the same trial at epochs 100, 300 and 500. The coefficient sizes were grouped in classes of width one and histograms plotted to show the distribution of the t_{ji} values at three different stages of training. The new iTPA algorithm gave average coefficient sizes at 100, 300, 500 epochs of 2.24, 2.68, and 2.79 respectively. The original algorithm gave 2.56, 3.33, and 3.81. Notice the left drift of the histograms produced by the new algorithm (see Fig 7). In contrast the histograms produced by the original algorithm tend to drift right, as expected (Fig 8). Further, these histograms have a long right tail which suggests that some weights are taking on a far more active role in the network. This might account for the worse generalization of the original algorithm on this problem.

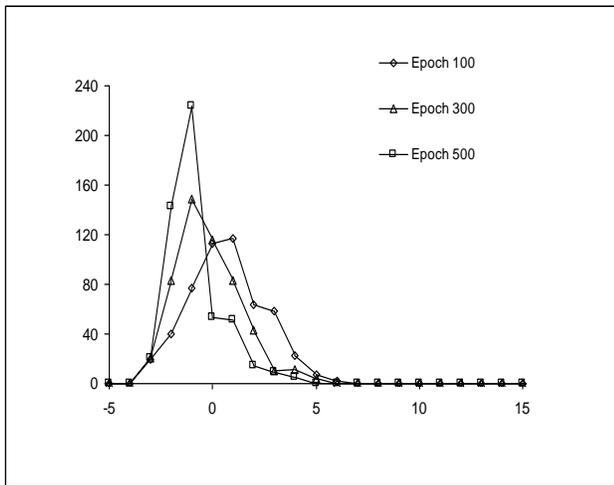


Fig. 7 Importance coefficient histograms for the new iTPA algorithm (non-linear dynamic plant). Horizontal axis: coefficient size grouped in classes of width 1. Vertical axis: absolute frequency of weights with this coefficient size

VI. COMPARISON OF THE DIFFERENT ALGORITHMS

In order to determine whether the difference in the results is statistically significant, we perform some hypothesis tests. The test used was a standard t -test with the sample of test errors from the iTPA algorithm compared with the corresponding sample from the original tangent plane algorithm for each dataset used in the study. A second test was carried out by comparing these test results with the ELM algorithm on the same set of problems.

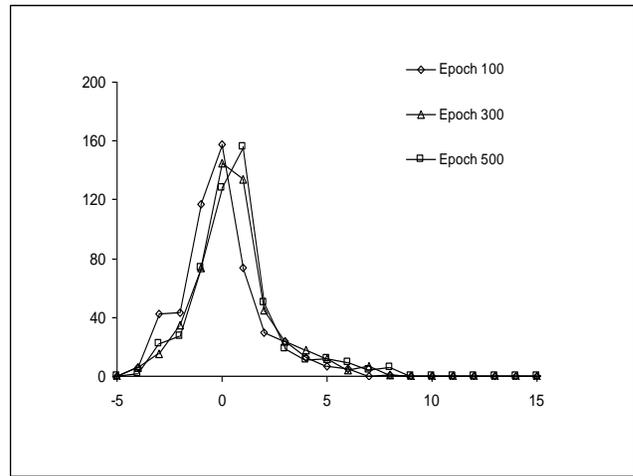


Fig. 8 Importance coefficient histograms for the original tangent plane algorithm (non-linear dynamic plant). Horizontal axis: coefficient size grouped in classes of width 1. Vertical axis: absolute frequency of weights with this coefficient size

The ELM algorithm requires the number of hidden units to be set which was found by grid search. For the correct application of the t -test, it was necessary to take the logarithm of the test errors (since the test errors have log-normal distribution) and remove any outliers, following the same procedure in [21]. The resulting samples were tested for normality using the Kolmogorov-Smirnov test.

TABLE II. RESULTS OF A T-TEST COMPARING THE MEAN TEST ERRORS OF THE DIFFERENT ALGORITHMS

Problem	Training samples	Test samples	Inputs	(a)	(b)	(c)
Spiral	194	192	2	L 6.51	-	E 7.03
Henon	100	100	4	-	-	-
Plant	150	103	13	L 5.37	L 11.98	-

Note: The entries show differences that are statistically significant on a 10% level and dashes mean no significance found. Column (a): iTPA algorithm (“L”) vs. original tangent plane algorithm (“T”). Column (b): iTPA algorithm vs. ELM algorithm (“E”). Column. (c): original tangent plane algorithm vs. ELM algorithm.

The results are tabulated in Table 2. Dashes mean differences that are not significant at the 10% level i.e. the probability that the differences are purely accidental. Other entries indicate the superior algorithm (e.g. iTPA algorithm - L, original tangent plane algorithm - T, ELM algorithm - E), and the value of the t statistic. Column (a) gives a comparison between the new iTPA algorithm and the original tangent plane algorithm. The results show two times L is better (spiral and non-linear dynamic plant) and once T is better (Henon map).

The new iTPA Algorithm performed better on the datasets that were more difficult to learn and so convergence speeds tended to be slower. Where convergence occurred quickly the original algorithm was the better method. Column (b) and (c) give comparisons between the new iTPA and original tangent plane algorithms, and the ELM algorithm. The results show 4 times no statistical difference, once L is better and twice E is better.

This suggests that the generalization performance of the tangent plane method is at least as good as the ELM algorithm, which is one of the best neural network classifiers. In situations where time varying signals are required, such as EEG and ECG signals, the sequential learning ability of the tangent plane algorithm might be the preferred method.

VII. CONCLUSIONS

A new variant of the tangent plane algorithm referred to as iTPA is proposed for feed-forward neural networks. This new algorithm includes two modifications to the existing algorithm. Firstly, a directional movement vector is introduced into the training process to push the movement in weight space towards the origin. This directional vector is built into the geometry of the tangent plane algorithm and implements a weight elimination procedure. Secondly, a high degree polynomial term is utilised to adjust the proportion of weights that receive an inwards push. Thus the algorithm can be tuned to decay specific weights to zero (which can help generalization).

Comparative tests were carried out using the new iTPA and original tangent plane algorithms, the gradient descent back-propagation algorithm and the Extreme Learning Machine. The results indicate that the new iTPA algorithm retains the fast convergence speed of the original method. However, the new iTPA algorithm is prone to problems with stability. Including a 50% reduction in step size often improves convergence behaviour without any diminution in learning speed. The results also show that the new algorithm gives improved generalization relative to the original algorithm in some problems, and has comparable generalization performance in yet others. Further, the generalization performance of the tangent plane method is at least as good as the Extreme Learning Machine, which is one of the best neural network classifiers.

VIII. FUTURE WORK

This paper shows that the newly developed improved tangent plane algorithm (iTPA) is at least as good as the extreme learning machine, which is one of the best neural network classifiers. In situations where time varying signals are required, such as EEG and ECG signals, the sequential learning ability of the improved tangent plane algorithm might be the preferred method.

REFERENCES

- [1] C.W. Lee, "Training feedforward neural networks: an algorithm giving improved generalization," *Neural Networks*, vol. 10, 1997 pp. 61-68.
- [2] P. Bartlett, "For valid generalization, the size of the weights is more important than the size of the network," *Advances in Neural Information Processing Systems 9*, Cambridge, MA: The MIT Press, 1997, pp. 134-140.
- [3] S. J. Nowlan, and G. E. Hinton, "Simplifying neural networks by soft weight sharing," *Neural Computation*, vol. 4, no. 4, pp. 473-493, 1992
- [4] A.S. Weigend, D.E. Rumelhart and B.A. Huberman, "Generalization by weight elimination with application to forecasting," *Advances in neural information processing (3)*, 1991, pp. 875-882.
- [5] P.M. Williams, "Bayesian regularisation and pruning using a Laplacian prior," *Technical report*, (312), 1994
- [6] P.O. Hoyer, "Non-negative matrix factorisation with sparseness constraints," *Journal of machine learning research*, (5): 1457 – 1469. 2004
- [7] Huiwen Zeng, "Dimensionality reduction using a mixed term penalty reduction," *IEEE workshop on machine learning for signal processing*, 2005
- [8] C.M. Ennett and M. Frize, "Weight elimination neural networks applied to coronary surgery morality prediction." *IEEE Trans Inf Technol Biomed*, 2003, 7(2):86-92.
- [9] R.M. Zur, Yulei Jiang, L. Pesce, and K. Drukker, "Noise injection for training neural networks: a comparison with weight decay and early stopping," *Med. Phys.* 2009, 36(10): 4810-4818.
- [10] J. Stoer, "Einführung in die numerische Mathematik," Springer, Vol. 1. S, 1976
- [11] W. Finnoff, F. Hergert, and H.G. Zimmermann, "Improving model selection by non-convergent methods," *Neural Networks*, vol.6, 1993, 771-783.
- [12] E.D. Karnin, "A simple procedure for pruning back-propagation trained neural networks," *IEEE trans neural networks*, vol. 1, no. 2, 1990, pp. 239-242.
- [13] Y.L. LeCun, J.S., Denker, and S.A. Solla, "Optimal brain damage," *Advances in neural information processing systems*, vol.2, 1990, pp. 598-605.
- [14] B. Hassibi, and D.G. Stork, "Second order derivatives for network pruning," *Advances in neural information processing systems*, vol.5, 1993, pp. 164-171.
- [15] S. Fahlman, and C. Lebiere, "The cascade correlation learning architecture," *Advances in neural information processing systems*, vol. 2. 1990, pp. 524-532.
- [16] R. Linder, S. Wirtz, and S.J. Poppl, "Speeding up backpropagation learning by the APROP algorithm," *Proceedings of the second international ICSC symposium on neural computation*, ICSC Academic Press, 2000, pp. 122-128.
- [17] K.L. Lang, and M.J. Witbrock, "Learning to Tell Two Spirals Apart," *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kaufmann 1988.
- [18] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Slew, "Extreme learning machine: Theory and applications," *Neurocomputing*, 70, 489-501, 2006
- [19] J.J.T. Lahnajärvi, M.I. Lehtokangas, and J.P.P. Saarinen, "Evaluation of constructive neural networks with cascade architectures," *Neurocomputing*, vol. 48, 2002, pp. 573-607.
- [20] K.S Narandra, and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE transactions on neural networks*, 1 (1), 4 1990.
- [21] L. Prechelt, "Connection pruning with static and adaptive pruning schedules," *Neurocomputing*, Volume 16, Issue 1, 1997, pp. 49-61

Exploiting the Role of Hardware Prefetchers in Multicore Processors

Hasina Khatoon

Computer & Info. Sys. Engg. Dept.
NED Univ. of Engg. & Technology
Karachi, Pakistan

Shahid Hafeez Mirza

Usman Institute of Engg. & Tech.
Karachi, Pakistan

Talat Altaf

Electrical Engg. Dept.
NED Univ. of Engg. & Tech.
Karachi, Pakistan

Abstract—The processor-memory speed gap referred to as *memory wall*, has become much wider in multi core processors due to a number of cores sharing the processor-memory interface. In addition to other cache optimization techniques, the mechanism of prefetching instructions and data has been used effectively to close the processor-memory speed gap and lower the memory wall. A number of issues have emerged when prefetching is used aggressively in multicore processors. The results presented in this paper are an indicator of the problems that need to be taken into consideration while using prefetching as a default technique. This paper also quantifies the amount of degradation that applications face with the aggressive use of prefetching. Another aspect that is investigated is the performance of multicore processors using a multiprogram workload as compared to a single program workload while varying the configuration of the built-in hardware prefetchers. Parallel workloads are also investigated to estimate the speedup and the effect of hardware prefetchers.

This paper is the outcome of work that forms a part of the PhD research project currently in progress at NED University of Engineering and Technology, Karachi.

Keywords—Multicore; prefetchers; prefetch-sensitive; memory wall; aggressive prefetching; multiprogram workload; parallel workload.

I. INTRODUCTION

Multicore processors are the mainstream processors of today with the number of cores increasing at a fast pace. A number of issues have emerged in these processors that are becoming more acute with the increasing number of cores. A large body of publications has accumulated in the last decade that has summarized these issues. Some of the challenges are presented in [1]. One of the main issues that directly impacts application performance is the large processor-memory speed gap referred to as memory wall by Wulf and Mckee [2] and elaborated by Weidendorfer [3]. Recent researches have sought solution to this problem through on-chip cache hierarchy [4] and novel architectural features like NUCA cache [5]. Other solutions include R-NUCA [6], Victim Replication [7], and Pressure-Aware Associative Block Placement [8]. A detailed summary of the publications related to the memory wall problem is presented in [9].

One of common solution to the memory wall problem is prefetching of instructions and data at every level of memory hierarchy. Prefetching is a latency hiding technique that access instructions and data from the next level of memory hierarchy before the demand for it is actually raised by the processor.

Prefetching was almost always beneficial in single core processors, even though there were some useless prefetches. As a result, prefetchers now form an integral part of most of the current generation processors. In multicore processors, all cores share chip resources that include on-chip memory hierarchy and the processor-memory interface. If all cores generate prefetch requests in addition to demand requests, a large amount of interference takes place causing contention for resources. This prefetcher caused contention may result in performance degradation in multicore processors, especially if prefetchers are used aggressively. Therefore, there is a need to investigate the effectiveness of prefetchers in multicore processors under different conditions and for all types of applications. The contribution of this paper is the analysis and quantification of the behaviour of applications in the presence and absence of prefetchers. The derived results provide guidelines for applications to activate prefetchers only when they are useful.

Recent research has focused on improving data prefetching mechanisms, especially for big data analysis and other streaming applications. Though prefetching pose degradation problems in multicore processors, especially when used aggressively, they remain the most effective mechanism to avoid stalls that are caused due to long latency accesses and contention based delays. This necessitates enhancements in the prefetcher designs that adapt to congestion and dynamically adjust their aggressiveness. Chen et al. in their publications [10, 11] have proposed storage efficient data prefetching mechanisms and power efficient feedback controlled adaptive prefetchers that are accurate and efficient. Other recent enhancements are discussed in Section II.

The rest of the paper is organized as follows. Section II gives a brief overview of related work. Section III outlines the experimental setup including test programs and specifications of the experimental platforms. Section IV presents the results and a brief analysis of the results and Section V concludes the paper.

II. RELATED WORK

Since prefetching is considered to be an important latency hiding technique, it has been used effectively in both single core processors and single core multiprocessors. Prefetching is performed in hardware, in software or in both. Software prefetching is supported by prefetch instructions and requires effort by the programmer or the compiler writer. Nataranjan et al. [12] have explored the effectiveness of compiler directed

prefetching for large data accesses in in-order multicore systems. Since the focus of this work is hardware prefetching, this section shall briefly describe some of the recent publications related to hardware prefetchers in the context of multicore processors.

Prefetchers are beneficial due to the principle of locality, an attribute of software. This is true most of the time in single core architecture, but as pointed out in [13], aggressive prefetching in multicore processors result in a large amount of interference giving rise to performance degradation. Ebrahimi et al. [13] have proposed more accurate prefetching with the use of local as well as global feedback by Hierarchical Prefetch Access Control (HPAC) to partially alleviate the above problem. Using coordinated prefetching, the authors compare the results of aggressive prefetching in multicore processors with that of a single core processor. With dynamic control of the prefetch aggressiveness using feedback directed control, they have shown that their technique improves the system performance by 14%.

Lee et al. [14] have identified degradation in performance due to congestion in the interconnection network especially due to prefetch requests in multicore processors. They have proposed to prioritize demand traffic by using TPA (Traffic-Aware Prioritized Arbiter) and TPT (Traffic-Aware Prefetch Throttling) to counter the negative effects of prefetch requests. Fukumoto et al. [15] have proposed the use of cache-to-cache transfer to reduce the overall congestion on the memory-bus interface.

Kamruzzaman et al. [16] have proposed a different way of using prefetching especially for applications like the legacy software that are inherently sequential in nature and cannot use all cores of the CMP. They have suggested the use of prefetch threads as *helper threads* to run on unused cores and make use of the injected parallelism for prefetching code and data. Using thread migration techniques, an overall improvement of 31% to 63% is shown for legacy software. The authors have concluded in their final analysis that the technique can also be used to enhance the performance of applications that are parallel.

Wu et al. [17] have proposed an automatic prefetch manager that estimates the interference caused by prefetching and adjusts the aggressiveness while programs are running. They have shown that this dynamic management improves the application performance and makes it more predictable. Verma et al. [18] have evaluated the effectiveness of various hybrid schemes of prefetching and have proposed to adaptively reduce the number of prefetches to reduce the interference. Lee et al. [19] identify the lack of parallelism that exists in DRAM banks, especially in multicore processor-based systems. They have proposed mechanisms to maximize DRAM Bank Level Parallelism (BLP) using BLP-aware Prefetch Issue (BAPI) with BLP-Preserving Multi core Request Issue (BPMRI) that helps improve the application performance with parallel servicing of requests. Ebrahimi et al. [20] have proposed mechanisms to exploit prefetching for shared resource management in multicore systems.

Nachiappan et al. [21] have suggested prefetch prioritization in the interconnection network on the basis of

the potential utility of the requests in order to reduce the negative effects of prefetching. Wu et al. [22] characterize the performance of the LLC (Last Level Cache) management policies in the presence and absence of hardware prefetching. They propose Prefetch-Aware Cache Management (PACMan) for better and predictable performance. Lee et al. [23] have proposed prefetch-aware on-chip networks and network-aware prefetch designs that is sensitive to network congestion. Manikantan and Govindarajan [24] have proposed performance-oriented prefetching enhancements that include *focused prefetching* to avoid commit stalls. The authors claim that this enhancement also improved the accuracy of prefetching.

A number of recent publications have proposed complex prefetching mechanisms that take into account various factors while prefetching code and data [18, 25]. Grannaes et al. [25] have proposed Delta Correlating Prediction Table (DCPT), a prefetching heuristics based on the table-based design of Reference Prediction Tables (RPT) and the delta correlating design of Program Counter/ Delta Correlating Prefetching (PC/DC) with some improvements. These complex prefetching techniques have overheads that cannot be ignored as prefetchers incur a significant burden on system resources. Since simple prefetchers have low overheads, they are used mostly in current generation processors. For example, the prefetchers used in our experimental platform are simpler [26] as compared to the prefetchers discussed in [18, 25].

III. THE EXPERIMENTAL SETUP

This section gives an account of the test programs, the experimental platforms and the hardware prefetchers present in these experimental platforms.

Although prefetching code and data have been significantly effective in single core processors, some of the recent publications have pointed out an anomaly that takes place when prefetchers are used in multicore processors. Use of aggressive prefetching cause interference and results in overall degradation of performance [13], demanding an adjustment in the prefetch strategy. In many instances, it has been observed that applications perform better without all the prefetchers used by default as these are built-in in all current generation processors. The designers of most of these processors have therefore provided mechanisms where applications may use prefetch manipulating techniques to selectively enable/ disable the built-in hardware prefetchers, whenever desired. This involves manipulation of Machine Specific Registers (MSRs) related to hardware prefetchers. The decision to enable/ disable prefetchers is left to the application designer. The application areas that benefit most due to cache locality and being prefetch sensitive may continue using the prefetchers, but these applications should also investigate the benefits, before using it by default.

A. Test Programs

Three types of benchmark programs are used to measure and evaluate performance with enabled/ disabled configurations of prefetchers: SPEC 2006 [27], the parallel Parsec Benchmark suite [28] and the concurrent matrix multiplication program [29]. SPEC 2006 is a commonly used

benchmark for evaluation of single and multiprogram workloads; the Parsec Benchmark suite is used to evaluate the performance for parallel workloads and the concurrent matrix multiplication program is used to evaluate the performance of concurrent workloads. A brief description about the three sets of benchmarks/ programs is given in the following subsections.

1) SPEC 2006 Workload

The first set of experiments are conducted to evaluate the effect on the performance of SPEC 2006 [27] benchmark programs when run on multicore processor with various configurations of the built-in hardware prefetchers. Since SPEC 2006 benchmarks are not inherently parallel, multiple copies of each benchmark are run in parallel as a multiprogram workload to keep all cores busy and observe the results with and without prefetchers. Most of the runs are *reportable* runs [27] and the results of *reference* input set is used to perform the analysis. The experiments performed on the example platform took between 9 to 19 hours each and in some cases, it was even higher. In case of multiprogram workloads, the time taken was up to 31 hours for complete execution.

2) The Parallel Benchmarks Suite

The Princeton Application Repository for Shared Memory Computers (PARSEC) is a parallel benchmark suite that is suitable to evaluate multicore processors [28]. It consists of 13 benchmark programs taken from several different application domains including financial analysis, *animation*, data mining, computer vision, etc. These are diverse and emerging multithreaded workloads focusing on desktop and server applications that are expected to be the eventual workloads for multicore processors. The number of threads of each program can be adjusted depending on the number of cores and the application requirements. A detailed description of the design and implementation of this benchmark suite is given in [30]. However, a brief description is given below.

Both current and emerging workloads from recognition, mining and synthesis (RMS) application areas are represented in this benchmark suite. Each of the applications has been parallelized fulfilling the requirements of multithreaded applications that can be run on parallel architectures like multicore processors. Using parallelization models of Pthreads, OpenMP and Intel TBB, these programs provide portability for various types of platforms. Some of the programs present in the suite are *dedup*, *blackscholes*, *facesim*, *fuidanimate*, etc., taken from the application areas of computer vision, data mining, visualization, media processing, animation, financial analysis, etc. Six different input sets with different properties are defined for each workload that can be used with variable number of threads. Out of the input sets of *test*, *simdev*, *simsmall*, *simmedium*, *simlarge* and *native*, the *native* input set is the largest and is closest to the actual inputs.

All 13 benchmark programs are run with the *native* input set using single thread and n threads, where n is chosen to be the number of cores for each of the experiments. A more detailed description about the use of this benchmark suite can be found in [31].

3) Matrix multiplication program

The third test program is the parallel matrix multiplication program. This program has been parallelized to run on multicore processors using SPC³PM (Serial, Parallel and Concurrent Core to Core Programming Model [29]), an algorithm developed at NED University by Ismail et al. for parallelization of programs. This programming model allows the user to specify any number of cores depending on the amount of parallelism and the available resources. More details about the model and algorithm can be found in [29].

B. The Experimental Platforms

Most of the experiments were conducted on a platform based on the 4-core Intel Core2 Quad processor running OpenSUSE 11.1 Linux 2.6.27.7 operating system. The main features of this machine are summarized in Table I as Platform No. 2. Both integer and floating point programs of SPECCPU2006 benchmark suite [27] and the Parsec Benchmark suite [28] were run on this platform using various combinations of the four built-in prefetchers per core in the multicore processor. A detailed description of the four prefetchers per core and a description of the Model Specific Register (MSR) to control them are given in the Intel Software Developers Manual [26].

Some experiments were also conducted on a 2-core and an 8-core machine to examine and validate some of the results obtained from the main platform. The salient features of these platforms are also listed in Table 1 as platform Nos. 1 and 3 respectively. The results of experiments conducted on platforms 1 and 3 are used as additional data for validation and testing of results and only a summary of the results are presented. The platform chosen to run the third test program is the dual-core Intel Xeon processor X5670 Series based SR1600UR server system having 24 cores. Other salient features of this platform are also listed in Table I as the specifications of platform No. 4.

C. Prefetchers in the main Experimental Platform

The example platforms 1 to 3 that are used to conduct most of the experiments in this study have four prefetchers per core, each performing the function of prefetching a specific set of information [26]. A brief description of the four hardware prefetchers in the experimental platforms is given in the following paragraph.

- The Instruction Prefetcher (IP), referred to as pf4 in this paper, prefetches instructions in the L1 instruction-cache based on branch prediction results.
- The Adjacent Cache Line (ACL) prefetcher, referred to as pf2, prefetches the next matching block in a cache block pair in to L2 cache.
- The Data Cache Unit (DCU) prefetcher, referred to as pf3, observes and detects the number of accesses to a specific cache block for a predetermined period of time and prefetches the subsequent block in the L1 D-cache.
- The Data Prefetch Logic (DPL) prefetcher, referred to as pf1, functions similar to the DCU prefetcher, except that the blocks are prefetched in L2 cache after it detects accesses to two successive cache blocks.

TABLE I. SPECIFICATIONS OF EXPERIMENTAL PLATFORMS

	Platform 1	Platform 2 (Main Platform)	Platform 3	Platform 4
Processor	Intel Core™ 2 Duo CPU @ 2.2 GHz	Intel Core™ 2 Quad CPU @ 2.66 GHz	Intel Core™ i7-2600 CPU @ 3.4 GHz	4 x Intel Xeon X5670@ 2.93CHz
No. of cores	2	4	8	4 x 6
Cache and System Parameters				
L1 D-Cache (per core)	32KB, 64B, 8-way associative	32KB, 64B, 8-way associative	32KB, 64B, 8-way associative	6x32KB
L1 I-Cache (per core)	32Kb, 64B, 8-way associative	32KB, 64B, 8-way associative	32KB, 64B, 8-way associative	6 x 32KB
L2 Cache	2MB, 64B, 8-way associative	4MB 64B, 16-way associative	4x256KB, 64B, 8-way assoc.	6 x 256 KB
L3 Cache	NA	NA	8MB, 64B, 16-way assoc.	12 MB
Main mem.	1GB	4GB	8GB	24 GB
Operating System	OpenSUSE 11.1 Linux Kernel 2.6.27.7	OpenSUSE 11.1 Linux Kernel 2.6.27.7	OpenSUSE 11.1 Linux Kernel 2.6.27.7	Windows 2008 Server (64-bit)

Each of the four prefetchers can be selectively enabled/disabled by putting On/Off individual bits in the Model Specific Register (MSR) number *0x1A0h* present in each core. This register can be accessed and the corresponding bits can be manipulated using assembly-level instructions. The tool used to manipulate the register bits for these experiments is *msr tools* [32] available as free software. In addition to hardware prefetchers, prefetch instructions are also provided in all current generation processors that can be used to program prefetching of data through software prefetching. It may be noted that the experimental platform No. 4 does not allow selective enabling/ disabling of its hardware prefetchers. It only allows *all* prefetchers to be either enabled or disabled.

The following paragraphs summarize the results of the experiments performed after selective enabling/disabling of prefetchers and the effect it has on the performance of multicore processors.

IV. RESULTS AND ANALYSIS

Table II gives a summary of the experiments conducted to deduce the following results on the main platform (platform 2). A number of experiments that were conducted on platforms 1, 3 and 4 are also discussed in this section (not given in Table II).

A. Benchmarks and Measurement Metrics

The experiments were conducted by running all 29 SPEC CPU2006 programs comprising of 12 integer benchmarks and 17 floating-point benchmarks, all 13

programs of Parsec Benchmark suite Version 2.1 and the concurrent matrix multiplication program. The effect of the use of prefetch inhibiting techniques on the overall performance of benchmark programs is illustrated through column charts. In addition, collected data is also presented in the form of tables that give more accurate information. Two separate sections present the results of SPEC2006 benchmarks as single program and multiprogram workloads. A third section presents the results of parallel benchmarks.

Some of the terms that shall be used to explain the results in this paper have been taken from [13] and are discussed here briefly. An application is said to have *cache locality* if the number of L2 cache hits per 1000 instructions is greater than five. If the L2 cache miss is greater than 1 per 1000 instructions (MPKI – Miss Per Kilo Instructions), the application is referred to as *memory intensive*. If the improvement in performance when a prefetcher is used is greater than 10% compared to no prefetching, the application is said to be *prefetch sensitive*.

B. SPEC CPU 2006 Results

Results of experiments 1 to 16 that were conducted with various prefetcher options are presented in this section.

1) SPEC CPU 2006 as Single Program Workload

Experiments 1 to 4 and 9 to 12 were conducted by using all SPEC programs as single program workload with various prefetcher options.

2) SPECint as Single Program Workload

In experiments 1 to 4, use of prefetchers mostly proved to be beneficial, because all resources were utilized by only a single core as SPEC benchmarks are not inherently parallel. Fig. 1 shows the performance in terms of execution time for SPECint 2006 benchmarks executed with and without the built-in hardware prefetchers in each of the cores. A number of experiments were conducted using various configurations of on-chip hardware prefetchers. Four of these experiments are listed in Table II. The data collected from the experiments is presented in Table III. An overall average degradation of 14.4% is observed in 10 out of 12 integer benchmarks when the DPL (pf1) prefetcher is disabled. This is because prefetching in L2 is more beneficial for most of the applications. The highest degradation of 54% is observed in *libquantum* benchmark. Since this benchmark consists of a library of software that simulates a quantum computer, it is expected to be prefetch sensitive and benefits most from any kind of prefetching mechanism. The other benchmark programs that are prefetch sensitive are *mcf*, *sjeng* and *xalancbmk*. Two of the benchmarks, namely, *hmmmer* and *omnetpp* give better performance when the prefetcher is off, with *hmmmer* giving an improvement of 7.3%. This is because *hmmmer* is database search software that searches for a gene sequence.

The experiments were again conducted by disabling two of the four prefetchers and a different set of results were obtained (experiment 3). When both DPL (pf1) and ACL (pf2) prefetchers are disabled, there is an average degradation of 13.5% in only 3 out of 12 integer benchmarks and 9 benchmark programs show an average improvement of 8.4%.

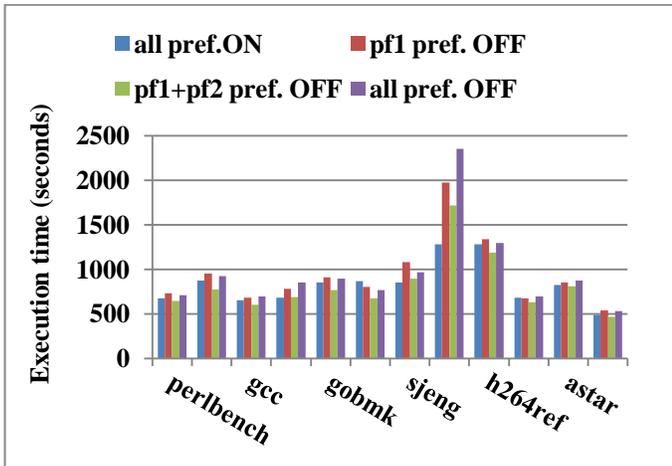


Fig. 1 Execution time of single copy of SPECint 2006 Benchmark programs with various prefetcher configurations

The highest degradation of 34.6% is observed in *libquantum* program because of the reasons mentioned before. The *hmmmer* program shows the highest improvement of 22.6%.

When all four prefetchers are disabled (experiment 4), the benchmarks show degradation in almost all SPECint programs with an average degradation of about 14%, with *libquantum* program suffering from the highest degradation of 84%. The reason for this behaviour has already been mentioned before. Only one of the programs namely, *hmmmer* shows an improvement of 11% when all the four prefetchers are off, because it does not benefit from prefetching.

The results show that prefetchers are beneficial for single SPECint programs in multicore processors. Most of the applications benefit from prefetchers because the interference between demand and prefetch requests is not significant as only a single core generates prefetch and demand requests.

a) SPECfp as Single Program Workload

The results of SPECfp benchmarks are shown in Fig. 2 (experiments 9 to 12). Compared to the integer benchmarks, only 6 out of 17 floating-point benchmarks suffer from an average degradation of 6.1% in performance when the DPL (pf1) prefetcher is disabled. There is only one SPECfp benchmark that is prefetch sensitive, namely *bwaves* which suffers from the highest degradation of 19.8%. *bwaves* is a computational fluid dynamics software that simulates blast waves in three dimensions. Such software tends to benefit from prefetching. An average improvement of 6.4% is observed in 10 out of 17 floating point benchmarks with as high as 13.9% improvement observed in *povray* benchmark program. This is an image rendering software that uses ray tracing to visualize an object.

TABLE II. LIST OF EXPERIMENTS WITH DIFFERENT PREFETCHER OPTIONS ON PLATFORM 2 WITH EXECUTION TIME

Benchmark	Execution Mode	Experiment No.	Prefetcher option	Execution Time in seconds
SPECint	Single Program Workload	1	All Enabled	32400
		2	DPL=Disabled	35520
		3	DPL+ACL= Disabled	31200
		4	All Disabled	35280
	Multi-program workload (4-copies)	5	All Enabled	44400
		6	DPL= Disabled	43860
		7	DPL+ACL = Disabled	47040
		8	All Disabled	46380
SPECfp	Single Program Workload	9	All Enabled	72120
		10	DPL=Disabled	70140
		11	DPL+ACL = Disabled	69960
		12	All Disabled	70680
	Multi-program Workload (4-copies)	13	All Enabled	104400
		14	DPL=Disabled	104520
		15	DPL+ACL= Disabled	106320
		16	All Disabled	112440
PARSEC Benchmarks	Single Thread Workload	17	All Enabled	6840
		18	DPL=Disabled	6780
		19	ACL =Disabled	7020
		20	IP =Disabled	6600
		21	All Disabled	7560
	Multiple Thread Workload (4 threads)	22	All Enabled	3480
		23	DPL=Disabled	3720
		24	ACL=Disabled	3960
		25	IP=Disabled	3360
		26	All Disabled	3720

The ray tracing programs do not benefit from prefetching. As a result of experiment 11, the behaviour of floating point benchmarks remains almost the same as with one prefetcher disabled, with only a small change that can be observed from Fig. 2.

When all the four prefetchers are disabled, 8 out of 17 benchmark programs suffer from an average degradation of 17% with the highest degradation of 31% seen in *GemsFDTD* program. This program benefits from prefetching because it is a computational electromagnetic application that comprises mostly of loops. All the above results indicate that there is anomaly even when a single copy of benchmarks is run and different applications show different behaviour with or without the use of prefetchers. Moreover, floating point benchmarks mostly perform better when prefetchers are disabled selectively as compared to integer benchmarks. A closer examination reveals that most of the SPECfp programs are not prefetch sensitive.

TABLE III. EXECUTION TIME OF SPEC2006 PROGRAMS AS SINGLE PROGRAM WORKLOADS

Benchmarks	Execution Time (in seconds) of Benchmark programs with selective enable/ disable of on-chip Prefetchers			
	All PFs enabled	pf1 disabled	pf1+pf2 disabled	All PFs disabled
Perlbench	670	728	644	712
bzip2	874	951	776	923
Gcc	652	684	601	692
mcf	679	778	686	849
gobmk	852	907	767	896
hmmer	866	803	670	768
sjeng	855	1085	895	968
libquantum	1279	1975	1722	2357
h264ref	1285	1342	1188	1296
omnetpp	682	673	633	692
astar	822	854	810	877
xalancbmk	486	540	469	531
bwaves	1124	1347	1283	1334
gamess	1995	1845	1751	1448
milc	899	907	904	1060
zeusmp	1119	1076	1073	1150
gromacs	1297	1116	1192	900
cactusADM	2263	2185	2199	2016
leslie3d	2046	2190	2165	2348
namd	914	914	915	778
dealII	752	730	728	686
soplex	704	756	811	920
povray	508	437	430	317
calculix	1938	1808	1894	1759
GemsFDTD	1768	1770	1822	2319
tonto	1274	1183	1199	967
lbm	1204	1170	1140	1197
wrf	1520	1536	1586	1763
sphinx3	1720	1685	1576	1802

3) SPEC CPU 2006 as Multiprogram Workload

In this section, we present the results of experiments 5 to 8 and 13 to 16, which were performed by running SPEC programs as multiprogram workload to keep all cores busy. Although multicore processors are more useful and powerful for parallel workloads, most of the software that runs on these processors today is not parallel. For all such software, the main advantage that can be gained with multicore processors

is higher throughput. Hence, studying the behaviour of multicore processors for multiprogrammed workload is also as important as for parallel workloads.

It was observed during these experiments, that there is a large increase in execution time with multiprogram workload. This is because a large number of memory requests, including demand and prefetch requests share the same limited bandwidth of the processor memory interface. The amount of memory required to run the programs also increases. In case of *mcf*, one of the integer benchmarks, the program becomes very slow and its progress almost stops on our experimental platforms because of the heavy usage of memory. This program is therefore not included in these measurements. For all other programs, the interference caused by multiple requests result in an overall degradation in performance as compared to the single program run on multiple cores. The execution time on the 4-core machine increases by an average of 50% for all integer benchmarks with the highest increase of 200% in the *libquantum* benchmark program. In case of floating-point benchmarks, multiprogram workload increases the execution time by almost 74% over the single program execution time, with the highest increase of 206% in the *lbm* benchmark program. In such a scenario it would not be fair to compare the performance of benchmarks when a single program is run with the performance of multiprogram workload with and without hardware prefetchers. The comparison is therefore made when a single program is run with and without prefetchers and when multiprogram workload is run with and without the hardware prefetchers.

The performance further degrades when multiprogram workload executes on an 8-core machine. The 8-core machine is an i7-based computer and the architecture of cores is similar to that of our main experimental platform. The gap between the execution time of single program and multiprogram is much wider than that of the 4-core machine. The average degradation in performance for 8-copies of integer benchmarks as compared to a single program run on the same machine is 150%, with *libquantum* suffering from the highest degradation of 490%. Fig. 3 shows the comparison of execution time of each SPECint benchmark program. When floating point benchmarks are run as multiprogram workload,

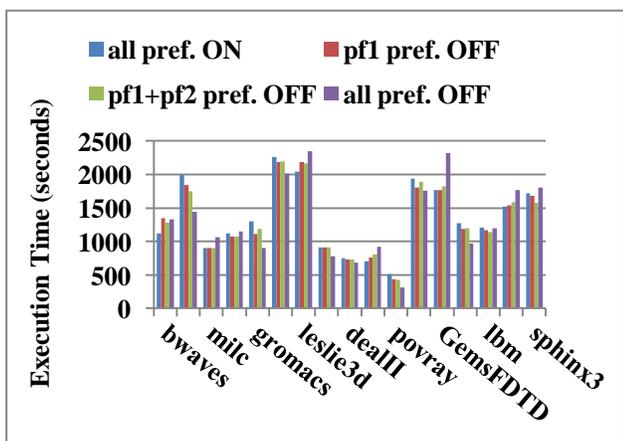


Fig. 2 Execution time of single SPECfp 2006 Benchmark programs with different prefetcher configurations

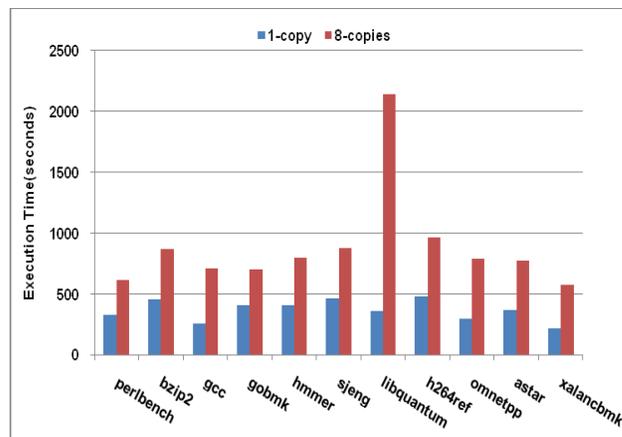


Fig. 3 Comparison of execution time of SPECint programs as single and multiprogram workloads on an 8-core machine (Platform No. 3)

the average degradation as compared to a single program is 173% with *lbm* suffering from the highest degradation of above 700%. The contention for resources is much higher in an 8-core machine because of a higher interference/ conflict between demand and prefetch requests giving rise to higher execution times. Similar results are obtained on a 2-core machine where two copies of benchmarks take much longer to execute as compared to a single program on the same machine.

b) SPECint as Multiprogram Workload

Keeping in view the objectives mentioned in the first part of section V.B.2 to compare multiprogram workloads separately, Fig. 4 summarizes the performance measurements of the benchmarks as a result of experiments 5 to 8. Table IV presents the data that were collected from these experiments. The increase in execution time is attributed to a number of factors including the interference that takes place between the demand and prefetch requests generated by all cores. The observations from Fig. 4 are summarized in the following paragraph.

Five out of 11 integer benchmarks suffer from an average degradation of 1.2% if the DPL (pf1) prefetcher is disabled (experiment 6), with the highest degradation of 2.2% observed in *h264ref* benchmark. This is video compression software that encodes video streams using two different parameter sets. 6 out of 11 integer benchmarks show an average improvement of 4.4% with the highest improvement of 12% observed in *omnetpp* benchmark. The *omnetpp* benchmark performs discrete event simulation by modelling a large Ethernet network on a campus. When both DPL (pf1) and ACL (pf2) prefetchers are disabled (experiment 7), 9 out of 11 integer benchmarks suffer from an average degradation of 10.7% with *sjeng* suffering from the highest degradation of 16.9%. Almost 25% of integer benchmarks perform better with an average improvement of 6.5%. When all the prefetchers are disabled, 9 out of 11 benchmark programs degrade in performance with an average degradation of 8% and the highest degradation of 15.2% is observed in *bzip2* program. *bzip2* is a compression software that benefits from prefetching. Two of the integer benchmarks improve in performance with an average improvement of 10.4%.

With multiprogram workload, disabling DPL prefetcher gives a better performance for most of SPECint benchmarks. This prefetcher belongs to L2 cache, which is interfaced to main memory.

a) SPECfp as Multiprogram Workload

The result of SPECfp programs (experiments 13 to 16) is illustrated in Fig. 5 with the data presented in Table IV. When the DPL(pf1) prefetcher is disabled, the SPECfp benchmarks show an average degradation of 3.3% in 10 out of 17 benchmarks with the highest degradation of 9% in *leslie3d* benchmark. This is a computational fluid dynamics program consisting of a large number of loops that benefit from

prefetching. There is an average improvement of 3.3% in 7 out of 17 floating point benchmarks with the highest improvement of 16% in *milc* benchmark.

Almost 71% SPECfp benchmarks suffer from an average degradation of 4.7% when both DPL(pf1) and ACL(pf2) prefetchers are disabled (experiment 15), with the highest degradation of 10.5% in *leslie3d* benchmark. On the other hand, 5 out of 17 benchmark programs show an average improvement of 4.8% with the highest improvement of 17.2% observed in *milc* program. SPECfp gives the best performance improvement when the ACL(pf2) prefetcher is disabled with an average improvement of 8.2% in all programs. The highest improvement of 14.3% takes place in *povray* program, which is a computer visualization program that renders images through ray tracing.

TABLE IV. EXECUTION TIME OF SPEC2006 BENCHMARKS AS MULTIPROGRAM WORKLOADS

Benchmarks	Execution Time (in seconds) of 4-copies of Benchmark programs with selective enable/disable of on-chip Prefetchers			
	All PFs enabled	pf1 disabled	pf1+pf2 disabled	All PFs disabled
perlbench	817	799	936	882
bzip2	1131	1144	1315	1303
gcc	948	908	953	977
gobmk	1023	1018	1159	1109
hmmmer	950	957	1067	1037
sjeng	1139	1141	1331	1244
libquantum	3771	3837	4002	4003
h264ref	1566	1601	1758	1687
omnetpp	1247	1098	1094	1037
astar	1269	1212	1261	1220
xalancbmk	794	768	825	835
bwaves	2129	2244	2276	2370
gamess	2341	2359	2427	2161
milc	2304	1930	1907	1819
zeusmp	1416	1445	1516	1498
gromacs	1430	1437	1474	1303
cactusADM	2660	2658	2705	2542
leslie3d	2698	2939	2980	3654
namd	1131	1121	1137	1053
deallI	915	951	967	1011
soplex	1536	1487	1487	1631
povray	559	564	568	513
calculix	2232	2227	2221	2184
GemsFDTD	3042	2984	2971	3301
tonto	1584	1618	1633	1645
lbm	3678	3650	3653	3646
wrf	2001	2124	2120	2650
sphinx3	2713	2778	2884	3827

C. Parallel Benchmarks Results and Analysis

Three sets of experiments were conducted using the parallel benchmarks of 'Parsec Benchmark suite'. The first and second set was run on platform number 2 and the third set of experiment was run on platform 3. The results are presented in the following paragraphs.

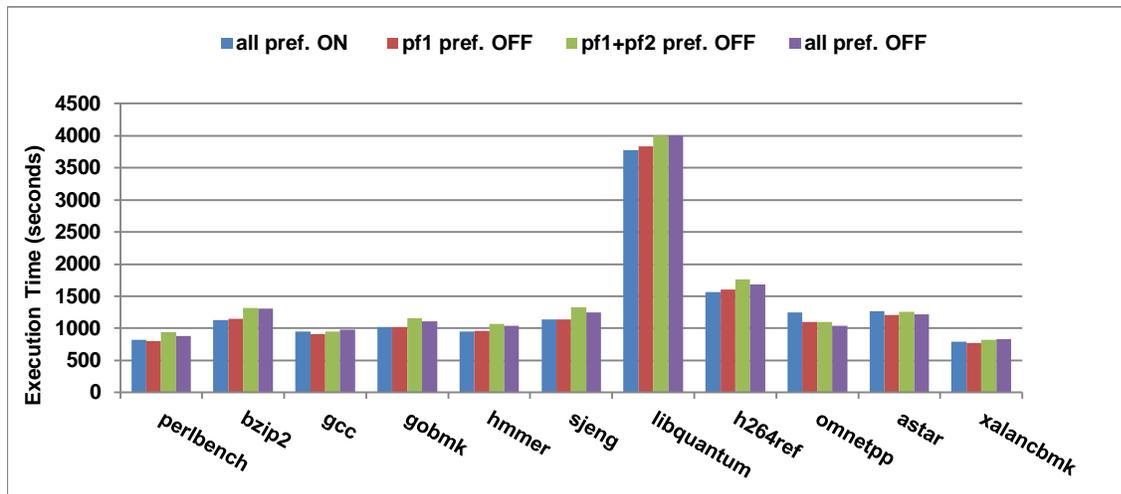


Fig. 4 Execution time of SPECint 2006 Benchmarks as multiprogram workloads (4-copies) with different prefetcher configurations

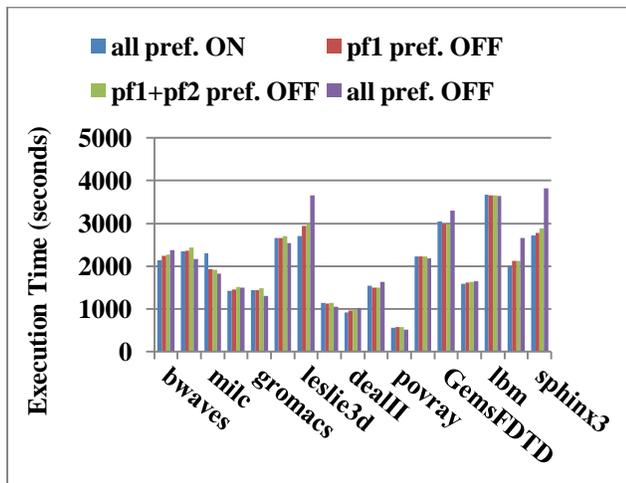


Fig. 5 Execution time of SPECfp 2006 Benchmarks as multiprogram workload (4-copies) with different prefetcher configurations

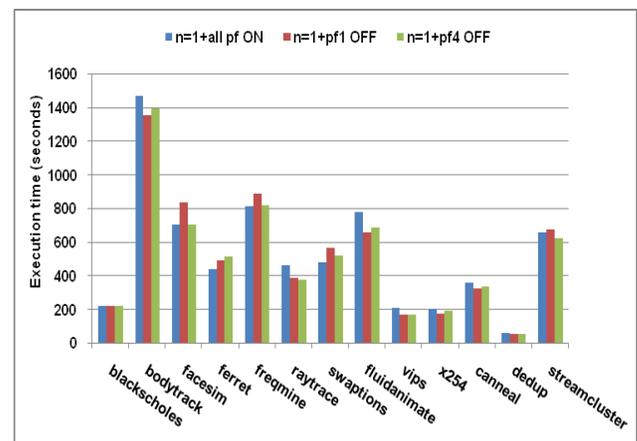


Fig. 6 Execution time of PARSEC Benchmarks with single thread and different prefetcher configurations

1) Parsec Benchmarks with Single Thread

Experiments 17 to 21 were performed with single thread and various configurations of hardware prefetchers. In experiment 18, 7 out of 13 benchmarks perform 9.5% better on the average, with the *vips* benchmark giving the highest improvement of 19.6%. The best results are obtained when the IP prefetcher is disabled (experiment 20), giving an average performance improvement of 5.3% in 9 out of 13 programs with the highest improvement of 20.6% in *vips* benchmark program. This is a media processing application that applies a series of transformations to an image. Other benchmarks that perform better with IP disabled are mostly image processing related applications. Fig. 6 gives the comparison of execution times when all prefetchers are enabled versus the DPL prefetcher disabled versus the IP prefetcher disabled respectively.

2) Parsec Benchmarks with n Threads

Experiments 22 to 26 were conducted using four parallel threads on a 4-core machine and eight parallel threads on an 8-

machine. As expected, there is an overall improvement in execution time with an average speedup of 2.2 over the single thread execution time on the 4-core machine with the highest speedup of 2.8 in *vips* benchmark program. Similarly, there is an average speedup of 3.3 over a single thread execution time on an 8-core machine. Fig. 7 shows the comparison between the execution times of Benchmark programs using a single and 8-threads on the 8-core machine.

The overall performance improves when prefetchers are enabled/ disabled for each of the benchmark programs. The best performance is achieved when the IP Prefetcher is disabled (experiment 25), where 11 out of 13 benchmark programs give an average improvement of 6.4% with the highest improvement of 19.3% in *streamcluster* program. This is a machine learning application that performs optimal clustering for a stream of data points. It is a prefetch sensitive application that benefits from prefetching into L1 cache. Fig. 8 gives the comparison of execution time of the benchmark programs when all prefetchers are enabled versus the IP prefetcher disabled versus all prefetchers disabled.

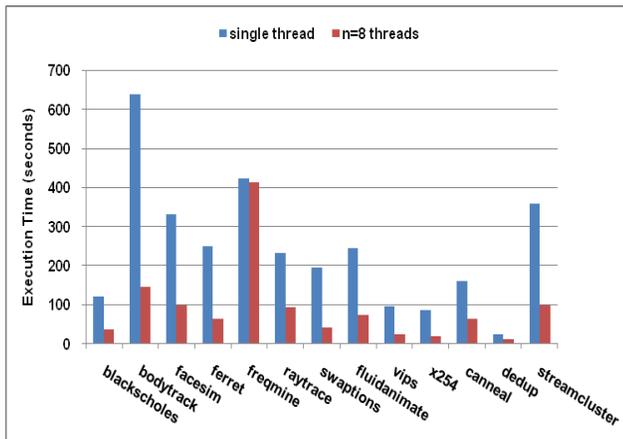


Fig. 7 Execution time of PARSEC Benchmarks with single and 8-thread on an 8-core machine (Platform No. 3)

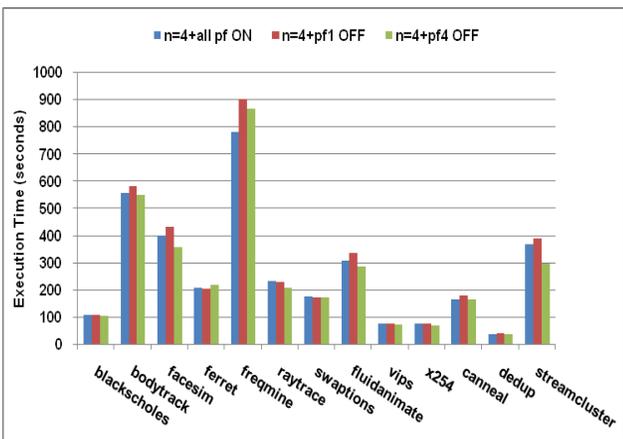


Fig. 8 Execution time of PARSEC Benchmarks with 4-threads on the 4-core machine (Platform No. 2)

D. Concurrent Matrix Multiplication

The platform used to run this program did not allow selective enabling and disabling of hardware prefetchers. This platform only allows all the prefetchers to be enabled/disabled. The experiments were conducted by varying the number of matrix elements from 100x100, 1000x1000, 2000x2000 to 10000x10000 for both integer and floating point operands and by varying the number of cores from 4, 8, 12 to 24. Some results of Matrix multiplication program for integer and floating point operands on the 24-core platform is given in Fig. 9 (a) and (b) respectively. There is a degradation observed in all cases when the prefetchers are disabled with an average degradation of 6.7% for integer operands and 5.95% for floating point operands, indicating that the use of prefetchers is beneficial for concurrent matrix multiplication program.

E. Observations and Analysis

The results presented in this paper give an insight into the effectiveness of hardware prefetching, one of the most commonly used cache optimization technique in multicore processors. We have carried out a detailed set of experiments to estimate the performance with and without the built-in hardware prefetchers in multicore processors on a number of

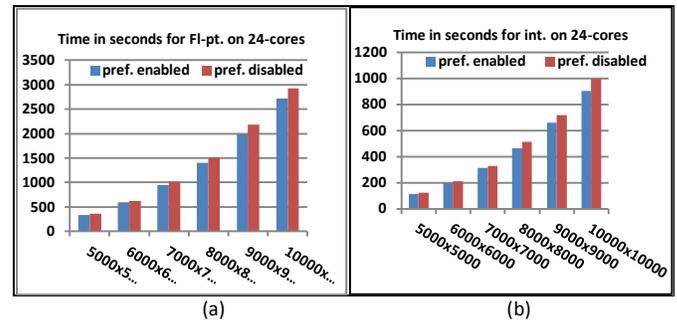


Fig. 9 Execution time of Conc. Matrix Mult. program for (a) Fl. Pt. (b) Integer

platforms. Both multiprogrammed and parallel workloads were used to study the effect. Two separate subsections briefly outline the observations and analysis of multiprogrammed and parallel workloads with various combinations of hardware prefetchers.

1) Multiprogrammed Workload

The results indicate that the effect of prefetching varies when an application is run as single program on a multicore processor compared to the case when the application is run as multiprogram workload. This is mainly because single programs suffer from less contention and interference. In general, most of the integer benchmarks benefit from prefetchers, whereas most of the floating-point benchmarks perform better without prefetchers.

Prefetching may be beneficial for applications that are prefetch sensitive. A larger number of integer applications are prefetch sensitive as compared to floating-point benchmarks. Even among integer applications, very few are prefetch sensitive, especially when run as multiprogram workloads. This is because the benefits of prefetching are overshadowed by the problems caused due to contention for resources and the interference between demand and prefetch requests generated by all cores. Some of the prefetches may also be useless. Prefetch to L1 cache by all cores cause redundant prefetches as multiple copies of the same block reside in multiple L1 caches. This also results in waste of cache space. In addition, all applications do not benefit from prefetching and do not exhibit the same behaviour for all types of prefetching. Database applications, image rendering through ray tracing, data mining applications and some image processing applications are some of the example areas that perform better with selective disabling of prefetchers.

The floating point benchmarks demonstrate a different behaviour pattern as compared to integer benchmarks. Most of these benchmarks perform better without prefetchers, especially when ACL prefetcher is disabled.

Another aspect that was explored was the performance of multicore processors for multiprogram workload. There is a significant increase in execution time as compared to single program workload. The average increase is 50% for a 4-core machine (4-copies) and 150% for an 8-core machine (8-copies) for SPECint benchmarks. Similar results are obtained for SPECfp benchmarks. The main reason attributed to this behaviour is the large amount of contention for resources,

which increases with the increasing number of cores. The proposed solution to this problem is that there should be a proportional increase in resources with the increase in the number of cores. This includes memory capacity, bandwidth of the interface between processor and memory and other components of the computer, as in case of conventional multiprocessors. This is not what is observed from the architecture of multicore-based computers. Even though it may be possible to write fully parallel software that concurrently uses all cores of a multicore processor, the performance may not be as good as expected because of the above-mentioned reasons.

2) Parallel Workload

Since most of emerging applications for multicore processors are parallel workloads, the results obtained from these experiments are significant. When all prefetchers are enabled, average speedup of 4-threads execution is 2.2 over single thread execution (experiments 17 and 22). The speedup improves for most of the applications when the hardware prefetchers are manipulated. For example, the highest speedup of 3.1 is obtained when all prefetchers are disabled and *vips* program is run with four threads on the four-core machine (experiment 26). The main reason for this improvement is that there is less contention and interference among threads when prefetchers are disabled. The prefetch sensitive parallel benchmarks degrade in performance when hardware prefetchers are disabled. For example, *freqmine* degrades in performance when prefetchers are disabled. This is a data mining application that identifies frequently occurring patterns in transaction databases. Fig. 8 gives an insight about other programs in this benchmark suite.

The use of prefetchers is beneficial for matrix multiplication program. The performance is better with prefetchers enabled because this is a data intensive application where the data access pattern is regular and predictable. Prefetching is considered to be suitable for such applications. The performance improves proportionately with the increase in the number of threads/ cores.

V. CONCLUSIONS

The role of hardware prefetchers have been exploited to examine their effectiveness in multicore processors with the goal of improving the overall system performance. Due to heavy sharing of on-chip resources including cache memory, there is degradation in performance when prefetchers are used aggressively, especially with multiprogram and parallel workloads.

The prefetchers need to be selectively enabled/ disabled depending upon the nature of the application and the type of prefetching. The selective use of prefetchers can control the interference of prefetch requests which interfere with demand requests due to extensive sharing of bandwidth at all levels of memory hierarchy and to the cache pollution caused due to useless prefetches. This results in better overall performance, thus effectively reducing the processor memory speed gap and lowering the memory wall.

Test results based on single program workload, concurrently running multiprogram workloads and parallel

workloads confirm that appropriate enabling/ disabling of prefetchers can be used by application programmers to improve the execution time of programs. Experimental results indicate that database applications, image rendering applications, animation and some data mining applications perform better when prefetchers are disabled selectively.

REFERENCES

- [1] J. Parkhurst, J. Darringer, B. Grundmann, "From Single Core to Multi-Core: Preparing for a new exponential", Proc. of ICCAD, 2006, p. 67-72
- [2] W. A. Wulf, S.A. McKee, "Hitting the Memory Wall – Implications of the Obvious", ACM SIGARCH Computer Architecture News, 1995. p. 20-24
- [3] J. Weidendorfer, "Understanding Memory Access Bottlenecks on Multicore", Mini Symposium – Scalability and Usability of HPC Programming Tools, ParCo2007, FZ Julich
- [4] L. Hsu, R. Iyer, S. Makineni, S. Reinhardt, D. Newell, "Exploring the Cache Design Space for Large Scale CMPs", ACM SIGARCH Computer Architecture News, 2007, Volume 33, Issue 4: 24-33
- [5] C. Kim, D. Burger, S. W. Keckler, "NUCA: A Non-Uniform Cache Access Architecture for Wire-Delay Dominated On-Chip Caches", IEEE Micro, November/December 2003. p. 99-107
- [6] N. Hardavellas, M. Ferdman, B. Falsafi, A. Ailamaki, "Near-Optimal Cache Block Placement with Reactive Non-uniform Cache Architecture", IEEE Micro, January/February (2010), p. 20-28
- [7] M. Zhang, K. Asanović, "Victim Replication: Maximizing Capacity while Hiding Wire Delay in Tiled Chip Multiprocessors", Proceedings of the 32nd International Symposium on Computer Architecture (ICSA-32), 2005. p. 336-345
- [8] M. Hammoud, S. Cho, R. G. Melhem, "A Dynamic Pressure-Aware Associative Placement Strategy for Large Scale Chip Multiprocessors", IEEE Computer Architecture Letters, Volume 9, No.1: January-June, 2010: 29-32
- [9] H. Khatoun, S. H. Mirza, "Improving Memory Performance Using Cache Optimizations in Chip Multiprocessors", Sindh University Research Journal (SURJ), Volume 43, Number 1A, June 2011: 43-50
- [10] Y. Chen, H. Zhu, H. Jin, X. Sun, "Algorithm-level Feedback-controlled Adaptive data Prefetcher: Accelerating data access for high-performance processors", Parallel Computing 38(2012) 533-551
- [11] Y. Chen, H. Zhu, H. Jin, X. Sun, "Storage-Efficient Data Prefetching for High Performance Computing", adfa, p.1, Springer-Verlag Berlin, 2012
- [12] R. Natarajan, V. Mekkat, W. C. Hsu, A. Zhai, "Effectiveness of Compiler Directed Prefetching on Data Mining benchmarks", Journal of Circuits, Systems and Computers, Vol. 21, No.2, 2012, 23 pages.
- [13] E. Ebrahimi, O. Mutlu, C. J. Lee, Y. N. Patt, "Coordinated Control of Multiple Prefetchers in Multi-Core Systems", Proceedings of the 42nd International Symposium on Micro-architecture (MICRO), Dec.2009, New York. p. 327-336
- [14] J. Lee, M. Shin, H. Kim, J. Kim, J. Huh, "Exploiting Mutual Awareness between Prefetchers and On-chip Networks in Multi-cores", 2011 International Conference on Parallel Architectures and Compilation Techniques (PACT 2011). p. 177-178
- [15] N. Fukumoto, T. Mihara, K. Inoue, K. Murakami, "Analyzing the Impact of Data Prefetching on Chip MultiProcessors", Proceedings of 13th Asia-Pacific Computer Systems Architecture Conference, 2008. p. 1-8
- [16] M. Kamruzzaman, S. Swanson, D. M. Tullsen, "Inter-core Prefetching for Multicore Processors Using Migrating Helper Threads", Proceedings of ASPLOS 2011, ACM. p. 393-404
- [17] C. J. Wu, M. Martonosi, "Characterization and Dynamic Mitigation of Intra-Application Cache Interference", Proceedings of IEEE International Symposium on Performance Analysis of System & Software (ISPASS 2011), p. 2-11
- [18] S. Verma, D. M. Koppelman, L. Peng, "Efficient Prefetching with Hybrid Schemes and Use of Program Feedback to Adjust Prefetcher Aggressiveness", Journal of Instruction-Level Parallelism 13 (2011): 1-14

- [19] C. J. Lee, V. Narasiman, O. Mutlu, Y. N. Patt, "Improving Memory Bank-Level Parallelism in the Presence of Prefetching", Proceedings of 42nd IEEE/ACM International Symposium on Micro-architecture (MICRO 2009), p. 327-336
- [20] E. Ebrahimi, C. J. Lee, O. Mutlu, Y. N. Patt, "Prefetch-Aware Shared-Resource Management for Multi-Core Systems", Proc.of ISCA,2011
- [21] N. C. Nachiappan, A. K. Mishra, M. Kandemir, A. Sivasubramaniam, O. Mutlu, C. R. Das, "Application-aware Prefetch Prioritization in On-Chip Networks", Proceedings of PACT, 2012
- [22] C. J. Wu, A. Jaleel, M. Martonosi, S.C. Steely Jr.,J. Emer, "PACMan: Prefetch-aware Cache Management for High Performance Computing", MICRO 2011
- [23] J. Lee, H. Kim, M. Shin, J. Kim, J. Huh, "Mutually Aware Prefetch and On-chip Network Designs for Multi-cores", IEEE Transactions on Computers, Preprint, 26 April 2013.
- [24] R. Manikantan, R. Govindarajan, " Performance-oriented Prefetch Enhancements Using Commit Stalls", Journal of Instruction-level Parallelism 13(2011) 1-28
- [25] M. Grannaes, M. Jahre, L. Natvig, "Storage Efficient Hardware Prefetching using Delta-Correlating Prediction Tables", Journal of Instruction-Level Parallelism 13 (2011)
- [26] Order Number 325462-040US. Intel® 64 and IA-32 Architectures Software Developer's Manual, Combined Volumes: 1, 2A, 2B, 2C, 3A, 3B and 3C. October 2011
- [27] SPEC CPU2006 Standard Performance Evaluation Corporation. Details can be found at the web site <http://www.spec.org/>
- [28] PARSEC (Princeton Application Repository for Shared-Memory Computers) website: address follows. Parsec v 2.1 Benchmark suite from the following website:
- [29] M. A. Ismail, S. H. Mirza, T. Altaf, "Concurrent Matrix Multiplication on Multi-Core Processors", International Journal of Computer Science and Security (IJCSS), Volume (5): Issue (2): 2011, p. 208-220
- [30] C. Bienia, "Benchmarking Modern Multiprocessors", PhD Thesis, Department of Computer Science, Princeton University, January 2011
- [31] M. Bhadauria, V. Weaver, S. Mckee, "Understanding PARSEC Performance on Contemporary CMPs", Proceedings of 2009 IEEE International Symposium on Workload Characterization, p. 98-107
- [32] msr tools and documentation from any of the following web sites
sourceforge.net/projects/msr
www.kernel.org/pub/linux/utils/cpu/msr_tools
Downloaded in April 2012
<http://parsec.cs.princeton.edu> downloaded in April 2012

Improving Assessment Management Using Tools

Shang Gao, Jo Coldwell-Neilson, Andrzej Goscinski
School of Information Technology
Deakin University
Waurin Ponds, Vic. 3217 Australia

Abstract—This paper firstly explains the importance of assessment management, then introduces two assessment tools currently used in the School of Information Technology at Deakin University. A comparison of assignment marking was conducted after collecting test data from three sets of assignments. The importance of providing detailed marking guides and personalized comments is emphasized and future possible extension to the tools is also discussed at the end of this paper.

Keywords—assessment management; WebCT Vista; Desire2Learn; CloudDeakin; marking guide; personalized comment; Markers Assistant; On-line Grades System

I. INTRODUCTION

Assessment management plays very important role at all educational levels. Good assessment management helps educators collect data about students' learning and their performance, as well as informing decisions about classroom instruction and curriculum content based on the collected data to personalize students' learning and maximise the outcomes.

The assessment process may consist of a number of tasks. First, creating assessment material, which can be essays, diagrams, drawings, programs, databases, spreadsheets and so on; then collecting the submissions after students complete the assessment; marking the submissions and recording grades before delivering feedback and results back to students. Students then review their results and if they feel they have been unfairly assessed, they can request a remark and marks adjusted accordingly.

This whole process is time consuming. Every assessor has their own customized methods and opinions of how to do the assessment. Also every assessor needs to provide accurate and meaningful evaluation feedback to students, which puts a lot of pressure on them.

To meet the assessment needs of both students and staff, the following factors should be considered [1]:

- The criteria for marking must meet the objectives of the course.
- The marking must provide a measure of the learning.
- The marking should provide effective feedback to the student.

II. BACKGROUND AND RELATED WORKS

There are many educational software development companies and institutions that have been providing solutions to improve student achievement. For instance,

CompassLearning Odyssey [2] provides assessment, curriculum, data management, and American state standards correlation engine. Its browser-based solution allows administrators and teachers to track student, class, school, and district data, aggregate and disaggregate the data, and produce detailed reports.

WebQuiz XP, developed by Smartlite [3], can be used to create online quizzes, tests, assessments and questionnaires. A custom explanation message can be shown if a wrong answer is chosen; users can even set a different message according to the answer given. WebQuiz XP also supports surveys or psychological tests, where questions do not have correct answers; this way, it can be used to collect data and display statistics.

Deakin University has been using WebCT Vista [4] as its primary on-line learning environment, called Deakin Studies Online (DSO) for a few years. Recently it is changed to Desire2Learn [5], called CloudDeakin. Both these learning environments are powerful teaching and learning platforms providing discussions, whiteboard, content reuse, performance reporting, on-line quizzes, easy-to-manage gradebook and assignment tools, etc. However, very few backend tools are available for assessment management in the WebCT Vista or Desire2Learn environment. They do not meet the many varied assessment needs of education today, such as creating a detailed marking guide, or returning students a result report with detailed comments and statistical information about the whole assignment.

For instance, in the School of Information Technology, students are usually required to submit their assignment attempts in electronic form. CloudDeakin provides dropboxes and a simple rubric editing/marking interface. But the majority of marking is still a manual process. As far as the authors are aware, many academics, especially in the mathematics discipline, simply download assignments from CloudDeakin, print them out, mark on paper and return to students. Alternatively they demand hard-copy submissions. Others may mark via the CloudDeakin provided interface directly, which is all right to provide comments, but it is hard to integrate a detailed marking guide or statistical information in the returned assessment report. Even using the rubric functionality in CloudDeakin has limitations. It is rather basic with only one grading scale accommodated across all marking criteria.

Some Computer Assisted Assessment tools support the creation of online quizzes with answers but do not provide the functionality or flexibility that we expect and need for effective reporting of feedback to students, especially the integration

with CloudDeakin. Colleagues in the School of Information Technology, Deakin University have developed two assessment tools, Markers Assistant (MA) [6] and On-line Grades System (OGS) [7].

The objectives of both these tools are to reduce the amount of time spent on administrative tasks associated with marking, increase the range of feedback that can be easily delivered to students, and provide an easy interface to access online submissions as well as deliver marks and feedback to students, thus allowing more time to be spent on working one-to-one with students to achieve good results.

III. ASSESSMENT TOOLS COMPARISON

In the following sections, we briefly introduce these two tools (Markers Assistant and On-line Grades System) and then use SIT104 Introduction to Web Development assignment marking as a case study to compare these assessment tools against the above assessment criteria. Conclusions and recommendations are also given at the end of this paper. SIT104 is a core unit in the Bachelor of Information Technology. Students generally complete this unit in the second half of their first year of study.

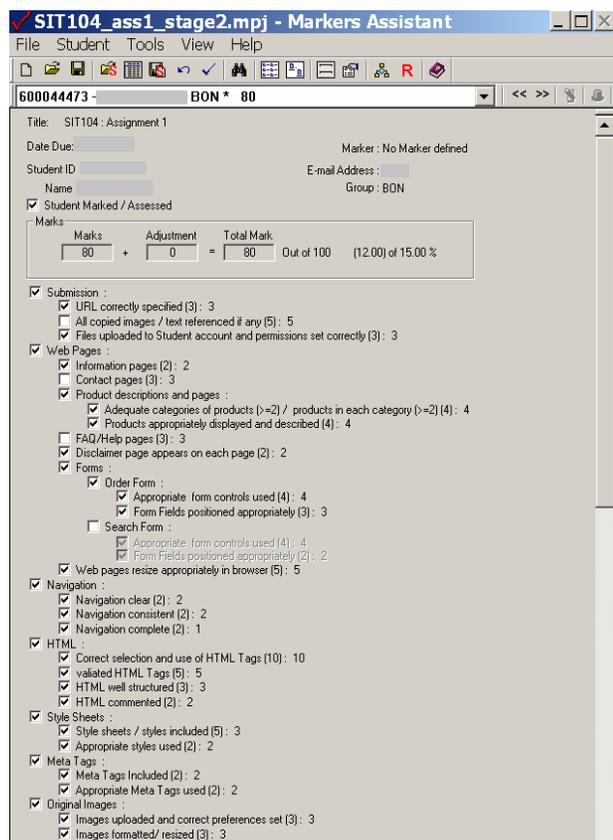


Fig. 1. Example MA marking guide

A. Markers Assistant

Markers Assistant (MA) [6] is a Windows based application developed to provide a flexible method of managing, assessing and delivering results to small and large numbers of students. It is designed to automate as many components of the marking process as possible, whilst

maintaining and improving the assessment feedback to the students within a reasonable timeframe [1]. It can:

1) provide automation that retrieves student submissions and presents them via a predetermined application

Individual students are identified by information such as student number or ID and an associated email address if results are to be delivered via email. Students can be found either by identifying them in a drop down menu via their ID or email address. A predetermined application is triggered automatically based on the format of the submission.

2) provide a flexible marking guide GUI that only requires the identification of the criteria and/or assessment of an item's value

The marking guide is displayed in the form of checkboxes and is organised in a hierarchical configuration where high level criteria must be met before other items within that criteria are considered, as shown in Fig. 1.

3) automate the calculation of results within the application

Each time a checkbox is ticked the marks are adjusted accordingly. Comments can be added, edited or deleted against an individual item in the marking guide or as a general comment that relates to the submission as a whole. All comments entered are stored in a general repository and can be reused for another student if required.

4) provide facilities to deliver student results via email

When the marking is complete, all results can be emailed to the students or exported in a file format that is recognised by WebCT or Desire2Learn.

5) provide facilities to collate final marks summaries

The result report is a text file containing the assessment criteria and the associated mark the student received for each item in the criteria. Any comments relating to the assessment are also included, as shown in Table. I. It also provides result summaries of all assessment performed, which is not possible in traditional methods of assessment. MA enables the following statistics [1]:

Marks summaries for each student:

- criterion mark; comment adjustments; total marks

Statistics for all students:

- number of students; number of students marked; number of zero marks; number of full marks; average mark; maximum mark; minimum mark

Item Statistics:

- item no; attempted; not attempted; average mark; description

Comments for all students:

- frequency

This analysis data helps students gain a better understanding of their strengths and weaknesses as well as provide them with an overview of all items within an assessment (Table. II).

TABLE I. EXAMPLE ASSESSMENT REPORT

SIT104 -Assignment 1 Marking Guide

Legend:	
# - Comments have been made, refer to SPECIAL COMMENTS by selecting the link	
Submission	.
URL correctly specified	3/3
All copied images / text referenced if any	0/5
Files uploaded and permissions set correctly	3/3
Web Pages	.
Information pages	2/2
Contact pages	0/3
Product descriptions and pages	.
Adequate categories of products (>=2) / products in each category (>=2)	4/4
Products appropriately displayed and described	4/4
FAQ/Help pages	0/3
Disclaimer page appears on each page	2/2
Forms	.
Order Form	.
Appropriate form controls used	4/4
Form Fields positioned appropriately	3/3
Search Form	.
Appropriate form controls used	0/4
Form Fields positioned appropriately	0/2
Web pages resize appropriately in browser	5/5
Navigation	.
Navigation clear	2/2
Navigation consistent	2/2
Navigation complete	1/2
HTML	.
Correct selection and use of HTML Tags	10/10
validated HTML Tags	5/5
HTML well structured	3/3
HTML commented	2/2
Style Sheets	.
Style sheets / styles included	3/5
Appropriate styles used	2/2
Meta Tags	.
Meta Tags Included	2/2
Appropriate Meta Tags used	2/2
Original Images	.
Images uploaded and correct preferences set	3/3
Images formatted/ resized	3/3
Relevant images used	2/2
Colors	.
Suitable colors used	2/2
Consistent use of color	2/2
Fonts	.
Font size and color suitable	2/2
Consistent use of Fonts	2/2
Sub Total	80
General Comments and Adjustments	
1. You should have demonstrated the use of images by using more than 2	0
Adjustments Sub Total	0
Total Mark	80 (Out of 100)
12.00 (Out of 15.00%)	

SPECIAL COMMENTS

Refer to the following for explanations on why you may have lost marks for individual items within the marking guide.

6) Enable all data collected to be saved and retrieved via a project file

Backup projects are automatically created in the background, as each student is marked and backup projects saved when the application is closed.

TABLE II. EXAMPLE STATISTICS DATA

----- Project Statistics -----										
Number of students	:	210								
Number of students marked	:	210								
Number of students NOT marked	:	0								
Number of zero marks	:	45								
Number of full marks	:	32								
Highest mark	:	100								
Lowest mark	:	0								
Average mark	:	68.05/100								
Average mark (no zero's)	:	86.61/100								
Variance (no zero's)	:	178.20								
Standard Deviation (no zero's)	:	13.35								
----- Item Statistics -----										
Item		Attempted		Not Attempted		Avg		Avg (no zero's)		Description
1		155		55		2.21/3		2.82/3		URL correctly specified
2		91		119		2.12/5		2.70/5		All copied images / text referenced if any
3		155		55		2.20/3		2.79/3		Files uploaded and permissions set correctly
5		153		57		1.45/2		1.84/2		Information pages
6		150		60		2.13/3		2.72/3		Contact pages
8		143		67		2.70/4		3.43/4		Adequate categories of products (>=2) / products in each category (>=2)
9		144		66		2.67/4		3.39/4		Products appropriately displayed and described
10		141		69		2.00/3		2.54/3		FAQ/Help pages
11		156		54		1.43/2		1.82/2		Disclaimer page appears on each page
14		122		88		2.26/4		2.87/4		Appropriate form controls used
15		122		88		1.70/3		2.16/3		Form Fields positioned appropriately
17		114		96		2.17/4		2.76/4		Appropriate form controls used
18		114		96		1.09/2		1.38/2		Form Fields positioned appropriately
19		164		46		3.90/5		4.96/5		Web pages resize appropriately in browser
21		152		58		1.42/2		1.81/2		Navigation clear
22		151		59		1.42/2		1.81/2		Navigation consistent
23		150		60		1.42/2		1.81/2		Navigation complete
25		165		45		7.84/10		9.98/10		Correct selection and use of HTML Tags
26		77		133		1.83/5		2.33/5		validated HTML Tags
27		165		45		2.36/3		3.00/3		HTML well structured
28		161		49		1.53/2		1.95/2		HTML commented
30		157		53		3.69/5		4.69/5		Style sheets / styles included
31		159		51		1.51/2		1.93/2		Appropriate styles used
33		163		47		1.55/2		1.98/2		Meta Tags Included
34		161		49		1.53/2		1.95/2		Appropriate Meta Tags used
36		156		54		2.22/3		2.83/3		Images uploaded and correct preferences set
37		158		52		2.26/3		2.87/3		Images formatted/ resized
38		156		54		1.49/2		1.89/2		Relevant images used
40		163		47		1.55/2		1.98/2		Suitable colors used
41		162		48		1.54/2		1.96/2		Consistent use of color

43	162	48	1.54/2	1.96/2	Font size and color suitable
44	163	47	1.55/2	1.98/2	Consistent use of Fonts

B. On-line Grades System

On-line Grades System (OGS) [7] is a Web based application developed to provide an easy way to access hyperlink-based assignments, assessing them and delivering the results within a browser.

It is designed to automate hyperlink access, to mark assignments via a GUI and to send feedback to students purely online. It can:

1) provide automation that retrieves student submitted hyperlinks and presents them in a browser.

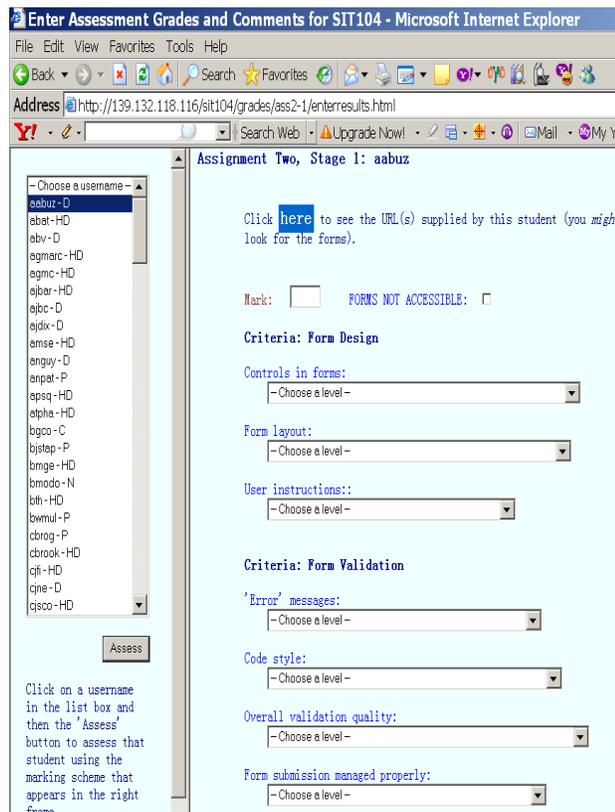


Fig. 2. Example OGS marking guide

Individual students are identified by an associated email address. Students can be selected from an email address selection menu.

2) provide an easy to use marking GUI that only requires the identification of the criterion and assessment value for an item after loading a HTML based marking guide.

Detailed marking guide items are displayed in the form of dropdown lists, as shown in Fig. 2.

3) automate the calculation of results within the application

Each time a dropdown list item is chosen, the mark is adjusted accordingly. Comments can be added, edited or deleted at the bottom of the marking guide page.

4) provide a separate interface to deliver student results online

When the marking is completed, all results can be viewed via a “lookup result” interface by typing the student ID and email address. The result report is a file stored in the underlying database containing the assessment criteria and the associated mark that the student receives for each item in the criteria. All the comments relating to the assessment are also included, as shown in Fig. 3.

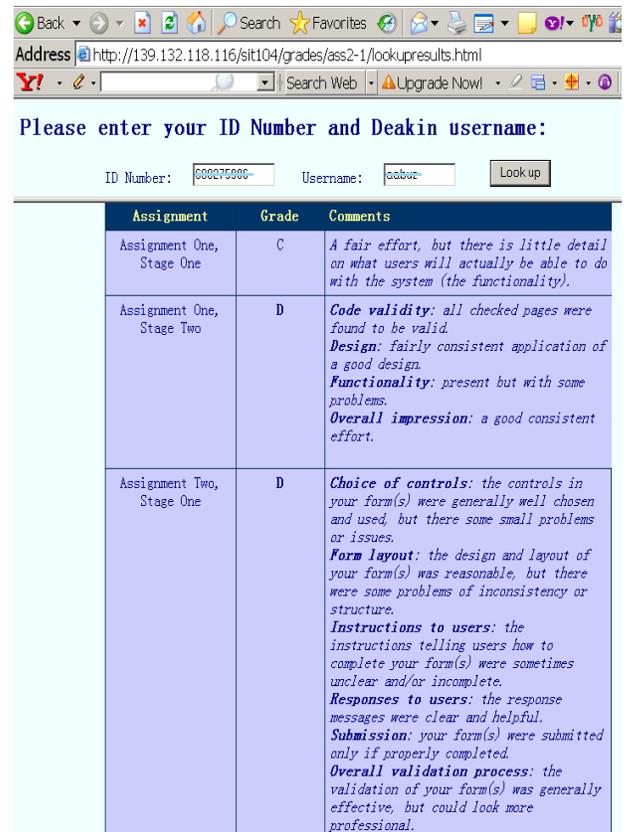


Fig. 3. Example OGS assessment report

5) enable all data collected to be saved and retrieved via a database

Backup has to be done manually to ensure the marking and comments are saved in the database.

IV. COMPARISON AND CONCLUSION

To compare these two tools, data was collected relating to the first and second assignments in the unit SIT104 Introduction to Web Development during trimester 2, 2012. The first two stages of the second assignment were included in the comparison. The unit SIT104 is delivered to students on three campuses, on-campus at Geelong (in regional Victoria), on-campus at Burwood (suburban campus in Melbourne) and off-campus (ie distance education).

A. Comparison

For assignment 1, both the MA and OGS were used to mark submissions made by the Geelong on- and off-campus students. The MA is used to provide just detailed comments to students but not the mark; the actual marking that was distributed to students was completed using the OGS.

For assignment 2 stage 1, the MA and OGS were both used for marking all submissions. To maintain consistency, only the OGS results were released to students, together with the comments created in MA.

For assignment 2 stage 2, the OGS is used for marking all student submissions. Comments were made as informative as possible on each submission in OGS.

Student feedback was collected via email query, in-class and after-class discussion.

By comparing these three different marking methods and feedback from students, the following similarities and differences are found:

- Similarities:
 - a) Both of the assessment tools support an easy-to-use interface with file loading automation;
 - b) Both of the assessment tools automate the calculation of results within the application;
 - c) Both of the assessment tools support data backup.
- Differences:
 - a) MA results are sent out via email; OGS results can be viewed online;
 - b) MA backups results in .txt file; OGS stores all the data in a database;
 - c) MA receives positive feedback on detailed marking criteria and personalised comments on each item;
 - d) MA receives positive feedback on providing results summaries and statistics information;
 - e) OGS's marking guide is considered somewhat "vague" and needs to be refined;
 - f) Well commented OGS marking results receive near "zero" marking complaints;
 - g) Although MA's comments can be reused, there is a low limit on the size of the comment editing field, whereas OGS provides enough text area to accommodate long comments;
 - h) OGS does not require additional training once the assessment environment is setup; MA requires additional training because of the availability of more powerful functionality.

B. Conclusion

The marking guide identifies the expectations of the assessment and the marks provide a measure of the student's success in meeting the expectations. In addition to the marking guide, it would be better to provide information for each marking item which explains what the expectation of each item is.

It was found that when students are given a correct answer they would not only identify where they went wrong for themselves but also discover what is required or is correct. From this point of view, no matter what kind of assessment tools we use, providing detailed and informative comments also help students improve their learning outcomes.

According to a recommendation of the University's Teaching and Learning Committee [8], feedback on assignments to students should:

- Be clearly linked to each published assessment criterion
- Assist learning, reward achievement, provide encouragement, explain results and enable students to improve their performance

The evidence provided above also suggests that a well written marking guide and informative comments does provide meaningful feedback to students.

V. FUTURE WORK

The above two assessment tools remove much of the complication from assessment by automating many of the tedious tasks, which allows assessors to focus on the assessment itself rather than the associated administrative processes.

Also the marking guides created using the tools are reusable and editable for future assessments.

Other features that will further improve the supporting environments and functionality of the two tools include:

For MA:

- Extend the system to allow annotated attachments to the results
- Provide database connectivity

For OGS

- Provide facilities to create and edit a marking guide within the application
- Add statistical output.

By combining the positive features of the two tools, a comprehensive, easy to use marking tool can be developed that not only minimises the tedious tasks of marking from an academic's perspective, but also provides students with comprehensive, detailed feedback that will encourage them to learn from their mistakes.

REFERENCES

- [1] J. Wells, W. Zhou, N. Paul, C. Brian, F. Joseph (2003), "Assessment management using software, Advances in web-based learning", International conference on web-based learning (ICWI 2003), 18-20 August 2003, Melbourne, AUSTRALIE 20031973, vol. 2783, pp. 411-422, ISBN 3-540-40772-3.
- [2] CompassLearning Odyssey, Compass Learning Website, available < <http://www.compasslearning.com/assessment/> >, accessed 24/05/2013.
- [3] Smartlite (2005), Smartlite software Website, available < <http://eng.smartlite.it/en2/products/webquiz/index.asp> >, accessed 24/05/2013.

- [4] webCT Vista, Blackboard Website, available < <http://www.blackboard.com/About-Bb/Overview.aspx>> , accessed 25/05/2013.
- [5] Desire2Learn, Desire2Learn website, available < <http://www.desire2learn.com/>>, accessed 31/05/2013.
- [6] Markers Assistant, Surreal Website, available <<http://www.surreal.com.au>>, accessed 25/05/2013.
- [7] On-line Grades System (OGS), Deakin University Website, available < <http://139.132.118.116/sit104/grades/ass2-1/enterresults.html>>, accessed 25/01/2013.
- [8] Deakin University, Appropriate levels and timing of feedback to students on assignments – operational plan target 1.3.6, Teaching and Learning Committee, Agenda paper 9.1, AB05/07/124, unpublished.

Data fusion based framework for the recognition of Isolated Handwritten Kannada Numerals

Mamatha.H.R

Department of Information Science
and Engineering
P.E.S. Institute of Technology
(West Campus)
Bangalore, India

Sucharitha Srirangaprasad.

Department of Computer Science
Indiana University, Bloomington,
Indiana, USA

Srikantamurthy K

Department of Computer Science
and Engineering
P.E.S. Institute of Technology
(South Campus)
Bangalore, India

Abstract—combining classifiers appears as a natural step forward when a critical mass of knowledge of single classifier models has been accumulated. Although there are many unanswered questions about matching classifiers to real-life problems, combining classifiers is rapidly growing and enjoying a lot of attention from pattern recognition and machine learning communities. For any pattern classification task, an increase in data size, number of classes, dimension of the feature space and interclass separability affect the performance of any classifier. It is essential to know the effect of the training dataset size on the recognition performance of a feature extraction method and classifier. In this paper, an attempt is made to measure the performance of the classifier by testing the classifier with two different datasets of different sizes. In practical classification applications, if the number of classes and multiple feature sets for pattern samples are given, a desirable recognition performance can be achieved by data fusion. In this paper, we have proposed a framework based on the combined concepts of decision fusion and feature fusion for the isolated handwritten Kannada numerals classification. The proposed method improves the classification result. From the experimental results it is seen that there is an increase of 13.95% in the recognition accuracy.

Keywords—feature selection; feature fusion; decision fusion; Curvelet transform; K-NN classifier; data fusion; isolated handwritten Kannada numerals; OCR;

I. INTRODUCTION

Achieving the best possible classification performance for a given problem domain has become the ultimate goal of designing the pattern recognition systems. This objective traditionally led to the development of different classification schemes for any pattern recognition problem to be solved. In the recent past, studies have been done to obtain the optimal feature set and classifier set.

Features play a very important role for any pattern classification task. A set of bad features can deteriorate the performance of a good classifier. With the increase in noise and dimensionality, feature selection becomes an essential step.

A feature that is having too much of confusing (contradictory) information than the rest of the set should be avoided as these features confuse the classifiers. To reduce the noise in the data, features which are weakly correlated to the class information should be removed. “Curse of dimensionality” is also a big motivation for feature selection.

Too many features increase the computational time without any significant change in the performance during the training phase [1].

In practical classification applications, if the number of classes and multiple feature sets for pattern samples are given, a desirable recognition performance can be achieved based on these sets of features using data fusion [2]. Fusion strategies are mainly classified into information fusion (low-level fusion/pixel level fusion), feature fusion (intermediate-level fusion), and decision fusion (high-level fusion) [3].

Information fusion combines several sources of raw data to produce new raw data that is expected to be more informative and synthetic than the inputs. Feature fusion deals with the selection and combination of features to remove redundant and irrelevant features. If two features have similar or nearly similar distribution, one of them is redundant. A feature is said to be irrelevant if it correlates poorly with the class information. The final set of features is fused together to obtain a better feature set, which is given to a classifier to obtain the final result. Feature fusion is an advancement of information fusion. Decision fusion uses a set of classifiers to provide a better and unbiased result. The classifiers can be of same or different type and can also have same or different feature sets.

There are different classifiers such as KNN, SVM, ANN etc., and a single classifier may not be well suited for a particular application. Hence a set of classifiers are merged together by various methods to obtain the final output [1]. It has been found that a consensus decision of several classifiers can give better accuracy than any single classifier [4]. Therefore, combining classifiers has become a popular research area during recent years. The goal of combining classifiers is to form a consensus decision based on opinions provided by different base classifiers. Combined classifiers have been applied to several classification tasks, for example to the recognition of faces, handwritten characters identification, and fingerprint verification [5, 6].

In this paper we have proposed a framework which combines both the concepts of feature fusion and decision fusion. First, a feature selection method is presented to find the best feature set from a set of features. Next the best feature set is applied on two training data sets with different sizes and of different complexity to know the effect of the training dataset size on the recognition performance of a feature extraction

method and the classifier used. After finding the best feature set, this feature set is combined with the other feature sets to form fused feature sets (union vector). These fused feature sets are classified using K-NN classifier and the fused feature set with highest recognition accuracy is chosen for the level. Lastly, the decision fusion method is used for a better classification results. Here, we have applied the proposed framework for the recognition of isolated handwritten Kannada character recognition. Results are presented using our own built handwritten Kannada numeral datasets.

Over the last few years, extensive research is being carried out on Handwritten Character Recognition (HCR) systems in the academic and production fields. A Handwritten Character Recognition system can either be online or offline. The process of finding letters and words present in a digital image of handwritten text is called off-line handwritten recognition. A number of methods of recognition of English, Latin, Arabic, Chinese scripts are excellently reviewed in [7, 8, 9, 10]. A HCR system has various applications such as being used as a reading aid for the blind, applications involving bank cheques, automatic pin code reading for sorting of postal mail.

A lot of work has been done on the recognition of printed characters of Indian languages. On the other hand, attempts made on the recognition of handwritten characters are few. Most of the research in this area is concentrated on recognition of off-line handwritten characters for Devanagari and Bangla scripts. From the literature survey it is seen that there is a lot of demand for character recognition systems for Indian scripts and an excellent review has been done on the OCR for Indian languages [11]. A Detailed Study and Analysis of OCR Research on South Indian Scripts can be seen in [12].

A method for the recognition of isolated Devanagari handwritten numerals based on Fourier descriptors has been proposed by Rajput and Mali in [13]. Another method proposed in [14] involves computing the zone centroid and further dividing the image into equal zones. The average distance from the zone centroid to each pixel present in the zone is computed. The aforementioned process is repeated for all the zones present in the image of the numeral. At last, n such features are extracted and considered for classification and recognition. F-ratio Based Weighted Feature Extraction for Similar Shape Character Recognition for different scripts like Arabic/Persian, Devnagari English, Bangla, Oriya, Tamil, Kannada, Telugu etc can be seen in [15].

The key factor in achieving high recognition rate in character/numeral recognition systems is the selection of a suitable feature extraction method. A survey on the feature extraction methods for character recognition is reviewed in [16].

Curvelet transform is used as one of the feature extraction methods in [17, 18, and 19]. Here the curvelet transform function is applied on the given image and the coefficients are obtained. The obtained coefficients are used in the feature vector for that particular image.

Literature survey shows that the automatic recognition of handwritten digits has been the subject of intensive research during the last few decades. Digit identification is very

important in applications such as interpretation of ID numbers, Vehicle registration numbers, Pin Codes, etc. In Indian context, it is evident that still handwritten numeral recognition research is a fascinating area of research to design a robust optical character recognition (OCR), in particular for handwritten Kannada numeral recognition.

The paper is organized as follows: in section II, we discuss the properties of the Kannada numerals and their complexity. Section III deals with generation of handwritten Kannada numeral datasets. The need for feature selection is presented in section IV. Section V details the proposed methodology used for the recognition. The experimental results and discussions are shown in section VI followed by the paper's conclusion in section VII.

II. KANNADA NUMERALS AND THEIR COMPLEXITY

Kannada or Canarese, the official language of the southern Indian state of Karnataka is described as 'sirigannada'. Kannada has now received the Classical Language status in India. It has a history of more than 1500 years and is also spoken in the neighboring states of Tamilnadu, Andhra Pradesh and Maharastra. The expatriate population of Kannada origin is also present in USA, Australia, Asia Pacific and Africa. The Kannada speaking population is of no more than 70 million. The script includes 10 different Kannada numerals of the decimal number system as in Table I.

Kannada characters have more complex structure and curved in shape. There are a large number of similar character groups (Table II).

The challenging part of Kannada handwritten character recognition is the distinction between the similar shaped components. Sometimes a very small part is the distinguishing mark between two characters or numerals. These small distinguishing parts increase the recognition complexity and decrease the recognition accuracy. The style of writing characters is highly different and they come in various sizes and shapes (Fig. 1). Same character may take different shapes and conversely two or more different character of a script may take similar shape. Kannada lacks a standard test bed of character images for OCR performance evaluation.

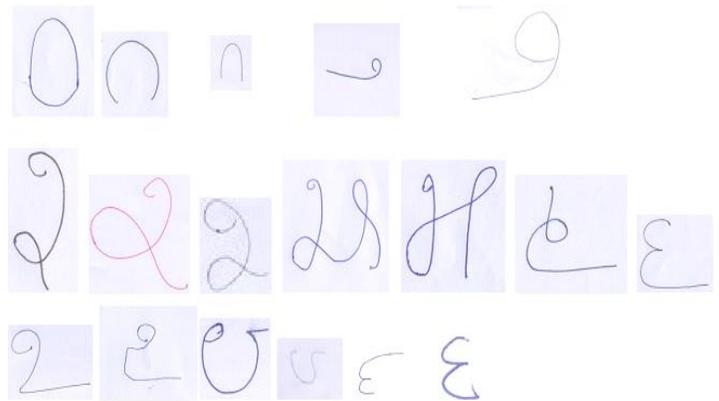


Fig. 1. Samples showing the style of writing characters with different size and shapes

TABLE I. KANNADA NUMERALS

Numeral	Printed sample of the numeral	Handwritten sample of the numeral
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		

TABLE II. EXAMPLES OF SOME SIMILAR SHAPED NUMERALS

Numeral	Handwritten sample	Printed sample	Handwritten sample	Printed sample
0 and 1				
6 and 9				
3 and 7				

III. GENERATION OF HANDWRITTEN KANNADA NUMERAL DATASETS

A major obstacle to research on handwriting character recognition of Indian scripts is the non existence of standard/benchmark databases. From the literature review, it can be seen that most of the experimentation is reported on the

basis of small databases collected in laboratory environments. Several standard databases such as NIST, MNIST, CEDAR and CENPARMI are available for Latin numerals [20]. But we can say to the best of our knowledge that, only two such standard databases namely databases for Bangla and Devanagiri scripts are available in the Indian Context until now. Hence we have made an attempt to create a database of our own for the experimentation in reference with [20].

The handwritten Kannada numeral database consists of two datasets, original and synthetic. Each dataset is randomly divided into respective training and test sets in the ratio of 8:2.

A. Generation of original dataset

Samples for the database have been collected using plain A4 paper and a tabular form designed for data collection purpose so that both constrained and unconstrained samples could be collected as shown in Fig.2. The only restriction imposed on the writers was to write one numeral in one box in case of the tabular form. These samples have been collected from a wide spectrum of population of various age groups which includes students from school and college, housewives and employees. Some of the samples were collected from the people with no knowledge of the Kannada language. There was no restriction on the type of pen and color of the ink used.

The collected documents were scanned at 300 dpi using a HP flatbed scanner and stored in jpg format. The individual numerals were extracted manually from the scanned documents and labeled. The images were not size normalized. Thus 100 different samples of each numeral were created with the total of 1000 samples. This dataset is considered as the dataset 1 for our experimentation.

B. Generation of Synthetic dataset.

Hand-printed patterns come from different writers and possess great variations. Recognition of hand-printed patterns is difficult when compared to machine-printed patterns. Some factors that complicate the recognition process in hand-printed character recognition in noise-less situations are discussed in [21]. Various strategies are followed in a recognition system in order to reduce the variability caused due to slant writing. Some of them are as follows: 1) the slanted word/character is normalized before recognition 2) the slant is compensated during training process by having a dataset covering as many as slant angles as possible 3) slant invariant feature extraction method is used.

In order to increase the dataset size with as many as slant angles, we generated synthetic data. Synthetic data was generated by subjecting the original data to the two transformations namely blurring and rotation, thereby increasing their number by a factor of 10[20].

In the first step all the samples in the original dataset were blurred by applying a Gaussian blurring kernel. Thus, the volume of the database was increased by 2(with and without blurring).

In the next step, both the blurred and original images were rotated by an angle of -5° , -10° , $+5^\circ$ and $+10^\circ$. Thus, the volume of the dataset was increased five times. Thus the total increase in volume was ten times the original number i.e., for each

original sample 9 synthetic images were generated (as shown in the Fig.3) and hence a total of ten images were obtained for each sample taking the total number of samples to 10,000. This dataset is considered as the dataset 2 for our experimentation.

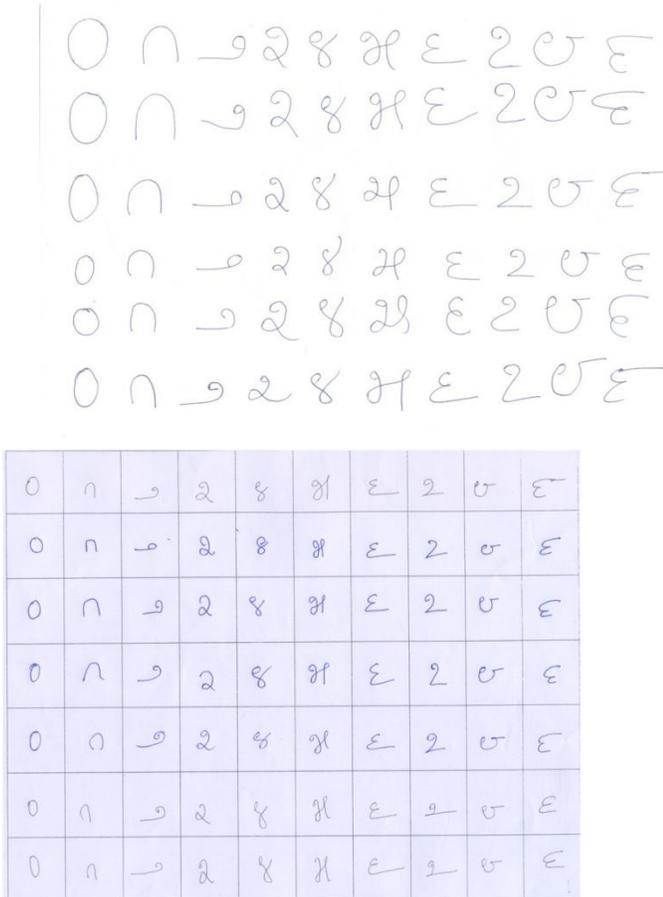


Fig. 2. Handwritten Kannada numeral samples (spaced discrete unconstrained and constrained handwriting)

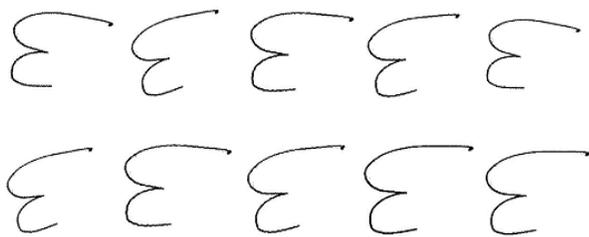


Fig. 3. Samples of synthetic data generated for numeral 9

IV. NEED OF FEATURE SELECTION

Feature selection is well-known problem and has been much explored specially in the areas of data mining and content based image retrieval. The problem deals with finding an appropriate subset of features from a given set, for some particular application domain, to improve the accuracy. This involves finding a minimal subset that represents the whole set, or to rank the features based on their importance, from the overall set.

Feature selection methods are mainly classified into filter method, wrapper method and hybrid method [22]. In the filter approach, the feature set is evaluated at once which is independent of any clustering algorithm or classifier. On the other hand the wrapper method calls the clustering algorithm or the classifier for each subset evaluation to find the final subset. While the filter method is unbiased and fast, the wrapper method gives better results for a particular clustering algorithm or classifier. Hybrid method is a fusion of both filter and wrapper methods. In this paper, we have used the wrapper based feature selection method.

V. PROPOSED METHODOLOGY

In this section, selection of feature set and recognition of isolated handwritten numerals using a framework based on the combined concepts of feature fusion and decision fusion is proposed as shown in the Fig.4. The proposed method consists of three stages. In the first stage, a framework for selection of a better feature set is proposed and in the second stage the fused feature vector is selected and in the last stage improvement in accuracy is shown using decision fusion approach applied on the fused feature vector.

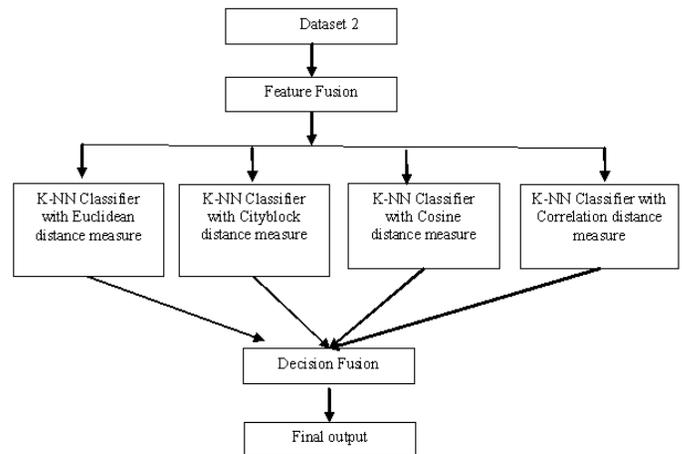


Fig. 4. A framework for the improvement of the recognition accuracy based on the combined concepts of feature fusion and decision fusion.

A. Framework for feature selection

The framework designed for the feature selection has various steps described as follows:

Dataset1 → Preprocessing → Feature extraction → Dimensionality reduction using standard deviation → Classification → Subset evaluation and selection of final feature set → final feature set

1) Preprocessing

Initially the color images were converted to gray scale and in turn the gray scale images were converted to binary using the global threshold method. Thinning was applied on the binary image. Thinning is an image preprocessing operation performed to make the image crisper by reducing the binary-valued image regions to lines that approximate the skeletons of the region. Region labeling was then performed on the thinned binary image of the numeral and a minimum rectangle bounding box was inserted over the numeral. The bounding

box image would be of variable size due to different style and size of numeral. Hence this image was resized to a 256*256 and thinning was applied again. These preprocessed samples are used in the next stage i.e., feature extraction.

2) Feature extraction

The feature extraction technique that we have used to extract the features of a numeral is the Curvelet transform. We have used the Curvelet Transform because it extracts features efficiently from images which contain a large number of C^2 curves (i.e. an image which has a large number of long edges) [23].

We have applied wrapping based discrete curvelet transform using Curvelab-2.1.2, a toolbox implementing the Fast Discrete Curvelet Transform, to find the coefficients of every 256*256 image in the database. These coefficients are used as the feature vectors for those images. In this experiment we have used the default orientation and 5 levels of discrete curvelet decomposition. Hence for an image of size 256*256, curvelet coefficients in five different scales were obtained. Thus we have five different feature sets.

3) Dimensionality reduction using standard deviation

The curvelet coefficients obtained for each sample are numeric. In this implementation, we have chosen wavelet in the finest level of curvelet transform. This is due to the fact that use of wavelet reduces the redundancy factor [24]. One subband at the coarsest and one subband at the finest level of curvelet decomposition are obtained after the application of curvelet transform in Scale 1 on the input. The numbers of subbands obtained at each level for the other scales of curvelet decomposition is different. The number of coefficients obtained after application of curvelet transform is very high. Hence if all the coefficients obtained are used in the feature vector, the size of the feature vector and the time taken for feature vector formation increases drastically. Therefore, for extracting the best features and also decreasing the size of feature vector for each sample, we use standard deviation as the dimension reduction technique [23].

The standard deviation of the coarsest and the finest levels are calculated first using the equation (1). Then, we calculate the standard deviation of the first half of the total subbands at each of the remaining scales.

We consider only the first half of the total subbands at a resolution level for feature calculation because; the curvelet at angle θ produces the same coefficients as the curvelet at angle $(\theta+\pi)$ in the frequency domain i.e. these subbands are symmetric in nature. Hence, considering half of the total number of subbands at each scale reduces the total computation time for the feature vector formation without the loss of information of the image. For the finest and the coarsest subbands the standard deviation calculated is used directly in the feature vector but for the other subbands the sum of the standard deviation is calculated and stored in the feature vector. It is seen that by applying standard deviation we can reduce the features as shown in the Table III.

TABLE III. THE REDUCTION IN THE NUMBER OF FEATURES

Scale	Obtained number of features after applying curvelet transform	Number of features obtained after applying standard deviation
1	29,241	171
2	94,777	426
3	1,64,953	348
4	1,80,601	322
5	1,84,985	316

$$s = \left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{1/2} \quad (1)$$

Where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, and n is the number of elements in the sample.

4) Classification

The classifier used in the proposed method is the k nearest neighbor classifier [25]. The Nearest Neighbor Classifier is an efficient technique to use when the classification problem has pattern classes that display a reasonably limited degree of variability. It considers each input pattern given to it and classifies it to a certain class by calculating the distance between the input pattern and the training patterns. It takes into account only k nearest prototypes to the input pattern during classification. Here, Cityblock measure is used as the distance and nearest used as the rule. The decision is generally based on the majority of class values obtained by classifying k nearest neighbors

5) Subset evaluation and selection of final feature set

A subset of features is evaluated based on its recognition accuracy i.e., its capability for class separability. The subset which gives highest recognition accuracy for the given dataset is selected as best feature set for the classification. The scale 1 feature (subset 1) has the highest recognition accuracy and is selected as the final feature set.

B. Combining classifiers (decision fusion)

A traditional approach in classification is the use of a single classifier, which assigns a class label for each feature vector describing the image content. The decision functions produced by different classification principles differ from each other. This makes the classification accuracy somewhat varied the feature patterns obtained from the real world images are non-homogenous, noisy and overlapping which may cause variations in the decision boundaries of different classifiers, due to these reasons different classifiers may classify the same image in different ways. As the features or classifiers of different types are able to complement one another in classification performance, the consensus decision of several classifiers can yield improved performance compared to individual classifiers [26].

1) Need for classifier combination

To improve the accuracy and efficiency of the classification system, a multi classifier system is preferred over a single classifier due to some of the following reasons like [1, 26]

- In certain applications, the volume of data to be analyzed is too large to be handled by a single classifier. Training a classifier with such a vast amount of data is usually not practical. A multi classifier system will be an efficient approach, where data is partitioned into smaller subsets, trained with different classifiers for different subsets and the outputs are combined.
- A single classifier cannot perform well when the nature of features is different. Using multiple classifiers with a subset of features may provide a better performance.
- Another reason for combining classifiers is to improve the generalization performance: a classifier may not perform well for a certain input when it is trained with a limited dataset. Finding a single classifier to work well for all test data is difficult. Instead multiple classifiers can be combined to give a better output than a single classifier. It may not necessarily out-perform a single best classifier, but the accuracy will be on an average better than all the classifiers.

2) Categories of multiple classifier systems

In a multiple classifier system, it is common that there are several base classifiers that are combined using a particular classifier combination strategy. It is obvious that a combination of base classifiers with identical errors does not improve the classification and hence, the base classifiers with decorrelating errors are preferred. Consequently, the base classifiers should differ from each other in some manner. This type of classifier combination can be achieved in one of the following ways: [1, 26 and 27]

- Variation of initial parameters of the classifiers: a set of classifiers can be created by varying the initial parameters, using which each classifier is trained with the same training data. For example, in K-NN classification, the value of k needs to be selected. By using different parameter values, it is possible to obtain differently behaving classifier.
- Variations of the training dataset of the classifiers: multi classifier systems can be built by training the same classifier with different training datasets. The type of training in the two level scheme can be either training the individual classifier and applying fusion or by training the individual classifier followed by training the fusion
- Variations in the number of individual classifiers used: training different types of classifiers like SVM, ANN, etc., with the same training dataset.
- Variations in the architectures: In several kinds of classifiers, the architecture can be selected. For example, the size of neural networks in the base classifiers can be varied.
- Variations in the feature sets: the base classifiers may use separate feature sets as their inputs. These feature sets may describe different properties of the object to be classified.

Once the base classifiers have been constructed, it is necessary to combine their opinions using some combination strategy. Classifier combination strategies are mainly classified into classifier fusion and classifier selection. In classifier fusion, every classifier is provided with complete information on the feature space, and the outputs from different classifiers are combined. Every classifier contributes to make a final decision whereas in classifier selection methods, every classifier is an expert in a specific domain of the feature space and the local expert alone decides the output of the ensemble. Classifier fusion is further categorized based on the output of classifiers and classifier selection is classified into dynamic and static classifier selection.

Some of the different classifier combination strategies are [26, 27 and 28]:

Strategies based on probabilities: These methods are also known as fixed combining rules. These strategies utilize the fact that the base classifier outputs are not just class numbers, but that they also include the confidence of the classifier.

Voting based strategies: The basic idea behind these methods is to make a consensus decision based on the base classifier opinions using voting. Hence, the class labels provided by the base classifiers are regarded as votes, and the final class is decided to be the class that receives the majority or most of the votes. The benefit of these methods is that the decision can be made solely on the basis of the class labels provided by the base classifier.

Strategies employing the class labels: In addition to the voting and the probability-based classifier combination methods, various classifier combination methods have been proposed that utilize the base classifier outputs in other ways than voting. In the most common case, these outputs are the class labels given by the base classifiers, though in certain cases methods such as probability distributions are employed. Some examples of these methods are class ranking, stacked generalization, error-correcting output codes.

C. Feature fusion

Once the best feature subset is obtained from the original set, we can use the derived set or can derive a new feature based on two or more of the selected features for the task of classification. Based on this concept, there are two existing techniques of feature combination: serial and parallel combination [2].

Feature combination (feature fusion) is the general technique where two features α and β are concatenated together [2]. If m and n are the weights of α and β , respectively, then according to the serial fusion, the combined feature is $[m \times \alpha ; n \times \beta]$. In parallel combination, a complex variable is used to combine the two features into a complex feature. The absolute value of the complex feature is taken as the final feature. Hence, if m and n are the weights of α and β , respectively, then the combined feature is set as $\|(m\alpha) + i(n\beta)\|$. In these cases, the weights m and n can be decimal or binary values. In the latter case, 0 as a weight for a particular feature denotes that the corresponding feature is discarded, while 1 denotes that the corresponding feature is selected for the final subset of features [1].

To improve the accuracy we have proposed a method based on feature fusion. Here we have used a serial based feature combination where the weights of the features are taken as 1. Features from the selected feature subset is serially combined with the features of the other extracted feature subsets to form a union vector. For example, selected feature set 1 is combined with feature subset 2 to form union vector (1, 2). Similarly, we have obtained four such union vectors (1, 2), (1, 3), (1, 4) and (1, 5).

VI. RESULTS AND DISCUSSIONS

The experiments were carried out in Matlab 7.5.0, on a 64-bit 2.67 GHz INTEL i5 processor, with 4 GB RAM. The curvelet transformation was done using the Curvelet 2.1.2 toolbox, available from <http://www.curvelet.org>. The morphological operations were performed using Matlab's Image Processing Toolbox.

The curvelet transform is used to extract the features from the numeral samples in the dataset1. All the five different subsets of features extracted (scale 1, scale 2, ..., scale 5) are applied on this dataset and classified. The recognition accuracy is calculated for each of the scales or the feature subset (as shown in Fig.5). The feature subset with the highest recognition accuracy is considered as the final or selected feature set for the next level of our methodology.

In our case scale 1 or (subset 1) has a highest recognition accuracy of 91% and is selected as the final feature set. The selected feature set (scale 1 features) from the feature selection framework is experimented on two datasets with a size of 1000 and 10,000 samples respectively using K-nearest neighbor with Cityblock as the distance measure. We obtained a 91% of recognition accuracy for the dataset 1 and 65.65% for dataset 2 proving that for any pattern classification task, an increase in data size affect the performance of any classifier [1].

To improve the accuracy we have proposed a method based on the concepts of decision fusion and feature fusion. Here we have built a classifier combination based on the variation of initial parameters of the classifier. K-NN classifier is used to build the multi classifier system. A set of differently behaving classifiers can be created by varying the initial parameters like K-value, the distance measures in the K-NN classifier.

For our experimentation we have varied the distance measure parameter of the K-NN classifier and built the multi classifier system. We have applied four distance measures- Euclidean, City block, Cosine and Correlation to build the base classifiers. A feature vector selected as the best features from the feature selection process that is scale 1 features (subset 1) was given to K-NN classifiers with different distance measures. The results of all these classifiers are combined and a vote was taken to see the class to which the sample was classified the maximum number of times and this was considered as the class to which the sample belonged to (plurality voting).

Multi classifier system built was experimented on the dataset2 and from the results we find that there was an increase in the accuracy for the dataset2.

An increase in the recognition accuracy is seen from the experimental results as shown in Table IV.

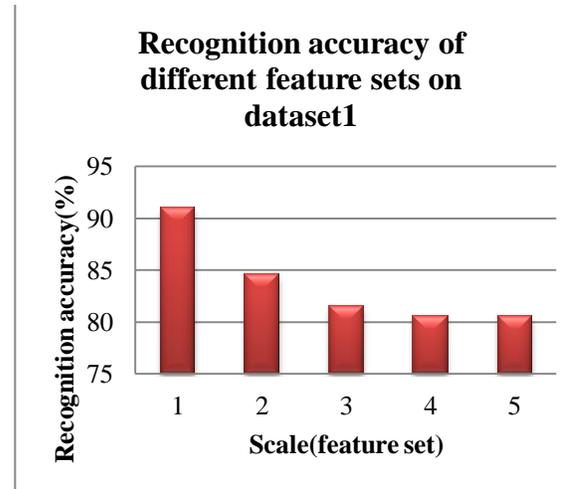


Fig. 5. Recognition accuracy of different feature sets on dataset1

Next, a method based on feature fusion was used. Features from the selected feature set are combined with the features of one of the subsets to form a union vector. In our case we have obtained four such union vectors (1, 2), (1, 3), (1, 4) and (1, 5). Using these union vectors, the experiment was repeated on the dataset 2 (synthetic dataset). Recognition accuracy for each of these union vectors is found (as shown in Fig 6).

Another important observation is that when the combination of scales are used for classification, the recognition rates appear better for the dataset2 and is the best (78.45%) with 0% rejection rate when the scales 1 and 5 are used together. One of the reasons of this result can be the using of image's information in different size's partitions and various scales. An increase in the recognition accuracy is seen from the experimental results as shown in Table IV.

Finally, we combined both the decision and feature fusion concepts and came up with a new framework. First we obtained the fused feature set which gave the highest recognition rate and this fused feature set was given to K-NN classifiers with different distance measures.

The results of all these classifiers are combined and a vote was taken to see the class to which the sample was classified the maximum number of times and this was considered as the class to which the sample belonged to (plurality voting). From the results it is seen that there is an increase in the recognition accuracy as shown in the Fig.7.

From the experimental results (Table V), we observe that the average time required for the recognition is very less and that is in seconds which is not going to affect the efficiency of the proposed method. This can be attributed to the fact that the entire co-efficient set obtained is reduced using standard deviation and this result in dimensionality reduction of the feature vector and hence reducing the time taken for recognition.

TABLE IV. IMPROVEMENT IN RECOGNITION ACCURACY (%) USING OUR PROPOSED FRAMEWORK

Dataset	Recognition Accuracy (%)				Improvement in accuracy (%)
	Before fusion	After decision fusion only	After feature fusion only	After feature and decision fusion	
Synthetic dataset(data set2)	65.65	70.2	78.45	79.6	13.95

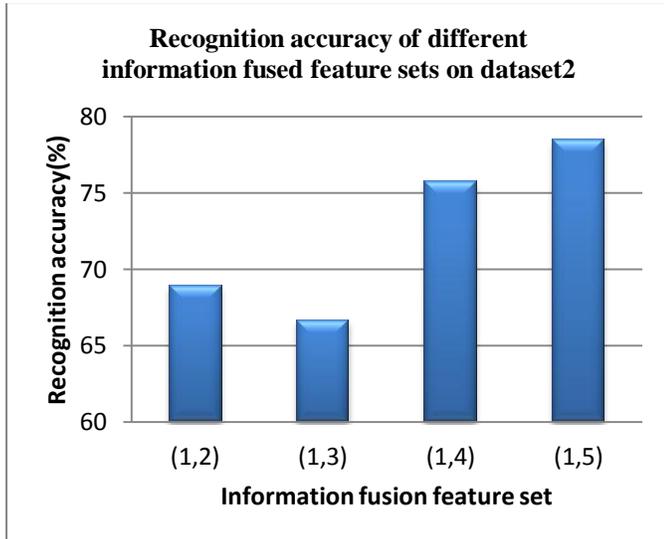


Fig. 6. Recognition accuracy of different feature fused feature sets on dataset2

TABLE V. AVERAGE RECOGNITION TIME

Dataset	Average recognition time in seconds			
	Before fusion	After decision fusion	After feature fusion	After decision and feature fusion
Synthetic dataset(data set2)	0.516	11.063	1.855	52.440

VII. CONCLUSION

In practical classification applications, if the number of classes and multiple feature sets for pattern samples are given, a desirable recognition performance can be achieved based on these sets of features using data fusion. Data fusion is an ever growing field with a wide scope of interdisciplinary research over the fields of computer science, mathematics, statistics and machine learning. In this paper, we have proposed a framework based on the combined concepts of decision fusion and feature fusion for the isolated handwritten Kannada numerals classification. The proposed method improves the classification result. From the experimental results it is seen that there is an increase of 13.95% in the recognition accuracy.

Recognition accuracy of Fusion methods

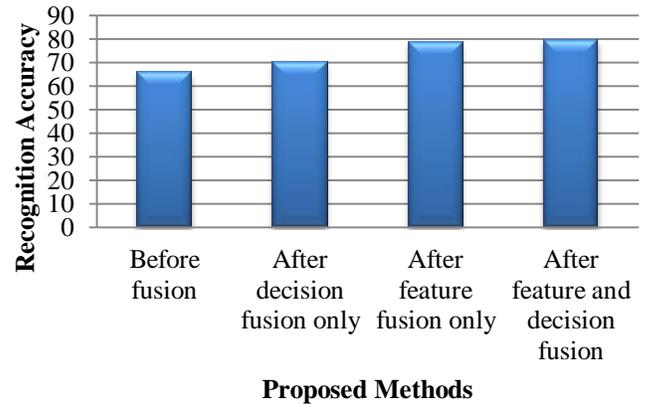


Fig. 7. Recognition accuracy of fusion methods

REFERENCES

- [1] Uttara Gosa Mangai, Suranjana Samanta, Sukhendu Das and Pinaki Roy Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification", IETE Technical review, vol 27, issue 24, 2010.
- [2] Jian Yang, Jing-yu Yang, David Zhang and Jian-feng Lu, "Feature fusion: parallel strategy vs. serial strategy", Pattern Recognition, 36 (2003), pp. 1369-1381, 2003.
- [3] B.V. Dasarthy, Decision Fusion, Computer Society Press, 1994.
- [4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers", IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(3) (1998), pp. 226-239, 1998.
- [5] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods for Combining Multiple Classifiers and their Applications to Handwriting Recognition", IEEE Trans. on Systems, Man and Cybernetics, 2(3) (1992), pp. 418-435, 1992.
- [6] A. K. Jain, S. Prabhakar, and S. Chen, "Combining Multiple Matchers for a High Security Fingerprint Verification System", Pattern Recognition Letters, 20(11-13) (1999), pp. 1371-1379, 1999.
- [7] P. Nagabhushan, S.A. Angadi and B.S. Anami, "A Fuzzy Statistical Approach to Kannada Vowel Recognition based on Invariant Moments", proceedings of NCDAR -2003, PESCE, Mandya, pp 275-285, 2003.
- [8] B.B. Chaudhuri and U. Pal, "A Complete Printed Bangla OCR System", Pattern Recognition, 31, pp 531-549, 1998..
- [9] U. Pal and B.B. Chaudhuri, "Indian Script Character Recognition: a Survey", Pattern Recognition, 37, pp 1887 - 1899, 2004.
- [10] M. Abdul Rahiman and M. S. Rajasree, "A Detailed Study and Analysis of OCR Research in South Indian Scripts", International Conference on Advances in Recent Technologies in Communication and Computing, pp.31-38, 2009.
- [11] T.V. Ashwin and P.S. Sastry, "A fonts and size-independent OCR system for printed Kannada documents using support vector machines". Sadhana 27, 2002, pp 35-58, 2002.
- [12] P. Nagbhushan and M. Pai Radhika, "Modified region decomposition method and optimal depth decomposition tree in the recognition of non-uniform sized characters - An experimentation with Kannada characters", Pattern Recognition Letter, 20, 1999, pp 1467-1475, 1999.

- [13] R. Kunte Sanjeev and Sudhaker Samuel, "Hu's invariant moments & Zernike moments approach for the recognition of basic symbols in printed Kannada text", *Sadhana* vol. 32, part 5, October 2007, pp. 521-533,2007.
- [14] R. Sanjeev Kunte and R.D. Sudhaker Samuel, "An OCR system for printed Kannada text using Two-stage Multi-network classification approach employing Wavelet Features", in the proc. of International Conference on Computational Intelligence and Multimedia Applications, pp 349-355,2007.
- [15] O.D. Trier, A.K. Jain and T. Taxt, "Feature Extraction Methods for Character Recognition—a Survey", *Pattern Recognition*, 29, 1996, pp 641–662,1996.
- [16] E. J. Candès and D. L. Donoho, "Curvelets –a surprisingly effective nonadaptive representation for objects with edges", In *Curves and Surfaces*, C. Rabut A. Cohen and L. L. Schumaker, editors, pages 105–120, Vanderbilt University Press, Nashville, TN,2000.
- [17] A.V. Narasimha Murthy, "Kannada Lipiya Ugama Mattu Vikasa",. Institute of Kuvempu Kannada Studies Publication, University of Mysore, 1975.
- [18] Ishrat Jahan Sumana, "Image Retrieval Using Discrete Curvelet Transform". Master Thesis. Gippsland School of Information Technology. Monash University, Australia, 2008.
- [19] J. L. Starck, E. J. Candès and D. L. Donoho, "The curvelet transform for image denoising", *IEEE Trans. Im. Proc.*, 11-6, 2002,pp 670–684,2002.
- [20] Ujjwal Bhattacharya and B.B. Chaudhuri, "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 31, No. 3, pp 444-457 ,March 2009.
- [21] Satish Kumar,"Study of features for Hand-printed recognition",World academy of science,engineering and technology,60,pp 1454-1466,2011.
- [22] H.Liu and L.YU,"Toward integrating feature selection algorithms for classification and clustering",*IEEE Trans.Knowledge and Data Engineering*,vol.17,pp.491-502,April 2005.
- [23] Farhad Mohamad Kazemi,Jalaleddin Izadian,Reihane Moravejjan and Ehsan Mohamad Kazemi, " Numeral Recognition Using Curvelet Transform", *ACS International Conference on Computer Systems and Applications*,pp 606-612,2008.
- [24] M.J. Fadili and J.L. Starck, "Curvelets and Ridgelets" , *Encyclopedia of Complexity and Systems Science* , Meyers, Robert (Ed.), Vol 3, pp 1718-1738, Springer ,New York, 2009..
- [25] S. Theodoridis and K. Koutroumbas,*Pattern Recognition*, Academic Press, New York,1999.
- [26] Leena Lepisto,"colour and texture based classification of rock images using classifier combinations",PhD Thesis,pp 49-51, 2006.
- [27] R. P .W Duin,"The combining classifier:to train or not to train",proceedings of 16th international conference on pattern recognition,Quebec,Canada,vol.2,pp.765-770,2002
- [28] Kuncheva,L.I.,"A theoretical study on six classifier fusion strategies",*IEEE transactions on pattern analysis and machine intelligence*,24(2),pp 281-286,2002.

AUTHORS PROFILE

Mamatha H R received her B E degree in Computer Science and Engineering from the Kuvempu University in 1998.and the M.Tech degree in Computer Networks and Engineering from the Visvesvaraya Technological University in 2006.Since 2008 she has been a Ph.D. student at the Visvesvaraya Technological University. She has 13 years of teaching experience. Currently she is working as an Associate Professor in the Department of Information Science and Engineering, P E S Institute of Technology. Her current research interests include pattern recognition and image processing. She has published 16 papers in various Journals and Conferences of International repute. She is a life member of Indian Society for Technical Education. She has mentored students for various competitions at international level.

Sucharitha S received her B.E degree in Computer Science and Engineering from Visvesvaraya Technological University in 2012. She is currently pursuing her Master of Science in Computer Science at Indiana University, Bloomington. Her current research interests are pattern recognition and image processing

Dr. Srikanta Murthy K received B.E (Electrical & Electronics Engineering). degree in 1986, M.Tech (Power Systems) in 1996 from National Institute of Engineering, University of Mysore and Ph.D. degree in Computer Science from University of Mysore in 2006. He joined the faculty of Electrical Engineering in NIE in 1987 and worked for K V G College of Engineering at various positions from 1991-2004. Presently he is working as Professor and Head, Department of CS&E, P E S Institute of Technology(south campus), Bangalore, India. He has served as Chairman Board of Examiners in computer science in Mangalore University and also the member of local inspection committee, Mangalore University and Visvesvaraya Technological University. Currently he is guiding 5 candidates towards the doctoral degree and also guided many projects at PG level. His current research interests are computer vision, image processing and pattern recognition. He has published 60+ papers in various Journals and Conferences of International repute. He is a life member of Indian Society for Technical Education, Indian Society of Remote Sensing, Computer Society of India, Society of Statistics and Computer Applications and associate member of Institute of Engineers

Designing a Markov Model for the Analysis of 2-tier Cognitive Radio Network

Tamal Chakraborty

Department of Electronics and Telecommunication
Engineering
Jadavpur University
Kolkata, India

Iti Saha Misra

Department of Electronics and Telecommunication
Engineering
Jadavpur University
Kolkata, India

Abstract—Cognitive Radio Network (CRN) aims to reduce spectrum congestion by allowing secondary users to utilize idle spectrum bands in the absence of primary users. However, the overall user capacity and hence, the system throughput is bounded by the total number of available idle channels in the system. This paper aims to solve the problem of limited user capacity in basic CRN by proposing a 2-tier CRN that allows another tier (or layer) of secondary users to transmit, in addition to the already existing set of primary and secondary users in the system. Markov Models are designed step-wise to map the interaction between primary and secondary users in both tiers by including suitable traffic distribution models and system parameters. Spectrum handoff is also incorporated in the developed Markov Models. Performance analysis is carried out in terms of SU transmission, dropping, blocking and handoff probabilities along with mathematical formulation of the overall SU throughput in 2-tier CRN. It confirms better spectrum utilization in spectrum handoff enabled 2-tier CRN over basic CRN with enhancement in quality of service for secondary users in terms of reduced dropping and blocking probabilities.

Keywords—Cognitive Radio Network; 2-tier; Voice over IP; Markov Model; Spectrum Handoff

I. INTRODUCTION

Wireless communication has witnessed increased popularity owing to rapid development of mobile and portable applications that have enabled users to communicate "anytime anywhere". This has led to formulation of admission control and network management policies to deal with the problems of scalability, fairness, synchronization and security, that arise with increased subscribers in wireless domain. Recent studies [1,2] have clearly demonstrated that while spectrum congestion hinders further growth in wireless communication, there are plenty of idle spectrum regions that are left unutilized. Cognitive Radio Network (CRN) [3-5] aims to create a common spectrum pool by including all such unused spectrum bands and allocate them to applications based on their requirements. It deploys opportunistic mode of communication where secondary or unlicensed users (SUs) transmit in the frequency slots when the corresponding primary or licensed users (PUs) are absent. However, practical implementation of CRN must address the issues of spectrum analysis, management and mobility, along with architectural specifications [6].

Extensive research work is being carried out to achieve higher spectrum utilization in CRN through formulation of

appropriate Medium Access Control (MAC) protocols [7, 8], handoff schemes [9], timing parameters [10, 11], etc. However, the system capacity in all these works has a maximum upper bound as derived in [12]. This paper introduces the concept of "2-tier CRN" that increases the capacity of CRN by admitting more number of SUs in the network. The SUs are categorized into two tiers. The SUs in the first tier are Voice over IP (VoIP) [13] users that transmit in the secondary transmission interval when the PUs are sensed idle. The second tier of SUs performs data transmission during the silence periods of VoIP SUs in the first tier. To the best of our knowledge, no such work has been reported on this issue so far as primarily, research has been carried out in CRN comprising of only one tier of SUs.

Markov Model serves as an effective tool to design CRN and has been implemented widely in recent works [14, 15]. The primary advantage of developing CRN with Markov Model is that it incorporates user-defined traffic distribution for PUs and SUs, along with customized network conditions and thereby, facilitates study of the complex interaction between PU and SU in CRN. Accordingly, the objective of this paper is to design Markov Models for basic and 2-tier CRN and analyze the increase in system capacity of 2-tier CRN over basic CRN with respect to SU dropping, blocking, handoff and transmission probabilities. A mathematical framework is also established that calculates the SU throughput for a complete spectrum handoff enabled 2-tier CRN.

The paper is organized as follows. The principle for 2-tier CRN is discussed in Section II. Markov Models for the first and second tier of 2-tier CRN are described in Section III along with spectrum handoff in these networks. Section IV provides mathematical model to calculate SU throughput in 2-tier CRN followed by performance analysis in Section V.

II. PRINCIPLE OF 2-TIER COGNITIVE RADIO NETWORK

The proposed 2-tier CRN consists of one tier of PUs and two tiers of SUs. PUs are allotted designated channels for transmission. When PUs are not transmitting, the idle channels are utilized by SUs. SUs are categorized into VoIP SUs and DATA SUs. VoIP SUs demand higher Quality of Service (QoS) and hence, have priority over the DATA SUs in accessing idle channels. Therefore, channels are utilized by SUs in the following manner.

- Whenever a licensed channel is sensed idle, VoIP SU occupies the channel and starts transmission. As VoIP

transmission occurs in talkspurts [16], there are idle periods of inactivity that are detected by codecs. Silence suppression [16] is performed, thereby making channel accessible to other users. At the onset of another talkspurt, the channel is reclaimed back by the VoIP SU for transmission. VoIP SU, therefore, occupies the first tier of 2-tier CRN and is denoted by SUTier1.

- DATA SU utilizes the channel during “off” period of SUTier1 and continues transmission until the channel is either reclaimed back by VoIP SU or is sensed busy at the end of secondary transmission time slot. These SUs implement queuing models to reduce packet loss when connection is terminated. DATA SU, thus, constitutes the second tier of 2-tier CRN and is denoted by SUTier2.

The principle of 2-tier CRN is depicted in a flowchart in Fig. 1.

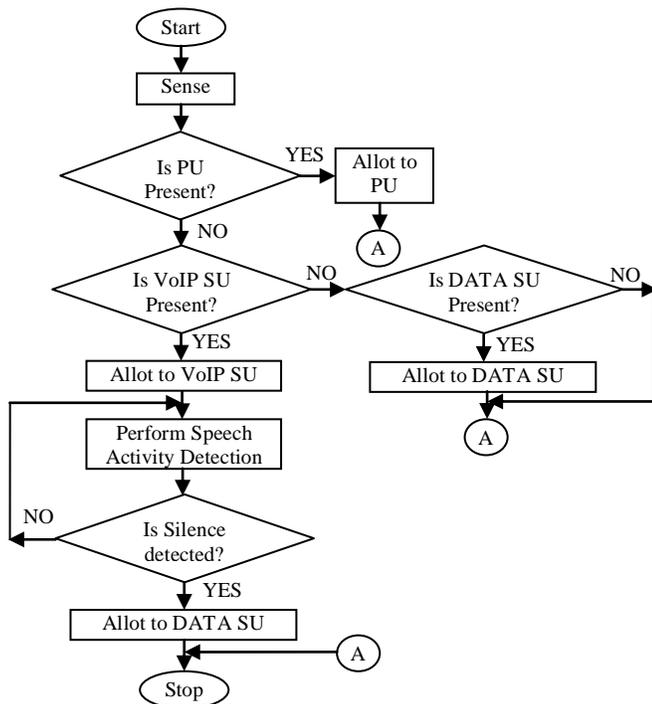


Fig. 1. Flowchart depicting the principle of 2-tier CRN

III. DESIGN OF MARKOV MODEL FOR 2-TIER CRN

This section deals with the design of Markov Models to study the interaction among PU, SUTier1 and SUTier2 under diverse channel conditions. Network is modelled as collection of states where each state denotes channel status with respect to PU, SUTier1 and SUTier2. Let the steady state probability for every such state be denoted by $P(i,j,k,l,m)$ where

- i = total number of active PUs transmitting in CRN,
- j = total number of active SUTier1 in CRN, that has arrived in the CRN,
- k = total number of active SUTier2 in CRN, that has been accepted by SUTier1,

- l =current “status” of SUTier1, and
- m =current “status” of SUTier2

The term “status” denotes the action taken by SU under different network conditions. The various status symbols along with their meanings are described in Table I.

Development of Markov model for 2-tier CRN is carried out incrementally in three phases. Initially, the first tier of CRN is modeled considering appropriate traffic distributions of PU and SUTier1. Secondly, SUTier2 is incorporated into the designed model following the principle of 2-tier CRN as discussed in Section II. Finally, spectrum handoff is incorporated for all SUs in the CRN.

A. Markov Model Design for first tier of CRN with Spectrum Handoff

Initially, the first tier of CRN is designed using Markov Model. It is obvious that in the absence of any further tier of SUs in the network, the first tier of CRN corresponds to the basic CRN comprising of PUs and a single set of SUs. It is considered that PU and SUTier1 arrive in CRN following Poisson distribution with mean rates λ_p and λ_s , respectively and have negative exponential service time distribution with mean rates $1/\mu_p$ and $1/\mu_s$ respectively. In order to design the Markov Model, $P(i,j,k,l,m)$ is calculated for every possible state. As SUTier2 is not present, $k=0$ for all $P(\bullet)$ in this scenario.

Spectrum handoff is implemented for SUTier1 such that on arrival of PU in current channel, SUTier1 shifts to the nearest available idle channel. It is to be noted that the implementation of spectrum handoff is dependent on several factors that include underlying MAC protocol, CRN architecture, handoff policies, etc. and hence, its discussion is beyond the scope of this paper. The generalized Markov model for CRN comprising of N channels is developed in Fig. 2 followed by the balance equations guiding the transmission of SUTier1.

TABLE I. STATUS SYMBOLS USED IN MARKOV MODEL

Status Value	Meaning	Definition
0	Transmission Mode	The SU has obtained access to a channel and is successfully transmitting.
	Null Mode	SU is not performing any transmission, handoff, blocking or dropping functions.
1	Handoff Mode	On PU arrival, the SU is performing spectrum handoff considering that an idle channel is available in the system. SU transmission is suspended temporarily during the handoff process.
2	Dropping Mode	SU transmission is suspended permanently as PU has arrived in the current channel and there are no idle channels available in CRN.
3	Blocking Mode	The incoming SU is not allowed to gain access to any channel for initiating transmission as there is no idle channel left in CRN.

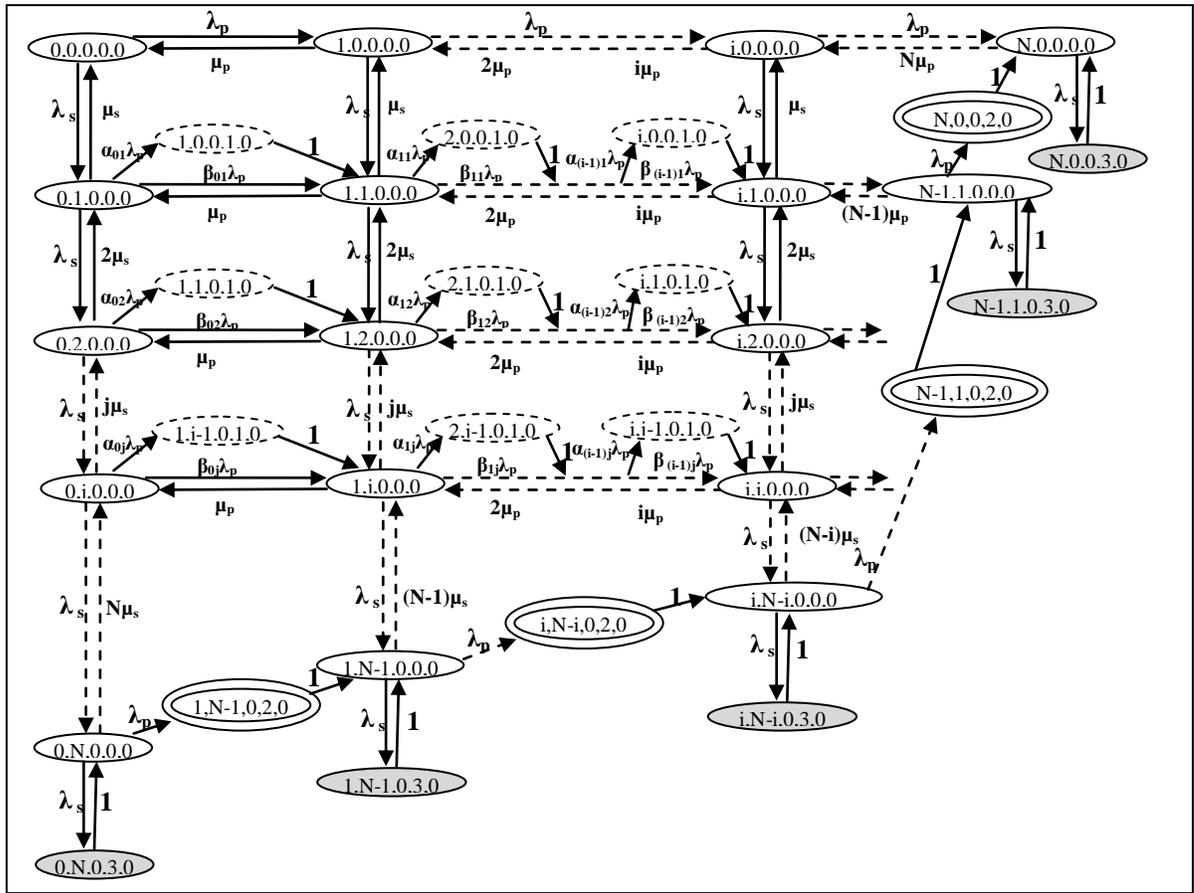


Fig. 2. Markov Model for the first tier of CRN

The balance equations governing the transmission of SUTier1 in the Markov Model for CRN are defined as follows.

$$i) \quad \boxed{i+j \leq N-1, i=0}: (\lambda_p + \lambda_s + j\mu_s)P(i, j, k, l, m) = \lambda_s P(i, j-1, k, l, m) + \mu_p P(i+1, j, k, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) \quad (1)$$

$$ii) \quad \boxed{i+j \leq N-1, i \neq 0}: (\lambda_p + \lambda_s + j\mu_s + i\mu_p)P(i, j, k, l, m) = \left\{ \frac{N-(i-1)-j}{N-(i-1)} \right\} \lambda_p P(i-1, j, k, l, m) + \lambda_s P(i, j-1, k, l, m) + P(i, j-1, k, 1, 0) + (i+1)\mu_p P(i+1, j, k, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) \quad (2)$$

$$iii) \quad \boxed{i+j = N, i=0}: (\lambda_p + \lambda_s + j\mu_s)P(i, j, k, l, m) = P(i, j, k, 3, 0) + \lambda_s P(i, j-1, k, 0, 0) \quad (3)$$

$$iv) \quad \boxed{i+j = N, i \neq 0}: (\lambda_p + \lambda_s + j\mu_s + i\mu_p)P(i, j, k, l, m) = \left\{ \frac{N-(i-1)-j}{N-(i-1)} \right\} \lambda_p P(i-1, j, k, l, m) + \lambda_s P(i, j-1, k, l, m) + P(i, j-1, k, 1, 0) + P(i, j, k, 2, 0) + P(i, j, k, 3, 0) \quad (4)$$

B. Markov Model Design for 2-Tier CRN with Spectrum Handoff only for SUTier1

In a 2-tier CRN, each SUTier1 allows SUTier2 to transmit during the silence periods as depicted in Fig. 1. Let SUTier2 arrive in CRN following Poisson distribution with λ_t as the mean rate and has negative exponential service time distribution with mean rate of $1/\mu_t$. Considering total number of PU and SUTier1 in the network at a certain time interval to be i and j respectively, the maximum number of SUTier2 admitted in CRN is j . The addition of SUTier2 by SUTier1 is depicted by a segment of the Markov Model in Fig. 3.

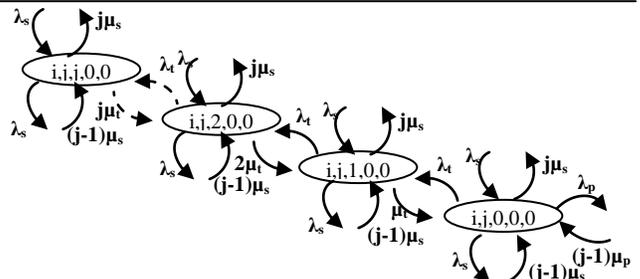


Fig. 3. Admission of SUTier2 in 2-tier CRN

The maximum system capacity in terms of users admitted in 2-tier CRN is given by,

$$C_{p_{max}} = PU + SUTier1 + SUTier2 = i + j + j = i + 2j \quad (5)$$

SUTier2 does not perform spectrum handoff in this model. Rather, it is dropped under three conditions namely, i) at a time when SUTier1 is dropped, ii) when SUTier1 performs spectrum handoff, and iii) after SUTier1 finishes transmission. Therefore, status of SUTier2, as denoted by m in $P(i,j,k,l,m)$, accepts values of 0, 2 and 3 depending on its transmission, dropping and blocking mode respectively. At any point of time, status combinations for SUTier1 and SUTier2 as represented by $\{l,m\}$ follow the conditions described in Table II.

Accordingly, the Markov Model for 2-tier CRN (where spectrum handoff is performed by SUTier1 only) is illustrated in Fig. 4 along with the balance equations for SUTier1 and SUTier2.

TABLE II. STATUS CONDITIONS FOR THE DESIGNED MARKOV MODEL OF 2-TIER CRN

Condition	Reason
$m = \{0,2\} \forall m \in \{l,m\}: l = 1$	SUTier2 is dropped when the corresponding SUTier1 implements spectrum handoff.
$l = 1 \forall l \in \{l,m\}: m = 1$	
$m = \{0,3\} \forall m \in \{l,m\}: l = 3$	The fact that SUTier2 is blocked from accessing the channel implies that SUTier1 is already blocked.
$l = 3 \forall l \in \{l,m\}: m = 3$	
$m = \{0,2\} \forall m \in \{l,m\}: l = 2$	Both SUTier1 and SUTier2 transmissions can be dropped on the arrival of PU. A special case occurs when SUTier2 transmission is dropped when the transmission time interval for SUTier1 is over and the channel is released.
$l = 0 \forall m \in \{l,m\}: m = 2$ when $j = k$	
$l = 2 \forall m \in \{l,m\}: m = 2$ when $j \neq k$	

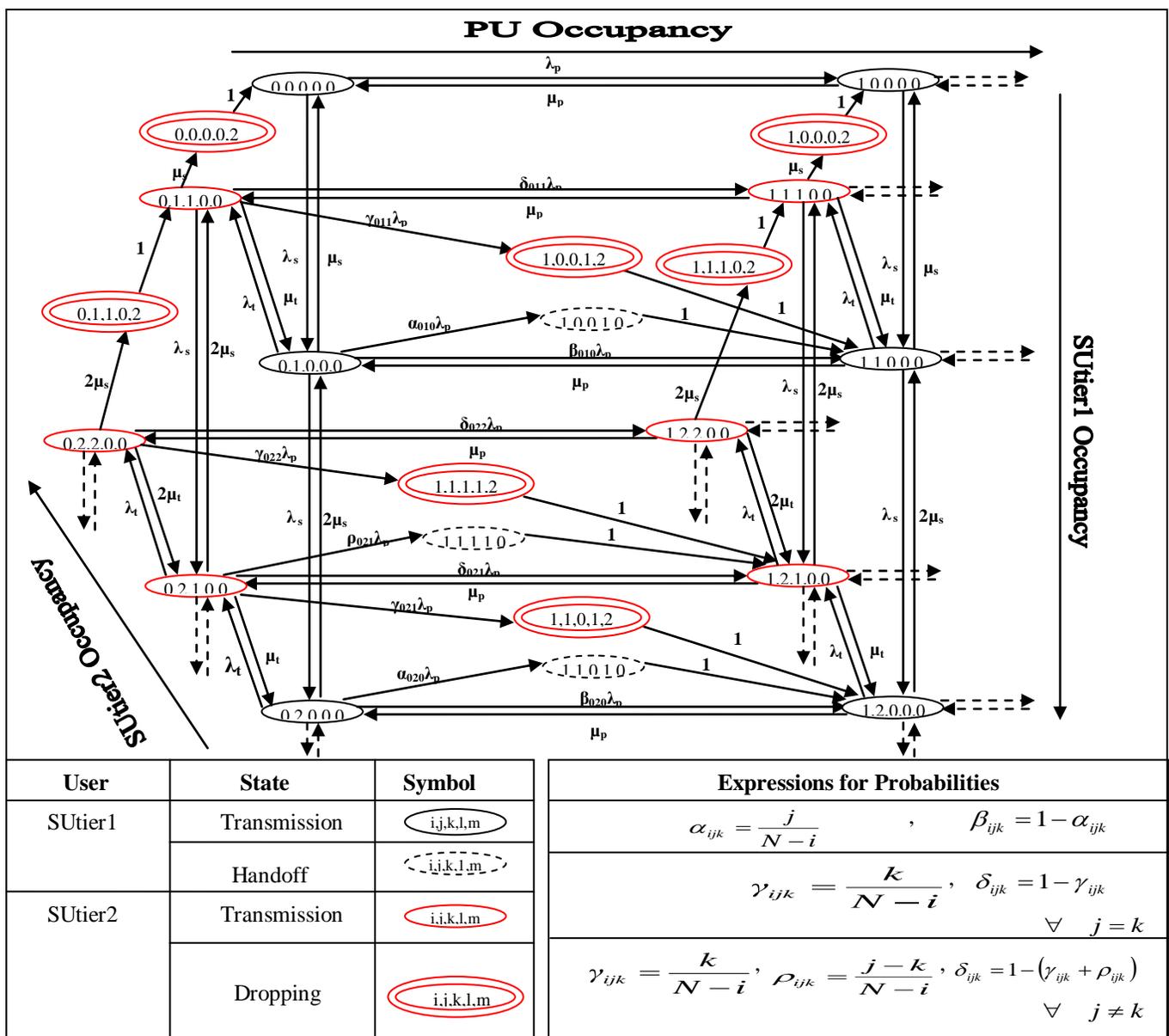


Fig. 4. Markov Model for 2-tier CRN with spectrum handoff only for SUTier1

The balance equations guiding the transmission of SUTier1 and SUTier2 in Markov Model for 2-tier CRN as per Fig. 4 are defined as follows.

CASE I: SUTier1

$$i) \boxed{i+j \leq N-1, i=0}: (\lambda_p + \lambda_s + \lambda_t + j\mu_s)P(i, j, k, l, m) = \lambda_s P(i, j-1, k, l, m) + \mu_p P(i+1, j, k, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) + \mu_t P(i, j, k+1, l, m) \quad (6)$$

$$ii) \boxed{i+j \leq N-1, i \neq 0}: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + i\mu_p)P(i, j, k, l, m) = \beta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + P(i, j-1, k, l, 0) + \mu_t P(i, j, k+1, l, m) + \lambda_s P(i, j-1, k, l, m) + (i+1)\mu_p P(i+1, j, k, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) + P(i, j-1, k, l, 2) \quad (7)$$

$$iii) \boxed{i+j = N, i=0}: (\lambda_p + \lambda_s + \lambda_t + j\mu_s)P(i, j, k, l, m) = P(i, j, k, 3, 0) + \lambda_s P(i, j-1, k, l, m) + \mu_t P(i, j, k+1, l, m) \quad (8)$$

$$iv) \boxed{i+j = N, i \neq 0}: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + i\mu_p)P(i, j, k, l, m) = \beta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + \lambda_s P(i, j-1, k, l, m) + P(i, j-1, k, l, 0) + \mu_t P(i, j, k+1, l, m) + P(i, j, k, 2, 0) + P(i, j, k, 3, 3) + P(i, j-1, k, l, 2) \quad (9)$$

CASE II: SUTier2

$$v) \boxed{i+j \leq N-1, k < j}: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + k\mu_t)P(i, j, k, l, m) = (k+1)\mu_t P(i, j, k+1, l, m) + \lambda_s P(i, j-1, k, l, m) + \mu_p P(i+1, j, k, l, m) + \lambda_t P(i, j, k-1, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) \quad (10)$$

$$vi) \boxed{i+j \leq N-1, k = j, i=0}: (\lambda_p + \lambda_s + j\mu_s + k\mu_t)P(i, j, k, l, m) = (j+1)\mu_s P(i, j+1, k, l, m) + \mu_p P(i+1, j, k, l, m) + \lambda_t P(i, j, k-1, l, m) + P(i, j, k, 0, 2) \quad (11)$$

$$vii) \boxed{i+j \leq N-1, k = j, i \neq 0}: (\lambda_p + \lambda_s + i\mu_p + j\mu_s + k\mu_t)P(i, j, k, l, m) = \delta_{(i-1)jk} P(i-1, j, k, l, m) + (i+1)\mu_p P(i+1, j, k, l, m) + \lambda_t P(i, j, k-1, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) + P(i, j, k, 0, 2) \quad (12)$$

$$viii) \boxed{i+j = N, k < j, i \neq 0}: (\lambda_p + \lambda_s + \lambda_t + i\mu_p + j\mu_s + k\mu_t)P(i, j, k, l, m) = P(i, j-1, k, l, 0) + P(i, j-1, k-1, l, 2) + \delta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + P(i, j, k, 2, 2) + P(i, j, k, 2, 0) + \lambda_s P(i, j-1, k, 0, 0) + \lambda_t P(i, j, k-1, 0, 0) + (k+1)\mu_t P(i, j, k+1, 0, 0) + P(i, j, k, 3, 0) \quad (13)$$

$$ix) \boxed{i+j = N, k = j, i \neq 0}: (\lambda_p + \lambda_s + i\mu_p + j\mu_s + k\mu_t)P(i, j, k, l, m) = \delta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + P(i, j, k, 2, 2) + P(i, j, k, 2, 0) + \lambda_t P(i, j, k-1, l, m) + P(i, j, k, 3, 3) \quad (14)$$

$$x) \boxed{i+j = N, k = j, i=0}: (\lambda_p + \lambda_s + j\mu_s + k\mu_t)P(i, j, k, l, m) = P(i, j, k, 3, 3) + \lambda_t P(i, j, k-1, l, m) \quad (15)$$

$$xi) \boxed{i+j = N, k < j, i=0}: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + k\mu_t)P(i, j, k, l, m) = \lambda_s P(i, j-1, k, l, m) + \lambda_t P(i, j, k-1, l, m) + (k+1)\mu_t P(i, j, k+1, l, m) + P(i, j, k, 3, 0) \quad (16)$$

C. Design of Markov Model for 2-Tier CRN with Spectrum Handoff for SUTier1 and SUTier2

In this section, Markov Model is designed for 2-tier CRN where both SUTier1 and SUTier2 perform spectrum handoff on sudden PU arrival. As PU arrives in the current channel and SUTier1 shifts to another channel, it sends information about the new channel to SUTier2. Thereafter, SUTier2 reorients its transceiver to frequency band corresponding to new channel and, thus, implements spectrum handoff. However, it must be noted that since admission of SUTier2 in CRN is completely governed by SUTier1, spectrum handoff can be performed by SUTier2 only when corresponding SUTier1 executes spectrum handoff and is represented by the following condition.

$$m = \{0, 1\} \forall m \in \{l, m\}: l = 1 \quad (17)$$

where l, m denote the status symbols in $P(\bullet)$.

Enabling handoff for all SUs in the network implies that as long as there are idle channels available in the system, the average system capacity is close to the maximum system capacity that is expressed in (5). Table III illustrates the conditions under which spectrum handoff can be performed by either only SUTier1 or both SUTier1 and SUTier2.

However, handoff mechanisms fail when all the idle channels are occupied by PUs and SUs. Mathematically, it is represented by, $N = i + j$ (18)

where N, i, j denote total number of channels, PU and SUTier1 in CRN respectively.

In this scenario, it can be ascertained from Table III that,

$$\delta = 1 - (\gamma + \rho) = N - (i + j) = 0 \quad (19)$$

Any further arrival of PU results in two cases.

- Case 1: $k < j, N > (i + j)$

Only SUTier1 is dropped as there is no SUTier2 in this channel. The probability of SUTier1 being dropped on PU arrival is given by,

$$\gamma = \frac{j}{N - i} \quad (20)$$

- Case 2: $k = j, N > (i + j)$

Both SUTier1 and SUTier2 are dropped on arrival of PU with probability as expressed in (21).

$$\rho = \frac{j-k}{N-i} \quad (21)$$

- Case 3: $k = j, N = (i + j)$

SUTier1 and SUTier2 are dropped as PU arrives with

probability = 1.

The complete Markov Model for 2-tier CRN with spectrum handoff implemented by all SUs is depicted in Fig. 5. Symbols as used in Fig. 4 are applied to denote the states in Fig. 5. The most significant balance equations for SUTier1 and SUTier2 corresponding to Fig. 5 are given as under.

The balance equations governing transmission of SUTier1 and SUTier2 in 2-tier CRN as per Fig. 5 are defined as follows.

CASE I: SUTier1

$$i) \quad i+j \leq N-1, i=0: (\lambda_p + \lambda_s + \lambda_t + j\mu_s)P(i, j, k, l, m) = \lambda_s P(i, j-1, k, l, m) + \mu_p P(i+1, j, k, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) + \mu_t P(i, j, k+1, l, m) \quad (22)$$

$$ii) \quad i+j \leq N-1, i \neq 0: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + i\mu_p)P(i, j, k, l, m) = \beta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + P(i, j-1, k, l, 0) + \mu_t P(i, j, k+1, l, m) + \lambda_s P(i, j-1, k, l, m) + (i+1)\mu_p P(i+1, j, k, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) + P(i, j-1, k, l, 2) \quad (23)$$

$$iii) \quad i+j = N, i=0: (\lambda_p + \lambda_s + \lambda_t + j\mu_s)P(i, j, k, l, m) = P(i, j, k, 3, 0) + \lambda_s P(i, j-1, k, l, m) + \mu_t P(i, j, k+1, l, m) \quad (24)$$

$$iv) \quad i+j = N, i \neq 0: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + i\mu_p)P(i, j, k, l, m) = \beta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + \lambda_s P(i, j-1, k, l, m) + P(i, j-1, k, l, 0) + \mu_t P(i, j, k+1, l, m) + P(i, j, k, 2, 0) + P(i, j, k, 3, 3) + P(i, j-1, k, l, 2) \quad (25)$$

CASE II: SUTier2

$$v) \quad i+j \leq N-1, k < j: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + k\mu_t)P(i, j, k, l, m) = (k+1)\mu_t P(i, j, k+1, l, m) + \lambda_s P(i, j-1, k, l, m) + \mu_p P(i+1, j, k, l, m) + \lambda_t P(i, j, k-1, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) \quad (26)$$

$$vi) \quad i+j \leq N-1, k = j, i=0: (\lambda_p + \lambda_s + j\mu_s + k\mu_t)P(i, j, k, l, m) = (j+1)\mu_s P(i, j+1, k, l, m) + \mu_p P(i+1, j, k, l, m) + \lambda_t P(i, j, k-1, l, m) + P(i, j, k, 0, 2) \quad (27)$$

$$vii) \quad i+j \leq N-1, k = j, i \neq 0: (\lambda_p + \lambda_s + i\mu_p + j\mu_s + k\mu_t)P(i, j, k, l, m) = \delta_{(i-1)jk} P(i-1, j, k, l, m) + (i+1)\mu_p P(i+1, j, k, l, m) + \lambda_t P(i, j, k-1, l, m) + (j+1)\mu_s P(i, j+1, k, l, m) + P(i, j, k, 0, 2) + P(i, j-1, k-1, l, 1) \quad (28)$$

$$viii) \quad i+j = N, k < j, i \neq 0: (\lambda_p + \lambda_s + \lambda_t + i\mu_p + j\mu_s + k\mu_t)P(i, j, k, l, m) = P(i, j-1, k, l, 0) + P(i, j-1, k-1, l, 1) + \delta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + P(i, j, k, 2, 2) + P(i, j, k, 2, 0) + \lambda_s P(i, j-1, k, 0, 0) + \lambda_t P(i, j, k-1, 0, 0) + (k+1)\mu_t P(i, j, k+1, 0, 0) + P(i, j, k, 3, 0) \quad (29)$$

$$ix) \quad i+j = N, k = j, i \neq 0: (\lambda_p + \lambda_s + i\mu_p + j\mu_s + k\mu_t)P(i, j, k, l, m) = \delta_{(i-1)jk} \lambda_p P(i-1, j, k, l, m) + P(i, j, k, 2, 2) + P(i, j, k, 2, 0) + \lambda_t P(i, j, k-1, l, m) + P(i, j, k, 3, 3) \quad (30)$$

$$x) \quad i+j = N, k = j, i=0: (\lambda_p + \lambda_s + j\mu_s + k\mu_t)P(i, j, k, l, m) = P(i, j, k, 3, 3) + \lambda_t P(i, j, k-1, l, m) \quad (31)$$

$$xi) \quad i+j = N, k < j, i=0: (\lambda_p + \lambda_s + \lambda_t + j\mu_s + k\mu_t)P(i, j, k, l, m) = \lambda_s P(i, j-1, k, l, m) + \lambda_t P(i, j, k-1, l, m) + (k+1)\mu_t P(i, j, k+1, l, m) + P(i, j, k, 3, 0) \quad (32)$$

TABLE III. DIFFERENT HANDOFF AND DROPPING CONDITIONS FOR SUTIER1 AND SUTIER2

Condition	PU Arrival Status	Probability	Handoff by SUTier1 only		Handoff by SUTier1 and SUTier2	
			Value of $\{l, m\}$	Remark	Value of $\{l, m\}$	Remark
$j \neq k$	PU arrives at a channel occupied by both SUTier1 and SUTier2. There are idle channels available in CRN.	$\gamma = \frac{j}{N-i}$	{1,2}	Handoff by SUTier1. SUTier2 is dropped.	{1,1}	Handoff by SUTier1 and SUTier2
	PU arrives at a channel that is used by SUTier1 only. There are idle channels available in CRN.	$\rho = \frac{j-k}{N-i}$	{1,0}	Handoff only by SUTier1. SUTier2 is unaffected.	{1,0}	Handoff only by SUTier1. SUTier2 is unaffected.
	PU occupies the channel not used by both SUTier1 and SUTier2.	$\delta = 1 - (\gamma + \rho)$	{0,0}	No handoff required	{0,0}	No handoff required
$j = k$	PU arrives at a channel occupied by both SUTier1 and SUTier2. There are idle channels available in CRN.	$\gamma = \frac{k}{N-i}$	{1,2}	Handoff by SUTier1. SUTier2 is dropped.	{1,1}	Handoff by SUTier1 and SUTier2
	PU occupies the channel not used by both SUTier1 and SUTier2.	$\delta = 1 - \gamma$	{0,0}	No handoff required	{0,0}	No handoff required

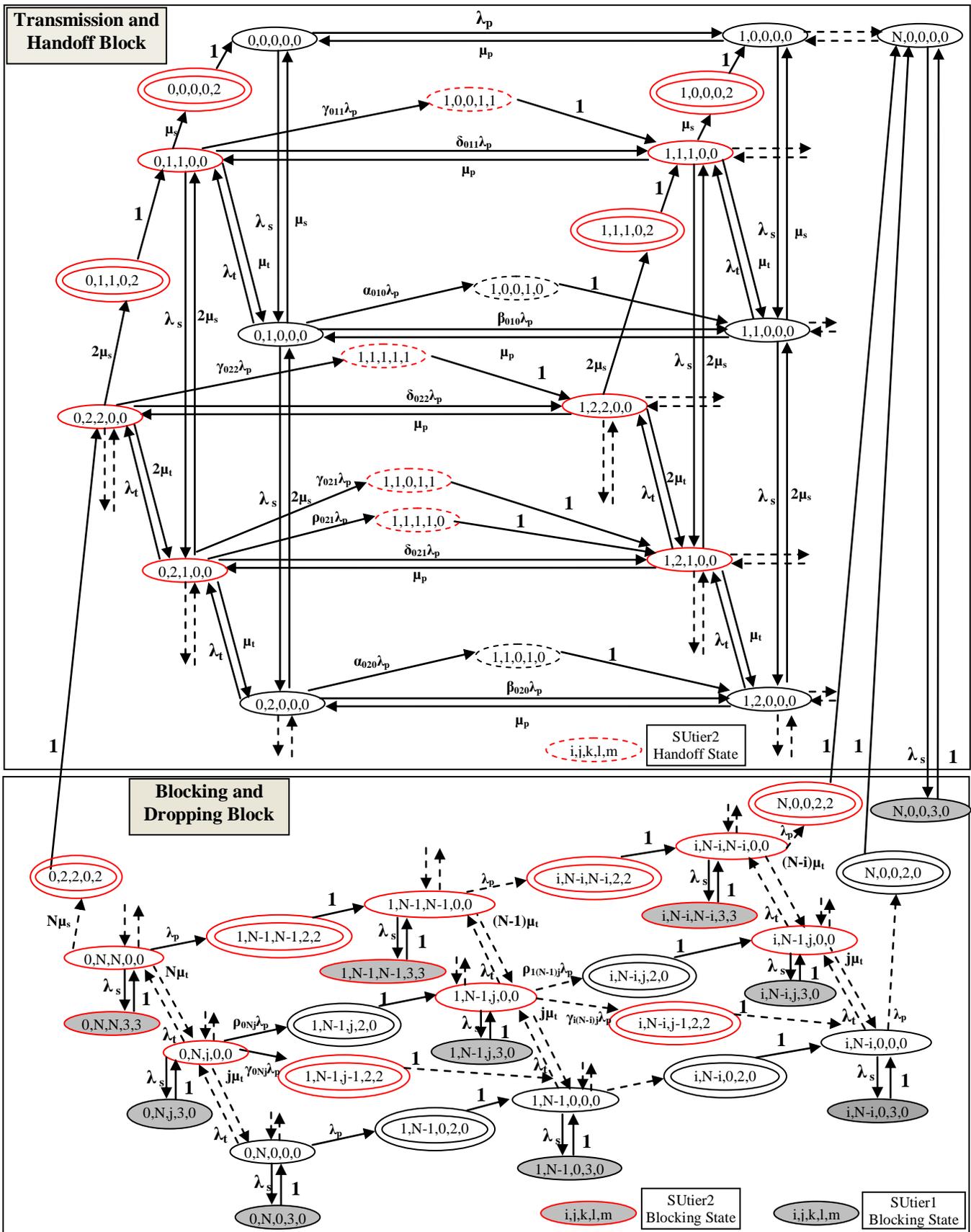


Fig. 5. Complete Markov Model for 2-tier CRN with spectrum handoff for both SUtier1 and SUtier2

IV. MATHEMATICAL FORMULATION OF SU THROUGHPUT
IN 2-TIER CRN

A mathematical expression is derived to obtain the throughput for the 2nd SUTier1 that arrives in the proposed 2-tier spectrum handoff enabled CRN. It must be noted that the number of available idle channels varies significantly depending on total number of spectrum handoff instances performed by existing SUs. Similarly, the number of SUs to get access to idle channels depends on the total number of SUTier2 supported by SUTier1. Let $m(t)$ be the overall number of available idle channels in CRN. Considering the effects of imperfect sensing by SUTier1 (false alarm and miss-detection), the total number of measured unoccupied channels as expressed in [17] is modified with respect to a particular time interval t and is defined as,

$$m'(t) = m(t) - m(t) \times p_f(\cdot) + (M - m(t)) \times (1 - p_d(\cdot)) \quad (33)$$

Let $Ph_b(a)$ be the probability of spectrum handoff performed by b th SU in tier 1 to shift from the current channel a' to a^{th} channel. Therefore, the throughput for the b^{th} SUTier1 having transmission rate $R_a(\text{VoIP})$ in a particular idle channel a' at a time interval t is given by,

$$C_b^{\text{SUTier1}}(t) = (1 - Ph_b(a)) \left\{ R_a(\text{VoIP}) \binom{m'(t)-1}{1} \left(\frac{1}{m'(t)-1} \right) \left(1 - \frac{1}{m'(t)-1} \right)^{j-1} \right\} \quad (34)$$

where j =total number of SUs in the system

Let $P_{ss}(\bullet)$ be a binary variable that defines whether SUTier2 is granted access by SUTier1 and is defined as follows.

$$P_{ss}(b) = 1 \text{ when } b^{\text{th}} \text{ SUTier1 allows SUTier2 to transmit} \\ = 0 \text{ otherwise.} \quad (35)$$

Accordingly, throughput of SUTier2 with transmission rate as $R_a(\text{DATA})$ in the a^{th} channel at time interval t is given by,

$$C_b^{\text{SUTier2}}(t) = P_{ss}(b) \left\{ R_a(\text{DATA}) \binom{m'(t)-1}{1} \left(\frac{1}{m'(t)-1} \right) \left(1 - \frac{1}{m'(t)-1} \right)^{j-2} \right\} \quad (36)$$

Combining (34) and (36), the total throughput for a particular set of SUTier1 and SUTier2 is expressed in (39).

There are several possibilities with respect to allotment of an idle channel to SUTier1 and is depicted as a three-layered tree in Fig. 6. For the 2nd SUTier1 in the system, the first layer determines whether the preceding SUTier1 grants access to SUTier2 or not. The second layer specifies the probabilities with which the 1st SUTier1 performs spectrum handoff in different channels. The third layer indicates the different spectrum handoff probabilities for the 2nd SUTier1.

A special case occurs when the 1st SUTier1 performs repeated handoff and finally occupies the penultimate channel. In this condition, the 2nd SUTier1 occupies only the last available idle channel and is dropped on the event of any further PU arrival as it cannot perform any spectrum handoff. Let $C_{2ndSUTier1}$ denote the throughput of 2nd SUTier1 corresponding to the second layer of the tree. The general expression for $C_{2ndSUTier1}$ is derived in (36). It is further modified to include the different conditions of spectrum handoff as per Fig. 6 and is expressed in (40).

Let C_{2ndSU} be the overall throughput of 2nd SUTier1 at the topmost layer of the tree and is given by,

$$C_{2ndSU}(t) = P_{ss}(1) C_{2ndSUTier1}(t) + (1 - P_{ss}(1)) C_{2ndSUTier1}(t) \quad (37)$$

V. PERFORMANCE ANALYSIS

This section analyzes the developed Markov models to establish the superiority of 2-tier CRN over basic CRN and also records significant performance improvement after incorporating spectrum handoff in 2-tier CRN. The key parameters that are used to analyze the performance improvement of 2-tier CRN over basic CRN include SU transmission, spectrum handoff, blocking and dropping probabilities and overall SU throughput.

Let P_L denote limiting probability of SU acceptance by available idle channel in CRN and is expressed as follows.

$$P_L = \sum_{\substack{i=0, \\ j=N-i}}^{N-1} P(i, j, 0, 0, 0) \quad (38)$$

$$C_{SUTier1}(t) = (1 - Ph_b(a)) \left[\left\{ R_a(\text{VoIP}) \binom{m'(t)-1}{1} \left(\frac{1}{m'(t)-1} \right) \left(1 - \frac{1}{m'(t)-1} \right)^{j-1} \right\} + P_{ss}(b) \left\{ R_a(\text{DATA}) \binom{m'(t)-1}{1} \left(\frac{1}{m'(t)-1} \right) \left(1 - \frac{1}{m'(t)-1} \right)^{j-2} \right\} \right] \quad (39)$$

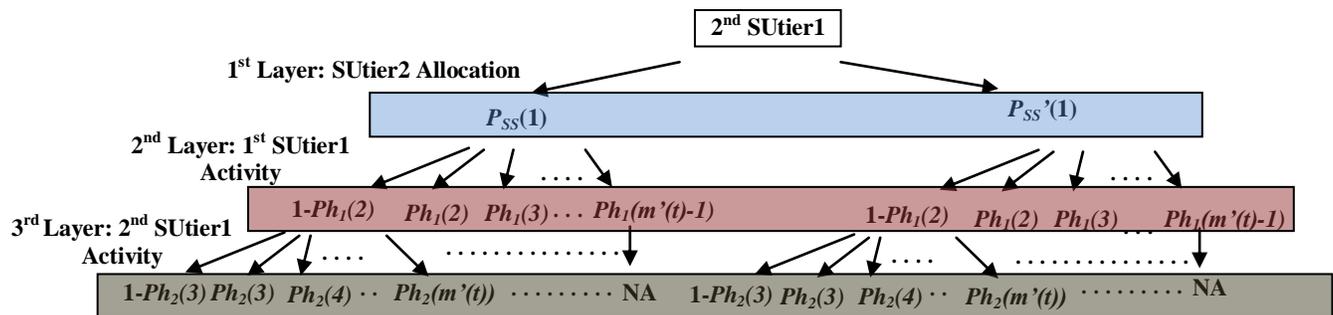


Fig. 6. Schematic Representation of all possibilities regarding channel allocation for 2nd SUTier1 on arrival in CRN

$$\begin{aligned}
 C_{2ndSUthrou}(t) = & \left[(1 - Ph_1(2)) \left[(1 - Ph_2(3)) \left[R_2(VoIP) \binom{m'(t)-1}{1} \left(\frac{1}{m'(t)-1} \right) \left(1 - \frac{1}{m'(t)-1} \right)^l + P_{ss}(2) R_2(DATA) \binom{m'(t)-1}{1} \left(\frac{1}{m'(t)-1} \right) \left(1 - \frac{1}{m'(t)-1} \right)^m \right] \right. \right. \\
 & + \left. \left[\sum_{k=3}^{m'(t)-1} Ph_2(k) \left[R_k(VoIP) \binom{m'(t)-k+1}{1} \left(\frac{1}{m'(t)-k+1} \right) \left(1 - \frac{1}{m'(t)-k+1} \right)^l + P_{ss}(2) R_k(DATA) \binom{m'(t)-k+1}{1} \left(\frac{1}{m'(t)-k+1} \right) \left(1 - \frac{1}{m'(t)-k+1} \right)^m \right] \right] \right. \\
 & + \left. \left[Ph_2(m'(t)) \left[R_{m'(t)}(VoIP) + P_{ss}(2) R_{m'(t)}(DATA) \right] \right] + \left[(Ph_1(2)) \left[(1 - Ph_2(4)) \left[R_3(VoIP) \binom{m'(t)-2}{1} \left(\frac{1}{m'(t)-2} \right) \left(1 - \frac{1}{m'(t)-2} \right)^l + P_{ss}(2) R_3(DATA) \binom{m'(t)-2}{1} \left(\frac{1}{m'(t)-2} \right) \left(1 - \frac{1}{m'(t)-2} \right)^m \right] \right. \right. \\
 & + \left. \left[\sum_{k=4}^{m'(t)-1} Ph_2(k) \left[R_k(VoIP) \binom{m'(t)-k+1}{1} \left(\frac{1}{m'(t)-k+1} \right) \left(1 - \frac{1}{m'(t)-k+1} \right)^l + P_{ss}(2) R_k(DATA) \binom{m'(t)-k+1}{1} \left(\frac{1}{m'(t)-k+1} \right) \left(1 - \frac{1}{m'(t)-k+1} \right)^m \right] \right] \right. \\
 & + \left. \left[Ph_2(m'(t)) \left[R_{m'(t)}(VoIP) + P_{ss}(2) R_{m'(t)}(DATA) \right] \right] + \dots + \left[Ph_1(m'(t)-1) \left[R_{m'(t)}(VoIP) + P_{ss}(2) R_{m'(t)}(DATA) \right] \right] \right] \\
 \text{where } & l = j - 3, m = j - 4 \forall P_{ss}(1), \\
 & l = j - 2, m = j - 3 \forall (1 - P_{ss}(1))
 \end{aligned} \tag{40}$$

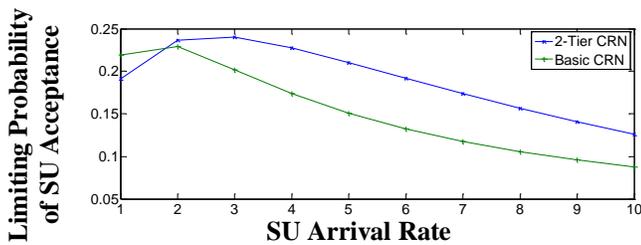


Fig. 7. Variation in limiting probability of SU acceptance by CRN with SU arrival rate

It is observed from Fig. 7 that 2-tier CRN provides higher probability of SU acceptance than basic CRN and thus reduces the overall blocking probability (denoted by P_B).

Let P_D define the steady state dropping probability that SU transmission is dropped before scheduled transmission interval is over. The expression for P_D is derived from [18] and is expressed in (41) as per the designed Markov Model.

$$P_D = \frac{\sum P_{drop}}{(1 - P_B) \lambda_s} \tag{41}$$

where $P_{drop} = \sum P(i, j, k, l, m) \forall (l, m) = (2, 2) | (2, 0) | (0, 2)$

Therefore, spectrum handoff must be performed by SUs to shift to available idle channels on PU arrival to reduce P_D . Let $P_{handoff_tier1}$ and $P_{handoff_tier2}$ be the probabilities of spectrum handoff performed by SUTier1 and SUTier2 respectively and are expressed as follows.

$$P_{handoff_tier1}(i, j) = \frac{\sum P(i, j, 0, 1, 0)}{(1 - P_B) \lambda_s} \tag{42}$$

$$P_{handoff_tier2}(i, j) = \frac{\sum P(i, j, 0, 1, 1)}{(1 - P_B) \lambda_s} \tag{43}$$

The dropping and handoff probabilities for SUs in 2-tier CRN are plotted in Fig. 8 and Fig. 9 respectively, for two scenarios that correspond to i) spectrum handoff by SUTier1 only, and ii) spectrum handoff by SUTier1 and SUTier2. It is

imperative that when SUTier1 performs spectrum handoff, SUTier2 is either dropped or else it must also perform handoff. This situation is clearly reflected in Fig. 8 where P_D is less for scenario 2 compared to scenario 1.

Therefore, reduction in blocking and dropping probabilities must increase SU throughput in spectrum handoff enabled 2-tier CRN. This is illustrated in Fig. 10 that plots the probability of successful transmission by SUs with increase in PU activity. It is observed from the figure that 2-tier CRN with complete spectrum handoff has the highest probability of transmission

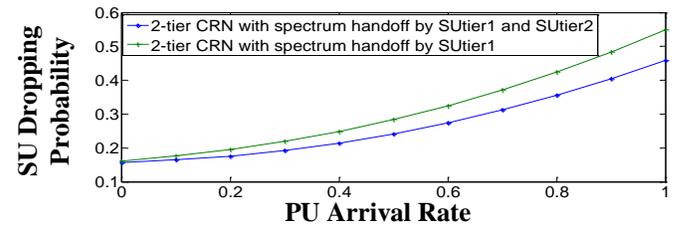


Fig. 8. Variation in SU dropping probability in CRN with PU arrival rate

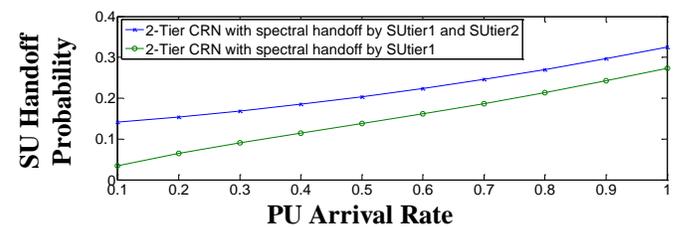


Fig. 9. Performance of 2-tier CRN with respect to SU handoff probability for varying PU arrival rate

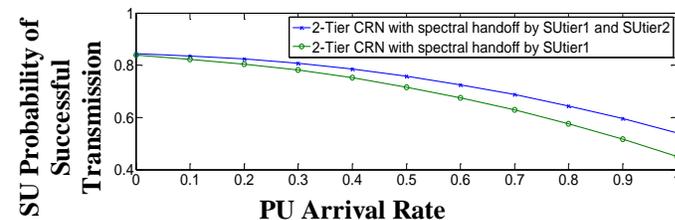


Fig. 10. Variation in probability of successful transmission by SU in CRN with PU arrival rate

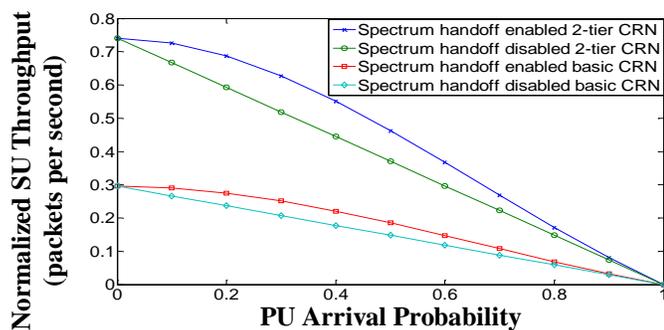


Fig. 11. Effect of spectrum handoff on normalized SU throughput with varying PU arrival probability for basic and 2-tier CRN

The normalized SU throughput as obtained from (37) is plotted in Fig. 11 for increasing probability of PU arrival on the current channel with respect to 2-tier CRN and basic CRN. As observed from the figure, spectrum handoff enabled 2-tier CRN provides the highest SU throughput compared to the other scenarios. In addition, CRN implementing spectrum handoff performs better as it records almost 25% enhancement in throughput (for 0.5 PU arrival probability) compared to spectrum handoff disabled CRN. However, as PU activity increases in CRN, number of idle channels reduces drastically. Under such circumstances, throughput for SUs supporting spectrum handoff decreases as reflected in Fig. 11.

Thus, observation from the designed Markov Model in Fig. 10 is validated using output derived from the mathematical model as represented in Fig. 11.

VI. CONCLUSION

This work has addressed the problem of limited system capacity in basic CRN by designing a 2-tier CRN that allows more number of SUs in the system. While the first tier of SUs involve in VoIP communication, the second tier of SUs exploit the silence periods in VoIP transmission to send data. Markov Models have been designed in this regard to highlight the difference between basic and 2-tier CRN. Spectrum handoff has also been incorporated in the developed Markov Model for performance enhancement. Analysis of the Markov models has recorded significant reduction in SU dropping and blocking probabilities in spectrum handoff enabled 2-tier CRN along with increase in successful transmission probabilities for SUs. A mathematical framework to study SU behavior has been formulated, that has recorded highest SU throughput after enabling spectrum handoff in 2-tier CRN and has, thus, confirmed the inference drawn from the Markov models. The 2-tier CRN is being studied presently to devise appropriate MAC protocols and spectrum handoff policies apart from architectural modifications and channel reservation schemes.

ACKNOWLEDGMENT

The first author acknowledges the support of INSPIRE Fellowship 2012 from DST, Govt. of India.

REFERENCES

[1] Federal Communications Commission, "Spectrum policy task force report," ET Docket No.02-135, 2002.

[2] U.S Department of Commerce, "United States Frequency Allocations: The Radio Spectrum", 2011.

[3] Federal Communications Commission, "Notice of proposed rule making and order: Facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies," ET Docket No. 03-108, February 2005.

[4] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," IEEE Communications Surveys and Tutorials, vol.11, no.1, pp.116-130, 2009, doi: 10.1109/SURV.2009.090109.

[5] J. Mitola III and G. Q. Maguire, Jr, "Cognitive radio: making software radios more personal," IEEE Personal Communications, vol. 6, no. 4, pp. 13-18, August 1999, doi: 10.1109/98.788210.

[6] F. Akyildiz, W. Y. Lee, M. C. Vuran and S. Mohanty, "NeXt generation / dynamic spectrum access / cognitive radio wireless networks: A survey," Computer Networks Journal (Elsevier) 50, pp. 2127- 2159, September 2006.

[7] C. Qian, L. Ying-Chang, M. Motani and W. Wai-Choong, "A Two-Level MAC Protocol Strategy for Opportunistic Spectrum Access in Cognitive Radio Networks," IEEE Transactions on Vehicular Technology, vol.60, no.5, pp.2164-2180, June 2011, doi: 10.1109/TVT.2011.2141694.

[8] T.T. Le and L.L. Bao, "Distributed MAC Protocol for Cognitive Radio Networks: Design, Analysis, and Optimization," IEEE Transactions on Vehicular Technology, vol. 60, no. 8, pp. 3990-4003, October 2011, doi: 10.1109/TVT.2011.2165325.

[9] L. Wang, C. Wang and C. Chang, "Modeling and Analysis for Spectrum Handoffs in Cognitive Radio Networks," IEEE Transactions on Mobile Computing, vol. 11, no. 9, pp. 1499-1513, September 2012, doi: 10.1109/TMC.2011.155.

[10] T. Chakraborty, I.S. Misra, and S.K. Sanyal, "Selection of optimal transmission time in cognitive radio network for efficient VoIP performance", Proc. of Fifth International Conference on Computers and Devices for Communication (CODEC), pp.1-4, India, 17-19 December 2012, doi: 10.1109/CODEC.2012.6509230.

[11] P. Wang, L. Xiao, Shidong Zhou and J. Wang, "Optimization of detection time for channel efficiency in cognitive radio systems," Proc. of Wireless Communications and Networking Conference, (WCNC 2007), pp.111-115, Hong-Kong, March 11-15, 2007.

[12] S. Srinivasa and S. Jafar, "How much spectrum sharing is optimal in cognitive radio networks?," IEEE Transactions on Wireless Communications, vol. 7, no. 10, pp. 4010-4018, October 2008, doi:10.1109/T-WC.2008.070647.

[13] B. Khasnabish, Implementing Voice over IP. Wiley-Interscience, John Wiley & Sons, Inc., 2003.

[14] Z. Xiaorong, S. Lianfeng and T.-S.P.Yum, "Analysis of Cognitive Radio Spectrum Access with Optimal Channel Reservation," IEEE Communications Letters, vol.11, no.4, pp.304-306, April 2007, doi: 10.1109/LCOM.2007.348282.

[15] Y.R. Kondareddy, N. Andrews, and P. Agrawal, "On the capacity of secondary users in a cognitive radio network," Proc. of IEEE Sarnoff Symposium (SARNOFF '09), pp.1-5, U.S.A., March 30 -April 1, 2009.

[16] I.A. Qaimkhani and E. Hossain, "Efficient silence suppression and call admission control through contention-free medium access for VoIP in WiFi networks," IEEE Communications Magazine, vol. 46, no. 1, pp. 90-99, January 2008, doi: 10.1109/MCOM.2008.4427236.

[17] H. Lee and D. Cho, "Capacity Improvement and Analysis of VoIP Service in a Cognitive Radio System," IEEE Transactions on Vehicular Technology, vol. 59, no. 4, pp. 1646-1651, May 2010, doi: 10.1109/TVT.2009.2039503.

[18] J. Martinez-Bauset, V. Pla and D. Pacheco-Paramo, "Comments on "analysis of cognitive radio spectrum access with optimal channel reservation"," IEEE Communications Letters, vol.13, no.10, pp.739, October 2009, doi: 10.1109/LCOMM.2009.090668.

A Fuzzy Rule Based Forensic Analysis of DDoS Attack in MANET

Ms. Sarah Ahmed

Research Scholar: dept. of Computer Science & Engineering
G. H. Raisoni College of Engineering, Nagpur
Maharashtra, India

Ms. S. M. Nirkhi

Assistance Professor: dept. of Computer Science &
Engineering
G. H. Raisoni College of Engineering, Nagpur
Maharashtra, India

Abstract—Mobile Ad Hoc Network (MANET) is a mobile distributed wireless networks. In MANET each node are self capable that support routing functionality in an ad hoc scenario, forwarding of data or exchange of topology information using wireless communications. These characteristic specifies a better scalability of network. But this advantage leads to the scope of security compromising. One of the easy ways of security compromise is denial of services (DoS) form of attack, this attack may paralyze a node or the entire network and when coordinated by group of attackers is considered as distributed denial of services (DDoS) attack. A typical, DoS attack is flooding excessive volume of traffic to deplete key resources of the target network. In MANET flooding can be done at routing. Ad Hoc nature of MANET calls for dynamic route management. In flat ad hoc routing categories there falls the reactive protocols sub category, in which one of the most prominent member of this subcategory is dynamic source routing (DSR) which works well for smaller number of nodes and low mobility situations. DSR allows on demand route discovery, for this they broadcast a route request message (RREQ). Intelligently flooding RREQ message there forth causing DoS or DDoS attack, making targeted network paralyzed for a small duration of time is not very difficult to launch and have potential of loss to the network. After an attack on the target system is successful enough to crash or disrupt MANET for some period of time, this event of breach triggers for investigation. Investigation and forensically analyzing attack scenario provides the source of digital proof against attacker. In this paper, the parameters for RREQ flooding are pointed, on basis of these parameters fuzzy logic based rules are deduced and described for both DoS and DDoS. We implemented a fuzzy forensic tool to determine the flooding RREQ attack of the form DoS and DDoS. For this implementation various experiments and results are elaborated in this paper.

Keywords—DoS and DDoS attack; DSR; Fuzzy logic; MANET; Network forensic analysis.

I. INTRODUCTION

Network forensics is still under active investigation by the research community, especially to address the issues in wireless networks [2]. Mobile Ad hoc network (MANET) a kind of wireless networks. It is the distributed systems having wireless mobile nodes that can freely and dynamically self-organise into arbitrary, temporary, and ad hoc network topologies, allowing connections within the network neither having pre-existing communication infrastructure nor centralized administered control management. As any network are having security vulnerabilities, so as MANET. However,

the MANET unique characteristics and features are advantageous, on the contrary can add up to security threats. One of the major types of problems in the network security is Denial of service (DoS) attacks because they are one of the most frequently used attack methods [6]. DoS are active attacks, which cannot be made stealth [5]. MANET are particularly susceptible to DDoS attack [1]. So, DoS/DDoS are easy to implement in MANET and to make it unrecognizable it is required to be done keenly. Flooding attack causes excessive volume of traffic to deplete key resources of the target legitimate users, since the system get congested so forth, there is denial of services. Flooding attack is a kind of denial of service attack in which the malicious node tries to inundate the victim by repeatedly sending redundant packets/data. The dynamic nature of MANET allows routing like dynamic source routing (DSR) and attackers can take the advantage of this dynamic source routing (categorized under the reactive routing) in which route is discovered on demand or when needed, for this interested node sends Route request message (RREQ) at discover phase and attacker can flood the network with RREQ causing denial of services.

Mission-critical applications demands technologies and methods for security incident investigation [2]. Network forensic is the act of capturing, recording, and analysing network audit trails in order to discover the source of security breaches or other information assurance problems [7]. Network forensics uncovers the facts of unauthorised or malicious activities. Forensic investigation aims to gain insight into and reach conclusions about critical questions of network security incident. The study of network forensics analysis for attacks in wireless network [2] and in MANET is considered still in progress. Forensic analysis can be done by unsupervised method it may require long iterations. Statistical methods like Cumulative sum (CUSUM), adaptive threshold, statistical moments. CUSUM or adaptive threshold methods main disadvantage is that require parameters for appropriate threshold value and statistical modelling method main problem is modelling the network traffic [6]. Modelling and estimating accurate threshold parameter for network traffic is a difficult problem. Security expert or forensic investigator analyses the network traffic using the empirical knowledge. Fuzzy logic deals with reasoning that is approximation and uncertainty assumption rather than exact value. This technique can be well implemented for analysis. Fuzzy based analysis system perform better for low and high intensity attack [6] and

reduce the time and cost of analysis [7]. This technique is efficient in complex analysis and it is rule based so easily modifiable, but requires fine tuning of rules.

This paper is organized as follows. Section II defines problem statement. Section III, IV points out various parameter of register request message in dynamic source routing is considered at general scenario and attack scenario. Section V describes the proposed fuzzy rules for network forensics analysis. Section VI determines the experiment and result for fuzzy forensic analysis tool. Section VII describe conclusion. Section VIII shows snapshots.

II. PROBLEM STATEMENT

Flooding attack is a kind of DoS attack. DoS are active attacks. DoS can be caused by an attacker to compromise one node or group of nodes in a network and DDoS can be caused by group of attacker nodes to compromise one node or group of nodes in the network. In our work, we have considered the DoS caused by an attacker to compromise group of nodes and DDoS caused group of attackers to compromise group of nodes.

MANET is mobile wireless network and requires ad-hoc settings, so routing is needed to find the path between source and destination. Dynamic routing eliminates the periodic routing updates and prevents nodes from unnecessary battery loss. In Dynamic source routing (DSR) [16], a node need to discover a route, it broadcasts a route request (RREQ) with a unique identification and the destination address as parameters. Any node that receives a route request, either if the node has already received the request it drops the request packet, or if the node recognizes its own address as destination the request reached target, otherwise the node appends its own address to the list of traversed hops in the packet and broadcasts this update route request. In MANET when attack is launched at routing (DSR routing), ROUTE REQUEST (RREQ) packet is broadcasted is sent again and again with address spoofing and without address spoofing (in our work we have considered non address spoofed RREQ packet) in a second violating of broadcast management causing DoS/ DDoS in the network. By doing this attacker tries to inundate the legitimate user by redundant RREQ packet. In DoS the rate of attack by a attacker is high as compared to DDoS attack to cause damage of same intensity. DoS attacks can be limited by enforcing the maximum route length that a packet should travel, source authentication, message integrity or using some other active approaches to trace the location of attacker by estimating signal strength, for this, nodes in MANET should have capabilities [1] to implement preventive measure as above which itself has own constrainer.

After attacks that had compromised the security of the entire network for a duration of time investigation. Forensic investigation uncovers the various facts related to attack by forensically analyzing the attack pattern.

In the proposed work, forensic analysis is done using fuzzy logic. Motivation of using fuzzy logic is that, through fuzzy logic more appropriate pattern analysis rules can be implemented for both DoS and DDoS due to RREQ flooding.

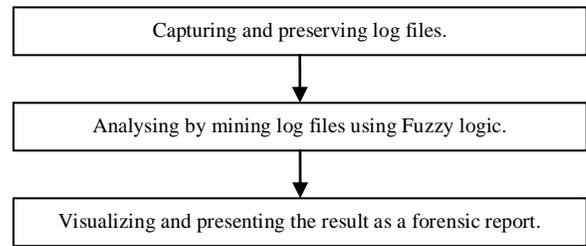


Fig. 1. Flow of work.

III. PARAMETERS BELONG TO GENERAL SCENARIO FOR ROUTE DISCOVERY

When a node wants to send a packet to the destination node, it first searches its Route Cache from a suitable route to the destination node. If no route from source node to destination node exists in source route cache, then source node initiates Route Discovery and sends out a ROUTE REQUEST message to find a route. When the message is first sent by a sender node, which is willing to find route of some destination node. The sender that is initiator node set the Initiator ID, the Target Id and the Unique Request Id in the ROUTE REQUEST message and then broadcasts the message. Nodes within the wireless transmission range receive this RREQ. The sender/initiator keeps a copy of the packet in a send buffer. The timestamps of message can be used to determine if it should send packet/message again. When any node receives a ROUTE REQUEST message it examines the Target ID to determine if it is the target/destination of the message. If the node is not the target it searches its own route cache for a route to the target. If a route is found it is returned. If not, the node's own id is appended to the Address List and the ROUTE REQUEST is broadcasted ahead to its neighbor.

TABLE I. FIELDS OF ROUTE REQUEST(RREQ) MESSAGE.

Fields	Explanation
Initiator ID	The address of the source node.
Target ID	The address of the target node.
Unique Request ID	A unique ID for identification of message.
Address List	A list of all addresses of intermediate nodes that the passes before its destination.
Hop Limit	The hop limit can be used to limit the number of nodes that the message is allowed to pass (varies from 1 to 255).
Acknowledgement bit	Option is set so that the destination node returns an acknowledgement when a packet is received.

If a node subsequently receives two ROUTE REQUESTs with the same Request id, it is possible to specify that only the first should be handled and the subsequent is discarded .If the node is the destination/target it returns a ROUTE REPLY message to the sender/initiator.

IV. PARAMETERS THAT BELONG TO ATTACK SCENARIO FOR ROUTE DISCOVERY

Various parameters which can be considered in attack situation for route discovery are:

- Total number of RREQ packet sent by the neighbor per unit time. In attacking scenario, attacker would not comply with broadcast management techniques adopted in current routing protocols such as limiting the maximum number of (continuous flow) RREQ packets sent per second. Therefore in attack scenario there is no limit, on the rate of RREQ message is considered by attackers for continuous flow.
- The hop count (TTL) is limitless in attacking scenario.
- The time duration for which targeted node is compromised, engaged in handling unnecessary routing load.
- In attacking situation the acknowledgement option is not emphasized and acknowledgement is not considered.
- To maintain a continuous flow by an attacker do attack with high rate and group of attackers do attack with low rate.

V. FUZZY RULES

Fuzzy forensic analysis tool uses fuzzy IF-THEN rules. A fuzzy IF-THEN rule is of the form, IF $X_1 = A_1$ and $X_2 = A_2$ and $X_n = A_n$ THEN Z_n , where X_i is linguistic variable description and A and Z are linguistic terms.

The ‘IF’ part is the antecedent or premise and the ‘THEN’ part is the consequence or conclusion.

TABLE II. LINGUISTIC VARIABLES AND DESCRIPTIONS.

Variables	Description
X1	The total number of RREQ continuous flow from a node in a second.
X2	For continuous flow the RREQ message length (Expanding ring searches on hop count).
X3	Time duration count of RREQ a node is targeted.
X4	Count of acknowledgement forwarded to initiator node of continuous flow, which is unattended (Route reply).
X5	The total number of RREQ continuous flow from group of nodes in a second.
X6	Count of Initiator nodes of continuous flow .

TABLE III. INPUT LINGUISTIC TERMS AND DESCRIPTIONS.

Name	Description
A1	greater than 1.
A2	greater than 255 hops.(for our work we considered 50)
A3	Difference in time stamp of continuous flow is greater than 1.
A4	greater than 0.
A5	greater than 1.
A6	greater than 1.

TABLE IV. OUTPUT LINGUISTIC TERM AND DESCRIPTION.

Name	Description
Z1	Flooding RREQ attack by a node.
Z2	Flooding RREQ attack by group of nodes.

TABLE V. RULES REPRESENTATION

Rule Name	Rule Representation
R1=DoS Attack	If($X_1=A_1, X_2=A_2, X_3=A_3, X_4=A_4$) then Z1
R2=DDoS Attack	If($X_2=A_2, X_3=A_3, X_4=A_4, X_5=A_5, X_6=A_6$) then Z2

Motivation of using fuzzy logic is that, through fuzzy logic more appropriate attack pattern analysis rules can be implemented [8].

For implementation of complete linguistic description of rule require compound rule structures which is implemented by disjunctive antecedents. One of the compound structures in our implementation is like X_1 , can be no attack (general normal scenario), lower rate attack and higher rate attack that is evaluated on:

$$\mu_{X1} = \begin{cases} 0 & \text{if } A1 < 2, \\ ((A1-2) \% 9) & \text{if } 2 \leq A1 \leq 11, \\ 1 & \text{if } A1 > 11 \end{cases}$$

The membership is considered as 0 when no attack is launched, membership is considered as $((X1-2) \% 9)$ when lower rate attack of RREQ flooding is launched, and 1 when higher rate attack of RREQ flooding is launched.

Same way, other compound structures $X_2, X_3, X_4, X_5,$ and X_6 are implemented for A_2, A_3, A_4, A_5, A_6 respectively.

The Mamdani Min type of fuzzy modeling is used to for composition of rules R1 and R2, using max-min rule of composition.

VI. EXPEREMENT AND RESULTS

In the experiment, for evaluation we implemented the various attack scenario and the trace files we considered as a log, is input to the forensic analysis tool which uses fuzzy logic for analysis to generate the forensic digital proof for each case having a unique hash value.

- Experiment 1st:

To implement the attack scenario we simulated the various attacks in NS-2 simulator. About 20 attack scenarios causing flooding of route request message on random 50 nodes with mobility speed 20ms, with varying simulation time between 60sec to 300sec, and varying number of attackers (one for DoS attack and three to seven for DDoS attack), at different rates per second (six to ten attack rate for DoS and three to five for DDoS attack) had been launched. In NS2 the connectivity is static.

After launching the attacks, the trace file (as log) is inputted to the fuzzy forensic analysis tool. The tool generates the case and attaches the hash value to the particular case. Then the proof report with particular hash value is deduced with details like attacker nodes, compromised nodes time duration of attack and rate of attack. The results are:

TABLE VI. RESULTS FOR DOS ATTACKS

Simulation time in seconds	Rate of Attack	Number of attacks with same rate	Time Duration in micro seconds	Detected Correctly
60/150	6	2	344/403	y/y
80/250	7	2	419/427	y/y
60/200	8	2	423/724	y/n
70/300	9	2	516/807	y/y
60/170	10	2	472/790	y/y

Number of attacker node: one
Number of Attack launched: Ten

TABLE VII. RESULTS FOR DDOS ATTACKS

Simulation time in seconds	Number of attacker nodes	Rate of Attack	Time Duration in micro seconds	Detected Correctly
60	3	5	448	y
150	3	4	394	y
80	4	5	406	y
250	4	4	412	y
60	5	4	386	y
200	5	3	401	y
70	6	4	256	y
300	6	3	427	n
60	7	4	229	y
170	7	3	236	y

Number of Attack launched: Ten

Fuzzy Forensic analysis tool is capable to find eighteen attack scenarios out of twenty attacks.

- Experiment 2nd:

To implement the attack scenario we simulated the various attacks using .Net technology for 50 nodes. In this the nodes as well as the attackers are randomly selected. The results are:

TABLE VIII. RESULTS FOR RANDOM ATTACKS

Random attacking scenario case	Randomly selected Attacker/ attackers	Detected Correctly
1	39	y
2	12,48,26	y
3	31,16	y
4	42	y
5	23,27,39,8	n

Fuzzy Forensic analysis tool is capable to find four attack scenarios out of five attacks.

Snap shots

VII. CONCLUSION

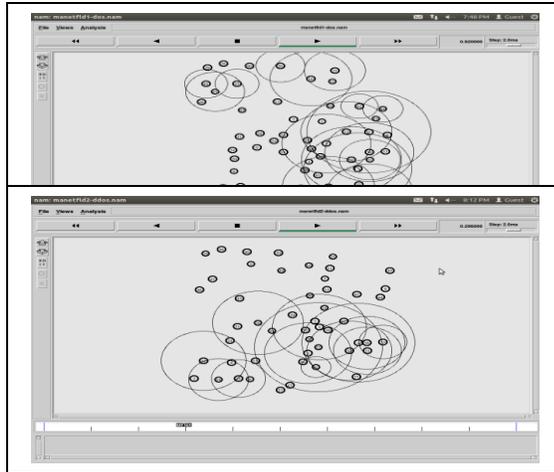
In this paper, we emphasized over the DoS/ DDoS carried due to flooding of RREQ routing packet while dynamic source routing in MANET. Flooding RREQ message per unit time without following broadcasting rule can easily implemented by attacker and for some duration attacker engage the network in unnecessary routing management leading to denial of services for some duration. There is requirement for gathering proof against attacker for this forensic analysis of attack trace evidence is needed to be done. Fuzzy forensic do this analysis using fuzzy logic.

For analysis we considered the various parameters of register request in dynamic source routing protocol at general scenario and in attacking scenario. According to these parameters the fuzzy analysis rules are generated and determined. Attack scenarios having varying simulation time and number of attackers is launched in NS2.

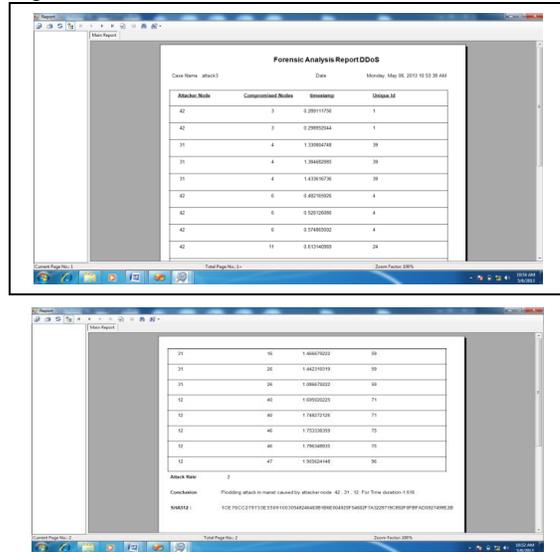
As well as in simulation for attack scenario using .Net having random attackers, random number of attackers is also implemented. Fuzzy forensic analysis tool provide protected reports with hash value with details like attacker nodes, compromised nodes time duration of attack and rate of attack. Our tool gives approximately 90% of correct detection for both DoS and DDoS RREQ flooding.

VIII. SNAPSHOTS

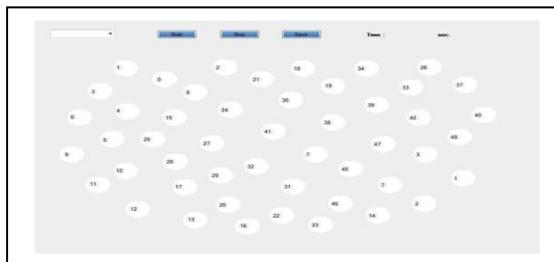
Attack simulation in NS2:



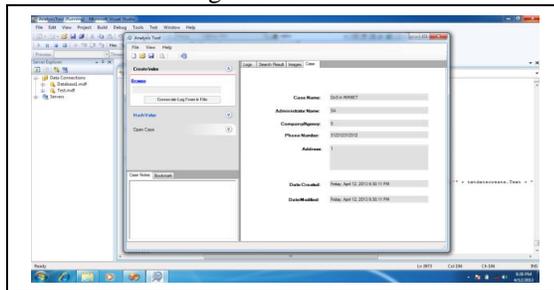
Report generation:



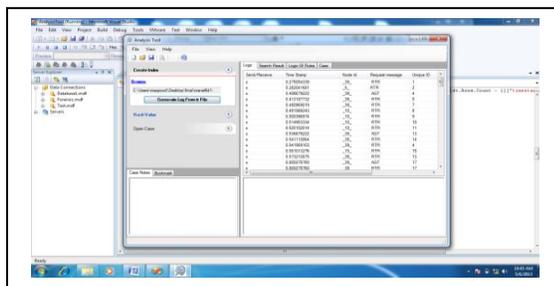
Attack simulation using .net environment:



Case with hash value generation:



Log evaluation:



REFERENCES

- [1] Yinghua Guo, Matthew Simon, "Network forensics in MANET: traffic analysis of source spoofed DoS attacks", Nov 2010 IEEE Fourth International Conference on Network and System Security.
- [2] Yinghua Guo, Matthew Simon, "Forensic analysis of DoS attack traffic in MANET", Nov 2010 IEEE Fourth International Conference on Network and System Security.
- [3] Ying Zhu, "Attack pattern discovery in forensic investigation of network attacks", IEEE journal on selected areas in communications, Vol 29, No. 7, August 2011..
- [4] Slim Rekhis and Nouredine Boudriga, "A Formal Rule-based Scheme for Digital Investigation in Wireless Ad-hoc Networks" 2009 Fourth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering.
- [5] Bing Wu, Jianmin Chen, Jie Wu, Mihaela Cardei, "A Survey on Attacks and Countermeasures in Mobile Ad Hoc Networks", Chapter 12, 2006.
- [6] Taner Tuncer Yetkin Tatar, "Detection SYN Flooding Attacks Using Fuzzy Logic", 2008 International Conference on Information Security and Assurance.
- [7] Jung-Sun Kim, Dong-Geun Kim, Bong-Nam Noh, "A Fuzzy Logic Based Expert System as a Network Forensics", July, 2004 IEEE.
- [8] Sarah Ahmed, S. M. Nirkhi, "A Fuzzy Approach for Forensic Analysis of DDoS Attack in MANET", Mar 2013, ICCSIT India.
- [9] R. Nichols and P. Lekkas, *Wireless Security-Models, Threats, and Solutions*, McGraw-Hill, Chapter 7, 2002.
- [10] Dhanant Subhadrabandhu, Saswati Sarkar, Farooq Anjum, "A Framework for Misuse Detection in Ad Hoc Networks", IEEE Journal on selected areas in communications, Vol. 24, No. 2, February 2006.
- [11] Q. Gu, P. Liu and C.H. Chu, *Tactical bandwidth exhaustion in ad hoc networks*, Proceedings of the Fifth Annual IEEE Information Assurance Workshop, PP. 257-264, 2004.
- [12] Jill Slay, Benjamin Turnbull, "The Need for Technical Approach to Digital Forensic Evidence Collection for Wireless Technologies", Proceedings of the 2006 IEEE workshop on Information Assurance United States Military Academy, NY.

- [13] Kevin P. Mc Grath and John Nelson, "A wireless Network Forensic System", ISSC June 2006, Dublin Institute of Technology.
- [14] H. Wang, D. Zang, K.G. Shin, " Change-Point Monitoring for the Detection of DoS Attacks", IEEE Transaction on Dependable and Secure Computing, vol:1 No:4, pp:193-208, 2004.
- [15] Y. Oshita, S. Ata, M. Murata, "Detecting Distrubuted Denial of Service Attacks by Analyzing TCP SYN Packets Statistically", pp:2043-2049 Globecom2004.
- [16] Jochen H. Schiller, "Mobile communication", Pearson education, chapter 8, 2008.
- [17] David B. Johnson, David A. Maltz, Josh Broch, "DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks".
- [18] Rashid Hafeez Khohar, Md Asri Ngadi , Satria Mandala,"A Review of Current Routing Attacks in Mobile Adhoc Networks", International journal of computer science and security, volume 2, No.- 3.
- [19] David Irwin and Ray Hunt, "Forensic Information Acquisition in Mobile Networks", IEEE 2009.
- [20] Shishir K. Shandilya, Sunita Sahu,"A Trust Based Security Scheme for RREQ Flooding Attack in MANET", International journal of computer applications, Vol. 5- No. 12, August 2010.

A comparative study of Image Region-Based Segmentation Algorithms

Lahouaoui LALAOU,

Laboratoire LGE département des électroniques Université de
M'sila 28000 city Ichbilila,
M'sila, Algeria

Tayeb MOHAMADI,

Départ Electronics Université Ferhat Abbas the Setif 19000
city Elmaabouda.
M'sila, Algeria

Abstract—Image segmentation has recently become an essential step in image processing as it mainly conditions the interpretation which is done afterwards. It is still difficult to justify the accuracy of a segmentation algorithm, regardless of the nature of the treated image. In this paper we perform an objective comparison of region-based segmentation techniques such as supervised and unsupervised deterministic classification, non-parametric and parametric probabilistic classification. Eight methods among the well-known and used in the scientific community have been selected and compared. The Martin's (GCE, LCE), probabilistic Rand Index (RI), Variation of Information (VI) and Boundary Displacement Error (BDE) criteria are used to evaluate the performance of these algorithms on Magnetic Resonance (MR) brain images, synthetic MR image, and synthetic images. MR brain image are composed of the gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) and others, and the synthetic MR image composed of the same for real image and the plus edema, and the tumor. Results show that segmentation is an image dependent process and that some of the evaluated methods are well suited for a better segmentation.

Keywords—Evaluation criteria; Martin's; Rand Index; Image Segmentation; Magnetic resonance image.

I. INTRODUCTION

Region-based segmentation methods are powerful tools for object detection and recognition. These methods aim at differentiating regions of interest (objects / background). Their objective is to divide the image into homogeneous zones to separate the different entities in the image. This is usually a first step in a more complex treatment chain involving pattern recognition. For example in medical imaging, segmentation is very important for representation and visualization as well as for the extraction of parameters and the analysis of images. Region based segmentation is a specific approach in which one seeks to construct surfaces by combining neighboring pixels according to a criterion of homogeneity. The nature of the considered images and the objective of the segmentation being multiple, there is no unique technique for image segmentation and segmenting an image into meaningful regions remains a real challenge [1]. According to Cocquerez et al.[2], the choice of a technique is related to the texture which is one of the important characteristics of an image. The purpose for region-based segmentation is to identify coherent regions of an image.

In order to compare the suitability of a segmentation method, we propose a comparative study between regions

based segmentation techniques. To correctly validate a result of segmentation of medical images, it is necessary to have the ground truth, which is quite difficult in this case of real images. The quality of imagery and the requirement of accurate segmentation are the crucial aspect in characterizing the performance of segmentation algorithms in brain images [3],[4]. Many image processing techniques have been proposed for brain MRI segmentation, most notably thresholding [5], region-growing [6], classifying [7], clustering[8], modelling [9], neural network based [10] and others.

As can be seen on **Error! Reference source not found.**, region based segmentation methods can be grouped into two famous families: deterministic based methods and probabilistic based classification methods. By the same way, each of these families can be subdivided into two groups. Deterministic classification family is composed of unsupervised and supervised methods. Whereas, probabilistic classification family contains parametric and non-parametric methods. In this paper, we present a comparative study of clustering based segmentation methods on synthetic and MR images. This paper is mainly devoted to study situations in which using different methods for the image segmentation. Its principal purpose is used five criteria and shows its suitability in unsupervised image segmentation. The performance of each technique is evaluated using the Martin's [11], Probabilistic Rand Index [12], Variation of Information [47] and Boundary Displacement Error criteria [53]. These measures compute the consistency degree between the regions produced by two segmentations. The evaluation of a segmentation algorithm consists in measuring the similarity between the reference algorithm and that obtained by this algorithm. The choice of an accurate measure is quite critical in order to provide a strict evaluation and reflect the real quality of an automatic segmentation with comparison to a manual one. The remainder of the paper is organized as follows: Section 2 presents the different region-based segmentation methods used for MR image analysis. The evaluation criteria are described in section 3. Section 4 describes the material and data used in this study. Experimental results on synthetic and real images are presented in section 5. Finally, a discussion concludes this paper in section 6.

II. REGION-BASED SEGMENTATION TECHNIQUES

A large number of segmentation approaches have been proposed in the literature [13, 14, 15, 16]. A good survey about their evaluation can be found in [17][18]. A list of

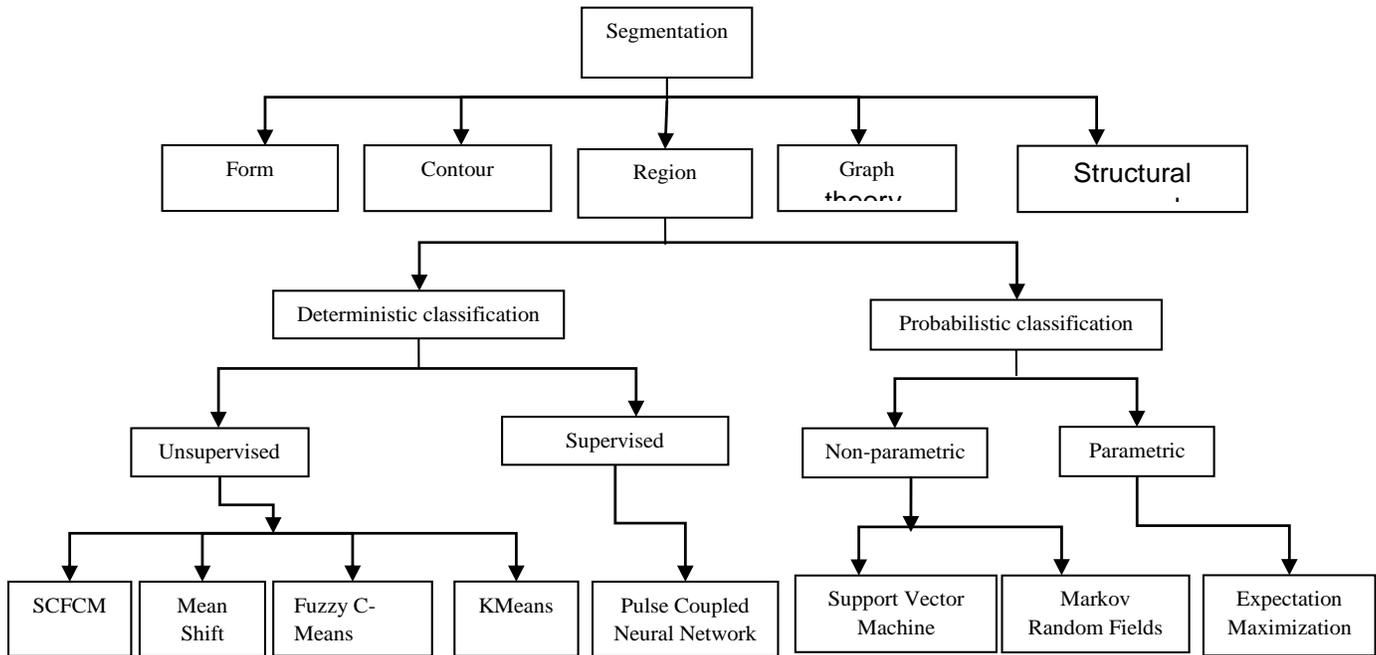


Fig. 1 Region-based segmentation methods

unsupervised, supervised, and non-parametric region based segmentation algorithms are presented in this section, such as Mean Shift (MS), Fuzzy C-Means (FCM), KMeans, Expectation Maximization (EM), Spatial Constraint Fuzzy C-Means (SCFCM), Markov Random Fields (MRF), Pulse Coupled Neural Network (PCNN), and Support Vector Machine (SVM). In the next subsections we will introduce briefly each of these techniques.

A. K-Means

K-Means algorithm is an unsupervised clustering algorithm that classifies the input data points into multiple classes based on their inherent distance from each other. The iterative K-Means clustering algorithm was first proposed by MacQueen [19]. The algorithm aims at partitioning the data set, consisting of ℓ expression patterns $\{x_1, \dots, x_\ell\}$ in an n -dimensional space, into k disjoint clusters $\{C_i\}_{i=1}^k$, such that the expression patterns in each cluster are more similar to each other than to the expression patterns in other clusters [20]. There are two popular partitioned clustering strategies: square-error and mixture modeling. The sum of the squared Euclidian distances between the samples in a cluster and the cluster center is called within-cluster variation. K-Means are widely used in many applications such as data extraction and image segmentation [21]. The K-Means method is an iterative algorithm that minimizes the sum of distances between each object and its cluster centroid.

B. Fuzzy C-Means (FCM)

Fuzzy C-Means (FCM) is an unsupervised fuzzy clustering algorithm [22]. Excerpted from the algorithm of C-means [23], it introduces the concept of fuzzy set in the definition of classes, each point in the data set belongs to each cluster with

a certain degree, and all clusters are characterized by their center of gravity. The FCM clustering algorithm was first suggested by Dunn [24] and later improved by Bezdek [25]. The FCM method proposes a fuzzy membership that assigns a degree of membership for each class by iteratively updating the cluster centers and the membership degrees for each data point. The cluster that has an associated pixel is one whose membership degree is highest. A novel approach called enhanced possibilistic Fuzzy C-Means clustering is proposed for segmenting MR brain image into different tissue types on both normal and tumor affected pathological brain images. FCM methods has been proposed for the segmentation of MR Images [26,27] and for the segmentation of major tissues in [28,29] and possible tumor on T1-weighted volumes. The FCM is often used in medical image segmentation [30, 31]. Chen et al. [32], have proposed an algorithm based on FCM for the correction of intensity in homogeneity and for segmentation of MRI images.

C. Fuzzy C-Means algorithm with Spatial Constraint (SCFCM)

Fuzzy C-Means algorithm with Spatial Constraint (SCFCM) is based on the clustering algorithm FCM described above, two kinds of information in image are used, the gray value, and space distributed structure. Based on the relevance of nearby pixels, the neighbors in the set should be similar in feature value. Its effectiveness contributes not only to introduction of fuzziness for belongingness of each pixel but also to exploitation of spatial contextual information. SCFCM clustering algorithm preserves the homogeneity of the regions better than existing FCM techniques, which often have difficulties when tissues have overlapping intensity. In order to reduce the noise effect during segmentation, the proposed

method incorporates both the local spatial context and the non-local information into the standard FCM cluster algorithm using a novel dissimilarity index in place of the usual metric distance.

D. Expectation Maximization (EM)

Expectation Maximization (EM) is one of the most common algorithms used for density estimation of data points in an unsupervised setting. The EM algorithm [33] is used to estimate the parameters of this model; the resulting pixel-cluster memberships provide a segmentation of the image. The EM algorithm can be considered as a variant of the K-Means algorithm where the membership of any given point to the clusters is not complete and can be fractional. An EM algorithm was proposed in [34] to model the homogeneities as a bias field of the image logarithm. This algorithm has been applied for the segmentation of brain MR image [35]. According to [36] the EM algorithm has demonstrated greater sensitivity to initialization than the K-Means or FCM algorithms.

E. Mean Shift (MS)

The Mean Shift (MS) [37] algorithm clusters an n -dimensional data set by associating each point with a peak of the data set's probability density. For each point, Mean Shift computes its associated peak by first defining a spherical window at the data point of radius r and computing the mean of the points that lie within the window. At each iteration the window will shift to a more densely populated portion of the data set until a peak is reached, where the data is equally distributed in the window. MS was successfully applied by Mayer et al. [38] in clustering, segmentation and filtering of natural resources in 2D images [39], using a paradigm adaptively to segment the brain MR images.

F. Markov Random Field (MRF)

The Markov Random Field (MRF) models are used for the restoration and segmentation of digital images. They can make up for deficiencies in observed information by adding a-priori knowledge to the image interpretation process in the form of models of spatial interaction between neighboring pixels. Hence, the classification of a particular pixel is based, not only on the intensity of that pixel, but also on the classification of neighboring pixels. The goal of segmentation is to estimate the correct label for each site. The segmentation is obtained by classifying the pixels into different pixel classes. These classes are represented by multivariate Gaussian distributions. A most of reference are cited, It can be viewed as a particular model selection problem, and different techniques have been proposed in the classical HMF case [40]. It has been used for brain image segmentation by modeling probabilistic distribution of the labeling of a voxel jointly with the consideration of the labels of a neighborhood of the voxel [41].

G. Support vector machine (SVM)

The Support Vector Machine (SVM) is a learning machine for two-group classification problems.

The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-

dimension feature space. SVM is a set of supervised learning techniques for solving problems of discrimination, regression and are particularly adapted to data process at very high dimensions [42]. The algorithm of the SVM is described as follows:

First specifies a small set of training pixels, such as a small part of an object and a small part of the background, as the clues. Then, fast SVM is applied to train the classifiers based on the training pixels. Finally, the remaining image, which is viewed as the test set, is subdivided into several regions by the classifier. A comparison between a segmentation method with SVM and FCM is applied in [43].

H. The Pulse-Coupled Neural Network (PCNN)

The Pulse-Coupled Neural Nets (PCNN) is a two-dimensional non-training neural network in which each neuron in the network corresponds to one pixel in an input image. The neuron receives its input as an external stimulus. These stimuli are combined in an internal activation system, and are accumulated until they exceed a dynamic threshold. This will result in a pulse output and through an iterative process. The algorithm produces a temporal series of binary images as outputs algorithm is based on the neurophysiologic models evolving from studies of small mammals. Depending on time as well as on the parameters, this dynamic output contains information, which makes it possible to detect edges, do segmentation, identify textures and perform other feature extractions. For the PCNN, the neurons associated with each group of spatially connected pixels with similar intensities tend to pulse together [44]. This is the basic principle of segmentation of the PCNN. In fact, there are many approaches for image segmentation with the PCNN. Generally, all the methods of segmentation can be classified into two kinds of schemes: common image segmentation and automatic image segmentation.

III. EVALUATION CRITERIA

The goal of this study is to perform a quantitative comparison between automatic segmentation and a set of ground truth segmentation (reference). We use the same methodology reported in, and an evaluation metric for image segmentation of multiple objects [45], where a quantitative predictive performance evaluation used full reference image quality assessment metrics has been conducted. In this section we present five criteria, the Probabilistic Rand Index, Global Consistency Error, Local Consistency Error, Boundary Displacement Error and Variation of Information.

A. The Probabilistic Rand Index (PRI)

In literature there are many criteria of nonparametric measures such as: Jaccard's index, Fowlkes, and Mallow's index [46], he is work by counting pairs of pixels that have compatible label relationships between the two segmentations to be compared. We consider two images reference and segmented respectively S_1 and S_2 of N points $X = \{x_1, x_2, x_3, x_4, \dots, x_N\}$; that assigned labels $\{b_i\}$ and $\{b'_i\}$ respectively to point x_i . The Rand Index can be computed as the ratio of the number of pairs of vertices or faces having the compatible label relationship in S_1 and S_2 . Can be defined as:

$$R(S_1, S_2) = \frac{1}{(2^N)} \sum_{\substack{i,j \\ i \neq j}} [I(l_i = l_j \wedge l'_i = l'_j) + I(l_i \neq l_j \wedge l'_i \neq l'_j)] \quad (1)$$

Where I is the identity function, and the denominator is the number of possible unique pairs among N data points. This gives a measure of similarity ranging from 1, when the two images reference and segmented respectively are identical, to 0 other wise. We first outline a generalization to the Rand Index, termed the Probabilistic Rand (PR) index, which we previously introduced in [47] The PR index allows comparison of test segmentation with multiple ground-truth images through soft non uniform weighting of pixel pairs as a function of the variability in the ground-truth set. The Rand index [47] counts the fraction of pairs of pixels whose labeling are consistent between the computed segmentation and the ground truth. This quantitative measure is easily extended to the probabilistic Rand index (PRI) [48] by averaging the result across all human segmentations of a given image. Consider a set of manually segmented (ground truth) images $\{S_1, S_2, \dots, S_K\}$ corresponding to an image $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$, where a subscript indexes one of N pixels. Let S test be the segmentation that is to be compared with the manually labeled set.

B. Martin Evaluation Criteria

Martin et al.[49] proposed two error measures to quantify the consistency between image segmentations of differing granularities, and used them to compare the results of algorithms to a database of manually segmented images. The Martin's similarity index which outperforms the others in terms of properties and discriminative power is employed for performance evaluation to compare the different region-based segmentation methods. The role of the test is to assess the quality of segmentation by transforming the measurements into a mathematical function called test. These criteria may be a test of homogeneity of a set of points of similarity, or any statistical test. Martin et al. [50] proposed an interesting error measure, which takes 2 images S_1 and S_2 as input, and produces a real-valued output in the range $[0, 1]$, the Martin's distance where 0 signifies no error and 1 worst segmentation, which the inverse for similarity 1 signifies no error and 0 worst segmentation. The measure is shown to be effective for qualitative similarity comparison between segmentations by humans, who often produce results with varying degrees of perceived details, which are all intuitively reasonable and therefore "correct". On the other hand, the Martin error measure is sensitive to qualitatively different segmentations. A segmentation error measure takes two segmentations S_1 and S_2 as input, and produces a real valued output. For a given pixel p_i consider the segments in S_1 and S_2 that contain that pixel. The segments are sets of pixels. If one segment is a proper subset of the other, then the pixel lies in area of refinement and the local error should be zero. If there is no subset relationship, then the two regions overlap in an inconsistent manner. In this case, the local error should be non-zero. If $R(S, p_i)$ is the set of pixels corresponding to the region in segmentation S which is the region that contains pixels p_i , the local refinement error, E , is defined as:

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) / R(S_2, p_i)|}{|R(S_1, p_i)|} \quad (2)$$

Note that this local error measure is not symmetric. It encodes a measure of refinement in one direction only: $E(S_1, S_2, p_i)$ is zero precisely when S_1 is a refinement of S_2 at pixel p_i , but not vice versa. There are two natural ways to combine the values into a measure of the error for the entire image. Global Consistency Error (GCE) forces all local refinements to be in the same direction. Local Consistency Error (LCE) allows refinement in different directions and in different parts of the image. Let n be the number of pixels:

$$GCE(S_1, S_2) = \frac{1}{n} \min\{ \sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i) \} \quad (3)$$

$$LCE(S_1, S_2) = \frac{1}{n} \sum_i \min\{E(S_1, S_2, p_i), E(S_2, S_1, p_i)\} \quad (4)$$

Although these error metrics are calculated by grouping pixels into objects first, they unfortunately tolerate over-segmentation and under-segmentation, as a consequence of their intended purpose for comparing human segmentations. As $LCE \leq GCE$, it is clear that GCE is a tougher measure than LCE.

C. Boundary matching (Boundary Displacement Error)

Several measures work by matching boundaries between the segmentations, and computing some summary statistic of match quality. The Boundary Displacement Error (BDE) measures the average displacement error of one boundary pixels and the closest boundary pixels in the other segmentation [48]. Work in [51] proposed solving an approximation to a bipartite graph matching problem for matching segmentation boundaries, computing the percentage of matched edge elements, and using the harmonic mean of precision and recall, termed the F-measure as the statistic. Furthermore, for a given matching of edge elements between two images, it is possible to change the locations of the unmatched edges almost arbitrarily and retain the same precision and recall score.

D. Information-based (Variation of Information)

The proposed metric measure is termed the variation of information (VI) and is related to the conditional entropies between the class label distributions of the segmentations. Work in [52] computes a measure of information content in each of the segmentations and how much information one segmentation gives about the other. Several measures work by counting the number of false-positives and false-negatives [53] and similarly assume existence of only one ground truth segmentation. Due to the lack of spatial knowledge in the measure, the label assignments to pixels may be permuted in a combinatorial number of ways to maintain the same proportion of labels and keep the score unchanged.

IV. MATERIEL

A. Data synthetic MR image

A large number of segmentation approaches have been proposed in the literature [54, 55, 56, 57]. A good survey about their evaluation can be found in [58][59] A list of unsupervised, supervised, and non-parametric region based

segmentation algorithms are presented in this section, such as Mean Shift (MS), Fuzzy C-Means (FCM), KMeans, Expectation Maximization (EM), Spatial Constraint Fuzzy C-Means (SCFCM), Markov Random Fields (MRF), Pulse Coupled Neural Network (PCNN), and Support Vector Machine (SVM). The image is constitute with: White matter(WM), gray matter (GM), cerebrospinal fluid(CSF), edema, tumor for the synthetic MR image are represented as a set of spacial probability maps for tissue and pathology shows in Figure.2. In our laboratory, from these different matters we have created the ground truth for each image.

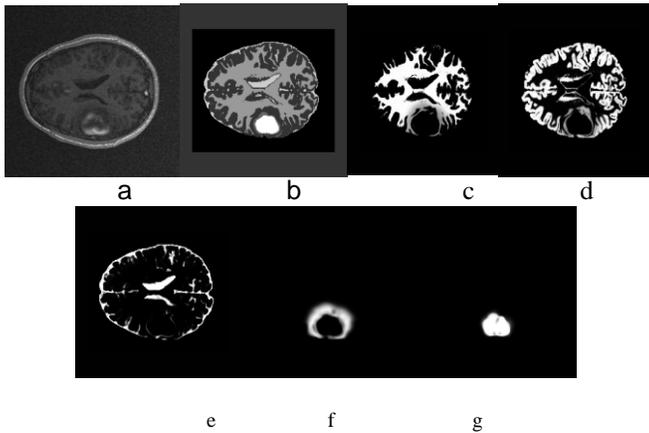


Fig. 2 From left to right : (a) image T1, (b) ground truth, (c) white matter, (d) gray matter, (e) CSF, (f) edema, (g) tumor.

In this section we compare the results of segmentation methods on synthetic MR image. In Fig 2 present an example of synthetic MR image with the display the images of different segmentation methods. And in the Table 2 below shows the values of evaluation criteria. These examples we allow to understand how to meet these criteria have different images segmentation. The values obtained in Table 2 shows the comparison between the automatic image segmentation and the ground truth image for synthetic MR images. The results of average and variance we applied for 25 synthetic MR image the results are given in following are varies between 0.3245 ± 0.0012 and 0.5021 ± 0.0013 for GCE criterion, the value between 0.124 ± 0.0034 and 0.3585 ± 0.0070 for LCE, the values between 0.4069 ± 0.0058 and 0.5912 ± 0.0067 for PRI, the values between 1.5021 ± 0.5871 and 5.2314 ± 1.2341 for VI, and 92.8908 ± 22.5487 and 3.7077 ± 0.6532 for BDE criterion.

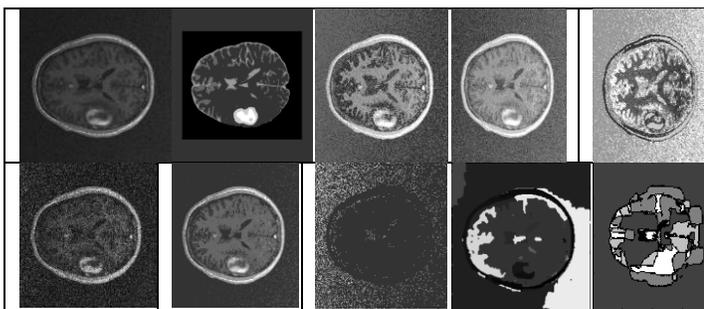


Fig. 3 Representative examples of results obtained with the different segmentation methods, (a) original image, (b) ground truth image, (c) FCM, (d) K-Means, (e) SCFCM, (f) MS, (g) EM, (h) MRF, (i) PCNN (j) SVM.

B. Real data

In this section, images are obtained from the IBSR (Internet Brain Segmentation Repository) database [60]. As described on the IBSR, the database is composed of three-dimensional coronal brain Magnetic Resonance Images (MRI). The coronal three-dimensional T1-weighted spoiled gradient echo MRI scans were performed on 2 different imaging systems.

The MR Brain data sets and their manual segmentations were provided by the center of morphometric analysis at Massachusetts general hospital and are available at IBSR. The voxels contain images segmented by experts for each sub-databases are the ground truth voxels. These databases are used by many that users' all around the world. It supplies brain MR images as well as the segmentation results that are performed by the trained experts in a manually guided manner. **Error! Reference source not found.**3 shows different images from the IBSR database. For our experiment, we used 25 test images from the IBSR database and the corresponding ground truth (segmented by the expert) to each image.

The different based segmentation methods are applied on each image and the Martin's criteria are used to evaluate the performance of each algorithm.

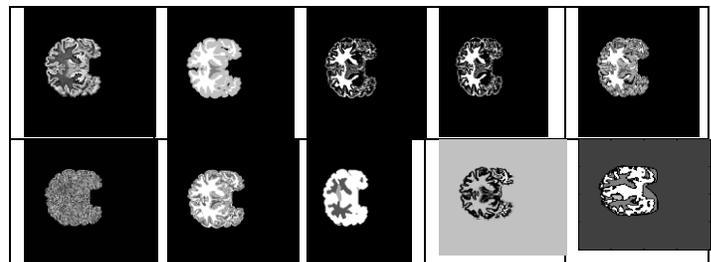


Fig. 4 Representative examples of results obtained with the different segmentation methods, (a) original image, (b) ground truth image, (c) FCM, (d) K-Means, (e) SCFCM, (f) MS, (g) EM, (h) MRF, (i) PCNN (j) SVM.

The analysis of the results of Fig. 4 demonstrates that some of the used algorithms generate as many classes as those generated by the laboratory. These findings are confirmed by the criteria reported in Table 3.

Comparison of the eight segmentation algorithms using LCE and GCE errors (mean) in the case of real images in these paper 25 images. The MS method performs better than the FCM, followed respectively by SVM, SCFCM, EM, K-Means, MRF, and PCNN. Accordingly, we compare the segmentation performance in brain tissue. To say that actual results are consistent with the results obtained on synthetic images.

For a quick interpretation of the results, Fig. 4 report the evolution of the martin's criteria. The best criterion values are obtained for the EM method (GCE criterion = 0.9268, LCE criterion = 0.9047, PRI=0.9724, VI = 0.4935, and BDE= 3.245) in average.

TABLE I. Averages and STD of: GCE, LCE, PRI, VI, and BDE mean values of the synthetic MR Dataset for the different segmentation methods.

criteria		FCM	K-Means	SCFCM	MS	EM	MRF	PCNN	SVM
GCE	Average	0.4152	0.4090	0.3795	0.3397	0.5021	0.4040	0.4021	0.3245
	STD	0.0052	0.0049	0.0050	0.0064	0.0013	0.0062	0.0063	0.0012
LCE	Average	0.3569	0.3443	0.1941	0.3149	0.3585	0.1284	0.3054	0.124
	STD	0.0041	0.0071	0.0564	0.0083	0.0070	0.0017	0.0064	0.0034
PRI	Average	0.5882	0.5826	0.5912	0.5552	0.5771	0.4165	0.4069	0.502
	STD	0.0077	0.0066	0.0067	0.045	0.0088	0.0062	0.0058	0.0084
VI	Average	3.1525	3.0997	3.2087	3.5489	2.3710	1.5021	1.5845	5.2314
	STD	0.9154	0.885	0.456	0.658	0.451	0.5871	0.1854	1.2341
BDE	Average	4.6365	4.7548	4.5193	4.5782	3.7077	92.8908	53.3238	24.2145
	STD	0.9124	0.7245	0.8546	0.8546	0.6532	22.5487	16.2354	5.123
criteria		FCM	K-Means	SCFCM	MS	EM	MRF	PCNN	SVM
GCE	Average	0.4152	0.4090	0.3795	0.3397	0.5021	0.4040	0.4021	0.3245
	STD	0.0052	0.0049	0.0050	0.0064	0.0013	0.0062	0.0063	0.0012
LCE	Average	0.3569	0.3443	0.1941	0.3149	0.3585	0.1284	0.3054	0.124
	STD	0.0041	0.0071	0.0564	0.0083	0.0070	0.0017	0.0064	0.0034
PRI	Average	0.5882	0.5826	0.5912	0.5552	0.5771	0.4165	0.4069	0.502
	STD	0.0077	0.0066	0.0067	0.045	0.0088	0.0062	0.0058	0.0084
VI	Average	3.1525	3.0997	3.2087	3.5489	2.3710	1.5021	1.5845	5.2314
	STD	0.9154	0.885	0.456	0.658	0.451	0.5871	0.1854	1.2341
BDE	Average	4.6365	4.7548	4.5193	4.5782	3.7077	92.8908	53.3238	24.2145
	STD	0.9124	0.7245	0.8546	0.8546	0.6532	22.5487	16.2354	5.123

TABLE II. Averages and STD of: GCE, LCE, PRI, VI, and BDE mean values of the dataset for the different segmentation methods.

criteria		FCM	KMeans	SCFCM	MS	EM	MRF	PCNN	SVM
GCE	Mean	0.9161	0.9169	0.7922	0.9180	0.9268	0.9028	0.8480	0.7856
	STD	0.0275	0.0285	0.1604	0.0317	0.0263	0.0210	0.0491	0.0210
LCE	Means	0.0829	0.8872	0.2308	0.8893	0.9047	0.8872	0.8893	0.7544
	STD	0.0117	0.0203	0.3141	0.0141	0.0154	0.0125	0.0141	0.0107
PRI	Mean	0.9315	0.9301	0.9221	0.6815	0.9724	0.9768	0.84786	0.8324
	STD	0.0164	0.0172	0.0083	0.0525	0.0093	0.0075	0.0248	0.0650
VI	Mean	0.6637	0.6668	0.7014	0.8679	0.4935	0.5478	0.7581	0.6847
	STD	0.1255	0.1280	0.0465	0.1141	0.1096	0.1245	0.0145	0.1128
BDE	Mean	0.6806	0.6731	0.7000	82.2160	3.245	8.5471	0.624	18.7584
	STD	0.0924	0.0943	0.0436	4.0190	0.8096	1.1211	0.0074	8.0874

Table 3 shows the output of the criteria; interval with a lower limit greater than 0 and high limited at 1, the values implies that the adaptive EM performs significantly better in segmentation than benchmark (the FCM, K-Means, SCFCM, MS, MRF, PCN or SVM). The GCE, LCE, and RI values of the EM method in Fig 12, for 25 brain images, which demonstrate the robustness of the method EM.

TABLE III. CPU time by different algorithms in Fig.11 in the same order.

	Image 256×256 pixels 524288 bytes	Image 180×180 pixels 259200 bytes	Image 172×158 pixels 217408 bytes
FCM (CPU time(s))	12.11	5.76	4.09
K-Means (CPU time(s))	2.27	1.82	1.01
SCFCM (CPU time(s))	23.41	10.45	7.46
MS (CPU time(s))	0.35	0.46	0.39
EM (CPU time(s))	20.86	9.08	7.38
MRF (CPU time(s))	1470.88	673.17	564.15
PCNN (CPU time(s))	10.76	6.09	6.05
SVM (CPU time(s))	18.12	8.21	7.25

4.5. Computational time

The processing time for segmenting images is presented in Table 4. We list the CPU time in segmenting images in Fig.10. It can be seen from Table 4 that the processing time for MRF are both higher than the other algorithms.

V. DISCUSSION AND CONCLUSION

This paper presents an objective comparison of region-based segmentation methods. Our study focuses on supervised and unsupervised deterministic classification, non-parametric and parametric methods probabilistic classification. Among the well-known and used techniques in the scientific community, we have selected eight techniques. These methods have been used on two different databases. The first composed synthetic MR images are available for download at www.ucinia.org, and the second composed of brain MR images from the IBSR database. For comparison, a ground truth is created in our laboratory for synthetic MR images and by an expert for IBSR database. To compare the different region based segmentation methods, we used the Martin's similarity indexes and Probabilistic Rand Index. Five criteria have been used: The global consistency error, the local consistency error, Probabilistic Rand Index, Variation of Information, and Boundary Displacement Error. At each time, the result of these criteria is the difference between the automatic segmentation and the ground truth. In this paper, we compared the performance of different region-based segmentation algorithms. Results show that the EM is outperforms the other seven algorithms in the three different dataset images. The analysis of the results of the five criteria demonstrate that except the EM, K-Means, SCFCM and the FCM algorithm, all the methods that we have tested perform well for the segmentation of images such those considered in this paper. Nevertheless, we are going to group them in two classes. The first class contained SCFCM, K-Means, FCM, and EM, the latter algorithm has a best performance with GCE = 0.6935, LGE = 0.4113 and PRI=0.8245 for the

synthetic data, GCE = 0.5021, LGE = 0.3585 and PRI=0.6067 for the synthetic MR data, and with GCE = 0.9268, LGE = 0.9047, and PRI=0.6067 for the IBSR data. This is consistent with what have been reported on the robustness of the MS algorithm for feature extraction and image segmentation. The MS algorithm is an unsupervised clustering-based segmentation method and needs no a priori information on the number and the shape of the data cluster. The FCM method takes advantage of local textual information and high inter-pixel correlation inherent. The second class, with a worst quality scores for the criteria groups decreasing: MRF, MS, PCNN, and SVM methods. The very high value of the five criteria for the EM method is due to known fixed segmentation parameters of the EM method estimated by optimizing the likelihood. The optimized requires no 'step size' parameters and will not oscillate around the optimum. However, there is no guarantee of global solutions. These results might be due to initialization the parameter for each algorithm. Last but not least, according to **Error! Reference source not found.** which reports the least values obtained for the GCE, LCE, PRI, VI, and BDE on the synthetic data it is shown that the demonstrated EM method is well adapted for any type of images synthetic MR, and MR images. In the second, by the FCM, K-Means, and SCFCM methods almost the same values of five criteria in different type of the three datasets. In this paper, the adaptive EM is outperforms the other seven algorithms in three dataset (synthetic MR images, and MR images).As a prospect to this study, we are actively working on 3D segmentation methods. In progress as well, a study to compare criteria for evaluation of the image segmentation methods.

References

- [1] G. R. Dattatreya, "Unsupervised context estimation in a mesh of pattern classes for image recognition", Pattern Recognition, Vol. 24(7), pp. 685-694, 1991.
- [2] Cocquerez et al, Analyse d'images : filtrage et segmentation, Ed Masson, 1995.
- [3] Herng-Hua Chang, Performance measure characterization for evaluating neuro image segmentation algorithms, NeuroImage 47 (2009) 122-135
- [4] Jifeng Ning et al. Interactive image segmentation by maximal similarity based region merging, Pattern Recognition 43 (2010) 445 - 456.
- [5] Salem Saleh Al-amri, N.V. Kalyankar and Khamitkar S.D , " Image Segmentation by using Thershod Techniques", Journal of Computing, vol2, pp 83-86,2010
- [6] R. Pohle and K. D. Toennies, "Segmentation of medical images using adaptive region growing," Proc. SPIE— Med. Imag., vol. 4322, pp. 1337-1346, 2001.
- [7] R.J. Schalkoff. Pattern recognition: statistical, structural and neural approaches. John Wiley and Sons, 1992.
- [8] S. Shen, W. A. Sandham, M. H. Grant, J. Patterson, and M. F. Dempsey, "Fuzzy clustering based applications to medical image processing," in Proc. IEEE EMBS 25th Annu. Int. Conf., 2003, pp. 747-750.
- [9] S.Z. Li. Markov random field modeling in computer vision. Springer, 1995.
- [10] L.O. Hall, A.M. Bensaid, L.P. Clarke, R.P. Velthuizen, M.S. Silbiger, and J.C. Bezdek. A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. IEEE T. Neural Networks, 3:672-682, 1992.
- [11] Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating algorithms and measuring ecological statistics, ICCV (2001) 416 - 423.
- [12] R. Unnikrishanan M. Hebert, "measures of similarity," proc. IEEE workshop computer vision applications, 2005.

- [13] yongxia et all, image segmentation by clustering of special patterns, pattern recognition letters, 28(2007) 1548-1555.
- [14] Kannan, S.R., 2008. A New segmentation system for Brain MR images based on fuzzy techniques. *J. Appl. Soft Comput.* 8(4), 1599-1606.
- [15] Zenaty, E.A., Aljahdali, S., Debnath, N., 2009. A Karnalized Fuzzy C-Means algorithm for automatic MRI segmentation. *Comput. Methods Sci.Eng.* 9, 123-136,doi 10.3233/JCM-2009-0241.
- [16] Chuang, K.,- S., et al., 2006.fuzzy C-Means clustering with spatial information for image segmentation. *Comput.Med. Imaging Graph.*3,9-15.
- [17] H. Zhang, J. E. Fritts, Image segmentation evaluation: A survey of unsupervised methods, *computer vision and image understanding* 110 (2008) 260-280.
- [18] S.Rian, C. Jqggim J. Xue, 'Brain Tissue classification of magnetic Resonance images using partial volume modeling', *IEEE Trans. On medical imaging*, vol.19, No.2, pp.1179-1187, 2000.
- [19] S. K. Khan and A. Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25:1293–1302, 2004.
- [20] M. Mignotte. Segmentation by fusion of histogram-based k-means clusters in different color spaces. *IEEE Transactions on Image Processing*, 17(5):780–787, 2008.
- [21] Zhang, Y. J. (2002a). "Image engineering and related publications" *International Journal of Image and Graphics*, (2002) 2(3), 441-452.
- [22] Pham D.L and Prince J.L. An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities. *Pattern Recognition Letters*. vol. 20, no.1, pp.57-68, 1999
- [23] Richard J. Hathaway, John W. Davenport and James C. Bezdek 'Relational duals of the c-means clustering algorithms' *Pattern Recognition*, Volume 22, Issue 2, 1989, Pages 205-212.
- [24] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57.
- [25] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. New York: Springer, 1999.
- [26] Claramunt CF., Sede M.H., Prelaz-Droux R., Vidale L., « Sémantique et logique spatio-temporelles ». *Revue internationale de géomatique*, volume 4, pp. 165-180, 1994.
- [27] R.L. Cannon, J.V. Dave, J.C. Bezdek. – Efficient implementation of the fuzzy c-mean clustering algorithms. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 8, n2, 1986, pp. 248–255.
- [28] Songcan Chen and Daoqiang Zhang. Robust Image Segmentation Using FCM With Spatial Constraints Based on New Kernel-Induced Distance Measure. *IEEE transactions on systems, man, and cybernetics—part b: cybernetics*, vol. 34, no. 4, august, 2004.
- [29] S. Ruan, B. Moretti, J. Fadili, and D. Bloyet. Fuzzy Markovian Segmentation in Application of Magnetic Resonance Images. *Computer Vision and Image Understanding*, 85:54–69, 2002.
- [30] D. Pham & J. L. Prince. Adaptive Fuzzy Segmentation of Magnetic Resonance Images. *IEEE Transactions on Medical Imaging*, vol. 18, no. 9, pages 737–752, September 1999.
- [31] S. Shen, W. Sandham, M. Granat & A. Sterr. MRI Fuzzy Segmentation of Brain Tissue Using Neighborhood Attraction With Neural-Network Optimization. *IEEE Transactions on Information Technology In Biomedicine*, vol. 9, no. 3, pages 459–467, September 2005.
- [32] Weijie Chen, Maryellen L. Giger. – A fuzzy c-means (FCM) based algorithm for intensity in homogeneity correction and segmentation of MR images. In : *Proceedings of IEEE International Symposium on Biomedical Imaging*, pp. 1307–1310, Arlington, VA, USA, 2004.
- [33] A. Dempster, N. Laird, and D. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Royal Statistical Soc., Ser. B*, vol. 39, no. 1, pp. 1-38, 1977.
- [34] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, pp. 429–442, Aug. 1996.
- [35] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 897–908, Oct. 1999.
- [36] Z.Yu, O.C, Au,R.Zou, W. Yu, J. Tian, an adaptative unsupervised approach toward pixel clustering and color image segmentation. *Pattern recognition*. 43(5)(2010)1889-1906.
- [37] D. Comaniciu and P.Meer,"Mean Shift: A robust approach toward feature space analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, May 2002, pp. 603-619
- [38] A. Mayer et H. Green span: Segmentation of brain MRI by adaptive mean shift. *International Symposium on Biomedical Imaging: Macro to Nano*, pages 319–322, avril 2006.
- [39] Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on* 24(5) (May 2002) 603–619
- [40] F. Forbes, N. Peyrard, Hidden Markov random field model selection criteria based on mean field-like approximations, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (9) (2003) 1089–1101.
- [41] Ali, A.A., Dale, A.M., Badaea, A, Johnson, G.A., 2005. Automated segmentation of neuroanatomical structures in multispectral MR microscopy of the mouse brain, *NeuroImage* 27 (2), 425–435.
- [42] V. N. Vapnik, *Statistical Learning Theory*, New York, wiley, édition, 1998.
- [43] C. Bielski and P. Soille. Order independent image compositing. *Lecture Notes in Computer Science*, 3617:1076-1083, September 2005.
- [44] Murali Murugavel, John M, Sullivan, Jr, 'Automatic cropping of MRI rat brain volumes using pulse coupled neural networks', *NeuroImage* 45 (2009) 845–854.
- [45] Marak Polak, Hong Zhang, An evaluation metric for image segmentation of multiple objects, *Image and Vision Computing. Image and Vision Computing*. 27 (2009) 1223–1227.
- [46] Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clustering. *J. Am. Stat. Assoc.* 78(383), 553-569 (1983).
- [47] R. Unnikrishnan, C. Pantofaru, and M. Hebert, , A Measure of objective evaluation of image segmentation algorithms," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop on Empirical Evaluation Methods in Computer Vision*, Jun. 2005, vol. 3, pp. 34–41.
- [48] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–944, Jun. 2007.
- [49] D. Martin, "An Empirical approach to grouping and segmentation", Ph.D.dissertation,2002, University of California, Berkeley.
- [50] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating algorithms and measuring ecological statistics, *ICCV* (2001) 416–423.
- [51] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cuff, "Yet Another Survey on Image Segmentation: Region and Boundary Information Integration," *Proc. European Conf. Computer Vision*, pp. 408-422, 2002.
- [52] M. Meila, "Comparing Clusterings: An Axiomatic View," *Proc. 22nd Int'l Conf. Machine Learning*, pp. 577- 584, 2005
- [53] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A Measure for Objective Evaluation of Image Segmentation Algorithms," *Proc. CVPR Workshop Empirical Evaluation Methods in Computer Vision*, 2005.
- [54] L. Szilagyi, Z. Benyo, S.M. Szilagyi, et al "MR Brain Image Segmentation Using an Enhanced Fuzzy C-Means Algorithm ", *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, Cancun, Mexico, Sep.17-21, 2003:724-726.
- [55] R. Moller., R. Zeipelt. "Automatic segmentation of 3D-MRI data using a genetic algorithm, *Medical Imaging and Augmented Reality*", 2001. *Proceedings. International Workshop on*, 10-12 June 2001:278 – 281.
- [56] G. Gerig, J. Martin, R. Kikinis, O. Kubler, M. Shenton, F.A. Jolesz, Unsupervised tissue-type segmentation of 3-D dual-echo MR head data, *Image and Vision Computing* 10 (1992) 349.

- [57] M. Bomans, K.-H. Höhne, U. Tiede, M. Riemer, 3-D segmentation of MR images of the head for 3-D display, *IEEE Transactions on Medical Imaging* 9 (1990) 177.
- [58] H. Zhang, J. E. Fritts, Image segmentation evaluation: A survey of unsupervised methods, *computer vision and image understanding* 110, (2008) 260-280.
- [59] S.Rian, C. Joggim J. Xue, 'Brain Tissue classification of magnetic Resonance images using partial volume modeling', *IEEE Trans. On medical imaging*, vol.19, No.2, pp.1179-1187, 2000.
- [60] Database of image, www.cma.mgh.harvard.edu/ibsr.

Automated Classification of L/R Hand Movement EEG Signals using Advanced Feature Extraction and Machine Learning

Mohammad H. Alomari, Aya Samaha, and Khaled AlKamha
Applied Science University
Amman, Jordan

Abstract— In this paper, we propose an automated computer platform for the purpose of classifying Electroencephalography (EEG) signals associated with left and right hand movements using a hybrid system that uses advanced feature extraction techniques and machine learning algorithms. It is known that EEG represents the brain activity by the electrical voltage fluctuations along the scalp, and Brain-Computer Interface (BCI) is a device that enables the use of the brain's neural activity to communicate with others or to control machines, artificial limbs, or robots without direct physical movements. In our research work, we aspired to find the best feature extraction method that enables the differentiation between left and right executed fist movements through various classification algorithms. The EEG dataset used in this research was created and contributed to PhysioNet by the developers of the BCI2000 instrumentation system. Data was preprocessed using the EEGLAB MATLAB toolbox and artifacts removal was done using AAR. Data was epoched on the basis of Event-Related (De) Synchronization (ERD/ERS) and movement-related cortical potentials (MRCP) features. Mu/beta rhythms were isolated for the ERD/ERS analysis and delta rhythms were isolated for the MRCP analysis. The Independent Component Analysis (ICA) spatial filter was applied on related channels for noise reduction and isolation of both artifactually and neutrally generated EEG sources. The final feature vector included the ERD, ERS, and MRCP features in addition to the mean, power and energy of the activations of the resulting Independent Components (ICs) of the epoched feature datasets. The datasets were inputted into two machine-learning algorithms: Neural Networks (NNs) and Support Vector Machines (SVMs). Intensive experiments were carried out and optimum classification performances of 89.8 and 97.1 were obtained using NN and SVM, respectively. This research shows that this method of feature extraction holds some promise for the classification of various pairs of motor movements, which can be used in a BCI context to mentally control a computer or machine.

Keywords—EEG; BCI; ICA; MRCP; ERD/ERS; machine learning; NN; SVM

I. INTRODUCTION

The importance of understanding brain waves is increasing with the ongoing growth in the Brain-Computer Interface (BCI) field, and as computerized systems are becoming one of the main tools for making people's lives easier, BCI or Brain-Machine Interface (BMI) has become an attractive field of research and applications, BCI is a device that enables the use of the brain's neural activity to communicate with others or to

control machines, artificial limbs, or robots without direct physical movements [1-4].

The term "Electroencephalography" (EEG) is the process of measuring the brain's neural activity as electrical voltage fluctuations along the scalp that results from the current flows in brain's neurons [5]. In a typical EEG test, electrodes are fixed on the scalp to monitor and record the brain's electrical activity [6]. BCI measures EEG signals associated with the user's activity then applies different signal processing algorithms for the purpose of translating the recorded signals into control commands for different applications [7].

The most important application for BCI is helping disabled individuals by offering a new way of communication with the external environment [8]. Many BCI applications were described in [9] including controlling devices like video games and personal computers using thoughts translation. BCI is a highly interdisciplinary research topic that combines medicine, neurology, psychology, rehabilitation engineering, Human-Computer Interaction (HCI), signal processing and machine learning [10].

The strength of BCI applications lies in the way we translate the neural patterns extracted from EEG into machine commands. The improvement of the interpretation of these EEG signals has become the goal of many researchers; hence, our research work explores the possibility of multi-trial EEG classification between left and right hand movements in an offline manner, which will enormously smooth the path leading to online classification and reading of executed movements, leading us to what we can technically call "Reading Minds".

In this work, we introduce an automated computer system that uses advanced feature extraction techniques to identify some of the brain activity patterns, especially for the left and right hand movements. The system then uses machine learning algorithms to extract the knowledge embedded in the recorded patterns and provides the required decision rules for translating thoughts into commands (as seen in Fig. 1).

This article is organized as follows: a brief review of related research work is provided in Section II. In Section III, the dataset used in this study is described. The automated feature extraction process is described in Section IV. The generation of our training/testing datasets and the practical implementation and system evaluation are discussed in Section V. Conclusions and suggested future work are provided in Section VI.

II. LITERATURE REVIEW

The idea of BCI was originally proposed by Jaques Vidal in [11] where he proved that signals recorded from brain activity could be used to effectively represent a user's intent. In [12], the authors recorded EEG signals for three subjects while imagining either right or left hand movement based on a visual cue stimulus. They were able to classify EEG signals into right and left hand movements using a neural network classifier with an accuracy of 80% and concluded that this accuracy did not improve with increasing number of sessions.

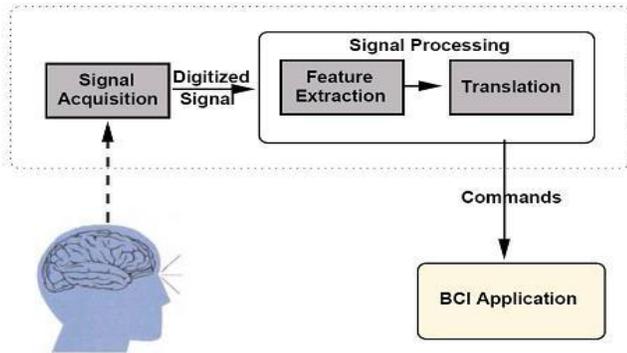


Fig. 1. Feature extraction and translation into machine commands

The author of [13] used features produced by Motor Imagery (MI) to control a robot arm. Features such as the band power in specific frequency bands (alpha: 8-12Hz and beta: 13-30Hz) were mapped into right and left limb movements. In addition, they used similar features with MI, which are the Event Related Desynchronization and Synchronization (ERD/ERS) comparing the signal's energy in specific frequency bands with respect to the mentally relaxed state. It was shown in [14] that the combination of ERD/ERS and Movement-Related Cortical Potentials (MRCP) improves EEG classification as this offers an independent and complimentary information.

In [15], a hybrid BCI control strategy is presented. The authors expanded the control functions of a P300 potential based BCI for virtual devices and MI related sensorimotor rhythms to navigate in a virtual environment. Imagined left/right hand movements were translated into movement commands in a virtual apartment and an extremely high testing accuracy results were reached.

A three-class BCI system was presented in [16] for the translation of imagined left/right hands and foot movements into commands that operates a wheelchair. This work uses many spatial patterns of ERD on mu rhythms along the sensory-motor cortex and the resulting classification accuracy for online and offline tests was 79.48% and 85.00%, respectively. The authors of [17] proposed an EEG-based BCI system that controls hand prosthesis of paralyzed people by movement thoughts of left and right hands. They reported an accuracy of about 90%.

A single trial right/left hand movement classification is reported in [18]. The authors analyzed both executed and imagined hand movement EEG signals and created a feature

vector consisting of the ERD/ERS patterns of the mu and beta rhythms and the coefficients of the autoregressive model. Artificial Neural Networks (ANNs) is applied to two kinds of testing datasets and an average recognition rate of 93% is achieved.

The strength of BCI applications depends lies in the way we translate the neural patterns extracted from EEG into machine commands. The improvement of the interpretation of these EEG signals has become the goal of many researchers; hence, our research work explores the possibility of multi-trial EEG classification between left and right hand movements in an offline manner, which will enormously smooth the path leading to online classification and reading of any executed movements, leading us to what we can technically call "Reading Minds".

III. THE PHYSIONET EEG DATA

A. Description of the Dataset

The EEG dataset used in this research was created and contributed to PhysioNet [19] by the developers of the BCI2000 [20] instrumentation system. The dataset is publically available at <http://www.physionet.org/pn4/eegmidb/>.

The dataset consists of more than 1500 EEG records, with different durations (one or two minutes per record), obtained from 109 healthy subjects. Subjects were asked to perform different motor/imagery tasks while EEG signals were recorded from 64 electrodes along the surface of the scalp. Each subject performed 14 experimental runs:

- A one-minute baseline runs (with eyes open)
- A one-minute baseline runs (with eyes closed)
- Three two-minute runs of each of the four following tasks:
 - The left or right side of the screen shows a target. The subject keeps opening and closing the corresponding fist until the target disappears. Then he relaxes.
 - The left or right side of the screen shows a target. The subject imagines opening and closing the corresponding fist until the target disappears. Then he relaxes.
 - The top or bottom of the screen. A target appears on either. The subject keeps opening and closing either both fists (in case of a top-target) or both feet (in case of a bottom-target) until the target disappears. Then he relaxes.
 - The top or bottom of the screen A target appears on either. The subject imagines opening and closing either both fists (in case of a top-target) or both feet (in case of a bottom-target) until the target disappears. Then he relaxes.

The 64-channels EEG signals were recorded according to the international 10-20 system (excluding some electrodes) as seen in Fig. 2.

B. The Subset used in the Current Work

From this dataset, we selected the three (two-minute) runs of the first task described above (opening and closing the left/right fist based on a target that appears on left or right side of the screen). These runs include EEG data for executed hand movements.

We created an EEG data subset corresponding to the first six subjects (S001, S002, S003, S004, S005, and S006) including three runs of executed movement specifically per subject for a total of 18 two-minute records.

IV. AUTOMATED ANALYSIS OF EEG SIGNALS FOR FEATURE EXTRACTION

A. Channel Selection

According to [6], many of the EEG channels appeared to represent redundant information. It is shown in [21, 22] that the neural activity that is correlated to the executed left and right hand movements is almost exclusively contained within the channels C3, C4, and CZ of the EEG channels of Fig. 2. This means that there is no need to analyze all 64 channels of data.

On the other hand, only eight electrode locations are commonly used for MRCP analysis covering the regions between frontal and central sites (FC3, FCZ, FC4, C3, C1, CZ, C2, and C4) [14]. These channels were used for the Independent Component Analysis (ICA) discussed later in the current section (Fig. 3).

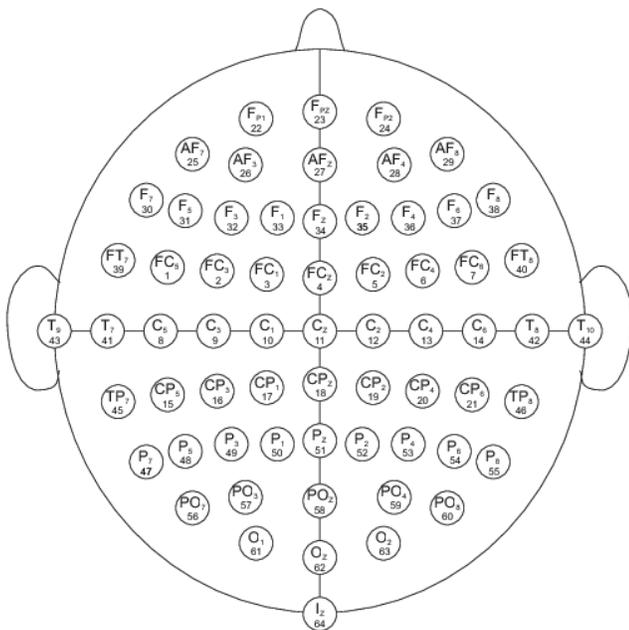


Fig. 2. Electrodes of the International 10-20 system for EEG

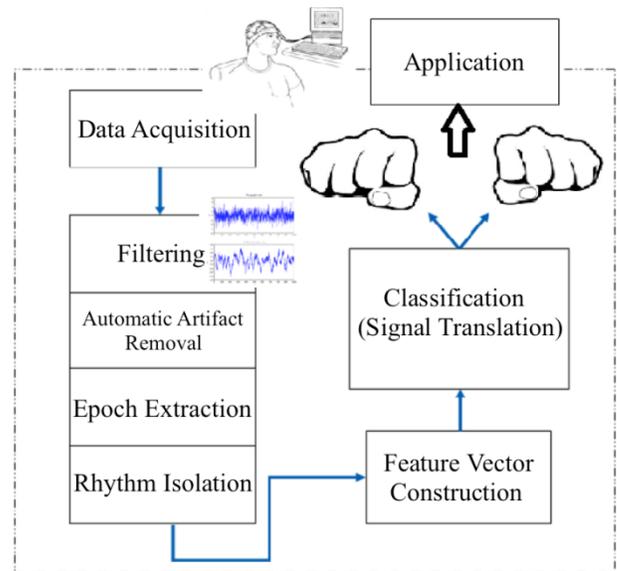


Fig. 3. Schematic diagram for the proposed system.

B. Filtering

Because EEG signals are known to be noisy and non-stationary, filtering the data is an important step to get rid of unnecessary information from the raw signals. EEGLAB [23], which is an interactive MATLAB toolbox, was used to filter EEG signals.

A band pass filter from 0.5 Hz to 90 Hz was applied to remove the DC (direct current) shifts and to minimize the presence of filtering artifacts at epoch boundaries. A Notch filter was also applied to remove the 50 Hz line noise.

C. Automatic Artifact Removal (AAR)

The EEG data of significance is usually mixed with huge amounts of useless data produced by physiological artifacts that masks the EEG signals [24]. These artifacts include eye and muscle movements and they constitute a challenge in the field of BCI research. AAR automatically removes artifacts from EEG data based on blind source separation and other various algorithms.

The AAR toolbox [25] was implemented as an EEGLAB plug-in in MATLAB and was used to process our EEG data subset on two stages: Electrooculography (EOG) removal using the Blind Source Separation (BSS) algorithm then Electromyography (EMG) Removal using the same algorithm [26].

D. Epoch Extraction (Splitting)

After the AAR process, the continuous EEG data were epoched by extracting data epochs that are time locked to specific event types.

When no sensory inputs or motor outputs are being processed, the mu (8–12 Hz) and beta (13–30 Hz) rhythms are said to be synchronized [4, 27]. These rhythms are electrophysiological features that are associated with the brain's normal motor output channels [4, 27]. While preparing for a movement or executing a movement, a desynchronization of the mu and beta rhythms occurs which is referred to as ERD and it can be extracted 1-2 seconds before onset of movement (as depicted in Fig. 4). Later, these rhythms synchronize again within 1-2 seconds after movement, and this is referred to as ERS.

On the other hand, delta rhythms can be extracted from the motor cortex, within the pre-movement stage, and this is referred to MRCP. The slow (less than 3 Hz) MRCP is associated with an event-related negativity that occurs 1-2 seconds before the onset of movement [28, 29].

In our experiments, we extracted time-locking events with type = 3 (left hand) or type = 4 (right hand) with different epoch limits and types of analysis:

- ERD analysis: epoch limits from -2 to 0 seconds.
- ERS analysis: epoch limits from 4.1 to 5.1 seconds.
- MRCP analysis: epoch limits from -2 to 0 seconds.

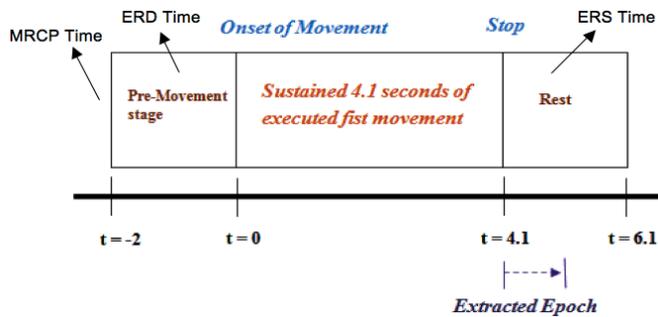


Fig. 4. Epoch Extraction (ERS/ERD and MRCP)

E. Independent Component Analysis (ICA)

After the AAR process, ICA was used to parse the underlying electrocortical sources from EEG signals that are affected by artifacts [30, 31]. Data decomposition using ICA changes the basis linearly from data that are collected at single scalp channels to a spatially transformed virtual channel basis. Each row of the EEG data in the original scalp channel data represents the time course of accumulated differences between source projections to a single data channel and one or more reference channels [32].

EEGLAB was used to run ICA on the described epoched datasets (left and right ERD, ERS, and MRCP) for the channels FC3, FCZ, FC4, C3, C1, CZ, C2, and C4.

F. Rhythm Isolation

A short IIR band pass filter from 8 to 30 Hz was applied on the ERD/ERS epoched datasets of the experiment for the purpose of isolating mu/beta rhythms. Another short IIR lowpass filter of 3 Hz was applied on MRCP epoched datasets for isolating delta rhythms. The result of this was 6 files for

each run: ERD/ERS and MRCP for both left and right hand movements for each subject.

V. PRACTICAL IMPLEMENTATION AND RESULTS

A. Feature Vectors Construction and Numerical Representation

After the EEG datasets were analyzed as described in the previous section, the activation vectors were calculated for each of the resulted epochs' datasets as the multiplication of the ICA weights and ICA sphere for each dataset subtracting the mean of the raw data from the multiplication results.

Then, the mean, power, and energy of the activations were calculated to construct the feature vectors. For each subject's single run, 6 feature vectors were extracted as <Power (8 features), Mean (8 features), Energy (8 features), Type (1 feature: ERS/ERD/MRCP), Side (1 target: Left/Right)> resulting in a 108×26 feature matrix.

The constructed features were represented in a numerical format that is suitable for use with machine learning algorithms [33, 34]. Every column in the features matrices was normalized between 0.1 and 0.9 such that the datasets could be inputted to the learning algorithms described in the next subsection.

B. Machine Learning Algorithms

In this work, Neural Networks (NNs) and Support Vector Machines (SVMs) algorithms were optimized for the purpose of classifying EEG signals into right and left hand movements. A detailed description of these learning algorithms can be found in [35] and [36].

The MATLAB neural networks toolbox was used for all NN experiments. The number of input features (25 features) determined the number of input nodes for NN and the number of different target functions (1 output: left or right) determined the number of output nodes. Training was handled with the aid of the back-propagation learning algorithm [37].

All SVM experiments were carried out using the "MySVM" software [38]. SVM can be performed with different kernels and most of them were reported to provide similar results for similar applications [6]. So, the Anova-Kernel SVM was used in this work.

C. Optimisation and Results

In all experiments, 80% samples were randomly selected and used for training and the remaining 20% for testing. This was repeated 10 times, and in each time the datasets were randomly mixed.

For each experiment, the number of hidden nodes for NN varied from 1 to 20. In SVM, each of the degree and gamma parameters varied from 1 to 10. The mean of the accuracy was calculated for each ten training-testing pairs.

The features that were used as inputs to NN and SVM are symbolized as follows:

- P: the power.
- M: the mean.
- E: the energy.

- X: the sample type (ERS/ERD/MRCP).

The results of the experiment are summarized in the Table I.

TABLE I. RESULTS FOR EXPERIMENT B.

Features	NN		SVM		
	Accuracy %	Hidden Layers	Accuracy %	Degree	Gamma
All	88.9	3	85.3	1	5
P, X	80.4	15	88.2	3	4
M, X	68.5	11	91.2	3	10
E, X	82.1	11	94.1	4	5
P, M, X	79.8	3	80.6	8	3
M, E, X	82.7	9	82.4	5	4
P, E, X	89.8	4	97.1	4	4

It is clear from the testing results that SVM outperforms NN in most experiments. An SVM topology of degree = 4 and gamma = 4 provides an accuracy of 97.1% if tested with the power, energy and type inputs of the experiment. A NN of 10 hidden layers can provide an accuracy of 86.5% if all features are used. These results clearly show that the use of advanced feature extraction techniques provides good and clear properties that can be translated using machine learning into machine commands.

The next best SVM performance (94.1%) is achieved using the energy and type features. In general, there has been an increase in the classification performance with the use of more discriminative features, such as the total energy, compared to the power and mean inputs.

VI. CONCLUSIONS AND FUTURE RESEARCH

This paper focuses on the classification of EEG signals for right and left fist movements based on a specific set of features. Very good results were obtained using NNs and SVMs showing that offline discrimination between right and left movement, for executed hand movements, is comparable to leading BCI research. Our methodology is not the best, but is somewhat a simplified efficient one that satisfies the needs for researchers in field of neuroscience.

In the near future, we aim to develop and implement our system in online applications, such as health systems and computer games. In addition, more datasets has to be analyzed for a better knowledgeable extraction and more accurate decision rules.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support received from Applied Science University that helped in accomplishing the work of this article.

References

[1] J. P. Donoghue, "Connecting cortex to machines: recent advances in brain interfaces," *Nature Neuroscience Supplement*, vol. 5, pp. 1085-1088, 2002.

[2] S. Levine, J. Huggins, S. BeMent, R. Kushwaha, L. Schuh, E. Passaro, M. Rohde, and D. Ross, "Identification of electrocorticogram patterns as the basis for a direct brain interface," *Journal of Clinical Neurophysiology*, vol. 16, pp. 439-447, 1999.

[3] A. Vallabhaneni, T. Wang, and B. He, "Brain-Computer Interface," in *Neural Engineering*, B. He, Ed.: Springer US, 2005, pp. 85-121.

[4] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, pp. 767-791, 2002.

[5] E. Niedermeyer and F. H. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*: Lippincott Williams & Wilkins, 2005.

[6] J. Sleight, P. Pillai, and S. Mohan, "Classification of Executed and Imagined Motor Movement EEG Signals," *Ann Arbor: University of Michigan*, 2009, pp. 1-10.

[7] B. Graimann, G. Pfurtscheller, and B. Allison, "Brain-Computer Interfaces: A Gentle Introduction," in *Brain-Computer Interfaces*: Springer Berlin Heidelberg, 2010, pp. 1-27.

[8] A. E. Selim, M. A. Wahed, and Y. M. Kadah, "Machine Learning Methodologies in Brain-Computer Interface Systems," in *Biomedical Engineering Conference, 2008, CIBEC 2008. Cairo, 2008*, pp. 1-5.

[9] E. Grabianowski, "How Brain-computer Interfaces Work," <http://computer.howstuffworks.com/brain-computer-interface.htm>, 2007.

[10] M. Smith, G. Salvendy, K. R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, "Machine Learning and Applications for Brain-Computer Interfacing," in *Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design*. vol. 4557: Springer Berlin Heidelberg, 2007, pp. 705-714.

[11] J. J. Vidal, "Toward Direct Brain-Computer Communication," *Annual Review of Biophysics and Bioengineering*, vol. 2, pp. 157-180, 1973.

[12] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimination between imagination of right and left hand movement," *Electroencephalography and Clinical Neurophysiology*, vol. 103, pp. 642-651, 1997.

[13] F. Sepulveda, "Brain-actuated Control of Robot Navigation," in *Advances in Robot Navigation*, A. Barrera, Ed.: InTech, 2011.

[14] A.-K. Mohamed, "Towards improved EEG interpretation in a sensorimotor BCI for the control of a prosthetic or orthotic hand," in *Faculty of Engineering. Master of Science in Engineering*, Johannesburg: University of Witwatersrand, 2011, p. 144.

[15] Y. Su, Y. Qi, J.-x. Luo, B. Wu, F. Yang, Y. Li, Y.-t. Zhuang, X.-x. Zheng, and W.-d. Chen, "A hybrid brain-computer interface control strategy in a virtual environment," *Journal of Zhejiang University SCIENCE C*, vol. 12, pp. 351-361, 2011.

[16] Y. Wang, B. Hong, X. Gao, and S. Gao, "Implementation of a Brain-Computer Interface Based on Three States of Motor Imagery," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS2007, 2007*, pp. 5059-5062.

[17] C. Guger, W. Harkam, C. Hertnaes, and G. Pfurtscheller, "Prosthetic Control by an EEG-based Brain-Computer Interface (BCI)," in *AAATE 5th European Conference for the Advancement of Assistive Technology*, Düsseldorf, Germany, 1999.

[18] J. A. Kim, D. U. Hwang, S. Y. Cho, and S. K. Han, "Single trial discrimination between right and left hand movement with EEG signal," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2003.*, Cancun, Mexico, 2003, pp. 3321-3324 Vol.4.

[19] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, pp. e215-e220, 2000.

[20] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1034-1043, 2004.

[21] L. Deecke, H. Weinberg, and P. Brickett, "Magnetic fields of the human brain accompanying voluntary movements: Bereitschaftsmagnetfeld," *Experimental Brain Research*, vol. 48, pp. 144-148, 1982.

[22] C. Neuper and G. Pfurtscheller, "Evidence for distinct beta resonance frequencies in human EEG related to specific sensorimotor cortical areas," *Clinical Neurophysiology*, vol. 112, pp. 2084-2097, 2001.

- [23] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics," *Journal of Neuroscience Methods*, vol. 134, pp. 9-21, 2004.
- [24] G. Bartels, S. Li-Chen, and L. Bao-Liang, "Automatic artifact removal from EEG - a mixed approach based on double blind source separation and support vector machine," in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2010, pp. 5383-5386.
- [25] G. Gómez-Herrero, "Automatic Artifact Removal (AAR) toolbox for MATLAB," in *Transform methods for Electroencephalography (EEG)*: <http://kasku.org/projects/eeg/aar.htm>, 2008.
- [26] C. Joyce, I. Gorodnitsky, and M. Kutas, "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation," *Psychophysiology*, vol. 41, pp. 313-325, 2004.
- [27] A. Bashashati, M. Fatourehchi, R. Ward, and G. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *Journal of Neural Engineering*, vol. 4, pp. R32-57, 2007.
- [28] A. Vuckovic and F. Sepulveda, "Delta band contribution in cue based single trial classification of real and imaginary wrist movement," *Medical and Biological Engineering and Computing*, vol. 46, pp. 529 – 539, 2008.
- [29] Y. Gu, K. Dremstrup, and D. Farina, "Single-trial discrimination of type and speed of wrist movements from EEG recordings," *Clinical Neurophysiology*, vol. 20, pp. 1596–1600, 2009.
- [30] J. T. Gwin and D. Ferris, "High-density EEG and independent component analysis mixture models distinguish knee contractions from ankle contractions," in 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, Boston, USA, 2011, pp. 4195-4198.
- [31] S. Makeig, A. J. Bell, T. P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," *Advances in Neural Information Processing Systems*, vol. 8, pp. 145-151, 1996.
- [32] A. Delorme and S. Makeig, "Single subject data processing tutorial: Decomposing Data Using ICA," in *The EEGLAB Tutorial*: <http://sccn.ucsd.edu/wiki/EEGLAB>, 2013.
- [33] M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson, "Machine learning-based investigation of the associations between cmes and filaments," *Solar Physics*, vol. 262, pp. 511-539, 2010.
- [34] R. Qahwaji, T. Colak, M. Al-Omari, and S. Ipson, "Automated machine learning based prediction of CMEs based on flare associations," *Sol. Phys.*, vol. 248, 2007.
- [35] R. Qahwaji and T. Colak, "Automatic Short-Term Solar Flare Prediction Using Machine Learning and Sunspot Associations," *Solar Phys.*, vol. 241, pp. 195-211, 2007.
- [36] R. Qahwaji, M. Al-Omari, T. Colak, and S. Ipson, "Using the Real, Gentle and Modest AdaBoost Learning Algorithms to Investigate the Computerised Associations between Coronal Mass Ejections and Filaments," in *Mosharaka International Conference on Communications, Computers and Applications (MIC-CCA 2008)*, Mosharaka for Researches and Studies, Amman, Jordan, 2008, pp. 37-42.
- [37] S. E. Fahlmann and C. Lebiere, "The cascade-correlation learning architecture," in *Advances in Neural Information Processing Systems 2 (NIPS-2)* Denver, Colorado, 1989.
- [38] S. Rüping, "mySVM-Manual ": University of Dortmund, Lehrstuhl Informatik 8, 2000.

Case Study of Named Entity Recognition in Odia Using Crf++ Tool

Dr.Rakesh ch. Balabantaray
Department of Computer Science
IIIT, BBSR

Suprava Das
Department of Computer Science
IIIT, BBSR

Kshirabdhii Tanaya Mishra
Department of Computer Science
IIIT, BBSR

Abstract—NER have been regarded as an efficient strategy to extract relevant entities for various purposes. The aim of this paper is to exploit conventional method for NER in Odia by parameterizing CRF++ tool in different ways. As a case study, we have used gazetteer and POS tag to generate different feature set in order to compare the performance of NER task. Comparison study demonstrates how proposed NER system works on different feature set.

Keywords—Named Entity Recognition; CRF++ Tool; Odia Named Entity

I. INTRODUCTION

NER is a subtask of information extraction that involves locating and classifying named entities such as person name, location name, organization name... etc. Besides information extraction, NER has applications in question answering (Toral et al., 2005; Molla et al., 2006), Machine translation (Babych & Hartley, 2003). In English language, recognition of named entity is easy with greater accuracy, but for Indian languages (especially for the language which are not morph analysed), recognition of named entity is challenge now. For Indian languages, many approaches have been applied for NE recognition. These approaches are: Rule based approach (krupka and Hausman, 1998) and Machine learning approach or hybrid approach Decision tree (Karkaletis et al. , 2000) , Hidden Markov model(Biker ,1997) , MEMM(Borthwick et al. ,1998) , CRF(Andrew McCallum and Wei Li , 2003)).This paper presents an overview of work done on locating named entity in a text for Odia language using conditional random field. We have used CRF++ (version 0.54) tool which is implementation of conditional random field, a machine learning approach for NE recognition. The statistical CRF model has been used for NER as it is more efficient to deal with Indian languages. Section-2 gives a brief description on conditional random field and section-3 gives brief description on Part of speech tag; section-4 describes preparation of training data and testing data for CRF based model followed by section 5 describes the features used for CRF framework, section 6 describes how CRF++ detects named entities and section 7 describes the result and accuracy. Conditional random field is a machine learning technique which overcomes the disadvantage of other machine learning approach like HMM and MEMM. In HMM, the words in input sequence are not dependant among each other. MEMM face label bias problem because of its stochastic state transmission nature. CRF overcomes these problems and gives a greater accuracy. Conditional random field are undirected graphical

model used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. As CRF is a discriminative, so the word identity feature is informative, this helps to label unseen words by exploiting the feature.

We have used the C++ based openNLP CRF++ package of version 0.54 (Taku Kudo, 2005). The CRF++ tool extracts the information from the training data and builds a CRF model according to weightage of information. When the test data presented with CRF model, the tool outputs the test data tagged with the labels that has been learnt.

II. CONDITIONAL RANDOM FIELD

Conditional random field is a machine learning technique which overcomes the disadvantage of other machine learning approach like HMM and MEMM. In HMM, the words in input sequence are not dependant among each other. MEMM face label bias problem because of its stochastic state transmission nature. CRF overcomes these problems and gives a greater accuracy. Conditional random field are undirected graphical model used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. As CRF is a discriminative, so the word identity feature is informative, this helps to label unseen words by exploiting the feature.

Conditional Random Fields can be defined as in [3] as follows: "Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G . Then (X, Y) is a conditional random field in case, when conditioned on X , the random variables Y_v obey the Markov property with respect to the graph:

$P(Y_v | X, Y_w, w \sim v) = p(Y_v | X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G ".

Here X might range over natural language sentences and Y denotes the label sequence.

What this means is that a CRF is an undirected graphical model whose nodes can be divided into exactly two disjoint sets X and Y , the observed and output variables, respectively; the conditional distribution is $p(Y|X)$ is then modelled. The aim of the CRF is to find out the label sequence $y \in Y$ that maximizes the conditional probability $p(Y|X)$ for a sequence X .

That is $y = \underset{y}{\operatorname{argmax}} p(Y|X)$

Thus, NER task can be considered as a sequence labeling task. Hence CRF can be used for NER task.

III. EXPERIMENTAL SET UP

A. Part Of Speech Tag

In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

There are large numbers of POS tagger available for English language which has got satisfactory performance but cannot be applied to Hindi language due to structural differences. For our experiment we have used POS-Tagger tool for Odia language which is implemented using conditional random field. The accuracy of this tool is not high but accuracy of tagging proper noun is quite high.

B. Gazetteer

We have prepared 4 different gazetteers. The words belongs to the person, location, organization are stored in 3 different gazetteers respectively. Another gazetteer contains only NE without any classification and it contains around 730 NEs. The named entities in gazetteer are arranged in dictionary order. For morph analysis we have used another gazetteer which

D. Corpus

A corpus for Odia language is collected which contains around 45000 tokens/words from the domain of health, tourism, general. This corpus contains about 1000 named entities of PERSON, LOCATION, and ORGANIZATION. This file is split into 2 sets, 80% of words are used for training data and 20% of them used for testing data.

E. Preparation Of Training Data

For case study training data needs to be prepared in 3 different ways for 3 different cases. To make CRF++ tool learns, training data should be in a particular format. So the training file needs to be pre-processed. We have taken 3 column format training data. 1st column remains same for all cases, but 2nd and 3rd column varies. 2nd column is generated using POS Tagger tool with POS tag for two cases and for one case it is tagged with set {YES, NO}. 3rd column of training data contains all of user generated annotations for named entities. For one case the Odia named entity tagged with tag set {B-name, I-name, 00}. The tokens which are not present in the gazetteer means which are not named entities are tagged as "00". And those which are named entities, if contains single token as NE, this tagged as "B-name", otherwise the 1st token is tagged as "B-name" and the rest tokens which are inside NE tagged as "I-name". For other two cases users are supported to label named entity by using the corresponding tags i.e. <PERSON>, <LOCATION>, <ORGANIZATION>.

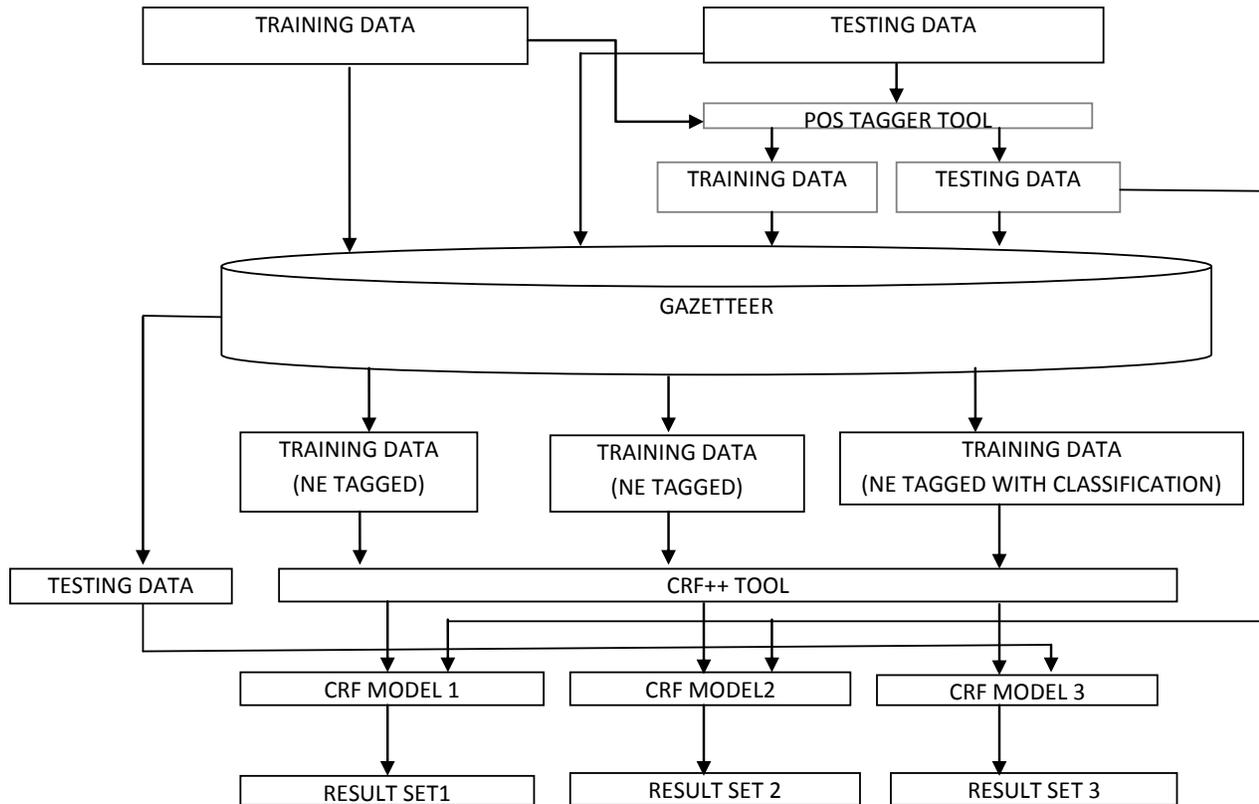


Fig. 1 [Work flow diagram]

F. Preparation Of Testing Data

Unlike the train data, the test data is in 2-column format. The test data is presented in same way as train data , only the difference is test data contains only tokens and corresponding POS tag (for two cases) and { YES, NO } tag (for one case).

The preparation of training data, testing data and analysis of NER system using CRF++ tool is schematically represented in FIGURE – 1.

IV. RESULT AND DISCUSSION

To evaluate the performance of NER in Odia language using CRF++ tool, we make use of 3 parameters i.e. precision, recall and f-measure.

Precision measures the percentage of correct NE tagged by CRF tool over the total number of NEs tagged by CRF tool.

$$precision = \frac{tp}{tp + fp}$$

Recall measures the percentage of NE tagged by CRF tool over the total number of NEs in the file tagged by gazetteer.

$$Recall = \frac{tp}{tp + fn}$$

F-measure is a measure that combines precision and recall is the harmonic mean of precision and recall.

$$F\ measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The comparative study for all the three cases has done. And result for these cases are given in the table below.

Measurement	Value
Precision	0.925
Recall	0.593
F – measure	0.71

TABLE-1: [Evaluation of NEs without classification]

Table – 1 show that our proposed feature sets can effectively identify Odia named entity from testing repository.

The table-2 describes the comparison between the cases where the classification of named entity is taken into consideration. For one case gazetteer is used to parameterize CRF++ Tool and for other case POS tag along with gazetteer is used to parameterize the tool, which causes generation of different sets of feature.

Table -3 shows the actual number of NEs present in training and testing repository and the number of named entity recognized by CRF MODEL. Based upon which the performance of the system is measured.

A. Comparison Graph

We have taken different dataset to measure the performance named dataset-1, dataset-2 and dataset-3.

Two classes of parameters are most important: the combination and selection of feature and tokenization of the text.

The impact of each feature (Gazetteer and POS tag) or group of feature (Gazetteer combined with POS tag) is computed. They are displayed in following graph.



Fig. 2 [Comparison of f measure of ORGANIZATION NEs using different dataset]

	GAZETTEER			All features (Gazetteer and POS tag)			F measure comparison
	P	R	F	P	R	F	
PERSON	0.87	0.81	0.84	0.97	0.44	0.63	25% decrease
LOCATION	0.88	0.82	0.85	0.75	0.50	0.60	18% decrease
ORGANIZATION	0.50	0.82	0.62	0.66	0.25	0.35	43% decrease

TABLE – 2: [Results for PERSON, LOCATION and ORGANIZATION using gazetteer and all features] P-Precision R-Recall F-F measure

	person		location		organization	
	TRN	TST	TRN	TST	TRN	TST
Gazetteer	382	180	248	175	183	70
Gazetteer and POS		109		121		43

TABLE – 3: [Calculation of total number of NEs for all cases]
TRN – Training Data, TST – Testing Data

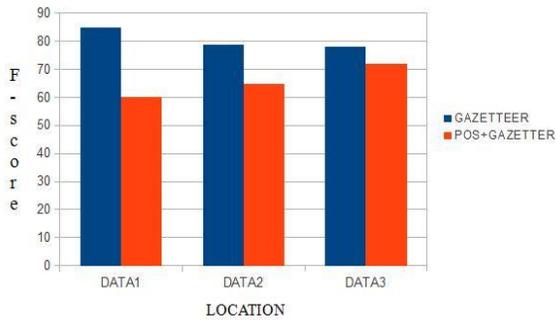


Fig. 3 [Comparison of f measure of LOCATION NEs using different dataset]

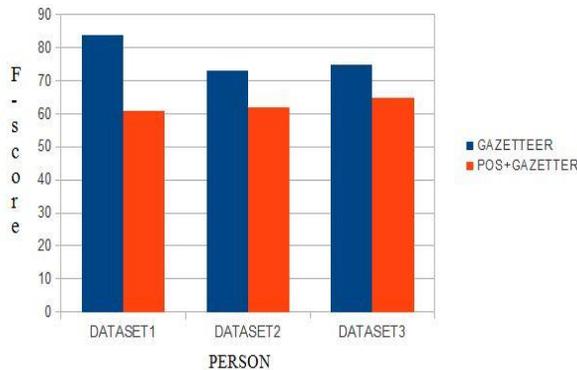


Fig. 4 [Comparison of f measure of PERSON NEs using different dataset]

VIII. CONCLUSION

In this paper we have shown a novel NER system based on conditional random field by generating various type of feature set. We have used CRF based POS tagger tool and gazette file to parameterize CRF++ Tool. The performance of the system is quite good when we experiment with individual case (f-measure for NEs only is 71% and f-measure for NEs with classification is 84% for PER, 85% for LOC and 62% for ORG). The performance of system decreases when we combine both POS tag and Gazetteer to generate feature. The reason for decrease in performance may be the average accuracy of POS Tagger tool. The accuracy may be increased if accuracy of POS Tagger tool is good. Morphological analysis has also shown a small contribution to the performance of the system. The current work is limited to recognizing the named entities which does not have nested structure.

REFERENCES

- [1] Wallach, H. M. 2004. Conditional random fields: An introduction, Technical Report MS-CIS-04-21, University of Pennsylvania, Department of Computer and Information Science, University of Pennsylvania.
- [2] Li Wei and McCallum Andrew. 2004. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper). In *ACM Transactions on Computational Logic*.
- [3] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data," *Proc. 18th Int'l Conf. Machine Learning*, 2001.
- [4] Sotirios P. Chatzis, Yiannis Demiris, "The Conditional Random Field Model for Sequential Data Modelling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 02 Oct. 2012. IEEE computer Society Digital Library.
- [5] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos," Learning Decision Trees for Named-Entity Recognition and Classification", *ECRAN*, 2000.
- [6] Colmenar, J.M., Abanades, M.A., Poza, F., Martin, D., Cuesta, A., Herran, A., Hidalgo, J.I., "On a generalized name entity recognizer based on Hidden Markov Models", *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference* , On page(s): 952 - 958
- [7] Andrew McCallum, Dayne Freitag, and Fernando Pereira," Maximum Entropy Markov Models for Information Extraction and Segmentation", *17th International Conf. on Machine Learning*, 2000, 591-598
- [8] CRF++. <http://crfpp.sourceforge.net/>

TX-Kw: An Effective Temporal XML Keyword Search

Rasha Bin-Thalab

Department of Information System
Faculty of Computers and Information
Cairo University, Egypt

Neamat El-Tazi

Department of Information System
Faculty of Computers and Information
Cairo University, Egypt

Mohamed E.El-Sharkawi

Department of Information System
Faculty of Computers and Information
Cairo University, Egypt

Abstract—Inspired by the great success of information retrieval (IR) style keyword search on the web, keyword search on XML has emerged recently. Existing methods cannot resolve challenges addressed by using keyword search in Temporal XML documents. We propose a way to evaluate temporal keyword search queries over Temporal XML documents. Moreover, we propose a new ranking method based on the time-aware IR ranking methods to rank temporal keyword search queries results. Extensive experiments have been conducted to show the effectiveness of our approach.

Keywords—temporal XML; Keyword Search; ranking

I. INTRODUCTION

The success of keyword search in IR has encouraged its emergence in XML [1-3] and databases [4, 5]. Although, temporal data are used commonly in historical applications (web logs, financial, scientific, and georeferencing applications), existing XML keyword search methods are not aware of temporal expressions in keywords. Temporal keyword refers to exploiting time dimension that is embedded inside the XML documents to provide alternative search methods and user experience.

A study made by Zhang et al. [6] showed that about 13.8% of queries have explicit time predicate and 17.1% of queries implicitly contain temporal intent. An example of a query with explicit time provided is "U.S. Presidential election 2008". An implicit time query example can be "Germany FIFA World Cup", here the time is not declared but the user is referring to the World Cup event in 2006. Furthermore, database applications contain information for long time periods, like, DBLP which keeps all publications that cover the years 1954 up till now. When searching in these documents archives, a temporal dimension plays an important role.

Keyword search in XML model is used to find nodes that contain keywords and checks the interconnections among them based on their lowest common ancestors (LCA) [1]. For example, query Q1 "Michael, Adams" in Fig. 1 returns node; 0.2.0.2.

However, LCA has a lot of drawbacks since it does not give a meaningful answer in all cases, e.g., as in query Q1, node 0.2.0.2 might not be the user intention. Another

drawback is that it does not consider ID/IDREF relationship between nodes, which may result in missing some relevant results. Recent approaches [4, 7, 8] preferred to model XML document as a set of interrelated objects rather than nodes. Each object is represented as a sub tree rooted by a representative node with its set of attributes. Also, ID/IDREF connections are considered in such approaches to increase relevant results.

In this paper, we integrate temporal constraints into keyword search approaches for temporal XML databases (TX-Kw). We perform semantic matching at object level rather than using traditional LCA techniques. There are three basic reasons that motivated us to use object-level in temporal keyword search over temporal XML documents. First, XML can be recognized as a set of real world objects, each of which has attributes and interacts with other objects through relationships in certain temporal intervals, for example player, team and coach entities in NBA DB as shown in Fig. 1 are considered objects in real world. Second, users aim to find a specific object information by typing a set of words and a specific time about such object. They do not aim to find if the information exists or not by retrieving the node that contains this information. Finally, temporal nature of nodes in temporal XML documents can be captured well in objects as long as their attribute values and relationships output change over time. Thus, object-level may be very helpful to give more relevant answers especially if we adapt ranking objects rather than ranking nodes by taking time dimension into account. In this case, keyword search results is either a single object which contains all keywords in the time specified or a set of interconnected objects that contain the keywords in that specified time.

Our objective in this paper is to effectively retrieve a single temporal object (STO) or a set of related temporal objects (RTO) that are the closest to user intention while considering ID/IDRef links. A brute force approach, which considers all result objects before the top-k scores, requires expensive processing time. We build several index structures for keywords and temporal objects to provide better performance. We propose an efficient algorithm based on these structures to get top-k results of temporal objects.

We summarize the contribution of this paper as follows:

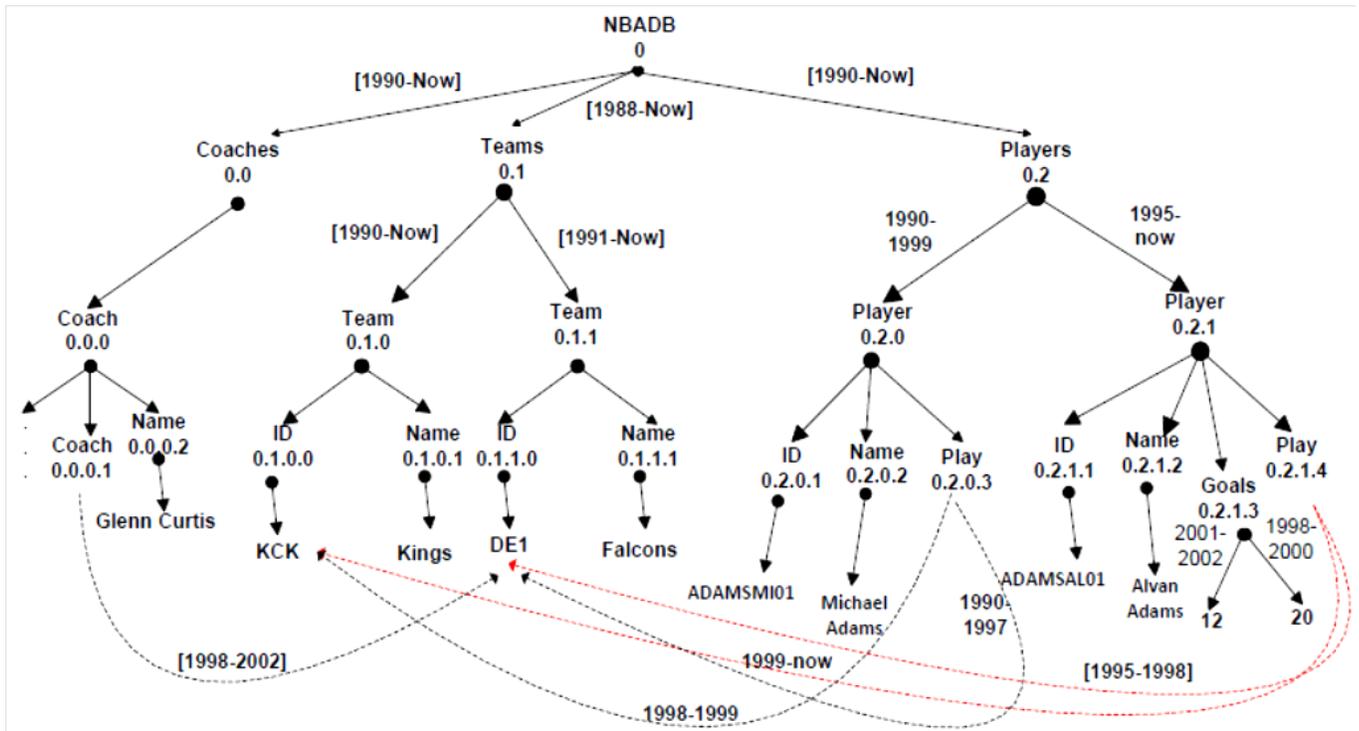


Fig.1. NBA DB portion (with Dewey IDs)

- Modeling XML into temporal objects based on their ID/IDRef attributes
- Build an efficient temporal keyword search algorithm for processing temporal keyword search queries over temporal XML documents.
- Design adaptive temporal ranking method for XML keyword search

The rest of this paper is structured as follows. Section II describes related work. Section III presents background and definition used through the paper. Section IV introduces semantic matching of objects to keywords. Section V addresses ranking functions used to rank result objects. Section VI presents structure indexes for enhancing performance. Keywords search algorithms are presented in section VII. We implement experiments to compare our algorithm to the state-of-the-art methods in Section VIII. Finally, Section IX provides concluding remarks and future works.

II. RELATED WORK

In literature, keyword search has been studied well in XML environment [9]. We categorize these approaches into two categories; tree and graph models. Approaches for XML graph model are more related to our work. On XML tree model, XSearch [10] and SLCA [2] provided an efficient way to calculate the smallest LCA (SLCA) XML node that contains all keywords. However, successive works were proposed to improve effectiveness like VLCA [11], and efficiency ELCA [3, 12]. Next, Bao [13] proposed an IR-style approach (called XReal) which basically utilizes the statistics

of underlying XML data to present a novel XML TF*IDF ranking strategy to rank the individual matches of all possible search intentions. On XML graph model, ObjectRank [4] is one of the earliest studies which designed a semantically meaningful ranking method using the authority transfer paradigm. However, ObjectRank results in only single objects and does not take a group or a cluster of objects as alternative results. On the other hand, several XML approaches were presented to implement graph model in XML keyword search result, in more effectiveness in trade of efficiency. XKeyword [8] provided an efficient keyword proximity queries for large XML graph databases. The authors adopt the concept that a keyword proximity query is a set of keywords and the results are trees of XML fragments (called Target Objects) that contain all the keywords. However, ranking of target objects is restricted to the distance between elements which leads to missing objects that are more related to the keywords, although they do not contain them. Recently, Bao [7] presented an object-level to retrieve effective results which are based on schema document. The result is either a single object or a set of interconnecting objects. In fact, the authors only considered object class but ignored object ID. Thus, they cannot discover duplicate objects and suffer the same problems as LCA-based approaches.

All the mentioned approaches did not take benefit from the temporal nature of temporal XML documents in retrieving results of temporal keyword search queries.

In information retrieval, several proposals addressed time-aware ranking of pages in www environment. They are divided into two categories: link based [14, 15], and content-

based [16, 17]. In link-analysis approach, Yu [18] modified the PageRank [19] algorithm by accumulating the weights of its citations, where each citation receives a weight that exponentially decreases by its age. Berberich [14] also extended PageRank to rank documents with respect to freshness. The difference is that this work defines freshness as a linear function that will give a maximum score when the date of document or link occur within the user specified period and decrease a score linearly if it occurs outside the interval. Second type of ranking methods is based on an analysis of document content [17]. Jatowt [20] presented an approach to rank a document by its freshness and relevance. A document is ranked high if it is modified significantly and recently. Diaz and Jones [21] used timestamp from document metadata to measure the distribution of retrieved documents and create the temporal profile of a query.

To the best of our knowledge, the first study of temporal keyword search in XML has been addressed by Manica [22]. They identified temporal constraints in a keyword query and intercepted the query processing, executed by a conventional XML search engine, in order to evaluate those constraints. However, temporal ranking results were not handled.

III. BACKGROUND, NOTIONS AND DEFINITIONS

In this section, we describe the concept of temporal objects (TOs), which we use in this paper. To define temporal objects in XML document, we combine the definitions of temporal object in [23, 24] and an object tree in [7] as follows.

Definition 1. A temporal object O_i represents a real-world entity or concept, each object has an object ID, attributes and lifespan. We define a temporal object in an XML document as a subtree (object tree) annotated with lifespan. Each object is represented by $\langle \text{OID}, \text{att_list}, \text{lifespan}, \text{OList} \rangle$

where "OID" is the object identifier. "att_list" is the list of attributes the object has. "lifespan" is the life time of the object in the system. "Olist" contains a list of $\langle \text{OID}, \text{Time} \rangle$ pairs denoting the objects that are connected to the object with the timestamp for that connection.

A set of objects with similar characteristics (attributes) are grouped into what is called a class. An object class (called class for brevity) consists of a signature that defines the object in reality.

How to identify the temporal objects is orthogonal to this work; here, we adopt the inference rules in XSeek [26] to help identify the object trees from XML Schema. In the case of no obvious schema, any other program for extraction schema is used. As we can see from Fig. 1, there are five temporal object instances for three classes; 2 objects for Player class, 2 objects for Team class, and 1 object for Coach class. A dashed line represents the ID/IDREF edges connecting objects. Note that nodes Players, Teams and Coaches are connection nodes which connect the node Players with the player objects, (the same for Team and Coach). In XML model, a real object class is distinguished in form of a subtree due to its hierarchal inheritance.

Another important concept is introduced, connections, which is used to define relationships between temporal

objects. We distinguish between two types of connections; containment and references. Two temporal objects have a containment connection if there is a containment edge annotated by timestamp connecting them (parent-child edge) denoting when this connection is active. On the other hand, reference connection is used if there is an ID/IDREF edge annotated by a certain timestamp between two temporal objects.

In this work, we apply a discrete notion of time and assuming the integers Z . The temporal expression T can refer to any time interval $[b, e]$ where $b \leq e$. Year is used as a time granularity for simplicity.

IV. TEMPORAL OBJECT MATCHING SEMANTICS

In context of XML keyword search, a temporal keyword query is a set of keywords attached with time (e.g., "Michael, Adams, 1995"). Usually, when a user issues his keyword search, he intends to get almost a single object that contains all the keywords he issued in the specific time domain or even a set of interrelated objects that contain all keywords and intersect at the time interval provided. These objects are more likely to have a well known relationship among them.

A. Single Temporal Object Matching Semantics

Definition 2. Given a temporal keyword query Q_i , a temporal object O_i is defined as a Single Temporal Object (STO) found in the document if it contains all the keyword(s) as part of its attribute's value or structure tag names, and its lifespan interval intersect with Q_i time.

One can conclude that STO plays LCA role in the temporal object oriented model. For example in Fig. 1, if we issue the query "Alvan, Adams, 2000-2005", STO returns the Player object rooted at node 0.2.1 rather than Name attribute at node 0.2.1.2 returned by traditional LCA.

B. Related Temporal Objects Matching Semantics

Keywords in Q_i keywords can be found in different objects rather than a single one. Furthermore, the time interval in the selected objects has to intersect with Q_i interval. For example, given a query Q_4 : "team, Curtis, [2000-2002]" in Fig. 1, here the user wants to know the team which is coached by Curtis in the interval [2000- 2002], the "team" keyword is contained in two objects rooted at nodes 0.1.0 and 0.1.1, and "Curtis" found in object Coach rooted at node 0.0.0. If we take the LCA based on their Dewey Id, the root node NBA 0 is returned. When considering the ID/IDREF into account, we can see that there is an ID/IDREF edge between objects team (0.1.1) and coach (0.0.0) during the interval [1998-2002]. Thus the result in this case will be both objects; team (0.1.1) and coach 0.0.0 since they are connected by a reference edge.

There are three types of connections that can exist between any two objects a and b . First, a and b can be connected via a lowest common object ancestor LCOA (e.g., in DBLP, one paper is an ancestor of multiple papers by cites relationship). Second, a and b can have a common object descendant COD (e.g., in Fig. 1, team 0.1.1 is COD for both objects player (0.2.1) and coach (0.0.0)). Moreover, a and b can be connected via an n-hop connections meaning, if there are n-1

intermediate distinct objects o_1, \dots, o_{n-1} such that there is a connection between each pair of adjacent objects and no objects of them are connected via ancestor-descendant relationship. For example, in DBLP, one paper may be connected with another paper through a set of intermediate paper citations although no direct connection between them.

Definition 3. *Related temporal objects result (RTO) is a tree structure composed of temporal objects having the query keywords and intersect with query temporal interval while having either ancestor/descendant or ID/IDREF relationships connection between them.*

In order to construct RTOs, we combine the use of schema structure of XML document and ID/IDRef edges. Fig. 2(a) shows schema of DBLP and NBA data sets. We can find that the set of possible objects could be connected to form RTOs.

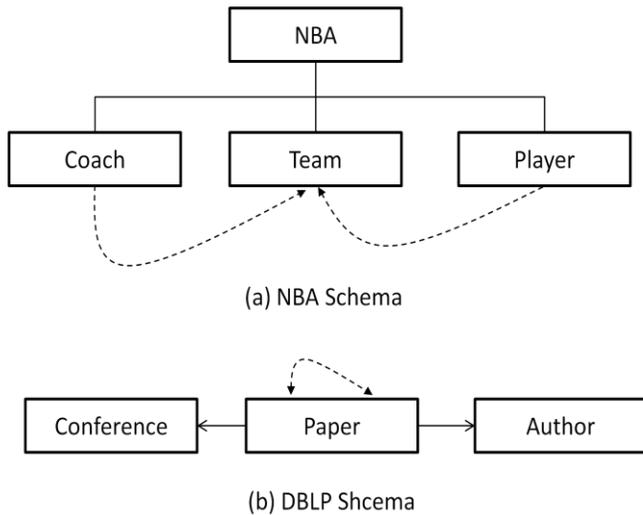


Fig.2. RTOs for NBA and DBLP Schemas

V. RANKING

The challenge of capturing effective results of temporal keywords is the selection of temporal objects that have the highest ranking scores (top-k). This problem is studied very well in IR [6], [20]. Incorporating time dimension into ranking models can significantly improve result effectiveness of user intention

In this section we introduce a ranking model to increase effectiveness for temporal queries. Note that pages in IR environment are mapped into objects in XML environment.

Thus we have an object granularity rather than a page. The object contains values of attributes each of which may have temporal data (like birth-date of employees) which corresponds to content time in IR, and the object has a lifespan interval against publication time of pages.

Lifespan interval starts when the object inserted into the system and its end time is determined by the time deletion of the object or left up to now. Object lifespan also is sensitive to the application semantics, such as in DBLP, publishing year of an article is considered as the start of the article's object lifespan.

A. Ranking Model

Temporal objects as well as temporal queries contain two integral information keywords and time. Here, we design a ranking model which is based on a mixture model used for IR [25] that linearly combines keywords similarity and temporal similarity and then we map it into XML objects. Given a single temporal object O_t and a temporal query Q_t , we compute the similarity degree ρ_s of O_t to Q_t using the following formula:

$$\rho_s(Q_t, O_t) = \alpha S_k(Q_t, O_t) + (1 - \alpha) S_t(Q_t, O_t) \quad (1)$$

where the mixture parameter α indicates the importance of keywords similarity $S_k(Q_t, O_t)$ compared to temporal similarity $S_t(Q_t, O_t)$. The higher the score the more relevant the object is. Next we define each similarity computation.

B. Keyword Similarity S_k

$S_k(O_t, Q_t)$ ranks the keywords of Q_t in object O_t . We compute this ranking by also taking timestamp of keyword into account. We extend the CT-rank of Jin [20] which addressed the ranking of keywords and its validity time in web pages. However, our granularity is an object not a document. An object also has attributes which taken into account in our ranking function. Let K_{time} denotes the frequency of keyword k with its timestamp t , $\langle k, t \rangle$, in O_t . K_{total} is the total number of all keywords in object O_t . N_O is the total number of objects in the XML document. N_k is the total number of objects that contain keyword k . $score(k, O_t)$ is used to compute rank of k in O_t . It is defined as follows:

$$score(k, O_t) = \frac{K_{time}(\langle k, t \rangle, O_t)}{K_{total}} \times \log\left(\frac{N_O}{N_k}\right) \quad (2)$$

Note that the score is based on TF/IDF [26] computation used in IR where mapping pages to objects. Since the object contains attributes, a keyword might occur more than once in the same attribute. K_{time} function is calculated in Equation 3.

$$K_{time}(\langle k, t \rangle, O_t) = \sum_{\forall a \in attr(O_t, \langle k, t \rangle)} (tf(a, \langle k, t \rangle)) \quad (3)$$

$tf(a, k)$ is the number of $\langle k, t \rangle$ pairs in the specified attribute. Finally, a contribution value cb is another factor to be considered in local score of an object, i.e. how many keywords there are in an object. For the whole keywords in the query Q_t , object O_t is ranked according to query keywords as:

$$S_k(Q_t, O_t) = \sum_{k \in Q_t \wedge k \in O_t} (score(k, O_t)) \times cb \quad (4)$$

We add the contribution factor cb to measure the whole keywords in Q_t that are contained in O_t . We compute cb as total number of query keywords in object O_t divided by the total number of Q_t .

C. Temporal Similarity S_t

Two temporal expressions can affect the ranking of an object; lifespan (e.g. publication time of an article) and links associated with an object (in and out edges). To compute temporal similarity of temporal object O_t for XML document we adapted Berberich' T-rank [27] approach by changing the granularity to be objects in an XML document. Given a single

temporal expression q_t in query Q , D is the document to be ranked, Berberich [27] equation defined in T-rank, as follows:

$$P(q_t|D) = \frac{1}{|D|} \sum_{T \in D} P(q_t|T) \quad (5)$$

As shown in “(5)”, the probability of generating the query temporal expression q_t from document D is an average of the probability of generating of time $P(q_t|T)$ divided by the number of intervals in document $|D|$. The probability $P(q_t|T)$ of generating a time interval q_t given a partition time T of a document can be defined in two ways; either by ignoring uncertainty or taking uncertainty into account. This will be illustrated later.

Now we adapt the equation above into temporal XML objects. Each web document D is mapped into a temporal object O_t in XML. Rather than considering temporal expressions of document D , we use temporal expressions which are labeled on edges of an XML graph. In turn, these edges are divided into content and reference edges as defined previously in Section III. Such temporal edges affect the whole similarity degree of the object O_t . We compute the temporal similarity S_t of O_t given query temporal expressions Q_t as shown below in (6).

$$S_t(Q_t, O_t) = \sum_{q_t \in Q_t} S'_t(q_t|O_t) \quad (6)$$

Consequently, S_t is computed as follow:

$$S'_t(q_t|O_t) = \left(\frac{1}{2}\right) \times \left(\frac{1}{|e_{cl}|}\right) \sum_{o_t \in e_{cl}} (score_t(q_t|o_t)) + \frac{1}{|e_{rt}|} \sum_{o_t \in e_{rt}} (score_t(q_t|o_t)) \quad (7)$$

Where $|e_{cl}|$ is the total number of temporal intervals on the containment edges of O_t , $|e_{rt}|$ is the total number of temporal intervals on the reference edges of O_t . To normalize temporal similarity into range [0-1], the score is divided by 2. $score_t$ is used to denote the probability of generated query q_t and object temporal T intervals which will be defined later. There are two ways to compute $score_t(q_t|o_t)$: uncertainty-ignore and uncertainty-aware as defined by Berberich [27].

Uncertainty-ignore mean that the time similarity is one if T and q_t are exactly the same. As shown below.

$$score_t(q_t|o_t) = \begin{cases} 1, & \text{if } q_t = o_t \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

However, temporal expressions can refer to the same time interval even they are not exactly equal, i.e. the relevance of a temporal object may change over time. For this purpose, uncertainty-aware may give approximate similarity. An object with its time partition is closer to q_t will receive a higher probability than an object with time far from q_t . On the other hand, when uncertainty is considered, $score_t(q_t, o_t)$ is defined in “9” as follows:

$$score_t(q_t|o_t) = \begin{cases} \frac{|q_t \cap o_t|}{|q_t| \times |o_t|}, & \text{if } q_t \cap o_t \neq \phi \\ \epsilon, & \text{otherwise} \end{cases} \quad (9)$$

We use ϵ as a very small value to overcome zero issue denoting no common time between the query and the object. The distance of a given interval t , denoted as $|t|$, is computed based on the values of begin and end interval as: $|t| = t.e - t.b + 1$ where $t.b$ and $t.e$ represent the start and end of interval respectively. Intuitively, this function gives a similarity that decreases proportional to the difference between q_t interval and time of object o_t . An object o_t with its time closer to q_t will receive a higher similarity than an object with its time interval far from q_t .

In summary, $S_k(k, O_t)$ gives the contents similarity of keyword with its time in object O_t and $S_t(O_t, Q_t)$ measures the active intervals of object O_t during the specified times in query Q_t .

D. Ranking multiple Objects

A set of RTOs are cooperated to contain all temporal query keywords with specified time while no single object contains all the keywords. Thus we need to compute the whole ranking of the participating objects. Given a query " w_1, \dots, w_m, t " where w represents the keywords and t represent the time predicate, and its corresponding set of interconnected temporal objects RTO (O_1, \dots, O_m). The related objects can be calculated as follows:

$$RTO(o_1, \dots, o_m) = \sum_{O_t}^m (\rho_s(Q_t, O_t)) \quad (10)$$

Where $\rho_s(Q_t, O_t)$ is defined in “(1)” and m is the number of objects.

E. Optimizing temporal ranking into XML environment

The extension of time-aware ranking approach in previous section maps only the flat structure in web pages to XML objects. One advantage of XML documents is its hierarchical nature. In temporal XML model, one object may contain other objects. As a result the object rank is affected by the ranking of its sub objects. To capture transferring of ranks we propose a recursive formula ρ_n to compute XML similarity between a temporal XML object and a temporal keyword search query. Hierarchical structure between objects is captured in “(11)” by distinguishing between two cases. The first (base case) computes the similarity if the object is a single object O , otherwise (recursive case), it recursively computes the similarities for its nested objects, based on the similarity (ρ_n) value of each child chd of O_t .

$$\rho_n(Q_t, O_t) = \begin{cases} \rho_s(Q_t, O_t) & , O_t \text{ is a single object} \\ \left(\sum_{c \in chd(O_t)} \rho_n(Q_t, c) \right) & , O_t \text{ is nested object} \end{cases} \quad (11)$$

VI. INDEX STRUCTURE

We pre-process the temporal XML document D by building a separate structure for each distinct keyword w . This is similar to a regular inverted index.

We precompute single keyword rank in object and combine them during run time.

A. Motivation

Many algorithms were proposed to evaluate XML keyword queries efficiently; SLCA [2] ELCA [3]. The basic idea is to utilize the document order of the nodes in the inverted lists to optimize the semantic pruning. Specifically, nodes in the XML tree are identified by Dewey id and an efficient eager stack algorithm is built. A graph model is used since the tree model does not effectively answer keywords queries. A graph model evaluates queries by finding the minimum connection tree MCT [28].

The XML document is parsed to build two structure indexes; keyword and object lists. The first is used to retrieve the objects whose attributes contain the keywords and the second to track the objects relationships. The indexes are explained in the next subsections.

B. Keyword List

Since a temporal XML database is modeled as a set of interrelated objects with timestamps associated on their containment edges and references (ID/IDREF), the inverted list is more complex than that in the traditional one. Keyword index is composed of tuple $\langle obj_id, K_{list}, TF_k \rangle$ where obj_id is the object identity (usually given in the data set or using dewey id generated automatically by the system) which contains the keyword. However, to efficiently join objects later, we map each obj_id signature with an ordered number. K_{list} is a list of attributes that contain the keyword which is composed of $\langle attr_name, time \rangle$ pairs by specifying the attribute name and the time validity interval of the keyword. TF_k is the term frequency of the keyword k in the object which is computed during the construction of index as shown in the first part of Equation 2. For example, the term frequency of keyword 'Adams' in the object rooted at node 0.2.0 is computed as: $1/7 = 0.14$, where 7 is the total number of keywords in the object. It is more efficient to compute keyword scores during the preprocessing time.

A B+ tree is built based on these objects ids and their time intervals to efficiently retrieve all the objects that contain the query keywords during a specific time.

Complexity size of keyword list is $O(K \times p)$ where K is the total number of keywords and p is the size of a tuple.

C. Object List

Object list is composed of tuple $\langle o_id, Olist \rangle$, where $Olist$ is a list of tuple $\langle I, LObj \rangle$ where I is the interval validity for the connection and $LObj$ is a list of connected objects ids. This list is built during the XML parsing when a reference attribute is encountered. Although this connection is placed in one hop, it is possible to connect objects via other objects. At end of traversing the XML document, a threshold τ is given to determine the number of hops to compute connected objects. Value depends on the application and user intention. For example, for NBA database, τ value will take value 2 to detect any relationships between coaches and players. Additionally, user may (may not) wish to see all the connected objects even these objects are far away.

Any required increase of τ value more than one is performed after finishing document traversal using breadth first search algorithm BFS. For example, player object at node 0.2.1 connects with team 0.1.0 at [1995-1998] by one hop, and with 0.0.0 at [1998, 1998] by two hops.

Temporal ranking of objects is computed on the run time during query processing stage.

Complexity of computing the object list is $O(N \times N)$ in the worst case, where N is the total number of objects in

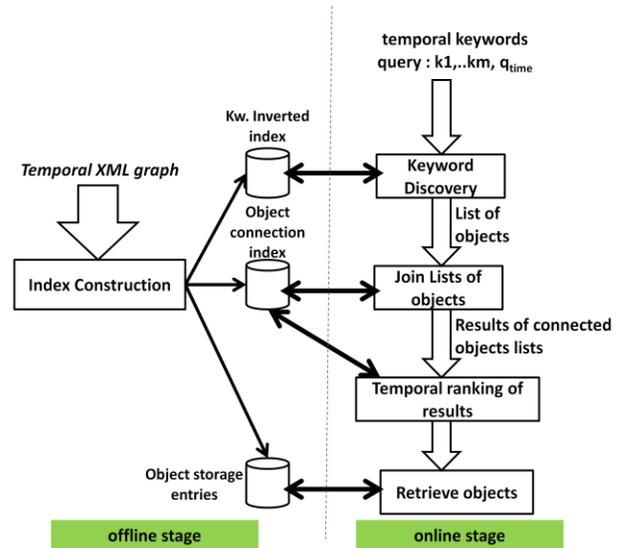


Fig.3. TX-Kw architecture

The document. Regarding the size of our indexes, we store only ids of objects which reduce storage size significantly. Also, storage entries are built to keep track the objects positions in the XML document.

VII. TEMPORAL KEYWORD SEARCH ALGORITHM

Fig. 3 illustrates TX-Kw architecture which consists of two stages: offline and online stages. In the offline stage, indexes are built for keyword and object lists as explained in VI. Query processing is performed in the online stage. When the user poses a temporal query, the query processor parses the query to extract the keywords and time intervals included. For each keyword, the inverted index is checked to extract objects that contain this keyword at the temporal interval specified.

For each list objects, a connection is performed between these objects. Resulting objects are ranked according to the Equation 10.

Algorithm 1 presents keyword searching and result ranking algorithm. qt is the time interval presented in the query. $LL[m]$ contains a list of objects that contain each keyword in the query. Recall that each object contains the object id and TF score of keyword within its object. When the query is issued, Algorithm 1 traverses lists in $LL[m]$ to extract all possible connected objects. First, we initialize the *Cont_Table* used to store the contribution percentage of an object to the query keywords, *RSTO* and *RRTO* used to store single and related temporal objects respectively, and *HT* used to store the

temporal ranking of objects. Then the Algorithm chooses the smallest list in LL as the starting list to join with other lists.

Two main steps the algorithm performs: *Compute_STO* and *Compute_RTO*. *Compute_STO* is called (line 4) which is defined in Algorithm 2 to compute single objects that contain all keywords within time query and stored in *RSTO* with its ranking score. Second, *Compute_RTO* is called (line 5) to extract related objects. Algorithm 3 returns *RRTO* which is a list of connected objects which cooperate to contain all query keywords. Finally we get top-k results by eliminating any repeated objects that exist in the lists and sorting *RSTO* and *RRTO* in descending order. Objects are retrieved from storage entry file to output to the user. Function *Compute_CB* details is omitted due to its simplicity.

Algorithm 3 shows the process to recursively join lists. The input is the *qt* as query time, and starting list *Ls*, *c* is the index of next list in *LL[m]* to be joined with *Ls*. The algorithm pushes each input object *o_l* into a path stack which stores the connected objects, *o_l* is joined with *LL[c]* using *match_obj* function, the returned list is a set of objects that are connected to *o_l*. The result is used in the next iteration to join with next list. Finally, when we finish traverse the lists for a given object, path stack computes their score by calling *CompRankL* and appends its score inside *RRTO*. *CompRankL* function computes the total rank of each object in a list based on two parts; keyword and time as explained in Equation 7, note that the keyword score is already computed while traversing the connected objects. To avoid re-computing temporal similarity, we use a hash table *HT* to store all computed scores of objects.

Algorithm 1 Temporal Kw Search (TX-Kw)

Input: qt:Interval time, LL[m]: object Lists, ObjIdx: object Index

Output: Ranked object(s) list: *RSTO* and

RRTO

- 1: Initialize HT, *RSTO*, *RRTO*, Cont_T able
 - 2: Sort LL[m]
 - 3: L1 ← LL[1]
 - 4: *RSTO* ← *Compute_STO*(L1,LL)
 - 5: *RRTO* ← *Compute_RTO*(L1,1, qt)
 - 6: Sort *RSTO* in descending order
 - 7: Sort *RRTO* in descending order
 - 8: Output *RSTO* and *RRTO*
-

Algorithm 2 Compute_STO

Input: Ls: smallest list, LL[m]: object lists, qt: interval

Output: List of STO object *RSTO*

- 1: for each object *o_l* in Ls do
- 2: cont_o ← *Compute_CB*(*o_l*)
- 3: if cont_o ==1 then /* *o_l* is STO */
- 4: score_{o_l} ← *STO_{rank}*(*o_l*)
- 5: Add <*o_l*, score_{o_l}> to *RSTO*
- 6: Delete *o_l* from LL[m]
- 7: end if
- 8: end for
- 9: function *STO_{rank}*(*o_l*:object)
- 10: score ← φ

- 11: for each list L1 in LL do
 - 12: *o₂* ← find *o_l.id* in L1
 - 13: score = score + *CompRank_o*(*o₂*)
 - 14: end for
 - 15: return score
 - 16: end function
 - 17: function *CompRank_o*(*o_l*:object *o_l*)
 - 18: *S_k* ← *o_l.TF* * log(N/N_k)
 - 19: if *o_l* is not in HT then
 - 20: *S_t* ← *Compute S_t* using *ObjIdx* (Eq. 7)
 - 21: add *o_l* to HT
 - 22: else *S_t* ← HT[*o_l*]
 - 23: end if
 - 24: Score_{o_l} ← *S_t* + *S_k*
 - 25: return Score_{o_l}
 - 26: end function
-

Algorithm 3 Compute_RTO

Input: L1: list, c: counter, LL[m]: object lists, qt: Interval

Output: *RRTO* :List of connected objects:

- 1: for each object *o_l* in L1 do
- 2: path ← push *o_l*
- 3: L2 ← *match_obj*(*o_l*, LL[c], qt)
- 4: if L2 = φ then
- 5: path pop
- 6: continue
- 7: end if
- 8: increment c by 1
- 9: if c ≤ m then
- 10: *Compute_RTO*(L2, c)
- 11: decrement c by 1
- 12: path pop
- 13: else
- 14: for each object *v* in L2 do
- 15: path ← push *v*
- 16: *sc* ← *CompRank_L*(path)
- 17: Add <*sc*, path> to *RRTO*
- 18: path pop
- 19: end for
- 20: decrement c by 1
- 21: path pop
- 22: end if
- 23: end for
- 24: function *match_obj*(object: *o*, list : *L_{in}*, interval: *qt*)
- 25: Retrieve *L_o* connected to *o* in *qt* from *objIdx*
- 26: *L_{des}* ← merge join *L_o* and *L_{in}*
- 27: return *L_{des}*
- 28: end function
- 29: function *CompRank_L*(*L_{obj}*)
- 30: score ← φ
- 31: for each object *o_l* in *L_{obj}* do
- 32: if *o_l* is not in HT then
- 33: *S_t* ← *Compute S_t* using *ObjIdx* (Eq. 7)
- 34: add *o_l* to HT
- 35: else *S_t* ← HT[*o_l*]
- 36: end if

```

37:     cb ← Compute_CB(oi)
38:     Sk ← oi.TF * log(N/Nk) × cb
39:     s ← Sk + St (Eq. 1)
40:     score ← score + s
41:   end for
42:   return score
43: end function

```

We use *match_obj* function to merge join lists to find the connected objects of an input object.

Complexity of Algorithm 1 is based on the complexity of Algorithms 2 and 3. Algorithm 2 traverse smallest list Ls to discover any possible STO objects and its complexity is $O(|L_s| \times \log(|LL|))$. In Algorithm 3, the complexity is $O(|L_s| \times m \log |L_x|)$ in the worst case where m is the number of keywords and L_x is the length of maximum list in LL. The overall complexity of Algorithm 1 is $O((|L_s|) \times (\log(|LL|) + m \log |L_x|))$

VIII. EXPERIMENTAL EVALUATION

In this section, we present different performed experiments to evaluate our approach. We compared our approach to two state-of-art approaches: SLCA [2] as an example of XML tree model and ISO_IRO [7] as an example of XML graph model.

We analyzed our results according to the three metrics that are used to analyze the results of any keyword search experiments; effectiveness, efficiency, and scalability.

A. Setup

We use two real data sets: DBLP [29] and NBA [30]. Table I shows the statistics of such data sets. DBLP contains the major conferences up to year 2002. We build the temporal XML indexes as follows: each conference is considered as a separate object and cite attribute is used to track the relationship between objects. NBA contains all information about players, coaches and teams in USA basketball starting from 1946 until 2008. It is a set of tables in a relational database converted into a single XML document where foreign keys are converted into id/idref attributes. Temporal intervals queries for conventional keyword search are executed for each instant in the interval. Here, we list sample queries for both data sets and its purposes:

User intention: List all coaches train player James Posey in [2005-2008]

TX-Kw: Coach James Posey [2005-2007]

Conventional Kw: Coach James Posey 2005, 2006, 2007

User intention: List all articles written by Elmasri in interval 1990-1993

TX-Kw: article Elmasri [1990-1993]

Conventional Kw: article Elmasri 1990, 1991, 1992, 1993

B. Efficiency

We use the query processing time as the main performance metric. For DBLP we test 7 queries, the first three are non-temporal queries and the rest are attached with one time instant.

The result of the log-scaled run time of tested approaches is shown in Fig. 4(a). We can see that our approach has better performance for temporal queries. However, ISO_IRO is slightly more efficient than TX-Kw for the first non temporal queries (Q1-Q3). The reason is that these queries retrieve the whole objects in the three algorithms but TX-Kw has an overhead computing for temporal ranking. Whereas SLCA is the more efficient keyword search index in XML tree models, its performance degrades in the XML graph model. This is

TABLE I. Data sets statistics

Data set	#nodes	#Kw	#object	Size	Idx size
DBLP	3329043	656387	328858	131MB	781MB
NBA	313251	3742	2517	6MB	23MB

TABLE II. Query keyword for efficiency in DBLP

Qid	DBLP
Q1	Agrawal Databases
Q2	Ling tok wang
Q3	VLDB Agrawal Abbadhi databases
Q4	Concurrency Control Algorithms Distributed Databases 1987
Q5	Ling Tok Wang 1993
Q6	XML Index 2002
Q7	book, java, 2001

TABLE III. Query keyword for efficiency in NBA

Qid	NBA
Q1	Player location Los Angeles 1990
Q2	pts position Robinson 1990
Q3	Coach Kareem Abdul-jabbar 1985
Q4	Colorado Dale Schlueter 1975
Q5	teams Robinson 1999

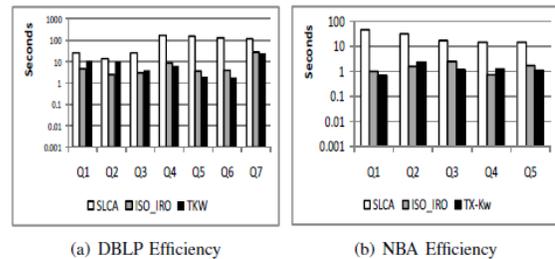


Fig. 4: Execution Time(log-scale)

This is because many keywords may be placed in one object which is easy for TX-Kw and ISO_IRO to distinguish them in the join process. On the other hand, SLCA has to search for each list of nodes that contain the keywords. Such number of nodes may be greater than the number of objects returned by the other algorithms. The worst case of SLCA is clear for temporal queries where more nodes are retrieved than that in TX-Kw.

For NBA, we test five temporal queries, the result of two of them (Q2 and Q4) are STO objects and the the result of the rest (Q1, Q3, Q5) are RTO objects. The performance of TX-Kw, as shown in Fig. 4(b), is better than ISO_IRO approach for RTOs queries since temporal constraints between objects are considered in the join process. ISO_IRO performance gets

slightly better than TX-Kw for STO queries. The reason is that TX-Kw spends time in retrieving objects to filter them according to time before performing the join process. On the other hand, SLCA keeps its worst performance as in DBLP.

C. Scalability

Here, the scalability is measured in two ways: changing the number of keywords and increasing the size of time intervals.

Figs 5(a) and 5(b) show the log-scaled run time as keywords increase with a constant time (one year) for DBLP and NBA. TX-Kw has the best performance overall other approaches. The performance varies according to the number of retrieved objects since some keywords may exist in the same objects and this might lead to reduction in processing time. However, the execution time of SLCA increases as the number of keywords increase.

The second method of measuring scalability is using time interval variation with a constant number of keyword. The performance is shown in Figs 5(c) and 5(d).

We can see that the performance of SLCA and ISO_IRO decreases with the increase of time interval size while it remains approximately constant for TX-Kw. This is because the conventional keywords search require traversing keywords index with each year in the interval while in temporal approach TX-Kw involves only one index traversal. However, there is a slight increase in processing time for TX-Kw depending on the number of retrieved objects as the validity interval increases.

D. Effectiveness

We evaluate quality of results using precision, recall which is heavily used in IR. Precision is the number of relevant objects retrieved divided by the total number of retrieved objects. Recall is the number of relevant objects retrieved divided by the number of relevant objects.

To calculate the precision and recall we manually reformulated tested temporal queries on NBA to be executed into the XML language XQuery and used the results as a basis for evaluation. For DBLP, we tested queries generated by 5 users. For each query the user wrote his attention in natural language and keywords query.

The queries are executed using our approach and ISO_IRO, Table IV shows the average results for precisions and recall for both approaches.

Recall has a high value for both data sets of TX-Kw against ISO_IRO approach since time intervals are not considered in ISO_IRO which lead to empty results in some cases. On the other hand, precision of TX-Kw in Table IV is higher since it detects all possible connections and temporal constraints while ISO_IRO returns more irrelevant results.

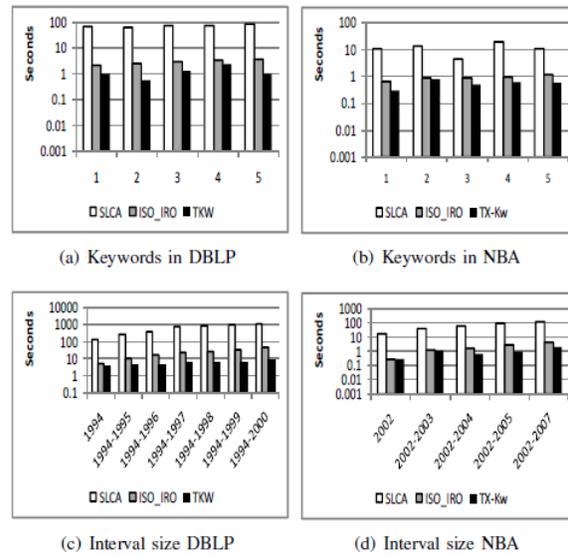


Fig. 5: Scalability Evaluation (log-scale)

Fig.5. Scalability evaluation

TABLE IV. Effectiveness performance

Data set	TX-Kw		ISO_IRO	
	Recall	Precision	Recall	Precision
NBA	0.9575	0.88	0.79	0.304
DBLP	0.978	0.729	0.778	0.432

TABLE V. Ranking performance

Data set	TX-Kw		ISO_IRO	
	R-rank	MAP	R-rank	MAP
NBA	0.906	0.88	0.467	0.3225
DBLP	0.820	0.774	0.695	0.581

We evaluate our proposed ranking method using two popular IR measurements [22]: Mean Average Precision (MAP) and Reciprocal rank R-rank. We use MAP to measure the overall precisions. Precision is computed for each relevant object step. Then we take the average of computed precisions. While R-rank is the inverse of the first rank of correct object retrieved.

We set the mixture parameter to 0.5 to identify weight temporal ranking and keywords ranking. Furthermore, we

Use uncertainty-aware method to compute MAP and R-rank measurements. Ranking performance is shown in Table V. We note that TX-Kw ranking method works very well in NBA where its values metrics are above 0.90 for R-rank and 0.88 for MAP respectively. ISO_IRO ranking has less effectiveness in ranking temporal constraints since its values metrics are below 0.5 for both metrics.

IX. CONCLUSION

We proposed a new approach, TX-Kw, which supports temporal keyword search queries over temporal XML documents. We model temporal XML as interconnected objects by considering containment and ID/IDREF edges. We also utilized time-aware ranking in IR by mapping it to temporal XML as well as providing an algorithm for temporal ranking that captures the hierarchical structure of XML document. An efficient algorithm is proposed to improve the performance retrieval. Finally we conducted experiments to evaluate and compare our approach against existing keyword search methods by measuring their effectiveness and efficiency. The experiments showed better performance of our approach TX-Kw against other state-of-the-art methods. As future work, we consider integrating both keyword and object indexes to enhance efficiency of our approach.

REFERENCES

- [1] Guo, L., et al., XRANK: ranked keyword search over XML documents, in Proceedings of the 2003 ACM SIGMOD international conference on Management of data. 2003, ACM: San Diego, California. p. 16-27.
- [2] Xu, Y. and Y. Papakonstantinou, Efficient keyword search for smallest LCAs in XML databases, in Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005, ACM: Baltimore, Maryland. p. 527-538.
- [3] Xu, Y. and Y. Papakonstantinou, Efficient LCA based keyword search in XML data, in Proceedings of the 11th international conference on Extending database technology: Advances in database technology. 2008, ACM: Nantes, France. p. 535-546.
- [4] Balmin, A., V. Hristidis, and Y. Papakonstantinou, Objectrank: authority-based keyword search in databases, in Proceedings of the Thirtieth international conference on Very large data bases - Volume 30. 2004, VLDB Endowment: Toronto, Canada. p. 564-575.
- [5] Ilyas, I.F., G. Beskales, and M.A. Soliman, A survey of top-k query processing techniques in relational database systems. ACM Comput. Surv., 2008. 40(4): p. 1-58.
- [6] Zhang, R., et al. Learning Recurrent Event Queries for Web Search. in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT State Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. 2010: ACL.
- [7] Bao, Z., et al., An effective object-level XML keyword search, in Proceedings of the 15th international conference on Database Systems for Advanced Applications - Volume Part I. 2010, Springer-Verlag: Tsukuba, Japan. p. 93-109.
- [8] Hristidis, V., Y. Papakonstantinou, and A. Balmin. Keyword Proximity Search on XML Graphs. in Proceedings of the 19th International Conference on Data Engineering. 2003.
- [9] Tian, Z., J. Lu, and D. Li, A survey on XML keyword search, in Proceedings of the 13th Asia-Pacific web conference on Web technologies and applications. 2011, Springer-Verlag: Beijing, China. p. 460-471.
- [10] Cohen, S., et al., XSEarch: a semantic search engine for XML, in Proceedings of the 29th international conference on Very large data bases - Volume 29. 2003, VLDB Endowment: Berlin, Germany. p. 45-56.
- [11] Li, G., et al., Effective keyword search for valuable lcas over xml documents, in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007, ACM: Lisbon, Portugal. p. 31-40.
- [12] Lin, R.-R., Y.-H. Chang, and K.-M. Chao, Improving the performance of identifying contributors for XML keyword search. SIGMOD Rec., 2011. 40(1): p. 5-10.
- [13] Bao, Z., et al., Effective XML Keyword Search with Relevance Oriented Ranking, in Proceedings of the 2009 IEEE International Conference on Data Engineering. 2009, IEEE Computer Society. p. 517-528.
- [14] Klaus, B., V. Michalis, and W. Gerhard. T-rank: Time-aware authority ranking, in Algorithms and Models for the Web-graph : Third International Workshop, WAW 2004. 2004. Berlin, ALLEMAGNE.
- [15] Dai, N. and B.D. Davison, Freshness matters: in flowers, food, and web authority, in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010, ACM: Geneva, Switzerland. p. 114-121.
- [16] Li, X. and W.B. Croft, Time-based language models, in Proceedings of the twelfth international conference on Information and knowledge management. 2003, ACM: New Orleans, LA, USA. p. 469-475.
- [17] Shaparenko, B., et al. Identifying Temporal Patterns and Key Players in Document Collections. in IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05). 2005: Springer.
- [18] Yu, P.S., X. Li, and B. Liu, On the temporal dimension of search, in Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. 2004, ACM: New York, NY, USA. p. 448-449.
- [19] Brin, S. and L. Page, The anatomy of a large-scale hypertextual Web search engine, in Proceedings of the seventh international conference on World Wide Web 7. 1998, Elsevier Science Publishers B. V.: Brisbane, Australia. p. 107-117.
- [20] Jatowt, A., Y. Kawai, and K. Tanaka. Temporal Ranking of Search Engine Results. in WISE 2005. 2005. Berlin Heidelberg: Springer-Verlag.
- [21] Diaz, F. and R. Jones, Using temporal profiles of queries for precision prediction, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004, ACM: Sheffield, United Kingdom. p. 18-24.
- [22] Manica, E., C.F. Dorneles, and R. Galante, Supporting Temporal Queries on XML Keyword Search Engines. Journal of Information and Data Management, 2010. 1(3): p. 471-486.
- [23] Bertino, E., E. Ferrari, and G. Guerrini, A Formal Temporal Object-Oriented Data Model., in EDBT, P.M.G.B. Apers, Mokrane & Gardarin, Georges, Editor. 1996, Springer. p. 342-356.
- [24] Bertino, E., et al., Extending the ODMG Object Model with Time, in Proceedings of the 12th European Conference on Object-Oriented Programming. 1998, Springer-Verlag. p. 41-66.
- [25] Kanhabua, N., Time-aware Approaches to Information Retrieval, in Department of Computer and Information Science. 2012, Norwegian University of Science and Technology. p. 187.
- [26] Salton, G., Automatic text processing: the transformation, analysis, and retrieval of information by computer. 1989: Addison-Wesley Longman Publishing Co., Inc. 530.
- [27] Berberich, K., et al., A language modeling approach for temporal information needs, in Proceedings of the 32nd European conference on Advances in Information Retrieval. 2010, Springer-Verlag: Milton Keynes, UK. p. 13-25.
- [28] Hristidis, V., et al., Keyword Proximity Search in XML Trees. IEEE Trans. on Knowl. and Data Eng., 2006. 18(4): p. 525-539.
- [29] <http://www.cs.washington.edu/>, Xml Data Repository. 2002.
- [30] <http://www.basketballreference.com/>, Basketball database 2.1. 2008.

OntoVerbal: a Generic Tool and Practical Application to SNOMED CT

Shao Fen Liang
Biomedical Research Centre,
NIHR GSTT and King's College London
London, SE1 3QD, UK

Robert Stevens
School of Computer Science,
The University of Manchester, Oxford Road,
Manchester, M13 9PL, UK

Donia Scott
School of Engineering and Informatics,
University of Sussex, Falmer,
Brighton, BN1 9QH, UK

Alan Rector
School of Computer Science,
The University of Manchester, Oxford Road,
Manchester, M13 9PL, UK

Abstract—Ontology development is a non-trivial task requiring expertise in the chosen ontological language. We propose a method for making the content of ontologies more transparent by presenting, through the use of natural language generation, naturalistic descriptions of ontology classes as textual paragraphs. The method has been implemented in a proof-of-concept system, OntoVerbal, that automatically generates paragraph-sized textual descriptions of ontological classes expressed in OWL. OntoVerbal has been applied to ontologies that can be loaded into Protégé and been evaluated with SNOMED CT, showing that it provides coherent, well-structured and accurate textual descriptions of ontology classes.

Keywords—ontology verbalisation; natural language generation; OWL; SNOMED CT

I. INTRODUCTION

Ontologies and terminologies are increasingly authored in languages based on Description Logics [1] such as the W3C-recommended Web Ontology Language, OWL [2], which support formal descriptions and definitions. This approach has two main benefits. First, the description of entities is explicit within the terminology – for example, what the authors mean by the concept of ‘heart disease’ can be made explicit and can be interpreted directly by software rather than depending on each user’s interpretation of the natural language label “*heart disease*”. Second, where no predefined term exists, new descriptions can be formed by composing expressions using existing classes and properties – e.g., a new class of heart complications caused by emerging diseases (SARS, AIDS, etc.) or new drugs or environmental agents. However, the benefit of using description logics comes at the cost of cognitive complexity and unfamiliar notation. For example, the rendering of the concept of ‘heart disease’ could be:

‘Heart Disease’ *EquivalentTo*
 (‘Disorder of Cardiovascular System’) *and*
 (is-located-in *some* ‘Heart Structure’)

Ontological descriptions can be much more complex than

this, comprising conjunctions of statements that themselves include other nested statements.

While such descriptions are explicit, they can be hard for humans to understand – even those who are trained ontologists.

Partly for this reason, some ontologies are annotated with natural language definitions associated to the logical definitions. These give an alternative view on the main entities of an ontology that avoids potentially impenetrable presentations since are easier to understand, especially when written in the style of natural language is that used by the community in question. For instance, the example above could be annotated with the following text:

A heart disease is a disorder of the cardiovascular system that is found in a heart structure.

Writing such natural language definitions by hand is, however, time consuming and there is no guarantee that the meanings of the natural language definitions are the same as the formal logical definitions in the description logic [3]. Also, writing definitions by hand only works for predefined (‘pre-coordinated’) definitions. When definitions arise from compositional use of pre-coordinated concepts (i.e., ‘post-coordinated’ definitions), there is no opportunity to write them by hand. Additionally, ontology authoring is heavily reliant on the intervention of ‘expert’ ontologists, who themselves require many months of experience and training.

One potential solution to this problem is to use existing technologies, such as Natural Language Generation (NLG) from computational linguistics, to produce these natural language descriptions automatically.

This paper describes OntoVerbal, a generic tool that generates automatically natural language descriptions of ontological definitions in ontologies written in OWL. Our aim is to develop a system to produce coherent, reasonably fluent natural language versions of a class’ axiomatisation. The potential advantages of OntoVerbal for ontology users are:

Familiarity: People may be able to access the content of the ontology without having to learn the (often complex)

This work is part of the Semantic Web Authoring Tool (SWAT) project (see www.swatproject.org), which is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/G032459/1, to the University of Manchester, the University of Sussex, and the Open University.

ontology-language in which it is written (e.g., various dialects of OWL, RDF* etc.);

Browsing: Navigating an ontology could be facilitated with a browser that provides the option for natural language presentations of selected parts;

Checking: Consistency- and error-checking are tedious and difficult tasks, relying on deep expertise in the domain that is being modelled and the ontology-language in use. Having access to natural language descriptions of selected parts of an ontology could make it easier to identify the peculiarities and errors in an ontology;

Flexible views: Ontologies are consulted for a range of purposes and by a range of users. Natural language generation provides unique opportunities for tailored presentations of a given ontology (or part thereof), whether by changing the focus of the content or indeed the specific natural language in which it is viewed;

Training: Taken together, the above facilities can provide a rich environment for training in ontology writing.

A key design choice in developing the OntoVerbal system was whether to develop (a) a specialised tool that would produce the best possible natural language for a specific ontology or (b) a generic tool that would produce useful but not necessarily perfect natural language for any OWL-EL ontology (the OWL profile that is more widely used in biomedical terminologies), or at least any ontology for a given field of interest. We chose to be generic, and thus to limit OntoVerbal to what is contained in the ontology itself, with minimum recourse to external linguistic resources or tailoring for specific ontologies. Inevitably, this choice means that the language generated is sometimes stilted and unnatural. OntoVerbal thus produces *acceptable* but not always *perfect* English for any OWL-EL ontology.

Similarly, in the representation of OWL's semantics, while we aim to capture the intuitive meaning of the axioms, it is not our intention to provide a 'tutorial' in OWL semantics. So, unlike ACEView [4], which will, for example, spell out the full implications of a transitive property, OntoVerbal does not verbalise all of OWL's semantics to 'explain' the axioms in its input.

Finally, we aim to produce verbalisations of the descriptions of individual classes rather than summarise an entire ontology. Ontology summarisation is an altogether different task [5], and in any case, it is hard to know what it would mean to 'summarise' an entire ontology of the size of many biomedical ontologies – tens or even hundreds of thousands of classes and millions of axioms.

Our chosen approach thus represents three sets of trade-offs: (a) generic applicability with a minimum of additional resources *vs* greater polish of the generated text, (b) comprehensibility of the generated text *vs* complete representation of the OWL semantics, and (c) class-by-class

verbalisation *vs* any attempt to summarise the ontology as a whole.

II. BACKGROUND

Several initiatives in the field of natural language generation (NLG) have addressed the problem of generating better, more coherent and easily processable texts derived from ontologies — for example, ILEX [6], M-PIRO [7], NaturalOWL [8] and Rabbit [9]. These systems make use of annotated data or users' interactions to construct sentences and paragraphs. However, using annotated data and user interaction for text generation does not necessarily help the overall task of ontology comprehension: manually annotating axioms with information to guide language generation works well, but it is time-consuming and requires skill and training; reliance on user input for tasks such as sentence-ordering presents similar issues. Our approach does not preclude the use of such resources to improve the language if so desired, but it does not require it.

Ontologies define the entities in a domain of interest. They are typically authored and presented in groups of axioms (known as 'frames,' sometimes referred to as 'concepts') relating to a single entity. Strictly speaking, the order of axioms in OWL is irrelevant to its meaning, and there is no formal notion of a frame. However, most OWL editors (e.g., Protégé [10], TopBraid Composer [11], Swoop [12] and the NeOn Toolkit [13]) group axioms together into frames as an organisational device to aid modelling and comprehension.

This suggests that grouping sets of axioms is a useful notion. The intuitive textual correlate of a single axiom is a sentence, and that of a frame is a paragraph. When authoring and reading ontologies in natural language, therefore, we focus on paragraphs: the coherence of a textual description of an ontology class, and consequently its comprehensibility, would be increased by grouping ontology (axiom-) sentences together in to topics or units of thought such as (frame-) paragraphs. Within such paragraphs, avoiding repetition by aggregating sentences that conform to regularities should also help.

On top of this, it would also be helpful to know and exploit the way in which many different axioms are asserted in an ontology, since these suggest what are natural or commonplace schemas in ontology construction and thus may provide useful indications of how the corresponding texts should be structured and ordered.

III. ONTOLOGY AXIOMS

In order to present textually the content of ontologies in a well-ordered, well-structured and fluent manner, we need a clear understanding of the groupings and orderings of axioms that are most typical of ontologies; in other words, we need to know how the ontology community use ontology axioms as a language for describing a domain.

A. Classifying axioms

If we read axioms from a natural language point of view, we could find that different axioms play different communicative roles. For example:

* <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/>

SubClassOf: indicates that one class is a sub class of another class; these axioms place classes into a taxonomic structure.

EquivalentClass: indicates that the classes listed are equivalent; these axioms provide definitions of classes.

DisjointClass: indicates that one class is different from other classes; these axioms describe the distinctiveness between classes.

ClassAssertion: indicates that an instance is a member of a class; these axioms provide an illustration of a class.

DisjointUnion: indicates that a set of classes are exhaustive and are all distinct from each other; these axioms present the alternatives within a given class.

Given the distinctiveness of axioms' respective communicative function, each category of axiom will obviously require different expressions when they are translated into natural language text. A further, orthogonal, classification to consider is the following:

Simple axioms: state relations between named classes. For example,

'Disease' (disorder) SubClassOf
Clinical finding (finding)

Complex axioms: contain not only named classes, but also properties, cardinalities or value restrictions, or combinations of named classes in anonymous class expressions, such as:

'Heart Disease' *EquivalentClass*
('Disorder of Cardiovascular System') and
RoleGroup some (Finding site some 'Heart Structure')

Simple axioms map quite naturally onto simple, declarative, natural language sentences, while complex axioms are likely to be reflected in sentences that are more syntactically and rhetorically complex, and which are often also longer.

B. Presenting axioms

The most common views for ontology editing tools to present each class are usage-based and frame-based views. The usage-based view will present to the user every axiom relating to a designated class, whereas tools that make use of the frame-based view will present axioms in distinct categories (e.g., equivalent classes, super classes, class members or disjoint classes). The usage-based view has the advantage of completeness, but the disadvantage of having an unstructured presentation (since axioms are presented in an unordered and unstructured manner). The frame-based view, on the other hand, while making use of a clear structure and ordering, does not provide complete information about the designated class, and the user is thus left to carefully check the complete set of axioms to achieve a full account of the designated class.

Our chosen approach is to model our natural language generation process using the best of these two worlds: for completeness we will follow the usage-based view, presenting all axioms relating to a given class; for comprehensibility, we

will follow the frame-based view, by providing a clear structure to the set of axioms.

C. Structuring axioms

Given the diversity, complexity and variety of communicative goals of the axioms in an ontology, the issue of how best to present them as a comprehensible text is not trivial. To guide this process, we have undertaken an extensive survey of how common axiom groups relate to a designated class.

TABLE I. LABELS USED TO CLASSIFY AXIOM GROUP

Axiom group	Simple	Complex
ClassAssertion	Ca	Car
DisjointClass	Dc	Dcr
EquivalentClass	Ec	Ecr
SubClassOf	Sc	Scr

We first developed a simple code for labelling axiom groups, assigning two characters to each simple axiom group and three characters to each complex axiom group (see examples in TABLE I). The two characters of each simple axiom group are the first and second capital letters from the axiom type; so for example:

ClassAssertion(John-Joe, Person) is labelled Ca.

The complex axiom groups have an additional character "r"; so for example:

ClassAssertion(John-Joe, Person (and has-Gender(male))) is labelled Car.

We focussed our survey on a set of 490 ontologies collected from the TONES Repository*, Swoogle†, and Ontology Design Patterns‡. Their Unified Resource Identifiers (URIs) were checked to ensure that our collection was a non-redundant set. Our task here was to determine how many of the axiom groups in this set belong to a designated class.

We therefore gathered, for each of the above-mentioned ontologies, all axioms that contained a designated class and grouped them according to their assigned categories. So, for example a designated class containing axioms of Ec, Scr and Ecr categories would be labelled as EcEcrScr (ordering the categories alphabetically).

This process led to the identification of more than 60 patterns from 268,969 classes of the 490 ontologies, showing that a class can be represented in a variety of structures ranging from a single simple axiom group to a combination of several simple and complex axiom groups. The most common pattern contains only Sc axioms, and is found in over 56% of the classes. The frequency drops sharply to 16.68% for individual Scr axioms and 13.59% for the combination ScScr. Together, these three patterns account for over 87% of

* <http://owl.cs.manchester.ac.uk/repository/>

† <http://swoogle.umbc.edu/>

‡ <http://ontologydesignpatterns.org/ont/>

ALTERNATIVE, etc. [16; 17]. They can also be signalled by punctuation (e.g., a colon for elaboration, comma between the elements of list, etc.).

Our exploration of RST has shown that some relations appear to map well to the characteristic features of ontology axioms. For example:

- the LIST relation captures those cases where a group of axioms in the ontology bear the same level of relation to a given class;
- the ELABORATION relation applies generally to connect different notions of axioms to a class (i.e., super-, sub- and defining- classes), in order to provide additional descriptive information to the class;
- The CONDITION relation generally applies in cases where an axiom has property restrictions.

Our experience and the evidence over many practical cases have indicated that the full set of rhetorical relations is unlikely to be applied to ontology verbalisation. In particular, the set of so-called presentational relations [18] are unlikely to apply, as ontology authors do not normally create comparisons or attempt to state preferences amongst classes. In addition, even within the set of informational relations, there are several that will not be found in ontologies. For example, since each axiom is assumed to be true, using one axiom as an evidence of another axiom would be redundant. Similarly, using one axiom to justify another axiom is not a conventional way of building ontologies.

A. From Axiom-sentences to Axiom-paragraphs

As mentioned earlier, the common approach for translating ontology axioms to natural language is to translate one axiom per sentence [19]. This approach often leads to repetitions and other infelicities; for example, a class can have many sub-classes and translating these sub-class statements into a string of sentences will lead to text that is not only inelegant through its repetition, but also tedious to read [20]. Psycholinguists have long shown that such texts impose an unnecessary cognitive overhead for the reader (see e.g., [21]). Therefore, combining sentences becomes an important issue for avoiding monotony and to aid ease of comprehension. This process is a core linguistic task for any natural language generator that aims to produce fluent text [22], and is thus one that we attempt to utilise in our work.

Transforming individual sentences or clauses into a single, complex sentence involves a process of compounding sentences that focuses on combining subjects, objects and verbs from component sentences or adding punctuation between clauses. In ontology axioms, we can find constructs that map directly onto the linguistic notions of subject, object and verb. For example, the axiom $A(X, P)$ has X as its subject and P as its object; A presents a predicate that holds between X and P , typically expressed in English through a verb. This allows us to apply linguistic operations of sentence (or clause) combining, commonly referred to within computational linguistics as *aggregation* [23; 24], to strings of axioms. Thus, if we have several axioms in the SubClassOf (Sc) group, then we can combine their subjects to generate a compound

sentence. Consider, for example, the following three axioms represented as individual sentences:

SubClassOf (X, P) → “X is a kind of P.”
SubClassOf (X, Q) → “X is a kind of Q.”
SubClassOf (X, R) → “X is a kind of R.”

Through the process of aggregation they can be combined to make a single sentence by keeping the same subject and removing the repetition between subject and object, then using a “comma” and an “and” (both discourse markers of the LIST relation that holds between the three axioms) to join objects as the following sentence:

“X is a kind of P, Q and R.”

We can extend this approach to produce a range of other types of expressions. For example, if the ontology also included the axiom

SubClassOf (Z, X) → “Z is a kind of X.”

that introduces an indirect relation to the subject X , we can use a simple linguistic operation (equivalent to making an active sentence into a passive) to swap subject and object and replace the predicate with its inverse and in doing so produce an appropriate textual expression:

SubClassOf (Z, X) → “A more specialised kind of X is Z.”

By iterating the process of aggregation with the complex sentence that we already produced above, we are able to derive a two-sentence paragraph with a consistent focus to cover all four axioms as:

“X is a kind of P, Q and R. A more specialised kind of X is Z.”

If we analyse the RST relation in the above example, we will find that its two sentences are in an ELABORATION relation. This is a simple example of the feasibility of transforming axiom-sentences to axiom-paragraphs using RST relations. However, it only illustrates the SubClassOf group. Thus, the next step is to build an RST schema to cover patterns that are required for our purpose.

B. Building an RST schema for ontology classes

Our first step in building an RST schema for describing an ontology class is to examine axiom groups in more detail to enable the process of combining multiple axioms into complex sentences (i.e., through aggregation). First of all, we refine the key axiom groups identified at the end of Section D by splitting them into their direct (signalled with a subscript 1) and indirect (signalled with a subscript 2) counterparts — all, that is, except axioms relating to ClassAssertion (Ca), since they are all in direct relations with the designated class (F). The resulting 11 advanced groups are then placed into our RST schema (

);

- the simple-direct category contains SubClassOf (Sc1), EquivalentClass (Ec1) and DisjointClass (Dc1) axioms in their simple and direct forms;

- the complex-direct category contains SubClassOf (Scr1) and EquivalentClass (Ecr1) axioms in their complex and direct forms;
- the simple-indirect category contains SubClassOf (Sc2), EquivalentClass (Ec2) and DisjointClass (Dc2) axioms in their simple and indirect forms;
- the complex-indirect category contains SubClassOf (Scr2) and EquivalentClass (Ecr2) axioms in their complex and indirect forms;
- ClassAssertion (Ca) axioms form a category of their own category.

The axiom groups belonging to each category follow the ordering of the axiom group suggested by the analysis in **Error! Reference source not found.** As we have illustrated before, SubClassOf axioms belonging to the simple-indirect category (Sc₂) can be converted to simple-direct (Sc₁) by swapping its subject and object; the same is true for DisjointClasses (Dc₂). For example, the axiom:

DisjointClass (A, B, C, F)

is in a direct relation to A, but in an indirect relation to F. If we want to change the focus to F, we can transform the axiom to become:

DisjointClass (F, A, B, C),
DisjointClass (F, B, C, A)

and so on. Such transformations do not affect the underlying meaning of the axiom, but they do change the focus of the resulting natural language expression, so that, for example:

“F is different from A, B and C.”

Has the same meaning as

“F is different from B, C and A.”

Turning now to our method of expressing axiom classes in English, we use a template-based NLG technique. Our choice in this is driven by an attempt to translate each axiom such that we preserve its meaning in the ontology and avoid introducing misleading information.

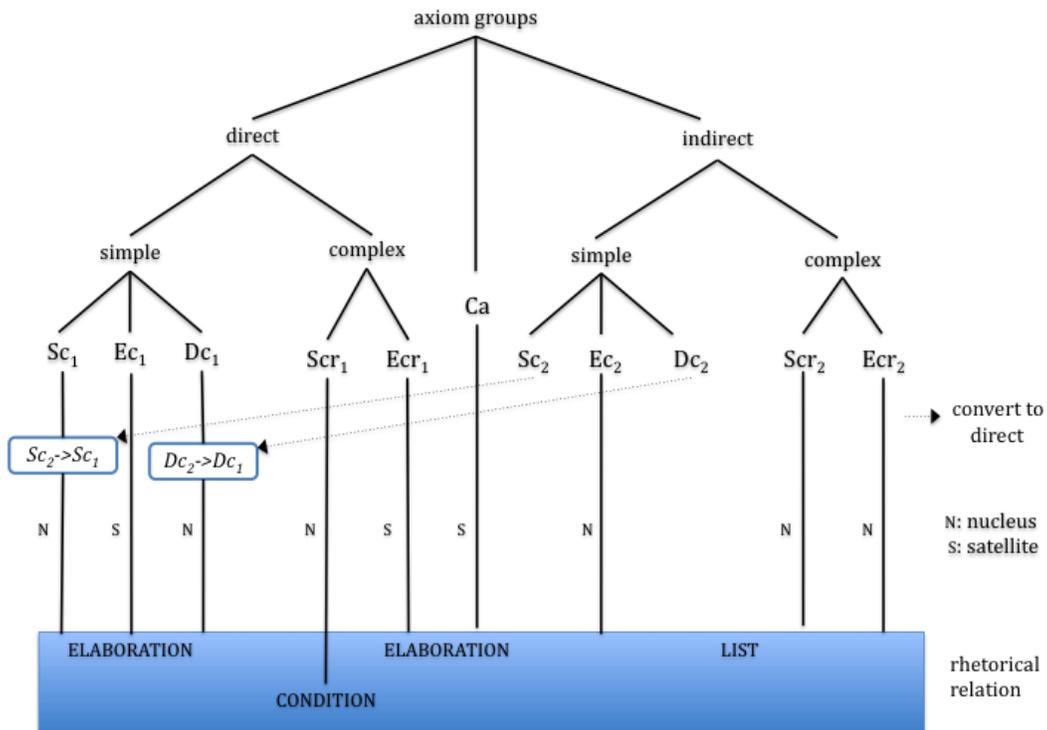


Fig.3. top level RST schema for ontology classes

For example, sub- and super-classes are usually in an ‘is a’ relation. However, translating them into the English string “is a” can lead to misunderstanding since that expression can be used in English to mean “equal to” or “is the same as”; clearly, though, a class is not equal to or the same as its super-class. In this context, a more accurate translation is “is a kind of” or “is a type of”.

C. The RST paragraph

In the simple-direct category, Sc₁ and the converted Sc₂ groups are always the starting description of a class and form the main Nucleus of the paragraph within a top-level elaboration relation. Ec₁ is the associated satellite, followed by Dc₁ and the converted Dc₂. When verbalised, the resulting text is along the lines of:

“F is a kind of X and Y. A more specialised kind of F is Z, and F is defined as P and Q. Also F is different from R.”

In the complex-direct category, Scr_1 is the Nucleus and the Ecr_1 is its Satellite within a condition relation. As the Ca group is on its own, it has been combined into this category as a satellite within an ELABORATION relation. Together, this would lead to a verbalisation such as:

“F is a kind of L that..., and is defined as M that..., and has members $N_1, N_2 \dots N_m$.”

We use “additionally” to connect these two categories to make sentences coherent, which leads to a verbalisation such as:

“F is a kind of X and Y.... Additionally, F is a kind of L that ... N_m .”

The last part is the simple-indirect and complex-indirect categories. All groups in this part are nuclei, and are in a LIST relation since each is an independent description. As these groups are in an indirect relation to the designated class, the subject in this part of the text is no longer the same as in the previous texts. Thus, we need to introduce some connection words to improve the coherency. For example:

“Another relevant aspect of F is ...” (for singular)

“Other relevant aspects of F are: ...” (for plural)

These groups may contain several indirect and complex axioms and without a clear boundary between these axioms, the text can be difficult to read. However, our use of RST, together with the Theory of Discourse Structure [25], allows us to introduce further discourse markers, and even layout elements such as bullet points to make the logical structure of the axiom class more apparent. So the overall content of a class could look like:

“F is a kind of X and Y.... Additionally, F is a kind of L that ... N_m . Other relevant aspects of F are:

- T is a kind of U that ... something to do with F...;
- V is defined as W that ... something to do with F...;
-”

V. ONTOVERBAL

We have implemented the above methods in a proof-of-concept system, OntoVerbal that takes as input OWL ontologies and produces as output a translation from the OWL into English*.

As mentioned earlier, OntoVerbal is a generic tool, and thus the output is not ‘pretty’ English, in that it can be stilted, pedantic and repetitive – characteristics that often come from the nature of the input itself.

For example, ontologies are inherently repetitious and often overly explicit compared to what we would expect of ‘good’ text (an ontology is likely to state, for example, that a primary school is a type of school or that a left hand is a type of hand); if we remain true to our goal of achieving fidelity

vis a vis the input, the generated text will necessarily contain these stylistic infelicities.

Another limiting factor is that, while clearly English, the names of ontological concepts are often technical telegraphese; for example, the names of concepts in the SNOMED CT ontology [26; 27] tend not to trip lightly off the tongue – names such as “bone structure of clavicle and/or scapula and/or humerus”, or “hypertensive heart and renal disease complicating and/or reason for care during the puerperium”.

A. OntoVerbal Paragraphs

We have tested OntoVerbal with a range of ontologies including:

a) The Travel ontology[†], which describes travelling modes, destinations, the boundaries of the destinations and so on.

b) A module of SNOMED CT[‡], which describes medical terminology covering most areas of clinical information such as diseases, findings, procedures, microorganisms, substances, etc.

c) The Experimental Factor Ontology (EFO)[§], which describes experimental variables (e.g. disease state, anatomy) based on an analysis of such variables used in the ArrayExpress database.

These examples of OntoVerbal’s input and output are shown in TABLE II. Although our textual paragraphs are not in perfect English, these ontology-entity-embedded-paragraphs are familiar to ontologists.

While our method allows for further improvement of the fluency of the generated paragraphs; the extent to which such improvements are desirable or necessary depends on individual needs. Nevertheless any improvements that do not require heavy additional resources would be welcome. In what follows we describe how we have achieved this, with reference to SNOMED CT.

B. Tuning OntoVerbal for SNOMED CT

SNOMED CT is a terminology used for coding in health records and is mandated for use for various purposes in numerous countries around the world**.

Although the native form of SNOMED CT is a description logic that predates OWL, it is available as OWL and conforms to the OWL EL profile (with the exclusion of disjointness). SNOMED CT provides a useful test-bed for OntoVerbal: it is axiomatically rich, it has no natural language definitions, it has a large user base, and it is employed at many stages by users who have limited experience in OWL.

[†] <http://swatproject.org/ontologies.asp>

[‡] For this we used the tool at <http://owl.cs.manchester.ac.uk/snomed/> with the signature ‘hypertension’.

[§] www.ebi.ac.uk/efo/

** <http://www.ihtsdo.org>

* In fact, the system is multilingual, producing both English and Mandarin, but we will focus only on the English here.

TABLE II. EXAMPLES OF ONTOVERBAL'S INPUT AND OUTPUT

Travel ontology	
Input	(Settlement SubClassOf AdministrativeDivision) (City SubClassOf Settlement) (Town SubClassOf Settlement) (Village SubClassOf Settlement)
Output	A settlement is a kind of administrative division. More specialised kinds of settlement are city, town and village.
SNOMED CT ontology	
Input	(Benign hypertensive renal disease (disorder) SubClassOf Hypertensive renal disease (disorder)) (Benign arteriolar nephrosclerosis (disorder) SubClassOf Benign hypertensive renal disease (disorder)) (Benign hypertensive heart AND renal disease(disorder) SubClassOf Benign hypertensive renal disease (disorder)) (Benign hypertensive renal disease (disorder) SubClassOf (Hypertensive renal disease (disorder) and (RoleGroup some (Finding site (attribute) some Kidney structure (body structure))))))
Output	Benign hypertensive renal disease(disorder) is a kind of hypertensive renal disease (disorder). More specialised kinds of benign hypertensive renal disease(disorder) are benign arteriolar nephrosclerosis (disorder) and benign hypertensive heart and renal disease (disorder). Additionally, benign hypertensive renal disease (disorder) is a kind of hypertensive renal disease (disorder) that rolegroup a finding site (attribute) a kidney structure (body structure).
Experimental Factor Ontology	
Input	(caudate nucleus SubClassOf cranial ganglion) (caudate nucleus SubClassOf (part of some basal ganglion))
Output	A caudate nucleus is a kind of cranial ganglion. Additionally, a caudate nucleus is a kind of part of a basal ganglion.

There are many versions of SNOMED CT. We tested OntoVerbal with the July 31 2010 version of the International SNOMED CT. Since the complete ontology is large (292012 classes, 62 object properties) we focussed on a module on hypertension (high blood pressure^{*}). Hypertension has many causes and effects, so this module contains a wide range of diseases and anatomic structures: it comprises 506 concepts, each corresponding to a separate OWL class[†]. In addition to its unique identifier which is a 64 bit integer, each concept/class has two associated names: a ‘fully specified name’ and a ‘preferred term’ which are natural language expressions such as those shown in TABLE III.

TABLE III. Examples of SNOMED CT names

SNOMED CT ID no.	Fully-specified name	Preferred name
118698009	Procedure on abdomen (procedure)	Procedure on abdomen
280129003	Disorder of soft tissue of thoracic cavity (disorder)	Disorder of soft tissue of thoracic cavity
63337009	Lower trunk structure (body structure)	Lower trunk structure
11511004	Hypertensive heart AND renal disease complicating AND/OR reason for care during puerperium (disorder)	Hypertensive heart AND renal disease complicating AND/OR reason for care during the puerperium

^{*} For this we used the tool at <http://owl.cs.manchester.ac.uk/snomed/> with the signature ‘hypertension’ .

[†] The OWL version includes a construct – RoleGroup – that is relevant to only a small number of concepts but has to be included in all for consistency. The Hypertension module was chosen, in part, because RoleGroups were always irrelevant, and they were ignored.

For simplicity, we make use of the preferred name. However, the naming conventions for SNOMED CT are complex — not least because this terminology is the result of combining two earlier terminologies, and many different editors and authors have been involved — and this poses particular challenges and limitations for the verbalisation process. As can be seen in TABLE III, names are often expressed in a kind of ‘telegraphese’, often involving missing articles; they can also be quite long and logically complex. Since concepts/classes obviously map onto nouns in natural language, when names such as these are used verbatim as part of the verbalisation process, they can lead to rather awkward text.

To achieve fully fluent verbalisation would require decoding names into their underlying semantics and re-generating them as more contextually appropriate nominal expressions; however, this would require making use of both linguistic and domain knowledge that is not available in the ontology itself.

For example, the process would have to model (a) the complex mapping between nouns and their articles (i.e., when to use “the”, “a”, “an”) and (b) human anatomy. Armed with the information that the body contains only one heart and one pelvis but several branches of the aorta and several arteries, the verbaliser would then be able to produce the appropriate expressions: “the heart”, “the pelvis”, but “a branch of the abdominal aorta” and “an artery”, etc. Given that our aim is to achieve a generic verbalisation tool, this is clearly beyond the scope of our work — although were such a translation module to exist, OntoVerbal could make use of it[‡]. In the absence of such a module, we have created a somewhat crude version by (a) relying on the use of articles as applied to anatomical terms in Wikipedia[§] and (b) consulting the official list of text definitions for SNOMED CT^{**} to find, for cases where the name of an anatomical entity in the ‘preferred’ name does not include an article, another instance of the same entity which does include one (e.g., “central nervous system” in one instance, but “the central nervous system” in another). When any missing articles were found, they were added to the SNOMED CT input. With the help of these adjustments, we are able to reduce to some extent the awkwardness of the verbalised concept names. In all other respects, though, the verbalisation of the SNOMED CT ontology makes use of the same natural language generation engine that applied to the other ontologies that we have tested.

VI. EVALUATION

OntoVerbal has now been implemented as a Protégé plugin^{††} that can offer an alternative textual view of a class

[‡] Such a process has been undertaken to great effect in the Spanish version of SNOMED CT.

[§] http://en.wikipedia.org/wiki/List_of_human_anatomical_features

^{**} Given in the file named sct1 TextDefinitions en-US 20110731.txt from the SNOMED CT Technical Implementation Guide that can be downloaded via http://ihtsdo.org/fileadmin/user_upload/doc/directory.html

^{††} OntoVerbal is available to downloaded from <http://swatproject.org/demos.asp>

alongside the Manchester syntax view and the graphical views provided by various plugins. We feel the need to undertake a formal evaluation to address two key issues:

Fidelity: Are the generated paragraphs faithful to their input? In other words, does the textual output of OntoVerbal convey a clear and true expression of that which is contained in the corresponding input? One way to test this is through a ‘round-trip’ study: given only the generated output, can a proficient ontologist re-create the semantics of the input from which it was derived?

Quality: Are they of reasonable quality? Ideally, the generated paragraphs should be of a standard that is not far from that which one would expect from a proficient ontologist given the task of rendering the input as text.

A. Experimental set up

We addressed these questions through an on-line experiment with OntoVerbal applied to SNOMED CT. To avoid making the study too easy or too hard by including very short or very long paragraphs, we narrowed the set under consideration to typical classes in the SNOMED CT ontology, which we found to contain between 3 and 5 axioms. From these we randomly selected 10 for the study.

We created two textual versions of each of the 10 selected classes (see Appendix A): one version was created by OntoVerbal; the other was written by an independent expert ontologist proficient in OWL (with Manchester syntax)[†], under instructions to “transform them into ‘fluent’ English paragraphs that (a) are semantically equivalent to the OWL and (b) another OWL expert could in principle use to re-create the original OWL”. The ontologists were instructed to use the SNOMED CT labels verbatim (through cut-and-paste); this means that if there was a missing article in the SNOMED CT labels (e.g., “artery of the abdomen” rather than “an artery ...” or “the artery ...”), they would reproduce this in the text – as would OntoVerbal.

We used these 20 texts to design two sets of materials: Set A contained verbalisations of classes 1–5 by the ontologist and 6–10 by OntoVerbal; Set B contained the verbalisations in the other order. With these materials, we conducted an on-line experiment, collecting data from 30 participants who were fluent in OWL EL with Manchester syntax. Half of the participants received Set A, and the others Set B. Each was shown all 10 verbalisations in random order (per participant), with instructions to write the equivalent OWL code. They were instructed to use the SNOMED labels (which were highlighted in the text) verbatim rather than attempt to transform them, and were allowed to cut-and-paste them directly into their code.

B. Results

We analysed participants’ responses by comparing the code they produced to the OWL input that led to each (machine and human) verbalisation. Since there will be a number of semantic equivalents to each SNOMED CT class

descriptions[†], for each of the 10 chosen class descriptions we created all their semantic equivalents. For each response to the presented paragraphs, we measured the Levenshtein distance [28] between the code produced by participants and each member of the set of semantically equivalent versions of the OWL input to that paragraph[‡]. To normalise for the length of class descriptions, we applied the following operation:

$$\text{Similarity} = (\text{Length of string} - \text{Levenshtein distance}) / \text{Length of string}$$

Which returns a value of 1 if the code produced is a perfect match to the input (or one of its semantic equivalents), and 0 if there is no match at all. Since the order of axioms in a class description is not a relevant factor, we treated each axiom independently, and registered the mean value over all axioms in the class. Finally, for each response by each subject, we recorded only the highest score received against all members of the set of semantically-equivalent versions of the OWL input. The results of this analysis are shown in TABLE IV.

1) Fidelity of OntoVerbal’s output

If the paragraphs produced by OntoVerbal are a clear and true expression of the OWL code that it receives as input, participants should be able to re-create the input code or some semantically equivalent version of it. This will be reflected in Similarity scores that are close to 1. Our results show that the mean score over the 10 class descriptions is 0.94. This is an extremely encouraging result, indicating that participants in the study were able to successfully ‘translate’ the paragraphs generated by OntoVerbal back into the OWL from which they were derived. From this we can conclude with confidence that the output of OntoVerbal is faithful to its input, in the sense that it conveys the correct semantics, since a ‘round-trip’ is clearly achievable.

TABLE IV. MEAN SIMILARITY SCORES FOR THE 10 CLASS DESCRIPTIONS

Class	OntoVerbal	Ontologist
1	0.94	0.69
2	0.93	0.80
3	0.92	0.88
4	0.95	0.84
5	0.95	0.78
6	0.94	0.86
7	0.89	0.81
8	0.93	0.95
9	0.96	0.83
10	0.94	0.85
Mean	0.94	0.83

[†] For example, the axiom “A SubClassOf B and C and D” would have 16 semantically equivalent versions.

[‡] We ignored all differences relating to layout (e.g., line breaks or indentation) or to case and punctuation (e.g., ‘SubClassOf:’ vs ‘subclass of’).

* The main author of the EFO ontology (www.ebi.ac.uk/efo).

2) Quality of the verbalizations

The high Similarity scores achieved for the output of OntoVerbal suggest that the texts it produces are of good quality for the purpose for which they are intended. Another strong test of the quality of the verbalisations produced by OntoVerbal is the extent to which they compare with verbalisations produced by the expert ontologist, given the same task; one would hope that the mean score for the two versions of each class description (i.e., machine vs human author) would be close. Our results show that although participants were able to translate successfully the paragraphs written by the ontologist (mean Similarity score is 0.83), their performance was consistently below that for the generated texts. Comparing the two statistically, the difference (human *versus* OntoVerbal) is highly significant (mean diff = -.106, t (two-tailed) = -8.025, p < .001).

In other words, the machine-generated verbalisations were better suited for a 'round-trip' than the (probably 'better English') human-written equivalents. As mentioned, the materials for the study were divided into sets, so that all participants saw all 10 paragraphs, but no participant saw both the human-written and OntoVerbal-generated versions of any given paragraph.

This was intended to reduce any bias arising from the style or naturalness of the texts presented, and indeed our statistical analysis of this show no significant effect of the set (mean diff = -.033, t (two-tailed) = -2.055, p < .06).

VII. CONCLUSION

The question we sought to address in this work is whether ontology classes can be transformed into readable and reasonably fluent natural language text by an automatic process that is itself reasonably generic. Our experience has provided positive proof that this is the case. We have shown that natural language generation (NLG) technology, enhanced by a discourse planner based on Rhetorical Structure Theory, can transform classes represented in OWL into coherent and fairly fluent paragraphs even in the face of strong constraints (e.g., retaining the textually awkward labels in the ontology and eschewing the use of special purpose linguistic resources). Our NLG architecture does not, of course, require these constraints to be in place: as more sophisticated processes and resources are added, the resulting texts will become closer to that normally found in everyday use. However, the question of just how 'natural' the generated text should be remains open. Indeed, our evaluation study has shown that in some contexts at least, 'natural' (as in human-generated) is not always best. One suspects that the preferred style of the generated text will depend on its intended use, for example, whether for ontology -checking, -browsing, -authoring or training. An advantage of the architecture we have developed for the NLG task of translating OWL classes into text is that it provides the flexibility to tune the style of the generated text.

We have tested our approach in a proof-of-concept system, OntoVerbal, which has also been applied to several ontologies that cover a number of domains, demonstrating

that coherent verbalisations can be produced for ontologies within the portion of OWL roughly corresponding to OWL EL. Elsewhere we have shown that the approach also works for other natural languages [29].

Most presentations of OWL ontologies take the form of an OWL syntax along with a visualisation of the ontology's graph or a (manually-written) textual summarisation of the ontology. Verbalisations of the classes in an ontology, such as those provided by OntoVerbal, offers another style of presentation and one that could be a useful counterpart to the overview afforded by a typical graphical presentation. Visualisation tools such as OWLViz* and OntoGraf† show the the classes in an ontology, but do not give much detail: OWLViz shows only the subclass hierarchy, and OntoGraf does not discriminate between the different quantifications on properties. A class verbalisation could fit neatly into this range of presentations: it gives a presentation of detail, but in a form familiar to users. It is possible to imagine a hybrid presentation with graphical overviews, textual summaries, and textual presentations of the classes' axiomatisation. Protégé, for example, supports many graphical visualisers (described in [30]) that are used to support a number of tasks.

Finally, although our effort has focussed on OWL, there is no a priori reason why it could not be extended to an arbitrary Resource Description Framework (RDF) graph: OntoVerbal is topic-centric (grouping axioms around a given topic) and RDF graphs have the same mechanism (grouping axioms on common URI), thereby making it possible to extract a graph on a topic and verbalise it. Mapping the RST roles onto triples in an RDF graph outside RDFS is, however, an open question.

REFERENCES

- [1] F. Baader, I. Horrocks, and U. Sattler. "Description logics as ontology languages for the semantic web". Lecture Notes in Artificial Intelligence, vol. 2605, pp. 228–248, 2005.
- [2] I. Horrocks, P.F. Patel-Schneider, and F.v. Harmelen. "From SHIQ and RDF to OWL: The making of a web ontology language". Journal of Web Semantics, vol. 1, pp. 7-26, 2003.
- [3] R. Stevens, J. Malone, S. Williams, R. Power, and A. Third. "Automating generation of textual class definitions from OWL to English". Journal of Biomedical Semantics, vol. 2(Suppl 2):S5, 2011.
- [4] T. Kuhn. "The understandability of owl statements in controlled english". Semantic Web journal, 2012.
- [5] E. Motta, P. Mulholland, S. Peroni, M. d'Aquin, J.M. Gomez-Perez, V. Mendez, and F. Zablith. "A Novel Approach to Visualizing and Navigating Ontologies". Proceedings of In International Semantic Web Conference, ISWC 2011, Springer-Verlag, 2011.
- [6] O.D. Michael, C. Mellish, J. Oberlander, and A. Knott. "ILEX: an architecture for a dynamic hypertext generation system". Natural Language Engineering, vol. 7, pp. 225-250, 2001.
- [7] A. Isard, J. Oberlander, C. Matheson, and I. Androutsopoulos. "Speaking the users' languages". IEEE Intelligent Systems Magazine, vol. 18, pp. 40-45, 2003.
- [8] D. Galanis, G. Karakatsiotis, G. Lampouras, and I. Androutsopoulos. "An open-source natural language generator for OWL ontologies and its use in Protégé and second life". Proceedings of 12th Conference of

* <http://www.co-ode.org/downloads/owlviz/>

† <http://protegewiki.stanford.edu/wiki/OntoGraf>

- the European Chapter of the Association for Computational Linguistics (EACL'09), pp. 17-20, 2009.
- [9] G. Hart, M. Johnson, and C. Dolbear. "Rabbit: developing a Control Natural Language for authoring ontologies". Proceedings of 5th Annual European Semantic Web Conference (ESWC 2008), pp. 348–360, 2008.
- [10] D.L. Rubin, N.F. Noy, and M. Musen. "Prote´ge´: a tool for managing and using terminology in radiology applications". Journal of Digital Imaging, vol. 20, pp. 34-46, 2007.
- [11] M. Erdman. "Ontology engineering and plug-in development with the NeOn Toolkit". Proceedings of 5th Annual European Semantic Web Conference (ESWC 2008), 2008.
- [12] A. Kalyanpur, B. Parsia, E. Sirin, B.C. Grau, and J. Hendler. "Swoop: a web ontology editing browser". Journal of Web Semantics, vol. 2, pp. 144–153, 2006.
- [13] D. Allemang, and I. Polikoff. "TopBraid, a multi-user environment for distributed authoring of ontologies". Proceedings of 3rd International Semantic Web Conference (ISWC 2004), Springer Verlag 2004.
- [14] W.C. Mann, and S.A. Thompson. "Rhetorical Structure Theory: toward a functional theory of text organisation". Text, vol. 8, pp. 243-281, 1988.
- [15] C. Mellish, A. Knott, J. Oberlander, and M. O'Donnell. "Experiments using stochastic search for text planning". Proceedings of 9th International Workshop on Natural Language Generation, pp. 98-107, 1998.
- [16] C.B. Callaway. "Integrating discourse markers into a pipelined natural language generation architecture". Proceedings of 41st Annual Meeting on Association for Computational Linguistics, pp. 264-271, 2003.
- [17] C. Sporleder, and A. Lascarides. "Using automatically labelled examples to classify rhetorical relations: an assessment". Natural Language Engineering, vol. 14, pp. 369-416, 2008.
- [18] M. Moser, and J.D. Moore. "Toward a synthesis of two accounts of discourse structure". Computational Linguistics, vol. 22, pp. 409–420, 1996.
- [19] A. Third, S. Williams, and R. Power. "Owl to english: a tool for generating organised easily-navigated hypertexts from ontologies". Proceedings of 10th International Semantic Web Conference (ISWC 2011), 2011.
- [20] S. Williams, and R. Power. "Grouping axioms for more coherent ontology descriptions". Proceedings of 6th International Natural Language Generation Conference (INLG 2010), pp. 197–201, 2010.
- [21] H.H. Clark. "Psycholinguistics". MIT Press. 1999.
- [22] C. Mellish, D. Scott, L.C.D. Paiva, R. Evans, and M. Reape. "A reference architecture for natural language generation systems". Natural Language Engineering, vol. 12, pp. 1–34, 2006.
- [23] H. Dalianis. "Aggregation as a subtask of text and sentence planning". Proceedings of Florida AI Research Symposium, FLAIRS-
- [24] M. Reape, and C. Mellish. "Just what is aggregation, anyway?". Proceedings of European Workshop on Natural Language Generation, 1999.
- [25] R. Power, D. Scott, and N. Bouayad-Agha. "Document structure". Computational Linguistics, vol. 29, pp. 211–260, 2003.
- [26] K.A. Spackman, and K.E. Campbell. "Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies". Journal of the American Medical Informatics Association, pp. 740-744, 1998.
- [27] M.Q. Stearns, C. Price, K.A. Spackman, and A.Y. Wang. "SNOMED clinical terms: overview of the development process and project status". Proceedings of AMIA Fall Symposium (AMIA-2001), Henley & Belfus, pp. 662-666, 2001.
- [28] V.I. Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals. ". Sov. Phys. Dokl., vol. 6, pp. 707-710, 1966.
- [29] S.F. Liang, R. Stevens, and A. Rector. " OntoVerbal-M: a Multilingual Verbaliser for SNOMED CT". Proceedings of 2nd International Workshop on the Multilingual Semantic Web (MSW 2011) in conjunction with the International Semantic Web Conference (ISWC2011), pp. 13–24, 2011.
- [30] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou. "Ontology Visualization Methods—A Survey". ACM Computing Surveys, vol. 39, 2007.

C. Appendix A: Verbalisations Of OWL Input By OntoVerbal And By An Expert Ontologist

Number	OWL input	OntoVerbal	Ontologist
1	pelvic structure SubClassOf: lower trunk structure lower trunk structure SubClassOf: structure of subregion of the trunk pelvic structure SubClassOf: the pelvis and lower extremities and the abdomen and the pelvis and lower trunk structure	A lower trunk structure is a kind of structure of subregion of the trunk. A more specialised kind of lower trunk structure is pelvic structure. Another relevant aspect of lower trunk structure is that a pelvic structure is defined as the pelvis and lower extremities, the abdomen and the pelvis and a lower trunk structure.	A lower trunk structure is a structure of the subregion of the trunk. A pelvic structure is a subtype of a lower trunk structure.
2	procedure on the brain SubClassOf: intracranial procedure intracranial procedure SubClassOf: procedure on the central nervous system intracranial procedure SubClassOf: procedure on the head intracranial procedure EquivalentClass procedure by site and has a procedure site some intracranial structure	An intracranial procedure is a kind of procedure on the central nervous system and procedure on the head. A more specialised kind of intracranial procedure is procedure on the brain. Additionally, an intracranial procedure is defined as a procedure by site that has a procedure site in an intracranial structure.	An intracranial procedure is a procedure on the central nervous system and a procedure on the head. Any procedure by site in which the procedure site is an intracranial structure is also an intracranial procedure. A procedure on the brain is a subtype of intracranial procedure.
3	abdominal vascular structure SubClassOf: abdominal and pelvic vascular structure abdominal and pelvic vascular structure SubClassOf: vascular structure of the trunk abdominal vascular structure SubClassOf: abdominal structure and abdominal and pelvic vascular structure	An abdominal and pelvic vascular structure is a kind of vascular structure of the trunk. A more specialised kind of abdominal and pelvic vascular structure is abdominal vascular structure. Another relevant aspect of abdominal and pelvic vascular structure is that an abdominal vascular structure is defined as an abdominal structure and an abdominal and pelvic vascular structure.	An abdominal and pelvic vascular structure is a vascular structure of the trunk. An abdominal vascular structure is a subtype of an abdominal and pelvic vascular structure.
4	chronic disease of the genitourinary system SubClassOf:	Chronic disease of the genitourinary system	A chronic disease of the

	<p>chronic disease</p> <p>chronic disease of the genitourinary system SubClassOf: disorder of the genitourinary system</p> <p>chronic hypertensive uraemia SubClassOf: chronic disease of the genitourinary system</p> <p>chronic disease of the genitourinary system</p> <p>EquivalentClass: chronic disease and disorder of the genitourinary system and has a finding site some structure of the genitourinary system</p>	<p>is a kind of chronic disease and disorder of the genitourinary system. A more specialised kind of chronic disease of the genitourinary system is chronic hypertensive uraemia. Additionally, chronic disease of the genitourinary system is defined as chronic disease that is a disorder of the genitourinary system, and has a finding site in a structure of the genitourinary system.</p>	<p>genitourinary system is a chronic disease and a disorder of the genitourinary system. Any chronic disease which is also a disorder of the genitourinary system and is found in the structure of the genitourinary system is also a chronic disease of the genitourinary system. A chronic hypertensive uraemia is a subtype of a chronic disease of the genitourinary system.</p>
5	<p>finding of the head and the neck region SubClassOf: finding of the body region</p> <p>head finding SubClassOf: finding of the head and the neck region</p> <p>finding of the head and the neck region EquivalentClass: finding of the body region and has a finding site some head and neck structure</p> <p>head finding EquivalentClass: finding of the head and the neck region and has a finding site some head structure</p>	<p>A finding of the head and the neck region is a kind of finding of the body region. A more specialised kind of finding of the head and the neck region is head finding. Additionally, A finding of the head and the neck region is defined as a finding of the body region that has a finding site in a head and neck structure. Another relevant aspect of finding of the head and the neck region is that a head finding is defined as a finding of the head and the neck region that has a finding site in a head structure.</p>	<p>A finding of the head and the neck region is a finding of the body region. Any finding of the body which is found in a head and neck structure is also a finding of the head and neck region. A head finding is a subtype of the finding of the head and the neck region.</p>
6	<p>nephrosclerosis SubClassOf: degenerative disorder</p> <p>degenerative disorder SubClassOf: disease</p> <p>arteriosclerotic vascular disease SubClassOf: degenerative disorder</p> <p>degenerative disorder EquivalentClass: disease and has an associated morphology some degenerative abnormality</p>	<p>Degenerative disorder is a kind of disease. More specialised kinds of degenerative disorder are nephrosclerosis and arteriosclerotic vascular disease. Additionally, degenerative disorder is defined as disease that has an associated morphology in a degenerative abnormality.</p>	<p>A degenerative disorder is a disease. Any disease which has an associated morphology of degenerative abnormality is also a degenerative disease. Nephrosclerosis and arteriosclerotic vascular disease are subtypes of degenerative disease.</p>
7	<p>kidney graft material SubClassOf: urinary tract material</p> <p>kidney graft material SubClassOf: solid organ graft material</p> <p>kidney graft material SubClassOf: urinary tract material and solid organ graft material</p> <p>transplant of the kidney EquivalentClass: kidney operation and solid organ transplant and renal replacement and has a method some surgical transplantation action and has a direct substance some kidney graft material and has an indirect procedure site some kidney structure</p>	<p>A kidney graft material is a kind of urinary tract material and solid organ graft material. Another relevant aspect of kidney graft material is that a transplant of the kidney is defined as a kidney operation that is a solid organ transplant, and is a renal replacement, and has a method in a surgical transplantation action, and has a direct substance in a kidney graft material, and has an indirect procedure site in a kidney structure.</p>	<p>Kidney graft material is a urinary tract material and a solid organ graft material. A kidney operation, solid organ transplant and renal replacement which has a method of surgical transplantation action, a direct substance of kidney graft material and an indirect procedure site of kidney structure is a type of transplant of the kidney.</p>
8	<p>graft SubClassOf: biological surgical material</p> <p>tissue graft material SubClassOf: graft</p> <p>tissue graft material SubClassOf: graft and body tissue surgical material</p>	<p>A graft is a kind of biological surgical material. A more specialised kind of graft is tissue graft material. Another relevant aspect of graft is that a tissue graft material is defined as a graft and a body tissue surgical material.</p>	<p>A graft is a biological surgical material. Tissue graft material is a subtype of graft as well as a body tissue surgical material.</p>
9	<p>benign essential hypertension complicating and/or reason for care during pregnancy SubClassOf: essential hypertension complicating and/or reason for care during pregnancy</p> <p>essential hypertension complicating and/or reason for care during pregnancy SubClassOf: essential hypertension in the obstetric context</p> <p>essential hypertension complicating and/or reason for care during pregnancy SubClassOf: pre-existing hypertension in the obstetric context</p> <p>essential hypertension complicating and/or reason for care during pregnancy SubClassOf: essential hypertension in the obstetric context and pre-existing hypertension in the obstetric context</p> <p>benign essential hypertension complicating and/or reason for care during pregnancy SubClassOf: benign essential hypertension in the obstetric context and essential</p>	<p>Essential hypertension complicating and/or reason for care during pregnancy is a kind of essential hypertension in the obstetric context and pre-existing hypertension in the obstetric context. A more specialised kind of essential hypertension complicating and/or reason for care during pregnancy is benign essential hypertension complicating and/or reason for care during pregnancy. Another relevant aspect of essential hypertension complicating and/or reason for care during pregnancy is that benign essential hypertension complicating and/or reason for care during pregnancy is defined as benign essential hypertension in the obstetric context and essential hypertension complicating and/or reason for care during pregnancy.</p>	<p>An essential hypertension complicating and/or reason for care during pregnancy is an essential hypertension in the obstetric context and a pre-existing hypertension in the obstetric context. A benign essential hypertension complicating and/or reason for care during pregnancy is a subtype of essential hypertension complicating and/or reason for care during pregnancy.</p>

	hypertension complicating and/or reason for care during pregnancy		
10	<p>procedure on artery of the abdomen SubClassOf: procedure on the abdomen</p> <p>procedure on artery of the abdomen SubClassOf: procedure on artery of the thorax and the abdomen</p> <p>abdominal artery implantation SubClassOf: procedure on artery of the abdomen</p> <p>procedure on artery of the abdomen EquivalentClass: procedure on artery and has a procedure site some structure of artery of the abdomen</p>	<p>A procedure on artery of the abdomen is a kind of procedure on the abdomen and procedure on artery of the thorax and the abdomen. A more specialised kind of procedure on artery of the abdomen is abdominal artery implantation. Additionally, a procedure on artery of the abdomen is defined as a procedure on artery that has a procedure site in a structure of artery of the abdomen.</p>	<p>A procedure on artery of the abdomen is a procedure of the abdomen and a procedure on artery of the thorax and the abdomen. Any procedure on artery which has a procedure site of structure of artery of the abdomen is also a procedure on artery of the abdomen. An abdominal artery implantation is a subtype of procedure on artery of the abdomen.</p>

Development of Copeland Score Methods for Determine Group Decisions

Ermatita *¹

Department of Information System,
Computer Science Faculty of Sriwijaya University
Indonesia
Jl. Palembang-Prabumulih, Ogan Ilir, INDONESIA

Sri Hartati *², Retantyo Wardoyo *², Agus Harjoko *²

Departement of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences
Gadjah Mada University, Indonesia

Abstract—Voting method requires to determine group decision of decision by each decision maker in group. Determination of decisions by group of decision maker requires voting methods. Copeland score is one of voting method that has been developed by previous researchers. This method does not accommodate the weight of the expertise and interests of each decision maker. This paper proposed the voting method using Copeland score with added weighting. The method has developed of considering the weight of the expertise and interests of the decision maker. The method accordance with the problems encountered of group decision making . Expertise and interests of decision makers are given weight based on their expertises of decision maker contribution of the problems faced by the group to determine the decision.

Keywords—Group Decision Support System; Copeland Score.

I. INTRODUCTION

Decision making is the selection process of various alternative actions that might be chosen through a specific mechanism to make the best decision. The decision maker is done in order to achieve certain goals or objectives for solving problems.

Organizational leaders rarely can solve the problem alone. Committees, working teams, project teams and task forces were formed in many organizations is approach to problem solving by group. GDSS is a computer-based interactive system to facilitate the achievement solution of problem by a group of decision makers. That is consistent with the statement of Turban (2005): A group decision support system (GDSS) is as interactive components of the facilities based system that solution of semi structured or unstructured problems by a group of decision makers in unstructured nature. GDSS was developed to address challenges to the quality and effectiveness of decision-making is done by more than one person (group of people). Issues that need to be highlighted in decision-making by a group of people, among others, is the number of decision-makers, the time should be allocated, and the increase the existing participants. GDSS provides support in solving the problem by providing a setting that supports communication for members who joined the group. The problem solving is done by a group of people who are members of the GDSS who need a voting method to obtain a group decision. Copeland score method is one method of voting to earn wages, which is a joint decision-making. So far, existing methods Copeland score considers that all of the decision

maker has the same weight, but sometimes the decision maker has a different weight in determining a joint decision. For it is necessary to develop a method of voting with respect to the weight of each decision maker based on the level of expertise and interests to the problem.

II. BACKGROUND THEORIES

A. GROUP DECISION SUPPORT SYSTEM (GDSS)

GROUP DECISION SUPPORT SYSTEM (GDSS) is an interactive computer based system that facilitates solution of some unstructured problems by a few (sets) of decision makers who work together as a group. GDSS can be applied to different groups of decision situations, which includes a review panel, task force executive meeting / board, remote workers, and so forth. The basic activities that occurred in any group who require support on a computer are:

- 1) *Calling information, involving the selection of data values from an existing database or calling simple information.*
- 2) *Information sharing, meaning the viewer displays the data on the screen to be viewed by groups.*
- 3) *Use of information, including application software technology, procedure, and group problem solving techniques to the data. [8]*

B. COPELAND SCORE BY WEIGHTING

Copeland score is one of the voting methods with a technique based on the reduction of the victory frequency with the defeat frequency by pair wise comparisons [1]. Examples of the determination of the method of paired comparisons copeland score can be seen in Figure 1.

Population	Preferences	Contest	Winner	Alternative	Copeland score
45%	a d b c	a vs. b	b	a	2 - 1 = 1
40%	b a d c	a vs. c	a	b	3 - 0 = 3 *
15%	c b a d	a vs. d	a	c	0 - 3 = -3
		b vs. c	b	d	1 - 2 = -1
		b vs. d	b		
		c vs. d	d		

Preference profiles

Pair-wise contests

Voting Results

Fig.1. Determination of the method of paired comparisons Copeland Score [1]

Many researchers studied used the Copeland Score method to problem solving in group decision maker. Research about the Copeland score method by [Faliszewski et al. (20080)] who use the electoral system Myhstic Ramon Llull and Copeland Election System. This method can be used to comprehensively control the electoral system. In addition, this study also shows that the integration of Llull and Copeland Voting preferences could overcome the irrationality of potential voters.

Furthermore, [Saari and Merlin (1994)] have been developed Copeland Method (CM) with Geometry approach. The study compared the relationship between Copeland Method (CM) and Positional Voting Methods. CM ranking is done in many ways that vary from voting in the election of the electoral system. The results show how the new CM has powerfull to vote.

Another research conducted by [Al-Sharrah (2010)] who performed a number of objects with the Copeland ranking Score. The study was conducted to rank objects (chemicals, projects, databases, etc.) when the number of available indicators provide different information. The results showed that the Copeland method was an effective and stable tool for ranking objects. The Copeland Score method has the advantage to facilitate the analysis of partial large collection of objects.

The Copeland Score method based on Weighting Score voting process is needed to determine which decisions can be recommended. Decision results of alternative ranking by each decision maker, must be processed to determine the decisions recommended by the group. To select an alternative decision-making group has been established from a variety of skills performed by using the method of Copeland Score [1] as a group decision. The results of each decision maker will be processed with voting by the Copeland Score method suppose the decision of each expert as the sequence shown in Table 1. It is assumed that there are three options, namely A_1 , A_2 , and A_3 . The process of copeland score method, all of the population who choose A_1 , A_2 , and A_3 in accordance with the table of preference profiles. Pair wise contests table shows that one option (e/g A_1) compared to the overall choice (A_2 , A_3). This pair wise comparison is done one by one and imposed on the overall participant choice. Pair wise comparisons between A_1 to A_2 so much to choose A_1 , A_1 pair wise comparisons were selected A_1 to A_3 . It turns out the pair wise contests of the table shows that A_1 is chosen twice. The alternative A_2 option does not appear whereas alternative A_3 options appears once.

The table 2 it can be shown A_1 has a winning choice as much as twice to A_1 and A_2 , and defeat one time to A_2 based on the pair wise contests. To determine whether choice of Alternative A_3 is the best option or not, then do the subtraction operation frequency with the frequency of wins versus defeat. We can see at Table 1 voting results showed that the choice of A_1 has the highest frequency. Based on the frequency, then the alternative voting A_1 was selected as the winner.

TABLE I. Result of decision Maker

Decision Maker	Alternative		
	P1	A2	A1
P2	A1	A3	A2
P3	A3	A1	A2

The results from each expert then contested in each element, so that the resulting such

TABLE II. Pairwise Contest

Pairwise Contest			
A1	Vs	A2	A1
A1	Vs	A3	A1
A2	Vs	A3	A3

Having obtained the results match the calculated value of copeland score. Copeland score results will be processed by using the weights of each DM and the weight of the place from the Copeland Score. Calculation results is shown Table 3. The highest score is the winner, in this case a group decision recommendation.

TABLE III. Result of Voting

Copeland score			
A1	2-0	0	Winner
A2	0-2	-2	
A3	1-1	0	

Voting results showed that the value of Alternative 1 (A_1) has a value of 2, which is the highest value. A_3 then the second highest, and the lowest is the A_2 . Winner of the sequences is $A_1 A_2 A_3$.

This sequence is not yet a final decision, the process is carried out based on the weight given by the DM. The weight refers to the determination of the agreement shows that the highest weight is an expert in the area of expertise that contributed most in decision-making. Suppose the weight of $DM1 = 5$, $DM2$ and $DM3 = 3 = 2$, the value of place based on the highest copeland is based on the amount of data. For example the data above there are 3 places the highest value is given to A_1 3, the value of 2 is given to A_3 and A_2 value is 1.

Top expert weight multiplied by the value of the place. The multiplication of the weights and the experts will place the weights in the ranking so as to produce the final decision recommendations. By calculations:

$$A_1 = 3 \times 3 = 9$$

$$A_3 = 2 \times 2 = 4$$

$$A_2 = 5 \times 1 = 5$$

Results of Copeland method calculation obtained by weighting the order: A1 A2 A3. This means the winner is alternative A1. The voting result of decision maker group is A1 as the recommendations of the group.

III. CONCLUDING REMARKS AND FURTHER WORKS

Shared decision making by some decision maker in a group requires voting methods. Voting method is implemented to accommodate the interests and expertise of the decision maker. To determine the group decision a decision maker requires specific weights of each decision maker. The weight determines how important decision maker with expertise have contributed to decision making. The copeland score method which has been developed by Garvish does not accommodate the weight of each decision maker. This paper developed methods of voting in decision making to accommodate the

weights based on the importance of expertise decisionmaker in the decision making process as a solution to the problem /

REFERENCES

- [1] Gavish, B. and Gerdes, J.H., 1997, Voting Mechanisms and Their Implications in A GDSS Environment, Annals of Operations Research Science Publisher.
- [2] Al-Sharrahm G., 2010, Ranking Using the Copeland Score: A Comparison with the Hasse Diagram, J. Chem. Inf. Model, 785-791.
- [3] Saari, D.G and Merlin, V.R., 1996, The Copeland Method* L: Relationships and the Dictionary, Economic Theory, 51-76.
- [4] Faliszewski, P and Hemaspaandra, E, Hemaspaandra, A. L. and Rothe, J., 2008, Copeland Voting Fully Resist Constructive Control, R.Fleisther and J.Xu (eds):AAIM 2008, 165:176.
- [5] Turban, E. and Aronson, E.J., 2005, Decision Support Systems and Intelligent Systems 7th-ed. jilid I (Sistem Pendukung Keputusan dan Sistem Cerdas), diterjemahkan oleh Dwi Prabantini, Andi, Yogyakarta.
- [6] Turban, E., Sharda, R., and Delen, D., 2011, Decision Support and Business Intelligence Systems, ninth Edition, Prentice Hall, new jersey, USA.
- [7] E. Turban, "Dicision Support and Expert Systems: Management Support Systems", Fourth Edition, Prentice-Hall, Inc., United State, 2005
- [8] Q. Zhang and J. Ma, "Determining Weights of Criteria Based on Multiple Preference Formats", online pada <http://www.is.cityu.edu.hk/Research/WorkingPapers/paper/0102.pdf> 12 Oktober 2004, 2004

New electronic white cane for stair case detection and recognition using ultrasonic sensor

Sonda Ammar Bouhamed

Computer Imaging Electronic System (CEM Lab)
University of Sfax
Sfax Engineering School
BP W, 3038 Sfax, Tunisia

Imene Khanfir Kallel, Dorra Sellami Masmoudi

Computer Imaging Electronic System (CEM Lab)
University of Sfax
Sfax Engineering School
BP W, 3038 Sfax, Tunisia

Abstract—Blinds people need some aid to interact with their environment with more security. A new device is then proposed to enable them to see the world with their ears. Considering not only system requirements but also technology cost, we used, for the conception of our tool, ultrasonic sensors and one monocular camera to enable user being aware of the presence and nature of potential encountered obstacles. In this paper, we are involved in using only one ultrasonic sensor to detect stair-cases in electronic cane. In this context, no previous work has considered such a challenge. Aware that the performance of an object recognition system depends on both object representation and classification algorithms, we have used in our system, one representation of ultrasonic signal in frequency domain: spectrogram representation explaining how the spectral density of signal varies with time, spectrum representation showing the amplitudes as a function of the frequency, periodogram representation estimating the spectral density of signal. Several features, thus extracted from each representation, contribute in the classification process. Our system was evaluated on a set of ultrasonic signal where stair-cases occur with different shapes. Using a multiclass SVM approach, recognition rates of 82.4% has been achieved.

Keywords—*Electronic white cane; ultrasonic signal processing; ground-stair classification ;temporal representation of ultrasonic signal; frequencial representation of ultrasonic signal*

I. INTRODUCTION

Domestic space is a complex environment that contains various obstacles of different types at different locations: right, left, top and bottom. Even for none visually impaired, the congestion of such obstacles, sometimes poses problems, so what about those with visual impairment? People with visual disabilities are often dependent on external assistance which can be provided by humans, trained dogs, or special electronic devices as support systems for decision making. Existing devices are able to detect and recognize objects that emerge on the floor, but a real risk is also coming from objects that are decreasing from the floor, as holes or descending stairs. Accordingly, we are motivated in this paper to develop an automatic vision tool to overcome these limitations.

Using a traditional white cane is a universal solution, allowing a less risky journey for blind people. Such a tool is used to explore the environment by a frontal sweep, or contact with the ground to detect the presence of an obstacle. However, this cane does not allow sufficient exploration of objects that are at the top or which are getting too closer. To

this end, the realization of an electronic cane automating the detection and recognition of fixed and mobile obstacles can offer more security and comfort to blind persons. This can be done through the integration of various specific sensors, which are designed to provide several types of information such as obstacles form, dimension, color and distance from the user. Some solutions are already exist on the market such as: Laser Cane [1], Teletact[2], UltraCanne [3], K Sonar cane [4], Smart Cane [5], Isonic [6], Guide cane [7], Palm Sonar [8], SmartWand [10], etc. These products help visual impairment people by collecting information through sensors and then, transmitting recommendations to them, through vibration or sound messages. A classification of these canes with respect to the type of sensors employed for obstacles detection is presented in [11].

The major disadvantages of these solutions are:

- 1) *They only detect obstacle existence and distance without specifying indication about their nature which is important for the user to know.*
- 2) *They are unable or inaccurate in detecting some obstructions that are not protruding but present potential threat such as descending stairs, holes, etc.*
- 3) *The system communicates its recommendations, through intensity or frequency variations. Thus feedback information is often sent to the user through vibration or sound signals. So a training course is needed to keep the user informed about how to understand and react in real time to alerts that are transmitted regarding the existence of obstacles as well as their recognition. On the one hand, such training can be sometimes more expensive than the product itself. On the other hand, it is often difficult and complex for the users to assimilate it properly. Furthermore, in the case, where information is transmitted as an acute sound, that may happens several times especially when the obstacle is very close, it may be embarrassing for the blind person when they are in public.*

Therefore, our interest is specifically focused, on the development of an electronic tool using two types of sensors which are ultrasonic sensors and monocular camera. Our choice of these sensors takes into account theirs area of operation and their performances. Our choice also depends on several other factors as: cost, type of scene, type of obstacle to be detected, detection range and desired precision of the measurements. The main idea consists in merging data

provided by the two sensor types to allow more accurate information, to be transmitted to the user via a Bluetooth module as a voice message specifying the object nature, characteristics and the distance between the detected obstacles and the device.

In this paper, we explore ultrasonic sensor potentials in object detection and mainly stairs recognition. This type of sensors has significant potential in robotic applications. Indeed, it has been widely used in collision avoidance systems and in localization and navigation of mobile robots. In addition to robotic application, ultrasonic sensors are used in many other applications in different fields such as echography in medical field, nondestructive testing of materials ... Advantages that encourage us to use ultrasonic sensors is the ease to obtain distance information from immediate objects without intensive processing which can considerably lighten the application. They are also able to perform under low visibility conditions making it ideal for night as well as day use. Thus, ultrasound sensor seems to be a good solution for our system to detect and recognize several objects. However, object recognition under different viewing conditions is still a challenge for autonomous systems. So, the motivation of our work is to challenge by applying only one ultrasound sensor for obstacle recognition taking into account the weaknesses of this sensor type.

Many features can be extracted from the ultrasonic signal, providing different information and descriptions that are used to describe the detected object.

The remainder of this paper is organized as follows: In section 2, we present the contribution in the literature and the flowchart for the study. Section 3 show the proposed system architecture design that includes software and hardware components, working principal and wearability performance requirements of proposed electronic cane system. The ultrasonic signal processing for obstacle detection as well as well-known approaches used in the literature for object recognition are shown in section 4. Then, section 5 present the related work of stair cases detection and recognition, before detailing proposed algorithm of ultrasonic signal preprocessing, feature extraction and SVM classification. The evaluation of our approach is discussed in section 6, and we finalize this paper with conclusions and perspectives.

II. CONTRIBUTION IN THE LITERATURE

After intensive study of blind needs in Tunisia through a large survey conducted with the Tunisian Blind Association (URA-Sfax) [45], it seems that the white cane presentation should not be replaced by other forms even when we look at improving it by making it intelligent with automatic obstacle detection tool and recognition options. In fact, the white cane is the clearest indicator to others, about blind person presence. In the literature, this idea was confirmed by some of researchers.

Indeed, in many studies related to the implementation of electronic cane, many authors considered to attach some components onto the white cane [4][6][10].

The design of our electronic white cane architecture, in this context, is as crucial task as the choice of the different hardware components. For example, it is necessary to satisfy the electrical conduction between sensors, microcontroller and

batteries. It is also necessary to determine the deviation angle of each sensor to be able to detect obstacles placed in front of the user.

To design a prototype of our electronic white cane Fig. 1 summarizes the different steps of our survey. We review in the next section related technologies regarding the visually impaired. The stair case detection task is also examined.

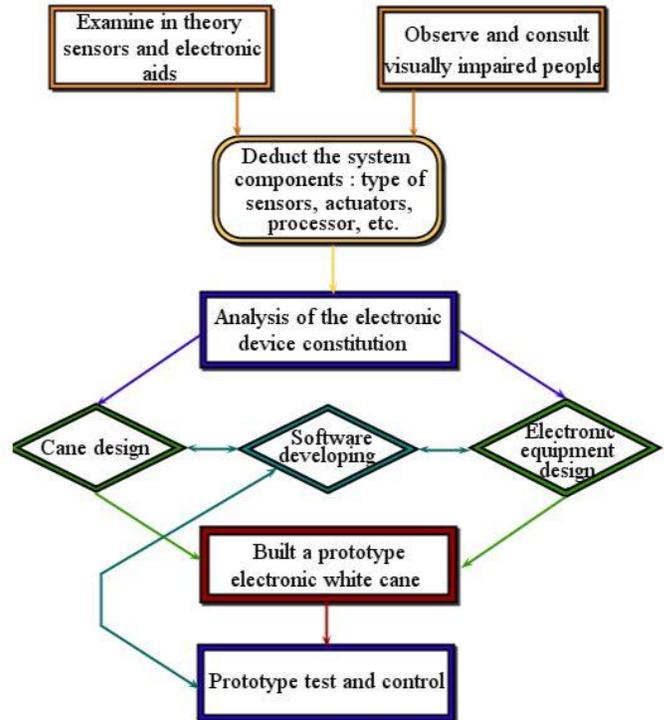


Fig. 1. Flowchart for the study

III. SYSTEM DESIGN

A. Sensors

1) Different sensor types

Sensors allow perception of the environment in more or less reliable way compared to the human eye. The use of different sensors is required, in different fields, to help the user in making a decision. Accordingly, we distinguish active and passive sensors.

A passive sensor measures a full energy provided by a physical phenomenon. In general terms, the sensors that use external energy sources to observe an object are called passive sensors. In the robotic world, the most used sensors are the Monocular cameras. They are inexpensive and efficient in terms of range, accuracy and amount of usable data.

Some systems use stereovision to detect and recognize objects. The principle is to infer information about the structure and distance in a 3D scene from two optical images taken from different viewpoints. It involves three stages: calibration, matching and triangulation. The mapping between the left and right images (registration) is the most crucial phase of processing.

An active sensor provides some kind of energy such as microwave, sound, light, etc., into the environment in order to

detect the changes that occur on the transmitted energy. That means it transmits and detects at the same time. In the robotics world, this type of sensors is very robust for near and far obstacle detection. In addition, it determines an accurate measurement of the distance to the obstacle. The most used active sensors are ultrasonic, laser, and radar. Ultrasonic sensors work well for close obstacles unlike laser ones, which operate well for distant obstacles. Radar sensors are very robust for near and far obstacle detection, but their medium accuracy doesn't allow them detecting small obstacles. It is important to note that the sensor characteristics differ from one to another but make each sensor meeting specific requirements. Therefore, to achieve the best choice, we propose in the following, sensor characteristics comparison.

2) Choice of the sensor

The sensors selection must take into account the area of operation of each one and its performance. Also, it depends on several factors: detection range, cost, desired precision of the measurements, type of obstacle and type of scene.

Several constraints eliminate the use of stereo-vision in our system. Indeed, the stereo-vision processing is directly dependent on the accurate positioning and calibration of two cameras. Thus, at any variation, the error will automatically affects on the result.

Such variations are common in cane movement, and thus, results can be less accurate. Moreover, our application must meet the constraints of computing time. Indeed, aiming at facilitating the movement of blind people, the running time of our system has to be as short as possible in order to meet real time system requirements. Such requirement risks to be not supported by using stereo vision system as it should generate twice more images than monocular camera system. Therefore, the choice of a monocular camera makes sense since we want to implement a technique that aims to be:

- Fast (real time).
- Low cost.
- Precise and with acceptable range of vision.

Although, providing the richest information allowing recognition of detected obstacles, an optical sensor use, doesn't only raise the processing time a lot, but get also truncated information of the real scene. Indeed, data get from the camera doesn't provide distance information, being a detail of extreme importance for such an application. Therefore the use of a depth sensor is an ultimate necessity.

The choice of an active sensor depends on the measurement range of the sensor, its response time, resolution, recognition reliability and finally the application requirements. For this end, a comparative survey is achieved and given in Table \ref{tab1}.

According to the survey results, shown in Table \ref{tab1}, the radar sensor is eliminated because it can neither detect small obstacles nor determine the distance to such objects. Thus, this sensor does not meet the requirements of our application.

TABLE I. GENERAL CHARACTERISTICS OF SOME ACTIVE SENSORS

	Laser	Radar	Ultrasound
Principle	Transmission and reception of light wave	Transmission and reception of electromagnetic wave	Transmission and reception of ultrasonic waves
Range	About 60 meters	About 250 m	From 3 cm to 10 meters
Accuracy	High (about 5 cm)	Medium (few meters)	Very high (5 mm)
Price	Very high	high	Low

The proposed tool does not require a very large extent, that's why an increase from 3 to 4 meters is more than sufficient. In addition, our goal is to offer not only an efficient and reliable cane, but also a low cost one. In this case, the best sensor, which is closest to our needs, is the ultrasound one.

3) Sensor system model

The objectives that we project to meet in the present paper, assuming that the blind people are navigating in environment autonomously, are to:

- Generate the "Ascending or descending Stair case" through found signal.
- Define the distance, between the blind and the staircase, to be transmitted to the user via a Bluetooth module as a voice message.

This work employed "LV-EZ0" ultrasonic sensor [25]. It can measure ranges from 0 inches to 254 inches (6,45-meters) and provides sonar range information from 6 inches up to 254 inches with 1 inch resolution. The interface outputs are pulse width output, analog voltage output, and serial digital output. We can choose one of the three sensor outputs. Ultrasonic sensors emit a high frequency pulse of 42 KHz. The packaging of the sensor is light and small enough (19,9 x 22,1 x 16,4 mm) to be fixed onto a cane without any inconvenience. The beam width of ultrasonic sensor is narrow enough so that the sensors do not interfere with each other while keeping their efficiency to detect any obstacles on the floor.

The used monocular camera is "LinkSprite JPEG Color Camera TTL Interface [12]. It has a small dimension 32mm x 32mm to allow its integration into the cane without any inconvenience. It can capture ranges from 10 to 15 meters with a maximum viewing angle of 120 Degrees and produce JPEG images whose resolution is adjustable up to 640 * 480. The monocular camera is powered from 3.3V or 5V and its power consumption varies between 80 mA and 100 mA.

The camera position is defined such that it can detect obstacles from the top to the bottom of the field of vision. Accordingly, it is placed almost in the middle of the cane.

B. Proposed system architecture

With the above specified components, the proposed electronic white cane system will work in such principles: Ultrasonic sensors and monocular camera allow scene acquisition through different data nature. Signals provided by sensors are processed in a signal processing unit. Data collected

from US sensor, is processed to provide depth information of the scene scanned according to a given direction. The resulting signal is then shared in as many segments as there are objects in the scene. For each segment we associate, not only, distance label, but also other specific labels, telling about some obstacle characteristics, that we extract from ultrasonic signal, such as form, situation, material consistency, etc. Otherwise, images captured by the monocular camera, are also segmented and each of their region is labelled.

Some of labels got from each one of the two sensors are gathered to ensure registration of ultrasonic and optical data. With such a design, our system determines the distance from the obstacle as well as some of its characteristics using data from both sensors: camera and US sensor.

All the collected information about the scene are then analysed to make a decision that is returned to the user as a voice message revealing the nature of the obstacle and the distance towards it. This message is transmitted from the SD card to a headset using Bluetooth module. The proposed system overview is shown in Fig. 3.

The electronic white cane design configuration is shown in Figure \ref{cane design}. The cane is designed to be adjustable in height, to suit its user size. This height is considered as an input parameter of our system, as well as the angle β between the cane and the horizontal. But, in order to simplify a little the task of testing and validation, we set, for our prototype, the cane length at 90 cm, and β at 113° , by fixing the cane on a carriage according to this inclination. We install our ultrasonic sensor that is used to detect on ground obstacles, at angle α with the cane allowing detection of an obstacle, or more specifically of the beginning of a staircase, at the distance of 2m. Such a distance shall guarantee more security for the user.

The box containing the monocular camera is placed at distance $d4$ from the cane handle. The distance between the cane and the eventual beginning of the staircase being set at 2 m, we can define the orientation angle γ of our camera to detect each obstacle on the floor.

The parameters used in the model are:

β : The angle between the cane and the carriage,

$d2$: The length of the cane,

$d3$: The distance between the cane and the eventual beginning of the staircase,

$d4$: The distance between the cane handle and the camera,

The following parameters are calculated using the parameters given above

$d1$: The distance between the tip of ultrasonic sensor and the floor.

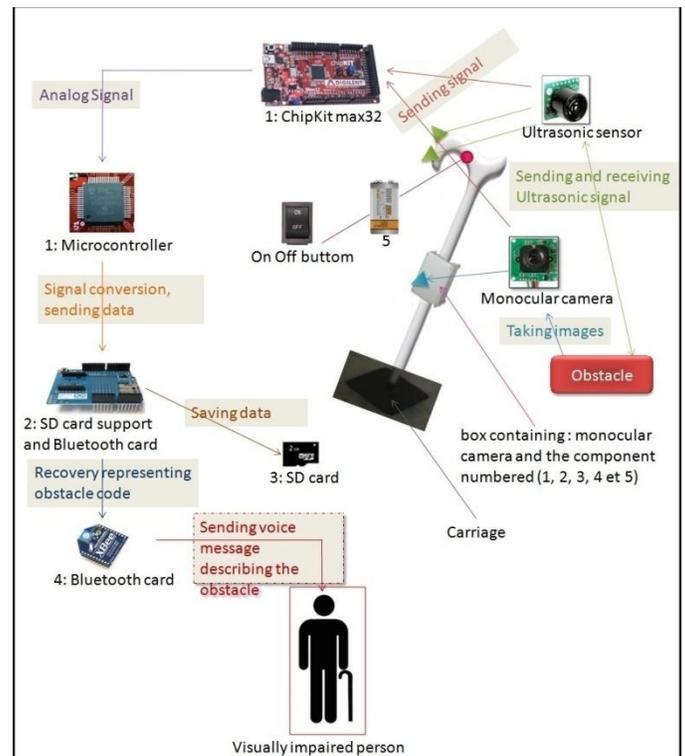


Fig. 2. Electronic cane system working principal

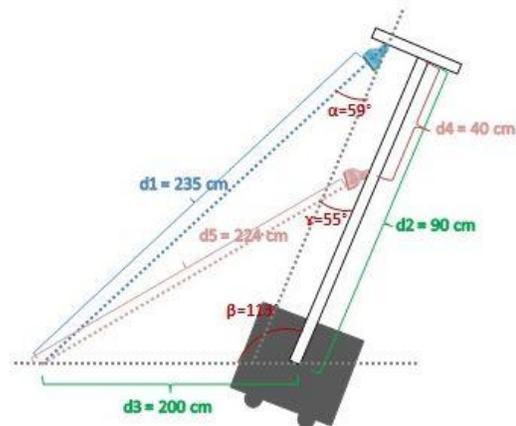


Fig. 3. Electronic white cane design configuration

$d5$: The distance between the tip of the camera and the floor

α : The angle between the cane and the ultrasonic sensor

γ : The angle between the cane and the camera.

To calculate the different sensor's angle inclination, we propose to use geometric rules within any triangle.

Let the triangle ABC shown in Fig. 4.

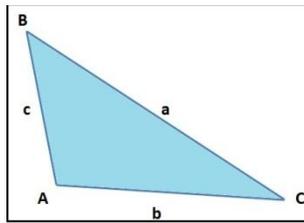


Fig. 4. Any triangle

To find the characteristics of the triangle, we use (1) to calculate the angles and (2) to determine the side lengths.

$$\cos(A) = \frac{b^2 + c^2 - a^2}{2.b.c} \quad (1)$$

$$a = b^2 + c^2 - 2.b.c.\cos(A) \quad (2)$$

we assume that the triangle has side dimensions $d1$, $d2$ and $d3$ and β is the angle between sides of dimensions $d2$ and $d3$. So, to calculate $d1$ we use (2) and we obtain :

$$d1 = 235 \text{ cm}$$

The same equation is used to calculate $d5$ and we obtain :
 $d5 = 224 \text{ cm}$

To calculate the angle γ , we assume that the triangle has side dimensions $d3$, $d5$ and $d6 = d2 - d4$. So, to calculate this angle, we use (1). We obtain $\gamma = 0.55^\circ$.

The same equation is used to calculate the angle α and we obtain : $\alpha = 0.59^\circ$.

C. Wearability performance requirements

The safety of visually impaired people imposes a reliable detection and recognition system. Many requirements have thus to be considered. Apart from electronic hardware and software concept, the wearability of the system is also a critical issue in our proposed system. The wearability requirements are:

1) Usability

The majority of electronic canes proposed in the literature requires training for their use. These courses are expensive and take long time. Aware of the importance of these details, the solution we propose saves on training costs due to its ease of use. It is our intention to provide essential information to the blind user with a simple tool that does not require any preliminary training.

2) Robustness

A System designed for people with visual impairments should be able to detect obstacles regardless variation of lighting conditions. Some risks may also occur during the use of the cane as his fall or it can be wet. To reduce the influence of these risks our cane should be anti-magnetic as well as being water resistant and shock resistant. Such a design has to be studied and managed.

3) Efficiency and precision

The system must reliably and precisely detect objects surrounding the blind regardless of their appearance, size and shape. Indeed, obstacles, missed in the detection step, expose the blind to a serious risk of accidents. Thus, our system is more effective as the number of errors, it might commit, is significantly reduced.

4) Real-time system

The term real time has several meanings depending on the context. In our context, we consider a system as 'real-time' one, if the information after its acquisition and processing remains relevant. The system must warn the visually impaired user so he can react in time.

5) The cost

The cost of the electronic white cane must be reasonable for all those requiring its use to help them in their day to day lives. The most popular electronic white canes proposed in the literature are very expensive since they use sophisticated sensors to have more efficiency. Unlike those products, our system relies on ingenious processing strategies, needing thus, only a single monocular camera, which is certainly less expensive than infrared camera or stereo vision systems as well as two ultrasonic sensors which the price is lower than the other active sensors such as radar and laser.

6) Lightweight

An embedded system on a white cane should not be cluttered with numerous sensors and large equipment which increases the weight of the cane. Several existing systems use many sensors which require a large box and increases the weight of the cane such as the Guide cane [8]. The small weight of the traditional white cane allows users to easily scan their environment. Electronic white cane, is certainly heavier, but it should not prevent the scan so that the user can feel at ease as with his traditional cane.

It can be seen from the previous points, that the proposed solution is a device similar to the basic traditional white cane but with a set of sensors, interacting with each other to obtain an intelligent and efficient electronic white cane.

IV. ULTRASONIC SIGNAL PROCESSING FOR OBSTACLE DETECTION

A. Ultrasonic signal

Use ultrasonic signal processing is frequently used in nondestructive testing (NDT) of materials, medical characterization of tissues [26], construction industry [27], alimentary industry [28], in robotic application, etc. The classification steps depends strongly upon the features extracted to represent the object. Many properties can be extracted from ultrasonic registers. Although, there are some conditions where only trivial signal processing is required, there are some other cases where extracting these properties is a complex task.

Several information can be obtained from signal amplitude, but that doesn't necessarily provide the best representation of the signal [39]. Sometimes, the signal's frequency is more significant when more specific information are hidden in frequency components. Fourier Transform (FT) is often used for transforming the collected signal from time based signal to

frequency-based one. The frequency-amplitude representation obtained by FT represents amplitude component for each frequency of the signal [30-40].

Numerous previous works have proposed various sets of ultrasonic features extracted from time and frequency domains and investigated the feasibility of using such parameters for an ultrasonic signal classification.

Time-domain ultrasonic features include principal components of signals [33] and estimates of the rectified signal envelope combined with various preprocessing methods: low-pass filtering, rectification, under-sampling, and mean-subtraction [34].

In previous researches the most used feature in time domain are time of flight (TOF) information [31], echo energy [37], maximum amplitude of the echo [36], and correlation [29]. If we want to use only TOF information, for the classification, we need multi sensor system [38]. The echo amplitude is inadequate information, if it is used alone. The works presented by Dror *et al.* [32] established that the echo representation in the frequency domain gives the best results.

The various features, extracted from frequency-domain, which have been proposed by previous works [32], are statistical parameters extracted from statistical moments of an ultrasonic frequency spectrum such as, coefficient of skewness, coefficient of kurtosis [35], mean and coefficient of variance.

Combinations of time-domain and frequency-domain features also have been readily used [41-42].

After feature extraction steps, we need effective feature selection schemes to reduce the redundancy features and optimize their set.

This previous research has commonly traced the general guidelines of feature extraction from various domains of ultrasonic data analysis, that suggest the following steps:

- Extract as many descriptors (features) as possible from various domains.
- Evaluate their discrimination power with respect to the concerned classification problem.
- Choose the best set of features for classification.

B. Object detection and recognition

Our interest is exclusively focused on the use of ultrasonic sensors in our tool. To the best of our knowledge and from the latest research, ultrasonic sensors are not yet used to detect and recognize descending and ascending stairs.

We find in the literature, in the robotic field, some works concerning the classification of targets (corner, plane, cylinder and edge) using one ultrasonic sensor.

Firstly, [19] uses an artificial neural network to recognize two or three-dimensional shapes (cube and tetrahedron) independently from orientation, based on the echoes of ultrasonic pulses similar to those used by an echolocating bat.

Secondly, [20] presents the results of detection and classification of simply shaped objects using ultrasonic

transducers. The subjects of object detection are an edge, a plan, a small cylinder and a corner using only one transducer and in indoor environment in mobile robotic applications.

Bozma and Kuc [22] introduced a concept for interpreting sonar TOF data obtained from specular surfaces. Physical properties of reflection and acoustic sensors are exploited to extract information about the environment in order to classify it in three ways (corner, plane and edge). This system uses a single mobile sensor for generating a sonar map.

Barat and Ait Oufroukh [21] developed statistical approaches for 2D target classification in an indoor environment using only the Time Of Flight (TOF), the maximum amplitude and 21 magnitudes to discriminate the different targets. This work classified targets in 4 ways (corner, plan, edge and small cylinder) using one transducer. \\

In [23], Pham *et al.* provides a new application to monitoring activities of people in smart environments. Several scenarios were developed in which ultrasonic sensors were used for patient and elderly monitoring. Trajectory-matching algorithms were devised to classify people movement trajectories in indoor environments.

In [24], authors present an intelligent approach based on a 3D model of the environment, where the emphasis is on the extraction of features. In fact, researches have been primarily focused on determining walls and corners using ultrasonic sensors. Walls are considered as the extension of a line segment lying on a plane, whereas, corners are considered as the intersection of two planes, being observed from inside the concave space.

Most of works listed above, use more than one ultrasonic sensor to detect and recognize obstacles. Some others use single mobile ultrasonic sensor, that can obtain data from different points of view of the objects.

V. STAIR CASES DETECTION AND RECOGNITION

A. Related Works

Many approaches were described in the literature using different algorithms to detect and recognize wall, holes, descending and ascending stairs. Such systems are often essentially based on a laser sensor, infrared sensor or monocular camera.

Yuan and Manduchi [13] developed a hand-held environment discovery tool for the blind that integrates a laser-based range sensor. The user receives local range information when he swings the system around him. The time profile of the range is analyzed, by means of an extended Kalman filter, to detect environmental features that are critical to mobility, such as curbs, steps and drop-offs. This filter is used to track the range data and detect environmental features. In other work, Yuan and Manduchi [14] describe a new virtual white cane based on a laser pointer and a camera. These authors present in their paper a surface-tracking algorithm based on a Jump-Markov model for automatic detection of geometric singularities. This algorithm describes the evolution of range data in different types of surfaces, for example the foot of a wall, a step or a drop-off, by moving the system around and pointing it at different areas of the environment.

Adams [15] introduced the new concept of an electronic cane for visually impaired people based on a combination of three infrared range sensors that were used to identify the terrain (even surface, ascending and descending stairs). The sensor system is close to the user's belt and it does not require swinging motion or any other movement by the user.

Lee and Lee [16] introduced the three infrared range sensor system for detecting, ascending and descending stairs. Decisions of the system are made based on current sensor readings. However, disturbance due to the user's movement was not considered.

Mihankhah [17] presents a theoretical analysis and implementation of autonomous staircase detection on a mobile robot. The robot is equipped with two laser sensors which scan the environment horizontally for the first and vertically for the second sensor.

Se and Brady [18] explain a distant stair case detection system, that uses optical camera or vision system camera to perceive outdoor environment.

Scherlen *et al.* [47] describe a new concept of Recognize Cane using a water detector, ambient humidity sensor and infrared sensors. This system can recognize the most common objects and environment clues like the soil humidity rate using a water detector and ambient humidity sensor. This system can also detect zebra crossings using brilliance sensor and a luminance sensor. The brilliance sensor is equipped with an infrared transmitter and receiver. This tool used two distant infrared sensors to recognize stairways or holes in the path of the user.

B. Range of ultrasonic sensor

The ultrasonic sensor provides four output formats which are pulse width output, analog voltage output, and serial digital output. The distance information d from the sensor tip to the obstacle can be obtained from the pulse width (PW) representation of range. Thereby, the distance value can be calculated using the scale factor of 147uS per inch. The sensor readings vary according to the terrain in our case, the floor or ascending or descending stair cases, as shown in Fig. 5.

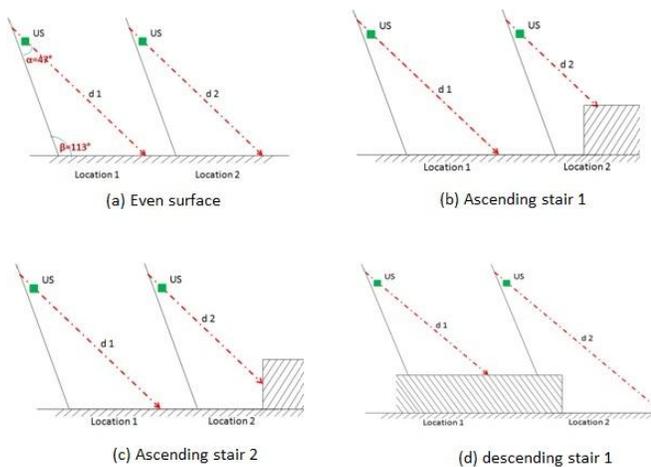


Fig. 5. Sensor system and environment

The three curves shown in Fig. 6 present the distance measures between the sensor and the nearest obstacle in three walking situations:

The top left curve shows the distance values when the user walks on a floor, without any change of floor state. The ultrasonic sensor outputs vary while the user walks, because the angle of incidence to the floor is large. So, it cannot provide accurate measurements.

The bottom curve shows the distance values when the user walks on a floor, then the cane detects an ascending stairs.

The top right curve shows the distance values when the user goes close to a descending stairs after an even surface.

Logically, the distance values must become larger (resp. smaller) than that obtained with a floor when the cane receives descending (resp. ascending) stairs. However, seeing curves of Figure\ref{sol}, it doesn't seem to be clear that the sensor readings change accordingly with floor states. Indeed, vibrations are common in cane movement resulting some errors in the ultrasonic output signal.

To separate the three cases experimental data identification rules of the floor state are developed in the following section.

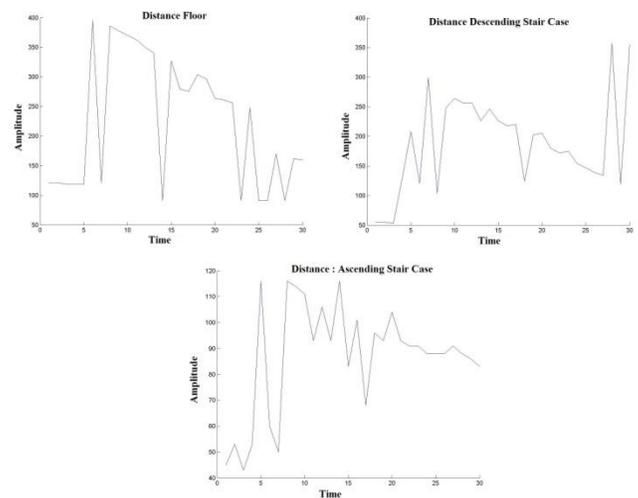


Fig. 6. Range sensor raw data – even surface (top left), ascending stairs (bottom) and descending stairs (top right)

C. Preprocessing and Feature Extraction

Since the ultrasonic sensor is attached to a cane, which is unstable due to the sweeping and tapping motions, enhancement of the ultrasonic sensor data is required. A low pass filter was used to filter ultrasonic registers. The use of the low pass filter has allowed eliminating the error in the ultrasonic signal.

Let us denote $x_i, i = 0, \dots, N$, an ultrasonic signal. Several features was extracted from filtered signal in different domains.

- mean:

$$\bar{x} = \frac{1}{N} \sum_{i=0}^N x_i \tag{3}$$

- Sample Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=0}^N (x_i - \bar{x})^2} \quad (4)$$

- Maximum.
- Minimum.
- The skewness (moment of order 3): is a measure of distribution symmetry around its mean.

$$\mu_3 = \frac{1}{N} \sum_{i=0}^N (x_i - \bar{x})^3 \quad (5)$$

- The kurtosis (moment of order 4): is a measure of whether the data peakedness is relative to a distribution.

$$\mu_4 = \frac{1}{N} \sum_{i=0}^N (x_i - \bar{x})^4 \quad (6)$$

- The root mean square (RMS): square root of the moment of order 2, being the variance that is given by (7). The RMS is a statistical measure of the varying quantity magnitude and it is given by (8).

$$V = \frac{1}{N} \sum_{i=0}^N (x_i - \bar{x})^2 \quad (7)$$

$$R = \sqrt{V} \quad (8)$$

In the frequency domain, numerous features were calculated from different filtered signal representations as shown in Fig.9: the spectrum, the spectrogram and the periodogram.

The features extracted from the spectrum were the same features computed from the filtered signal in time domain. The spectrogram is a time-frequencial representation.

This time-frequency transform decomposes the signal x over a family of time-frequency atoms $A_{t,f}$ where t and f are the time and the frequency localization indices. The resulting atom coefficients can be written as follows:

$$F[t, f] = \sum_{i=0}^{N-1} x[i] A_{t,f}^* [i] \quad (9)$$

where $*$ denotes the conjugate and the Short-time Fourier atoms $A_{t,f}$ shall be written as follows:

$$A[i] = w[i - tu] \exp\left(\frac{i2\pi ki}{K}\right) \quad (10)$$

where $w[i]$ is a Hanning window of support size K .

The time-frequencial representation provides a good domain for signal representation and classification. In fact, this

type of representation contains some details that cannot be seen in the temporal representation of ultrasonic signal.

The texture of the spectrogram representations contains distinctive patterns that capture different characteristics of the ultrasonic signals.

The time-frequencial representation provides an image that is used to extract Haralik's texture features [44] which are:

- Angular Second Moment:

$$f_1 = \sum_i \sum_j p(i, j)^2 \quad (11)$$

Where i and j are two different gray level. p is obtained by calculating the number of times when a pixel with value i is adjacent to a pixel with value j .

- Contrast:

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\}, |i - j| = n \quad (12)$$

Where N_g is the gray level number in the 2D image.

- Correlation is given by :

$$f_3 = \frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (13)$$

Where μ_x , μ_y , σ_x and σ_y are the means and std. deviations of, respectively, p_x and p_y being partial probability density functions.

- Sum of Squares or Variance:

$$f_4 = \sum_i \sum_j (i - \mu)^2 p(i, j) \quad (14)$$

- Inverse Difference Moment that is given by :

$$f_5 = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j) \quad (15)$$

- Sum Average that is performed as follows:

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \quad (16)$$

Where x and y are the input coordinates (row and column) in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$.

- Sum Variance:

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8) p_{x+y}(i) \quad (17)$$

- Sum Entropy:

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} \quad (18)$$

- Entropy:

$$f_9 = -\sum_i \sum_j p(i, j) \log(p(i, j)) \quad (19)$$

- Difference Variance:

$$f_{10} = -\sum_{i=0}^{N_g-1} i^2 p_{x+y}(i) \quad (20)$$

- Difference Entropy:

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\} \quad (21)$$

- Information measure of correlation 1:

$$f_{12} = \frac{HXY - HXY_1}{\max\{HX, HY\}} \quad (22)$$

- Information measure of correlation 2:

$$f_{13} = (1 - \exp[-2(HXY_2 - HXY_1)])^{1/2} \quad (24)$$

Where

$$XHY = -\sum_i \sum_j p(i, j) \log(p(i, j))$$

HX, HY are the entropies of p_x and p_y

$$HXY_1 = -\sum_i \sum_j p(i, j) \log\{p_x(i)p_y(j)\} \quad (25)$$

$$HXY_2 = -\sum_i \sum_j p_x(i)p_y(j) \log\{p_x(i)p_y(j)\} \quad (26)$$

The features calculated from the filtered signal spectrum, were also extorted from the periodogram, in addition to other features that have been extracted from this representation:

- the variance performed according to (7);
- the bias (the moment of order 1):

$$B = \frac{1}{N} \sum_{i=0}^N (x_i - \bar{x}) \quad (27)$$

The whole frequency features constitute a 57 component feature vector.

D. SVM Classification

The performance of an obstacle categorization system depends on obstacle representation as well as on classification algorithm. In our system, we choose to apply SVM classifier in the classification task. SVM consists in a group of supervised learning methods that can be applied in classification. SVMs

are used in many real-world applications such as text categorization, hand-written, character recognition, image classification, etc., and they are now established as one of the standard tools for machine learning and data mining [46]. The use of SVM classifier is interesting because it minimizes the bound taking into account empirical error and classifier complexity at the same time. In this way, SVMs are able of learning in sparse, high dimensional spaces with relatively few training examples [43]. They used an optimal hyper-plane as a decision function (Cf, Fig. 7). Thus, the optimal separating hyper-plane is used to classify an unlabeled input data, by using the following decision function:

$$f(X) = \text{sign}\left(\sum_{x_i \in SV} (y_i \alpha_i K(x_i, X) + b)\right) \quad (28)$$

where SV is the set of support vector items x_i , b is the offset value, K is the kernel, α_i are the optimized Lagrange parameters and y_i is the label of x_i , y_i may be 1 may be -1.

The optimal separating hyper-plane is the one that maximizes the distance between itself and the nearest data point of each class as shown in Fig. 7.

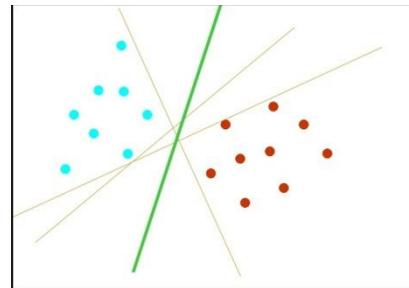


Fig. 7. The optimal separating hyper-plane

Different types of kernel can be used, RBF, Polynomial, etc... The kernel type affects the performance of SVM classifier.

In our system, we use RBF kernel which is defined as:

$$K_{RBF}(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

Where x_i is the support vector, x_j is the testing data point and γ determines the area of influence this support vector has over the data space.

We consider three classes of indoor environment objects: even surface, descending stairs and ascending stairs.

For each category of features, an SVM classifier is trained to separate these classes by using one-against-one strategy.

VI. EVALUATION OF THE PROPOSED APPROACH

Two raw data sets are constructed, the first is for the estimation of the optimal separating hyperplane parameters and the second for generalization, using the estimated optimal separating hyperplane.

Fig. 10 and Fig. 11 show the signal preprocessing procedure, feature extraction and classification in, respectively, time domain and frequency domain.

Tab.II and Tab. III show the classification performances by use of ultrasonic signal in time domain for, respectively, the training data set and the generalization data set.

TABLE II. CLASSIFICATION RATES OBTAINED BY TRAINING DATA SET USING THE CONSIDERED TIME FEATURES

(Input) Known as :	(Output) Classified as:		
	Even floor	Ascending stair cases	Descending stair cases
Even floor	0.46%	0.43%	0.11%
Ascending stair cases	0.05%	0.95%	0%
Descending stair cases	0.13%	0.2%	0.67%
Accuracy	70.73%		

TABLE III. CLASSIFICATION RATES OBTAINED BY GENERALIZATION DATA SET USING THE CONSIDERED TIME FEATURES

(Input) Known as :	(Output) Classified as:		
	Even floor	Ascending stair cases	Descending stair cases
Even floor	0.45%	0.40%	0.15%
Ascending stair cases	0.09%	0.91%	0%
Descending stair cases	0.08%	0.5%	0.42%
Accuracy	60.26%		

The results show a high ambiguity between even floor and ascending stair cases in the time domain. This problem is illustrated in Fig. 8 and Fig. 9 which represent the separating power of, respectively, "mean" and "maximum" features. We can clearly note the problem of no distinction between this two classes.

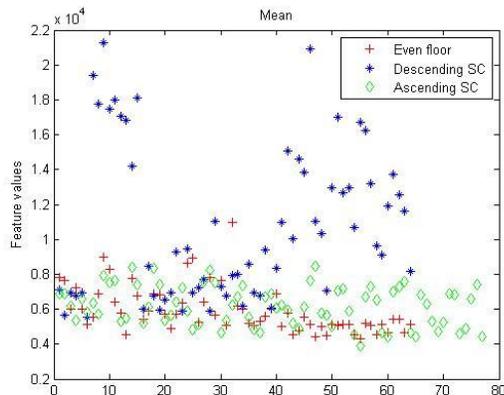


Fig. 8. The separating power of the 'mean' feature

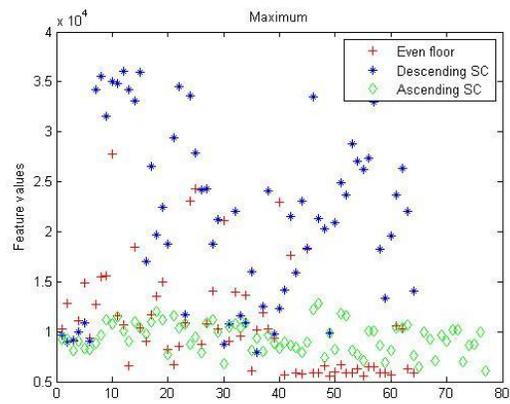


Fig. 9. The separating power of the 'maximum' feature

Such confusion is resolved by using the frequency domain features.

Tab.IV and Tab.V show classification performances of ultrasonic signal in frequency domain for the training data set and respectively for the generalization data set.

TABLE IV. CLASSIFICATION RATES OBTAINED BY TRAINING DATA SET USING THE CONSIDERED FREQUENCY FEATURES

(Input) Known as :	(Output) Classified as:		
	Even floor	Ascending stair cases	Descending stair cases
Even floor	0.75%	0.19%	0.06%
Ascending stair cases	0.10%	0.90%	0%
Descending stair cases	0.19%	0.04%	0.77%
Accuracy	80.97%		

TABLE V. CLASSIFICATION RATES OBTAINED BY GENERALIZATION DATA SET USING THE CONSIDERED FREQUENCY FEATURES

(Input) Known as :	(Output) Classified as:		
	Even floor	Ascending stair cases	Descending stair cases
Even floor	0.89%	0.11%	0%
Ascending stair cases	0.26%	0.71%	0.03%
Descending stair cases	0.32%	0.12%	0.56%
Accuracy	80.97%		

The results obtained from classification in frequency domain show that the confusion between even floor and ascending stair cases are really weakened and it is clearly seen from these results that the best classification is obtained while projecting raw data on different representation of ultrasonic in signal frequency domain.

Each representation allows having information that cannot be obtained from the others representations. The fusion of this information reduces the imperfection of the data and improves the system's performance.

Coarse to fine strategy

In our system, we look at meeting not only efficiency and precision, but it is also necessary to ensure optimal processing time. Hence, we are motivated to reduce the running time of the system processing.

When a blind people navigates in his environment, he needs to be alerted only when the cane detects ascending or descending stair cases. So, it is not necessary to classify the environment on three classes every time. Indeed, we propose a strategy which is based on two levels of classification in order to speed up the classification process, without compromising recognition performance. The first level is ensured by a strong SVM classifier that classifies the environment on two classes Even floor and Not Even Floor. Meanwhile, the second level is optional as it is only used if the decision of the first level's classifier is "Not Even Floor" (Cf. Fig. 12).

Tab.VI and Tab.VII show the classification performances by use of ultrasonic signal in frequency domain, on, respectively, the training data set and the generalization data set, while based on the new approach.

Our second strategy allows not only to decrease the time processing but also it provides significant improvement of classification performances. We can thus deduce that solutions based on multiple classifiers are more general than those based on one classifier.

TABLE VI. CLASSIFICATION RATES OBTAINED BY TRAINING DATA SET WITH TWO LEVELS OF CLASSIFICATION

<i>(Input) Known as :</i>	(Output) Classified as:	
	<i>Even Floor</i>	<i>Not Even Floor</i>
Even Floor	0.63%	0.37%
Not Even Floor	0.10%	0.90%
Accuracy	80.97%	

TABLE VII. CLASSIFICATION RATES OBTAINED BY GENERALIZATION DATA SET WITH TWO LEVELS OF CLASSIFICATION

<i>(Input) Known as :</i>	(Output) Classified as:	
	<i>Even Floor</i>	<i>Not Even Floor</i>
Even Floor	0.85%	0.15%
Not Even Floor	0.18%	0.82%
Accuracy	82.76%	

Tab.VIII and Tab.IX show the classification performances, in ascending and descending stair cases, by use of ultrasonic signal in frequency domain, on, respectively, the training data

set and the generalization data set, while basing on the new approach.

TABLE VIII. CLASSIFICATION RATES OBTAINED BY TRAINING DATA SET WITH TWO LEVELS OF CLASSIFICATION

<i>(Input) Known as :</i>	(Output) Classified as:	
	<i>Ascending stair cases</i>	<i>Descending stair cases</i>
Ascending stair cases	0.99%	0.01%
Descending stair cases	0.06%	0.94%
Accuracy	96.45%	

TABLE IX. CLASSIFICATION RATES OBTAINED BY GENERALIZATION DATA SET WITH TWO LEVELS OF CLASSIFICATION

<i>(Input) Known as :</i>	(Output) Classified as:	
	<i>Ascending stair cases</i>	<i>Descending stair cases</i>
Ascending stair cases	0.94%	0.06%
Descending stair cases	0.16%	0.84%
Accuracy	89.83%	

VII. CONCLUSIONS AND PERSPECTIVES

Blinds and visually impaired people need some aid to interact with their environment with more security. Accordingly, a multi-sensor system that scans floor surfaces and detects the presence of stairs was developed.

In this paper, we have presented a new electronic tool that incorporates two ultrasonic sensors and one monocular camera, intended for visually impaired assisting. Only one ultrasonic sensor was used to detect and identify three floor states, even floor, ascending stair case and descending stair case. To this end, we developed an approach for detection as well as identification of floor states. Such performances are challenging, since no existing solutions has proposed detecting stairs. Besides, most of existing tools aiming to detect objects basing on ultrasonic measurements make use of a series of ultrasonic sensors. The recognition result is estimated to 82.7% for detecting stair presence and 89.8% for precising if it consists in either ascending or descending type.

The results of this study allowed us to prove how much using one ultrasonic sensor to recognize the floor state is interesting. The recognition result is not perfect, as it doesn't reach the zero error performance, that is crucial for the tool that we are developing, but it is sufficiently satisfactory to contribute in the decision.

Our future works will focus on this topic. Indeed, we are working on merging data captured from two different sources of knowledge, precisely ultrasonic sensor and monocular camera, to improve the system's performances.

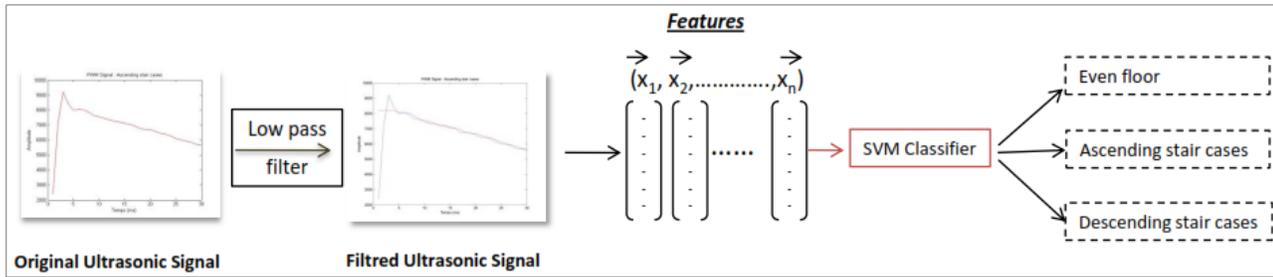


Fig. 10. Proposed strategy : Feature extraction and classification in time domain

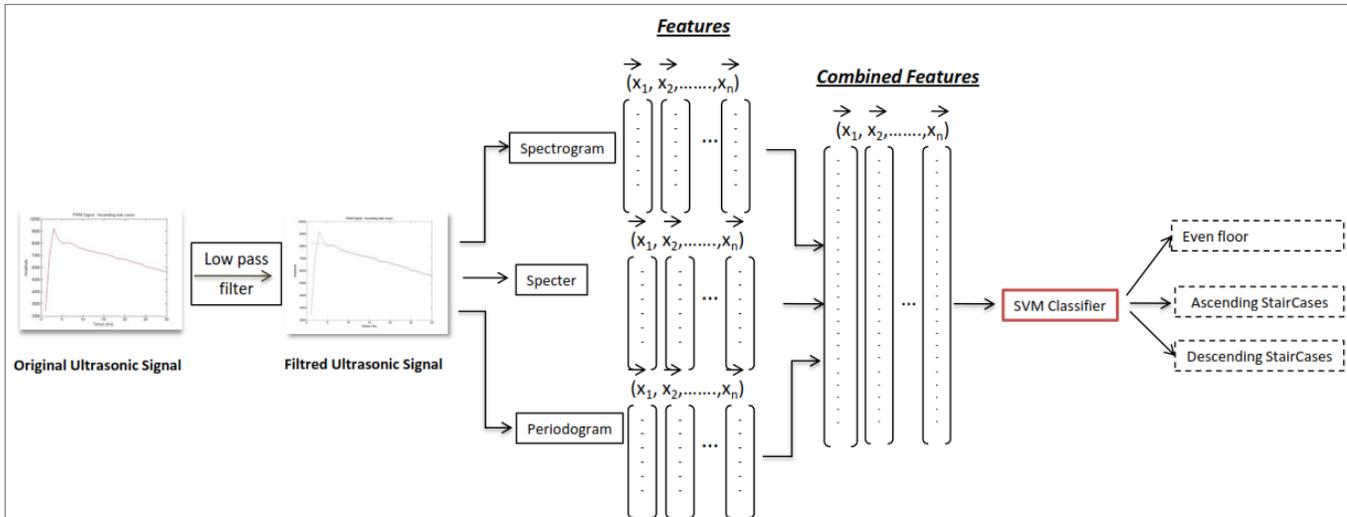


Fig. 11. Proposed strategy : Feature extraction and classification in frequency domain

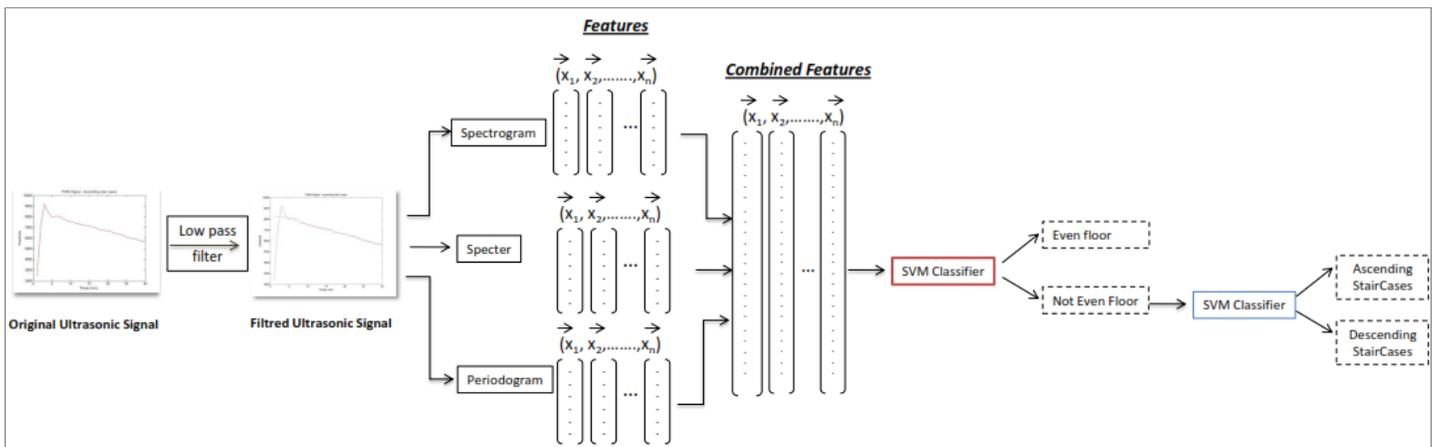


Fig. 12. Proposed strategy : two levels of classification in frequency domain

REFERENCES

- [1] J. M. Benjamin, N. A. Ali, and A. F. Schepis, "A Laser Cane for the Blind" Proceedings of the San Diego Biomedical Symposium, vol. 12, pp. 53--57, 1973.
- [2] R. Farcy, R. Leroux, R. Damaschini, R. Legras, Y. Bellik, C. Jacquet, J. Greene and P. Pardo, "Laser Telemetry to improve the mobility of blind people: report of the 6 month training course", ICOST 2003 1st International Conference On Smart homes and health Telematics Independent living for persons with disabilities and elderly people, Paris, pp. 24--26, September 2003.
- [3] B. Hoyle, D. Withington and D. Waters, "UltraCane", Available from: "<http://www.soundforesight.co.uk/index.html>", June 2006
- [4] T. Terlau and W. M. Penrod, "K'Sonar Curriculum Handbook", Available from: "<http://www.aph.org/manuals/ksonar.pdf>", June 2008
- [5] L. Whitney, "Smart cane to help blind navigate", Available from: "http://news.cnet.com/8301-17938_105-10302499-1.html", 2009.
- [6] E. Kee, "iSONIC cane for the virtually impaired", Available from: "<http://www.ubergizmo.com/2011/01/isonic-cane-for-the-virtually-impaired/>", 2011.
- [7] J. Borenstein and I. Ulrich. "The GuideCane A Computerized Travel Aid for the Active Guidance of Blind Pedestrians". Proceedings of the IEEE International Conference on Robotics and Automation, Albuquerque, NM, pp. 1283--1288, Apr. 21-27, 1997.
- [8] K. Takeuchi, "The Palm Sonar", Available from: "<http://www.palmsonar.com/>", 2010.
- [9] MaxBotix, "LV-MaxSonar-EZ4 Data Sheet", Available from: "http://www.maxbotix.com/documents/MB1040_Datasheet.pdf", 2005.
- [10] S. Park, L. Kim, S. Ha, H. Cho and S. Y. Lee, "An Electronic Aid for a Visually Impaired Person Using an Ultrasonic Sensor", International Conference on Coastal Engineering (ICCE 2012), Las Vegas, pp. 10--14, Jan. 2009.
- [11] S. A. Bouhamed, J. F. Elleuch, I. K. Kallel, D. S. Masmoudi, "New electronic cane for visually impaired people for obstacle detection and recognition", IEEE International Conference on Vehicular Electronics and Safety (ICVES), Istanbul, pp. 416--420, 2012.
- [12] linksprite, DataSheet: "LinkSprite JPEG Color Camera Serial UART Interface", 2005.
- [13] D. Yuan and R. Manduchi, "A Tool for Range Sensing and Environment Discovery for the Blind", Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04), 2004.
- [14] D. Yuan and R. Manduchi, "Dynamic Environment Exploration Using a Virtual White Cane", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, pp. 243--249, 2005.
- [15] M. D. Adams, "On-Line Gradient Based Surface Discontinuity Detection for Outdoor Scanning Range Sensors", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1726--1731, 2001.
- [16] M. Lee and S. Lee, "Design and analysis of an infrared range sensor system for floor-state estimation", Journal of Mechanical Science and Technology, pp. 1043--1050, 2011
- [17] E. Mihankhah, A. Kalantari, E. Aboosaeed, H. D. Taghirad, S. Ali and A. Moosavian, "Autonomous Staircase Detection and Stair Climbing for a Tracked Mobile Robot using Fuzzy Controller", Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics, Bangkok, Thailand, pp. 1980--1985, 2009
- [18] S. Se and M. Brady, "Vision-based Detection of Stair-cases", Fourth Asian conference on computer Vision, ACCV 2000, Vol. I, pp. 535--540, 2000.
- [19] I. E. Dror, M. Zagaeski and C. F. Moss, "Three-dimensional target recognition via sonar: a neural network model", Neural Networks, Vol. 8, No. 1, pp. 149--160, 1995.
- [20] J. M. Oria and A. M. G. Gonzalez, "Object recognition using ultrasonic sensors in robotic application", IECON, 19th Annual Conference of IEEE Industrial Electronics, pp. 1927--1931, 1993.
- [21] C. Barat and N. Ait Oufroukh, "Classification of indoor environment using only one ultrasonic sensor", IEEE Instrumentation and Measurement Technology Conference, 2001.
- [22] O. Bozma and R. Kuc, "Building a sonar map in a specular environment using a single mobile sensor", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. Vol. 13, NO. 12, pp. 1260--1269, 1991.
- [23] V. T. Pham, Q. Qiu, A. A. P. Wai and J. Biswas, "Application of ultrasonic sensors in a smart environment", Pervasive and Mobile Computing, Vol. 3, pp. 180--207, March 2007.
- [24] G. N. Marichal, A. Hernández, L. Acosta and E. J. González, "A Neuro-Fuzzy System for Extracting Environment Features Based on Ultrasonic Sensors", Sensors, Vol. 9, pp. 10023--10043, 2009.
- [25] MaxBotix, "LV-MaxSonar-EZO High Performance Sonar Range Finder", Available from: "http://www.maxbotix.com/documents/MB1000_Datasheet.pdf", 2005.
- [26] K. K. Shung and G. A. Thieme. "Ultrasonic Scattering in Biological Tissues". CRC, N.W. Boca Raton, Florida, 1992.
- [27] L. Vergara, R. Miralles, J. Gosalbez, J. V. Fuente, U. L. Gomez, J. JAnaya, M. G. Hernandez and M. A. Izquierdo. "On estimating concrete porosity by ultrasonic signal processing techniques". In 17th ICA, Roma, September 2001.
- [28] M. J. W. Povey. "Rapid Determination of Food Material Properties, Ultrasound in Food Processing". Blackie Academic & Professional, London, Wenheim, New York, 1997.
- [29] L. Kleeman and R. Kuc, "Mobile robot sonar for target localisation and classification", The international Journal of Robotics Research, Vol. 14, No. 4, pp. 295--318, 1995.
- [30] D. DeFatta, J. Lucas, and W. Hodgkiss, "Digital Signal Processing". Wiley, 1988.
- [31] B. Barchan and R. Kuck, "Differentiating sonar reflections from corners and planes by employing an intelligent sensor", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 12, No. 6, pp. 560--569, 1990.
- [32] I. E. Dror, M. Zagaeski and C. F. Moss, "Three-dimensional target recognition via sonar: a neural network model", Neural Networks, Vol. 8, No. 1, pp. 149--160, 1995.
- [33] S. Bae, L. Udpa, and S. Udpa, "Classification of Ultrasonic Weld Inspection Data Using Principal Component Analysis", Review of Progress in Quantitative Nondestructive Evaluation, vol. 16, pp. 741--748, 1997.
- [34] D. Berry, L. Udpa, and S. S. Udpa, "Classification of Ultrasonic Signals via Neural Networks", Review of Progress in Quantitative Nondestructive Evaluation, vol. 10A, pp. 659--666, 1991.
- [35] L. M. Brown and R. DeNale, "Classification of Ultrasonic Defect Signatures Using An Artificial Neural Network", Review of Progress in Quantitative Nondestructive Evaluation, vol. 10A, pp. 705--712, 1991.
- [36] H. Hamadene, and E. Colle, "A method based on neural networks for the recognition of the environment scanned by ultrasonic sensor", EUFIT'96 Fourth European Congress on Intelligent Techniques and soft Computing, Vol 1, Aachen, Germany, pp. 249--254, September 2-5, 1996.
- [37] G. Lindstedt, and G. Olsson, "Using ultrasonics for sensing in a robotic Environment", IEEE, 1993.
- [38] H. Peremans, J. M. K. Audenaert and V. Campenhout, "A high resolution sensor based on tri-aural perception", IEEE Transaction on Robotics and Automation, Vol. 9, N°. 1, pp. 36--48, 1993.
- [39] K. R. Rao, "Discrete Transforms and their Applications". Van Nostrand Reinhold, New York, 1985.
- [40] R. Kuc, "Introduction to Digital Signal Processing". McGraw-Hill, 1988.
- [41] J. B. Santos and F. Perdigao, "Automatic defects classification - a contribution", NDT & E International, vol. 34, pp. 313--318, 2001.
- [42] S. J. Song, H. J. Kim, and H. Cho, "Development of an intelligent system for ultrasonic flaw classification in weldments", Nuclear Engineering and Design, vol. 212, pp. 307--320, 2002.
- [43] V. Vapnik, "Statistical Learning Theory, wiley Interscience publication", 1998.
- [44] R. M. Haralick, K. Shanmugam, and I. Dinstein. "Textural features for image classification". IEEE Transactions on Systems, Man and Cybernetics, Vol. 3, No. 6, pp. 610--621, Nov 1973.
- [45] URA-Sfax, Available from: "<http://www.ura-sfax.org/fr/regional.php>", 2006.
- [46] G. H. John, "Enhancements to the Data Mining Process", thesis, 1997.
- [47] A. C. Scherlen, J. C. Dumas, B. Guedj, A. Vignot, "RecognizeCane": The new concept of a cane which recognizes the most common objects and safety clues", Proceedings of the 29th Annual International Conference of the IEEE EMBS, France, pp 63566359, 23-26 August 2007.

Watermarking in E-commerce

Peyman Rahmati, and Andy Adler

Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada

Thomas Tran

School of Information Technology and Engineering
University of Ottawa
Ottawa, Canada

Abstract—A major challenge for E-commerce and content-based businesses is the possibility of altering identity documents or other digital data. This paper shows a watermark-based approach to protect digital identity documents against a Print-Scan (PS) attack. We propose a secure ID card authentication system based on watermarking. For authentication purposes, a user/customer is asked to upload a scanned picture of a passport or ID card through the internet to fulfill a transaction online. To provide security in online ID card submission, we need to robustly encode personal information of ID card's holder into the card itself, and then extract the hidden information correctly in a decoder after the PS operation. The PS operation imposes several distortions, such as geometric, rotation, and histogram distortion, on the watermark location, which may cause the loss of information in the watermark. An online secure authentication system needs to first eliminate the distortion of the PS operation before decoding the hidden data. This study proposes five preprocessing blocks to remove the distortions of the PS operation: filtering, localization, binarization, undoing rotation, and cropping. Experimental results with 100 ID cards showed that the proposed online ID card authentication system has an average accuracy of 99% in detecting hidden information inside ID cards after the PS process. The innovations of this study are the implementation of an online watermark-based authentication system which uses a scanned ID card picture without any added frames around the watermark location, unlike previous systems.

Keywords—Data hiding; geometric distortion; watermarking; print-and-scan; and E-commerce Introduction

I. INTRODUCTION

In E-commerce, one clear concern of content owners is unauthorized reproduction of digital products. Copyright owners seek methods to control and detect such reproduction, and therefore research on digital product copyright protection has significant practical significance for E-commerce. Most electronic commerce systems use cryptography to secure the electronic transaction process [1]. Encryption provides “data confidentiality, authentication, data integrity, and in some cases authentication of the parties involved” [1, 2]. Copyright protection involves the authentication of the ownership and can be used to identify illegal copies. To detect reproduction of a digital product, a digital watermark created from information about the relationship between the product and its owner can be used. This information may be perceptible or imperceptible to the human senses. In 2009, Hirakawa and Iigima evaluated the effectiveness of using digital watermark technology for E-commerce website protection; and they reported a 60% reduction in the quantity of unauthorized content on E-commerce websites when protected by Digital watermarking technology [3]. In 2008, Sherekar et al.

recommended that the watermarks for images in e-governance and e-commerce applications should be invisible for human eyes and robust for possible attacks, such as geometric attack, and compression attack (JPEG or other image compression formats) [4].

A number of watermarking algorithms have been proposed over twenty years [5, 6]. Friedman proposed a trusted digital camera, which embeds a digital signature for each captured image [7]. With the digital signature, one can verify that the image is not changed as well as identify a specific camera that pictured the image [7]. Yeung and Mintzer proposed an authentication watermark that uses a pseudo random sequence and a modified error diffusion method to protect the integrity of images [8]. Lin and Chang proposed a scheme to insert authentication data in JPEG coefficients so that the authentication watermark has resilience against JPEG compression [9]. Wong and Memon proposed a secret and public key image watermarking schemes for grayscale image authentication [10]. Digimarc Corporation developed a search engine, MarcSpider, to search web sites for images that contain Digimarc watermarked images. When watermarked images are found, the information is reported back to the registered owners of the images [11]. In [12, 13], it has been shown how documents can be marked so that they can be traced in the photocopy process.

One of the most common attacks for watermarked multimedia products is Print-Scan (PS) process as the watermark can be degraded by the PS operation used once or several times [14]. The robustness of watermarking algorithm against PS attack for the online authentication system is a new, important challenge in multimedia communication security as well as E-commerce [15]. The progress in Print-Scan resilient watermarking will ease promoting watermark-based E-commerce and provide the ground for copyright tracking to prevent any illegally copying after selling a digital watermark product. This study proposes a watermark-based E-commerce model designed for online, secure ID card submission. The proposed model in comparison with preceding models has five new preprocessing blocks in the decoder with the role of providing robustness for watermarking algorithm against PS distortions (figure 1 and figure 3).

The applications of the proposed online ID card based authentication system are where 1) a seller needs to check the identity of a buyer before successfully completing the trade through the internet and 2) an applicant needs to electronically submit his/her Passport/ ID card to a high security organization. For example, a company for authentication purposes may ask customers to upload a scanned picture of

their watermarked Passport or watermarked ID card to continue a trade with them through the internet. The watermark extracted from the uploaded ID card image, which is already scanned from the hardcopy of the ID card, determines the genuineness of the hardcopy.

This paper is organized as follows: In the next section, we review related work and discuss their drawbacks; Section III is to explain the proposed ID card authentication method and also to detail the design of the five proposed preprocessing blocks in the decoder; Section IV discusses the achieved experimental results; and finally conclusion and suggestion for the further research are offered in section V.

II. RELATED WORKS

In the print-scan resilient data hiding area, distortion parameters quantification due to print-scan operation is challenging. There are several papers that model Print-Scan distortions [16, 17]. Generally, we can divide these distortions into three parts [18]: 1) "Randomness": The printed and scanned image is highly different than the original digital image. 2) "Man-dependency": the setting of the printer and the scanner may change; and also the paper orientation in printer's paper input tray and on flatbed of scanner may change during the PS process. 3) "Indistinguishability": the distortion of PS process is an accumulation from both printer and scanner. The regained watermark from the inspected data is applied for authentication in different ways, such as localizing the occurred distortions [19-20] or recognizing the type of attacks performed [21].

The main challenge in online authentication system is to overcome the print-scan distortions, which is considered as a combination of different attacks [22-24]. Longjiang Yu [18] proposes a print-and-scan model so that his work is realized in the presence of an added rectangular frame around the watermark location. This rectangular frame around watermark location makes it easier to find the geometric distortion along the print-scan process, and to localize the watermark location. The main drawback of using a frame around the watermark location is that in various types of authentication applications either the presence of this frame is not allowed or not favorable.

For example, it might not be allowed in important documents, such as passport, driving license, and ID cards. Solanki et al. proposed a print-scan resilient data hiding algorithm analyzing halftone effect (intensity shift) occurred after print-and-scan operation for the sake of presenting a model of print-and-scan process [17].

The main drawback of their method is that halftone analyzing in PS operation is severely dependent to the hardware features of Printer and Scanner, which are variable from one commercial brand to another.

This paper proposes a new online ID card authentication model which is different, compared with the preceding models [17, 18], in this way that it does not offer any added rectangular frame around the watermark location, and also there is no need to model halftone effect occurred after PS process.

III. METHOD

In this work, we used ID cards as the required document needed to be submitted through the internet for online authentication purposes. The proposed model establishes a linkage between the ID card holder's photo and his/her personal identification number, considered as the watermark. Also, we used a simple block-based watermarking algorithm in spatial domain, and proposed five preprocessing blocks in the decoder to remove the PS distortions (figure 1). The proposed online ID card based authentication system for E-commerce has the same security design similar to [25]; however, improves its robustness against PS attack/ distortion. In [25], Ingemar et al. proposed a new Security Architecture of a watermark-based E-commerce model which involves watermarking as an extra security for the online authentication system along with cryptography (figure 2). In Figure 2, the original data (payload) is first encrypted and watermarked in encoder (transmitter) and then sent to a decoder (receiver) through the internet to be decrypted and extracted. In this architecture, watermarking has been used as a security layer to add extra personal information about the user (customer) to the original data (payload) to increase the security of online authentication system in E-commerce. In the following, we will see the design of the proposed ID card based authentication system.



Fig.1. Overall schematic of the proposed ID card based authentication system.

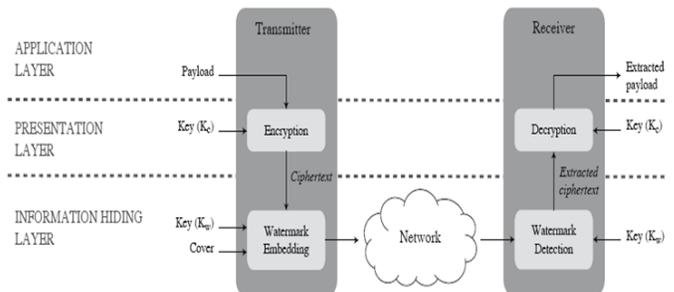


Fig.2. Representing the security architecture of a watermark based authentication system, reproduced from [25].

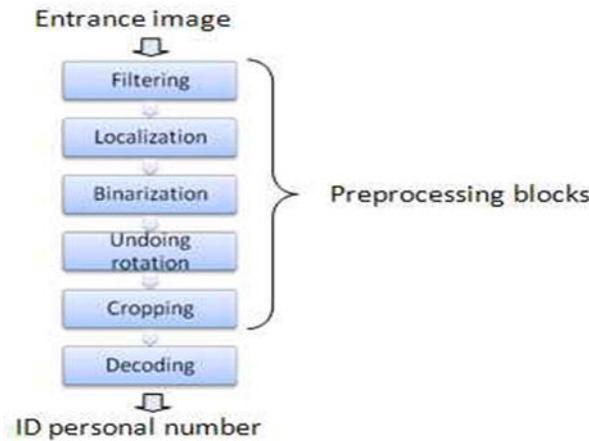


Fig.3. Representing the name and the sequence of applying the five proposed preprocessing blocks in the decoder to remove PS distortions.

A. Encoder

In the first step of authentication system, we need to hide the personal information of user/customer into the original digital ID card image. We use block-based embedding due to its simplicity in implementation, providing low computing time for real time application [26]. Several approaches have been suggested to achieve digital data hiding: Some in the spatial domain [27, 28], some in the frequency domain [29, 30], some on the basis of quantization [31, 32], and some based on spread spectrum methods [33, 34]. In this work, a simple block-based watermarking method using Hadamard patterns in spatial domain is introduced. The personal information, which is ID card personal number, is embedded into the ID card's holder photo place in the encoder. Two Hadamard patterns (f_0 and f_1) with small changes in their intensities, lower frequency than the frequency of the intensity changes in the original image, are applied to embed the data stream (ID card personal number) into the original image. First, the original image is divided into blocks with dimension of $N \times N$ and each bit of the data stream is assigned to each block. Then embedding procedure follows up this rule: if the bit of data stream assigned to a block is 0, f_0 will add up to that block; if it is 1, f_1 will add up to the block, see figure 4. The correlation between patterns (f_0 and f_1) and the blocks (B) is zero. Suppose that I is an original image with the dimensions $N_1 \times N_2$ which is divided into blocks, with the dimensions $N \times N$. Since a bit of all desired data bits is embedded into each block, so, $(N_1 \cdot N_2) / N^2$ bits can be hidden into the original image. Note that $f_0(k,l), f_1(k,l); 0 \leq k, l \leq N-1$ are indicators of the Hadamard patterns, which has property of $f_0 = -f_1$, and $B(k,l)$ indicates the blocks. We can write the binary bit ($W(i, j)$) to be hidden as follows:

$$W(i,j) \in \{0,1\}; 0 \leq i \leq (N_1/N)-1, 0 \leq j \leq (N_2/N)-1 \quad (1)$$

The embedding algorithm will start by converting the designed patterns to an image matrix. Therefore, the watermarked image is:

$$I_w(m,n) = I(m,n) + \lambda \cdot f_w \left(\left\lfloor \frac{m}{N} \right\rfloor, \left\lfloor \frac{n}{N} \right\rfloor \right) (m \text{ Mod } N, n \text{ Mod } N), \quad (2)$$

where $0 \leq m \leq N_1-1, 0 \leq n \leq N_2-1$, $I_w(m, n)$ is the watermarked image, λ is Inductance Coefficient, $[x]$ is the largest integer that is smaller or equal to x , and Mod is residue of an integer division. The above equation when $f_0 = -f_1$ can be shortened as:

$$I_w(m,n) = I(m,n) + \lambda \cdot \left(2W \left(\left\lfloor \frac{m}{N} \right\rfloor, \left\lfloor \frac{n}{N} \right\rfloor \right) - 1 \right) \cdot f(m \text{ Mod } N, n \text{ Mod } N) \quad (3)$$

Inductance Coefficient (λ) compromise between visual quality of the watermarked image and resistance of the used method against attacks. The bigger the Inductance Coefficient, the lower the quality of the watermarked image will be. In (3), if the bit in the binary watermark placed at the position (i, j) of the original image is zero, the block related to (i, j) from the original image will induce with the pattern f_0 , and reversely, if the bit at the position (i, j) is one, the block related to (i, j) from the original image will induce with the pattern f_1 . Finally, the security of the algorithm can obtain by a watermark key is fed to the encoder and decoder blocks.

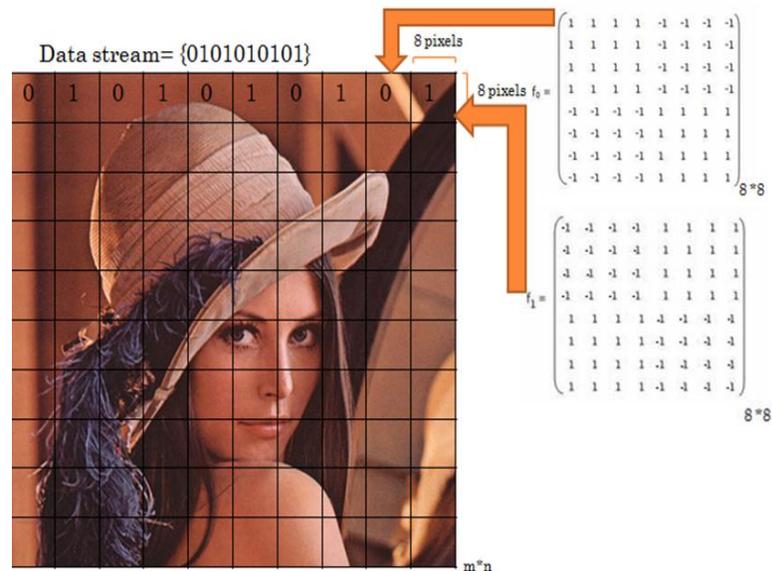


Fig.4. Representing an example of embedding the data stream with 10 bits into Lena image using two Hadamard patterns (f_0 and f_1), which are low frequency and $f_0 = -f_1$.

B. The Watermark localization in the decoder

The next step after embedding information in the encoder is decoding the information after attacking the authentication system by the PS operation. To provide robustness for PS distortions, five preprocessing blocks are proposed in the decoder. Figure 3 shows the name and the sequence of applying these preprocessing blocks in the decoder. As shown in the figure 3, the first block is named filtering block to remove the possible noise on the entrance image. Gaussian filter is used as a low pass filter to denoise the entrance image. Then, a localization block is proposed with the duty of estimating the location of watermark region (ID card's holder photo place). In this block an approximate watermark region is achieved and the remaining area out of this region is omitted.

Whereas, we embed the information into a rectangular frame, belonging to the ID card holder's photo place, in the encoder, therefore, we should look for a rectangular frame (watermark region) in the decoder. To localize the rectangular watermark region, we put a rectangular mesh over the entrance image to the decoder, see figure 6. The dimension of the rectangular elements inside the mesh can be calculated by having the maximum occurred rotation angle after PS operation. This maximum rotation angle (MRA) can be written by two parameters as:

$$MRA = \theta_{max} = \theta_i + \theta_u \quad (4)$$

Where θ_i is maximum rotation angle created by the printer, and θ_u is maximum probable rotation angle that may occurs by user in the scanner. Now, the maximum dimension of the rectangular frame can be evaluated as:

$$\begin{aligned} Height &\approx L \cdot \cos(\theta_{max}) + W \cdot \sin(\theta_{max}) + H0 \\ Width &\approx W \cdot \cos(\theta_{max}) + L \cdot \sin(\theta_{max}) + W0 \end{aligned} \quad (5)$$

Where L is the approximate height of the ID card's holder photo place (watermark region), and W is the approximate width of watermark location. Both of these parameters are known by having "dots per inch" (dpi), which is adjustable in printer and scanner setting. H0, W0 are the additional parameters to qualify our approximation, selected by user. Figure 5 represents equation (5). After applying the mesh over the entrance image, the rectangular watermark region is achieved by the rule: The rectangular element inside the mesh with the biggest width in its histogram specification is the one has the watermark region inside. This rule comes from this fact that the watermark is embedded in the photo place of ID card's holder which has biggest width in its histogram compared with other regions of ID card. Note that, the output of the localization block is an estimate of the watermark region and we still need to have next blocks to get the exact watermark location.

C. Binarization algorithm in the decoder

In the previous block an estimate of the watermark region, including regions without watermark, achieved. The binarization block, located after localization block in figure 3, is proposed with the duty of discriminating the exact watermark location from the other region without any watermark inside.

The proposed binarization method is based on the thresholding and tries to find the exact watermark location using histogram specification of the estimated watermark region, achieved from the localization block. Figure 7 shows the histogram specification of the estimated watermark region in localization block. The circular sign in this figure corresponds to the region without any watermark, which should be removed in this block. To remove the region without watermark, we need to use a threshold value to make a binary image from the entrance image.

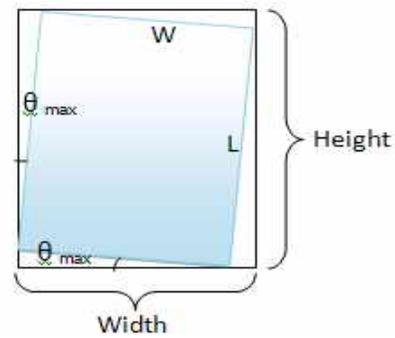


Fig.5. A schematic to get a frame with the maximum dimension based on the maximum probable rotation angle θ_{max} .



Fig.6. Outline of how a rectangular grid is applied on the entrance image to localize the watermark location.

The first local minimum (star sign in figure 7) around the circular point can be the initial guess of threshold value to make a binary image. As it is shown in Figure 8, if we consider the star point as the threshold value, the earned binary image will not show our desired watermark region, a rectangular frame. However, we consider this point as our initial threshold value. Looking at the histograms depicted in Figure 7 which belongs to the image in Figure 10(a) after print-scan operation, we can find out the existence of the gray levels shifting (halftone effect), appeared as several small peaks around the circular sign in figure 7. This is because of histogram distortion occurred after the print-scan operation.

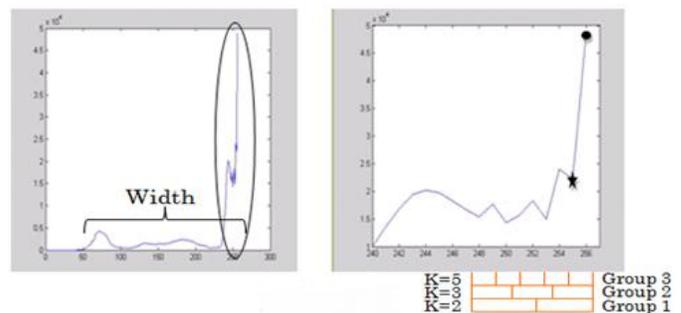


Fig.7. Histogram specification of the output image of localization block, which is an estimated image of watermark location.

In the left panel, histogram of the estimated watermark location which has the biggest width. In the right panel, a closer view of the black oval to show the occurred histogram distortion (halftone effect) after print-scan operation. The star sign on the right curve indicates the initial threshold value in binarization block, and K shows the number of the subdivisions.

A hierarchical algorithm to get the best threshold value which discriminates exactly the watermark location from other regions is proposed in this stage. Defining parameter P over the histogram as start points for searching the threshold value, and also parameter D that expresses the distance of the search, P-D is the last search point. We divide the distance D into K equal subdivision, and it is supposed that the bin with the least local minimum in each subdivision is our desired threshold value (figure 7).

The local minima are achievable by differentiating from the histogram curve. The number of the binary images is equal to the number of our subdivision. By considering several K for a fixed D, we will establish several groups with different number of subdivisions, see Figure 8. The more number of groups, the more precision and the more computing time to select the optimum threshold value will be. In each group a truth criterion (dmin) for the binary images, earned based on the number of K used in each group, is considered as:

$$D_k = \sum_{m=1}^M [(V_{k,m} - V_{I,m})^2] ; d_{min} = \min [D_k] \quad (6)$$

where D_k is vector distance, $k=1,2,\dots,K$, and K is the number of subdivisions used in each group, and V_I is our ideal feature vector that includes M specified features and may be expressed as: $V_I = [V_{I,1}, V_{I,2}, \dots, V_{I,M}]^T$. As an example, the ideal feature vector can be chosen to include:

The number of pixels in the watermark location, the perimeter of the watermark location, the aspect ratio of the watermark location, and the ratio of the number of pixels in the watermark location (black rectangle in figure 10(d)) to the number of pixels out of watermark location (white region in figure 10(d)). Figure 9 shows the flowchart of the iterative binarization algorithm for a single group in figure 8. The binary image with the least distance criterion, dmin, in each group is considered as the output binary image in that group. The hierarchical process (figure 8) will be terminated in a group if the condition $dmin < THD$ in that group is met, where THD is an arbitrary number selected by user. In the end, the output binary image of the group with the least dmin is selected as the final binary image.

Figure 10 shows the experimental results of applying the proposed binarization block for different values of K. As it is shown in Figure 10, the higher the number of the subdivisions (K), the more the accuracy in estimating the watermark location (black region in figure 10 (d)) will be.

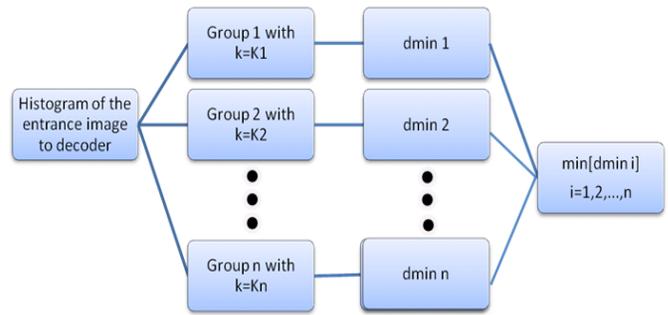


Fig.8. Hierarchical representation of finding the best threshold value from histogram specification of entrance image to the binarization block.

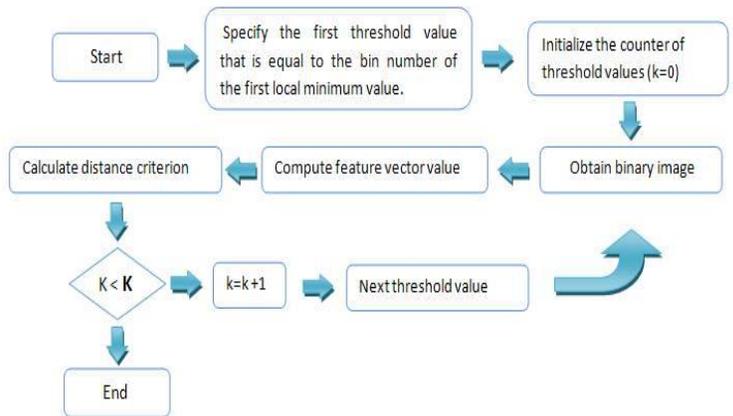


Fig.9. Iterative binarization algorithm for a single group in the hierarchical representation shown in figure 8.

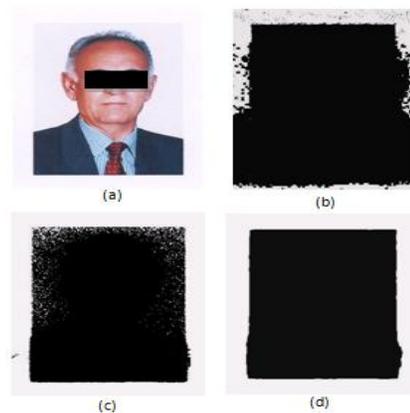


Fig.10. Example of three binary images earned from three different threshold values in an ID card with a white background. (a) The entrance image to the binarization block after print-scan operation, with a histogram specification drawn in Figure 7. (b) The obtained binary image after applying the first local minimum (star point in figure 7) as initial threshold value (Note that, the binary image does not show the watermark location accurately). (c) Obtained binary image by choosing a Threshold value achieved in group 3 with 3 subdivisions (Note that, the number of tested binary images is equal to $K=3$, and also the search distance was $D=10$). (d) Obtained binary image in group 5 with 5 subdivisions, which is the best achieved binary image.

D. Undoing the rotation

In this stage, we need to eliminate the occurred rotation angle on the obtained watermark location (black region in figure 10(d)) after print-scan operation. Several approaches have been proposed to undo the rotation [17, 18]. In [18], a template is produced by transforming the original image to a rotation space without the interpolation operation. By template matching, pixels of the original coordinate in the rotated image are defined and interpolation points are eliminated. After this process, pixels of the original coordinate are rotated back during the rotation restoration process; and non-integer coordinate is rounded into integer on some points. The benefit of this method is its capability in eliminating interpolation operation, that which preceding methods needed it for undoing rotation. The main drawback of [18] is its complexity in implementation which prohibited me to apply this method for undoing rotation. In this work, the method used for undoing rotation was Radon transform. The Radon transform is the projection of the image intensity on a radial line directed at a specific angle. We can estimate the rotation angle of the rectangular frame (watermark location) in the output image of binarization block using the following rule: the angle with the biggest projection value in its radon transform corresponds to the rotation angle of the rectangle frame. Applying the Radon transform, we can simply undo the rotation of the rectangle frame (watermark location), shown in black in figure 10 (d).

E. Cropping criteria

This phase of proposed authentication system is one of the most important units. This is because we need to crop the derotated image from the previous section at its optimum edges to achieve an accurate rectangular watermark location, where the personal information is hidden. Any mistake in estimating the optimum edges will result in the loss of hidden information inside the watermark location. The proposed algorithm to find the optimum edges uses two criterions: Average and Similarity criterion. We define two different regions: **Non-transition region** which is rows and columns with no intensity variation when we go from one row/column to its immediate adjacent row/column, and **Transition region** which is rows and columns with intensity variation when we go from one row/column to its immediate adjacent row/column, see figure 11. To reach to the optimum edges to crop the image, we need to first remove the non-transition region and then find the optimum edges within transition region. To remove the non-transition region, we define the average criterion (AVC) as follows:

$$\begin{aligned} AVC &= |AV(i) - N_i|; & \text{For rows} \\ AVC &= |AV(j) - N_j|; & \text{For columns} \end{aligned} \quad (7)$$

where N_i is average of the gray levels of the most outer pixels of the rectangular watermark location, and $AV(i)$ and $AV(j)$; $i=1,2,\dots,P$; $j=1,2,\dots,Q$ are, respectively, the average of the pixels in each row and column in non-transition region. Also, P and Q are the number of rows and columns of the watermark location respectively. The rows and columns having AVC lower than T , a threshold value selected by user, have to be removed. By doing so, we would remove the rows and columns without any transition. This criterion is done in a small distance from the most outer edges of the rectangular

watermark location, see figure 11(b). Note that, the existence of the region without any transition in Fig. 11(b) depends on the application and the value of rotation angle created by PS process. In cases with small applied rotation angle, we do not have any non-transition region. In the following, we need to choose the optimum edges within the transition region to crop the rectangular frame. To do so, we consider the homogeneity criterion for each one of rows or columns within the transition region. The homogeneity criterion (HMC) for rows and columns within the transition region is evaluated as:

$$HMC = \frac{VAR}{AVG}, \quad (8)$$

where VAR is the variance of gray levels of pixels in each row or column, and AVG is the average of the gray levels of the pixels in the same row or the column. It is supposed that the optimum edges within the transition region are located between two columns/rows with the highest similarity in intensity. Now, the similarity between the sequential rows/columns can be evaluated by taking the difference between the homogeneity values of those rows/ columns. Therefore, the similarity criterion (SCR) can be written as:

$$\begin{aligned} SCR &= |Hmc(i \pm 1) - Hmc(i)|; & \text{For rows} \\ SCC &= |Hmc(j \pm 1) - Hmc(j)|; & \text{For columns} \end{aligned} \quad (9)$$

Note that, the movement direction to compute the above equation is always from outer edges toward inner edges (see the direction of arrows in Figure 11). This difference (the similarity criterion) is the least amount at the optimum edges within the transition region. Therefore, we will consider a row or column as an optimum edge to crop the image if the SCR in that row or column is lower than a critical value CV , selected by user. We can write the cropping criterion (CRC) as follows:

$$CRC|_{i,j} = SCR|_{i,j} < CV \quad (10)$$

Finally, we deliver the cropped rectangular frame to the next decoding block to extract the hidden data inside it.

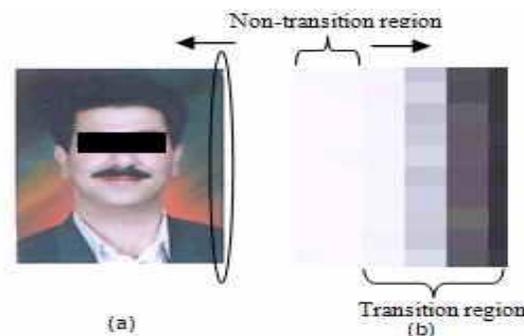


Fig.11. Example of the transition and non-transition regions in a test image after print-scan operation. (a) Representing a derotated rectangular watermark location. (b) A close view of the black oval in figure 11(a) to represent the transition region, and non-transition region. (Note the direction of drawn arrows in figure 11(a) and figure 11(b)).

F. Decoder

The final stage for the proposed authentication system is the decoding block. The decoding process is according to this

property that the original image has a minimum similarity to the Hadamard patterns (f_0 and f_1), which are already used in the encoder. This means that correlation between the blocks (B) and Hadamard patterns is always zero, i.e. $Corr(B, f_x)=0$ where B is the blocks inside the original image in figure 4 and f_x can be either f_0 or f_1 . A decision function for extracting a bit of the hidden data at the position (i, j) can be written as:

$$d(i, j) = Corr(B_{i,j}^W(k, l), f_1(k, l)) - Corr(B_{i,j}^W(k, l), f_0(k, l)), \quad (11)$$

where $B_{i,j}^W(k, l)$ is the block in the watermarked image. Since $B_{i,j}^W(k, l) = B_{i,j}(k, l) + \lambda \cdot f_x(k, l)$, and also $Corr(B, f_x)=0$; therefore, we can write:

$$d(i, j) = \lambda \cdot Corr(f_1, f_x) - \lambda \cdot Corr(f_0, f_x) = \begin{cases} +\lambda, & \text{if } x = 1 \\ -\lambda, & \text{if } x = 0 \end{cases} \quad (12)$$

where x is the unknown hidden bit in the block $B_{i,j}^W(k, l)$, and $Corr(X, W)$ is define as:

$$Corr(X, W) = \frac{\sum \sum (x - \bar{x})(w - \bar{w})}{\sqrt{(\sum (x - \bar{x})^2) \cdot (\sum (w - \bar{w})^2)}} \quad (13)$$

In the end, the decision function can be written as:

$$\hat{W}(i, j) = sgn(d(i, j)) = \begin{cases} +1 & \text{if } x = 1 \\ -1 & \text{if } x = 0 \end{cases} \quad (14)$$

Where $\hat{W}(i, j)$ is a bit of the binary hidden data at the position (i, j).

IV. EXPERIMENTAL RESULTS

The proposed ID card based authentication algorithm was tested on a Pentium IV (PC), Intel 3.0 GHz, with Windows XP Professional, 3.0 GB RAM, in MATLAB 8.0 (Mathworks, Natwick, USA). After several testes on the Hadamard patterns a low frequency template with the dimension 8×8 was selected.

In our experiments, we used typical printer and scanner with commercial brands: HP Photosmart 8450 and Canon L9950F, respectively. A database of 100 different ID cards with a wide range of possible colors, as background colors of the ID cards, was used. The original digital ID card image was printed with the resolution of 300 dpi, and then scanned with the resolution of 600 dpi. We embedded the ID card personal number, including 12 characters, inside the ID card's holder photo place in the encoder. The proposed ID card based authentication algorithm was applied to the whole of the database, including 100 ID cards, and an average accuracy criterion (ACC) was defines as:

$$ACC = 1 - \left(\frac{EB}{HB} \right) \quad (15)$$

Where EB is the number of detected bits with error, and HB is the number of all hidden bits. The average accuracy criterion has been depicted as a diagram in Figure 12 for different parameters introduced in the proposed ID card based authentication system. As it is obvious from the drawn diagram, the more the number of subdivisions (K) in the

binarization block, the higher the average accuracy of detecting the hidden data will be. In Figure 12, the average accuracy is increased from 80% to 86% when increasing the number of the used groups in binarization block from $K=1$ to $K=2$. In Figure 12, the ideal feature vector, VI , was applied to get the accuracy results with two features were: the number of pixels inside the watermark location, and the aspect ratio of the watermark location. Also in the case of applying four features to achieve the average accuracy in figure 12, the used features were the same ones mentioned in the binarization section. The threshold value (THD) was set to 0.05 and the number of the groups in binarization block was different between 3 and 5 over our database.

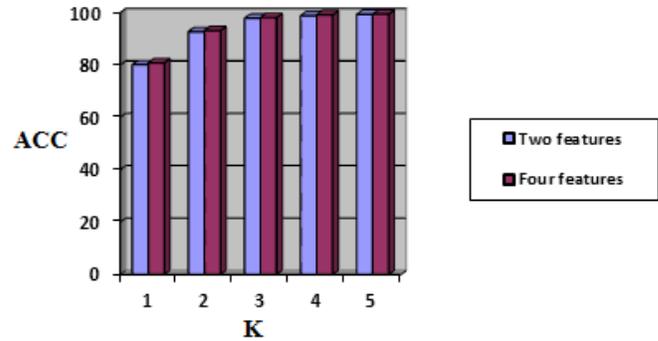


Fig.12. Representing a diagram to compare the average accuracy criterion for different values of subdivision number (K), and also different number of used features. (Note that, the selected value for distance of search, D , was equal to 10).

Figure 13 shows several case studies selected from our database before watermarking (figure 13(a)) and after watermarking and PS operation (figure 13(b)). Figure 13(c) depicts the Personal information, including 12 characters, embedded into the ID card's holder photo place.



Fig.13. Experimental results of applying the proposed authentication system over five case studies selected form our database, including 100 ID cards. (a) The original image of ID card's holder before watermarking. (b) The Image of ID card's holder after watermarking and PS operation, including ID card's holder personal information. The background of the study cases in this figure are different from one case to the other, ranged from light, plain background to the background with busy texture.

Figure 14 represents the effect of changing the Inductance Coefficient (λ) over the Signal to Noise Ratio (SNR) of the watermarked images in our database. The higher the inductance coefficient, the lower the SNR will be. This means that to preserve the visual quality of an image after watermarking, we need to select an appropriate value for λ .

TABLE I. Results of watermarked image quality (PSNR).

Methods	Images					
	Image (a)	Image (b)	Image (c)	Image (d)	Image (e)	Image (f)
Proposed method	44.3324	43.1223	40.5634	39.8912	45.6554	42.1342
Tsai's method	37.4323	38.2123	38.5654	38.2134	42.6723	39.3251

In all our experiments, the Inductance Coefficient after several times of examination was set to a fixed value of 8. With this value for λ , the watermarked image will be fairly imperceptible for human eyes, see figure 13(b).

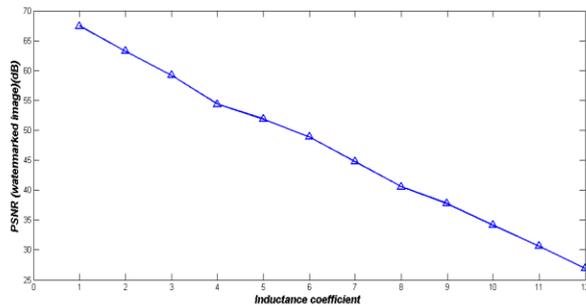


Fig.14. The effect of changing Inductance Coefficient (λ) over the visual quality of the watermarked images, calculated by using Signal to Noise Ratio (SNR) in decibel (dB), in our database. (Note that as the λ increases the SNR decreases, which means the watermark inside the image is more perceptible for human eyes).

Figure 15 is to assess the average accuracy of the proposed authentication model in detecting the hidden information into the watermark location when we do several Print-Scan operations in sequence. As the number of PS operations increases, the average accuracy of the authentication system decreases so that it reaches to 81% with $\lambda=8$ after doing PS operations for 6 times in sequence, which is still an acceptable accuracy (see figure 15).

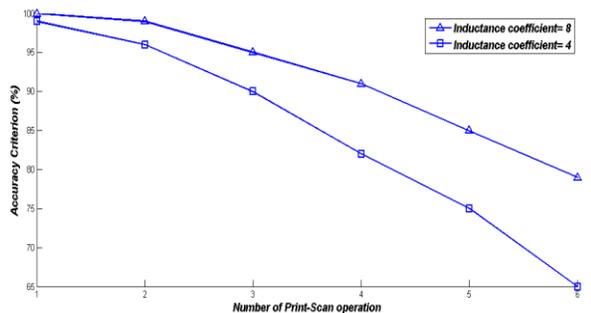


Fig.15. The average accuracy of the proposed authentication system for different values of λ when doing PS operations for several times in sequence.

In order to verify the effectiveness of the proposed watermark scheme, the method proposed by Tsai et al. [35] was also simulated for performance comparison. We used peak signal-to-noise ratio (PSNR) to evaluate the quality of the watermarked images using the two watermark methods. The PSNR is defined as:

$$MSE = \left(\frac{1}{m \times n}\right) \sum_{i=1}^m \sum_{j=1}^n (I(i, j) - I_w(i, j))^2$$

$$PSNR = 10 \log_{10} \frac{255 \times 255}{MSE} \text{ dB} \tag{16}$$

where I is the original image, I_w is the watermarked image, and $m \times n$ is the number of pixels in I . The PSNR values of each test image using the proposed method, and Tsai's method are summarized in Table 1. According to our experiments, if the PSNR value is greater than 36 dB, the watermark is almost invisible to the human eyes. To test the robustness of the proposed watermarking method, geometric distortions and common attacks including additive Gaussian noise, rotation, scaling, and cropping was performed to attack the watermarked images. Table 2 shows the experimental results of the average error ratio of the extracted watermarks of different attacks. The error ratio is defined as the number of extracted bits with error divided by the number of all hidden bits. In all cases, the results of our approach are better than the comparison method.

TABLE II. Results of average error ratio (%) of the extracted watermark at different attacks using the proposed method, and Tsai's method [35].

Attacks	Methods	
	Proposed method	Tsai's method
Additive uniform noise	11.14	12.67
Removed 1 row and 3 columns	6.19	6.34
Removed 3 row and 8 columns	17.98	19.47
Cropping ratio 90%	18.58	18.76
Cropping ratio 75%	31.24	32.36
Linear geometric transform (1.020,0.015,0.010,1.021)	6.75	8.91
Rotation 5°	1.12	3.45
Rotation 20°	6.13	7.23
Rotation 5° + cropping ratio 75%	30.93	34.75
Rotation 20°+ cropping ratio 90%	24.12	25.32

V. DISCUSSION AND CONCLUSION

This paper proposed a watermark-based E-commerce model to provide an online secure ID card authentication system which uses scanned picture of customer's ID card as identity. With the popularization of the internet and E-commerce and the expansion of E-government services, a variety of recorded data and documents that are relevant to such transactions and services are constantly created and exchanged electronically. In such situations, it is important to preserve the reliability of electronic data and documents by ensuring that the content cannot be altered. A watermark-based E-commerce model can provide us with an online secure ID card authentication system so that it is possible to learn whether the content of digital documents/data has been altered or not. Digital watermark technology embeds the user/customer's personal information into the digital content and makes it hard for criminals to abuse a content-based electronic business. In this work, the user/customer takes the scan of hardcopy of his/her ID card and then uploads the scanned picture of ID card through the internet for authentication purposes to fulfill an online trade/transaction. The proposed authentication system extracts the watermark inside the ID card's holder photo place in the decoder and then checks it out with the ID card personal number. If the extracted watermark and the ID card personal number are the same, the identity of the user/customer will be verified; otherwise, the identity will be denied. The main attack for the proposed authentication system is PS operation which imposes several distortions on the watermark location. To remove the PS distortions, five preprocessing blocks in the decoder are proposed. According to the experimental results, the proposed ID card authentication system has an average accuracy of 99% in finding correctly the hidden information into the 100 ID cards after PS operation. Unlike a preceding ID card authentication system [17], the proposed authentication method does not need to add a rectangular frame around the watermark location, which makes it applicable for online passport based authentication system. Moreover, the proposed authentication method outperforms the preceding proposed authentication system [18] in this way that it does not need to model the PS distortion (halftone effect), which is *variable* for printers and scanners from one brand to another, to remove the PS distortion on the watermark location. As the future work, we can use scanned picture of Passport as identity for the proposed authentication system. Also, this work is extendable where a frequency based watermarking algorithm is applied in the encoder and the decoder, in anticipation of achieving high quality watermarked images and higher average accuracy in finding correctly hidden information into the watermark location after PS operation.

REFERENCES

- [1] Schneier, B., Applied Cryptography, Second ed. John Wiley & Sons, New York, 1996.
- [2] Ford, W., Baum, M., Secure Electronic Commerce. Prentice Hall, Upper Saddle River, NJ, 1997.
- [3] Hirakawa, M., and Iijima, J. "Validating The Effectiveness of Using Digital Watermarking Technology for E-commerce Website Protection" The 9th Asian eBusiness Workshop, pp. 127-132, Japan, 2009.
- [4] Role of Digital Watermark in E-governance and E-commerce" IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 1, 2008.
- [5] Podilchuk C.I., Delp E.J.: Digital watermarking: algorithms and applications, IEEE Signal Processing Magazine, Vol. 18 (2001) 33-46.
- [6] Lee S.J., Jung S.H.: A survey of watermarking techniques applied to multimedia, Proc. of IEEE International Symposium on Industrial Electronics (ISIE), vol. 1 (2001) 272-277.
- [7] Friedman G.L.: The trustworthy digital camera: Restoring credibility to the photographic image, IEEE Trans. Consumer Electron., vol. 39 (1993) 905-910.
- [8] Yeung M.M., Mintzer F.: An invisible watermarking technique for image verification, Proc. ICIP (1997) 680-683.
- [9] Lin C.Y., Chang S.F.: A robust image authentication method surviving JPEG lossy compression, Proc. SPIE, vol. 3312 (1998) 296-307.
- [10] Wong P.W., Memon N.: Secret and public key image watermarking schemes for image authentication and ownership verification, IEEE Trans. Image Processing, vol. 10 (2001) 1593-1601.
- [11] Digimarc Corporation, PictureMarcTM, MarcSpiderTM, <http://www.digimarc.com>
- [12] Brassil, J., Low, S., Maxemchuk, N., O'Gorman, L., Electronic Marking and Identification Techniques to Discourage Document Copying. In Infocom94, 1994.
- [13] Brassil, J., O'Gorman, L., Maxemchuk, N., Low, S., Document Marking and Identification using both Line and Word Shifting. In Infocom95, Boston, MA, April 1995, 853-860.
- [14] K. Solanki, U. Madhow, B. S. Manjunath, S. Chandrasekaran and I. El-Khalil, "Print and scan" resilient data hiding in images," *IEEE Trans. Information Forensics and Security.*, vol. 1, no. 4, pp, 464- 478, Dec. 2006.
- [15] Hyejoung Yoo, Kwangsoo Lee, Sangjin Lee, and Jongin Lim, "Off-Line Authentication Using Watermarks" Springer-Verlag Berlin Heidelberg, ICICS 2001, LNCS 2288, pp. 200-213, 2002.
- [16] C. Y. Lin and S. F. Chang, "Distortion modeling and invariant extraction for digital image print-and-scan process," presented at the Int. Symp. Multimedia Information Processing Dec. 1999.
- [17] K. Solanki, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Estimating and undoing rotation for print-scan resilient data hiding," presented at the ICIP, Singapore, Oct. 2004.
- [18] L.Yu, X. Niu and S. Sun, "Print-and-scan model and the watermarking countermeasure," in Image and Vision Computing., May 2005, vol. 23, pp. 807- 817.
- [19] R. B. Wolfgang and E.J. Delp, "Fragile watermarking using the VW2d watermark", Proceeding of the SPIE/IS&T International Conference on Security and Watermarking of Multimedia Contents, vol. 3657, pp. 204-213, Jan. 1999.
- [20] J.Hu, j. Hunang, D.Hunang and Y. Q. Shi, "Image fragile watermarking based on fusion of multi-resolution tamper detection," IEE Electronic Letters, vol. 38, no. 24, pp 1512-1513, Nov. 2002.
- [21] D. Kundur and D.Hatzinakos, "Digital watermarking for telltale temper-proofing and authentication," proceedings of the IEEE Special Issue on Identification and Protection of Multimedia Information, vol. 87, no. 7, pp. 1167-1180, July 1999.
- [22] A. T. S. Ho, J. Shen, H. P. Tan, and J. Woon, "Security-printing authentication using digital watermarking," Electronic Imaging, vol. 13, no.1, Jan. 2003.
- [23] J. Mercer, Authentication News, 5 (9/10), 2001.
- [24] C.-Y. Lin and S.-F. Chang, "Distortion modeling and invariant extraction for digital image print-and-scan process," Intl. Symp. on Multimedia Information Processing, Taipei, Taiwan, Dec. 1999.
- [25] Ingemar J. Cox1, Gwena'el Do'err, and Teddy Furon, "Watermarking Is Not Cryptography" 2006.
- [26] M. Swanson, B. Zhu, and A. Tewfik, "Data hiding for video in video," presented at the IEEE Int. Conf. Image Processing, 1997.
- [27] M. Utku-Celik, G. Sharma, E. Saber, and A. Murat-Tekalp, "Hierarchical watermarking for secure image authentication with localization," *IEEE Trans. Image Processing*, vol. 11, pp. 585-595, Jun. 2002.
- [28] M. Ramkumar, "Data Hiding in Multimedia-Theory and Applications," Ph.D., New Jersey Inst. Technol., Newark, 2000.

- [29] C.-Y. Lin and S.-F. Chang, "Watermarking capacity of digital images based on domain-specific masking effects zero-error information hiding capacity for digital image," in *Proc. IEEE Int. Conf. Information Technology: Coding and Computing*, Las Vegas, NV, Apr. 2001.
- [30] C.-T. Hsu and J.-L. Wu, "A DWT-DFT composite watermarking scheme robust to both affine transform and JPEG compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 776–786, Aug. 2003.
- [31] F. Perez-Gonzalez, F. Balado, and J. Martin, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Trans. Signal Processing*, vol. 51, pp. 960–980, Apr. 2003.
- [32] N. Liu and K. P. Subbalakshmi, "Vector quantization based scheme for data hiding for images," in *Proc. SPIE Int. Conf. Electronic Images '04*, San Jose, CA, Jan. 2004.
- [33] I. Cox, J. Kilian, T. Leighton, and T. Shamoan, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, pp. 1673–1687, Dec. 1997.
- [34] D. F. H. Malvar, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Processing*, vol. 51, pp. 898–905, Apr. 2003.
- [35] P. Tsai, Y.C. Hu, C.C. Chang, "A color image watermarking scheme based on color quantization," *Signal Process.* (2004) 95–105.

A Novel Software Tool for Analysing NT[®] File System Permissions

Simon Parkinson and Andrew Crampton
School of Informatics
University of Huddersfield
HD1 3DH, UK
Email: s.parkinson@hud.ac.uk

Abstract—Administering and monitoring New Technology File System (NTFS) permissions can be a cumbersome and convoluted task. In today's data rich world there has never been a more important time to ensure that data is secured against unwanted access. This paper identifies the essential and fundamental requirements of access control, highlighting the main causes of their misconfiguration within the NTFS. In response, a number of features are identified and an efficient, informative and intuitive software-based solution is proposed for examining file system permissions. In the first year that the software has been made freely available it has been downloaded and installed by over four thousand users¹.

I. INTRODUCTION

Controlling access permissions to a given file system is an important aspect of data security. Having a secure and flexible way of viewing and managing access control should be a standard requirement of all modern file systems. This should certainly be true of the New Technology File System (NTFS), since NTFS is currently the most common file system in use. This is mainly due to Microsoft's dominance of computing operating systems. Surprisingly, however, no such flexibility exists for the NTFS and the process for determining access controls is cumbersome at best.

The NTFS implements access control with the use of Access Control Lists (ACLs). Each file system object (folder or file) will have an associated ACL for controlling access. An ACL contains a list of ACEs (Access Control Entities). Each ACE contains information regarding the interacting user or group, and the level of access that they will be granted.

It is well reported that from observing an ACE that the following information can be established [1]–[3]:

- 1) The user or group that the ACE applies to.
- 2) The level of granted permission for a user or group.
- 3) Information regarding the prorogation of the permission down the directory hierarchy

The way in which users are required to interact with ACEs and ACLs in the NTFS results in the following peculiarities:

- 1) Permissions are interacted with on a per object level, rather than per user [4]. This does not allow for the

administrator to evaluate user permission across a whole directory structure.

- 2) Interacting with a single ACL using Windows Explorer as seen in Figure 1 requires the traversal of four different interfaces. Interacting with multiple ACLs soon becomes a cumbersome task, which could ultimately result in permissions being overlooked.
- 3) Not only is the administrator required to examine users or groups within the ACL, they have to remember, or explore, group association to evaluate the inheritance of permissions from different groups.

It is well reported that these time-consuming peculiarities result in the potential for errors to occur, which could ultimately result in users being denied access, or in the worst case, the possibility for unwanted access to occur [3]–[7].

Previous efforts to provide a solution to the identified problems [4] have been mostly successful, however, since their production the NTFS has evolved to allow for the specification of fine- and -coarse grained file system permissions [8]. This brings additional complexity as not only can the standard six permission levels be granted, there is the possibility to create 'special permissions' which are constructed from any combination of the possible fourteen permission attributes.

Microsoft provide a variety of command line utilities [9]–[11] and third-party solutions are also available [12] to examine permission allocation. However, the shortcomings of these utilities make none of them serve as a single solution. These shortcomings can be summarised as the inability to:

- 1) Show both fine- and coarse-grained permissions.
- 2) Examine permissions on multiple folders at once.
- 3) Evaluate permissions per user rather than per object.

There is insufficient literature available to suggest that freely available tools have been developed to significantly aid with the administration and reporting of NTFS permissions [1]–[3], as well as providing detailed information regarding the low-level implementation NTFS access control [13]. There are few research papers aimed at understanding NTFS access control [14], [15] and how it can be improved through better administration [8]. One author has provided a formal model of NTFS access control, describing fundamentals of rigorous implementation [16], but there is no indication of the production of any tools that make this available for system administrators.

One paper provides the results for an alternative management interface for NTFS permissions [7]. Through careful con-

¹Available at: <http://eprints.hud.ac.uk/9743>
and http://download.cnet.com/NTFSPermissionsExplorer/SnapIn30002094_4-75325639

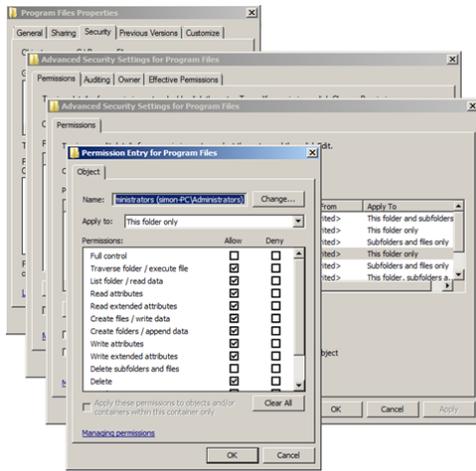


Fig. 1. Analysing NTFS file system permissions using Windows Explorer

sideration to human and computer interaction, an application was designed where they could performed administration tasks significantly faster, whilst reducing potential errors. However, the work is restricted to only viewing file system permissions for a single directory at any one time. Since the work was been published, there is no evidence that the tool has been made available in the public domain. Other work includes using novel ways to represent security policies [17]. This work is also concerned with temporal aspects of managing file system permissions, whereas the work in this paper is also concerned with providing useful features to aid the quality of the analysis and help to reduce misconfiguration.

This paper starts by giving a detailed description of how NTFS implements file system permissions, highlighting complexities that result in misconfiguration. A design is then provided, detailing how a software tool can be used to help overcome the complexities, reducing misconfiguration. The next section discusses the functionality of the produced piece of software. This section describes how the functionality can be used to overcome the highlighted complexities by using real-world examples where possible. Finally, we conclude by discussing the beneficial impacts that the solution can bring, and suggest future developments.

II. NTFS ACCESS CONTROL

In this section we describe the inner-workings of the NTFS as regards to permission management. It is necessary to investigate the following aspects to motivate the designed solution.

A. Access control structure

The NTFS follows in the footsteps of Microsoft's object-oriented approach to implementation. This means that the file system is made up of multiple file and folder objects, and any subject within the operating system (user or process) can request operations on the objects.

To control access to file system objects, the NTFS implements Access Control Lists (ACLs) by applying an ACL to each object within the file system. Each ACL will contain a Security Identifier (SID) which is a unique key that identifies

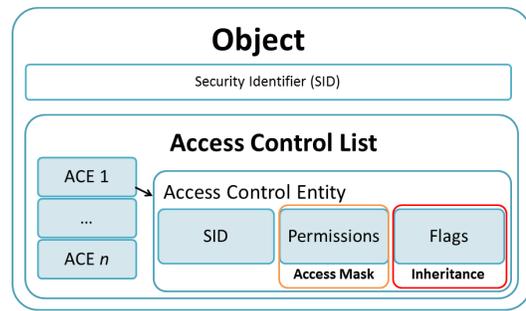


Fig. 2. Access Control List illustration

TABLE I. BIT MASK

Bit / Bit range	Description	Example
0-15	Object specific access rights	Read Data, Execute, Append Data
16-22	Standard security access rights	Delete ACE, Write ACL, Write owner
23	Access to ACL	Access System Security
24-27	Reserved	n/a
28	Generic all	$29 \cup 30 \cup 21$
29	Generic Execute	All needed to execute
30	Generic Write	All needed to write to a file
31	Generic Read	All needed to read a file

the owner of the object and the primary associated group. The structure of the ACL is a sequential storage mechanism which contains access control entries (ACEs). An ACE is an element within an ACL which dictates the level of access given to the interacting subject. The ACE contains a SID that identifies the particular subject, an access mask which contains information regarding the level of permissions and the inheritance flags. Figure 2 illustrates the logical structure of an ACL and associated ACEs.

B. Access Mask

An ACE within the NTFS is made up of a combination of fourteen individual permission attributes. The NTFS provides six levels of standard coarse-grained permission that consist of a combination of predefined attributes. It is also the case that NTFS allows for the creation of special coarse-grained permissions which consist of any combination of the fourteen individual attributes [3].

The access mask is represented by a thirty-two-bit vector. Table I identifies the use of each bit within the vector. It is evident from the table that the standard coarse-grained permissions are represented as follows;

Fine-grained special permissions are represented by using the bits within the range of zero to fifteen. Creating a special permission for most is a very useful feature; however, it can often be a source of confusion as it requires the complete understanding of the authority that each attribute holds [18].

A good example of having to use special permissions is when you wish to assign a group of users the standard privilege elevation of modify for all the contents of a shared folder.

TABLE II. STANDARD COARSE GRAINED PERMISSION BITS

Coarse-grained level	Set bit(s)
Read	bit31
Write	bit30
List folder contents	bit31 \cup bit29
Read and execute	bit31 \cup bit29
Modify	bit31 \cup bit29 \cup bit30
Full control	bit28

TABLE III. PROPAGATION AND INHERITANCE

Bit	Name	Use
1	container inherit ace	Applies the ACE to all the children objects
2	no propagate inherit ace	Propagates the ACE to the child object without bit 1 being set, therefore, stopping propagation at the first level.
3	inherit only ace	The ACE only applies to children objects. (i.e. does not apply to container)

However, creating an ACE with the modify permission on the folder explicitly will result in the user being able to delete the folder itself rather than the child objects (Table I). To get around this problem we would simply assign the group or user the default permission level of Modify, and then go and modify the permissions' attributes turning it into a special permission so that only subfolders and files can be deleted.

C. Propagation and Inheritance

It is necessary to discuss the different mechanisms behind the way that NTFS permissions can propagate throughout the directory structure. Within the ACL there are two types of ACE; (1) Explicit and (2) Inherited. Explicit entries are those that are applied directly to the objects' ACL, whereas inherited are those that are propagated from their parent object. The type of ACE allows to determine whether the permission was assigned directly to the directory in question (explicit) or if it was inherited from the directory that it resides within (inherited).

This mechanism is controlled by the bit-flag within each ACE as seen in Figure 2. Table III shows the standard three coarse-grained levels of propagation and explains their use.

Furthermore, the creation of fine-grained special file system permissions also allows for the creation of custom fine-grained inheritance rules. Special inherited permissions can be different depending on whether the ACE has the container inherit ace bit flag set which controls whether the ACE is applied to all the children objects or not. The creation of fine-grained propagation rules can easily be overlooked and can ultimately result in the unintended propagation of access.

One of the main difficulties with access propagation with the NTFS is correctly evaluating the effective propagation rules. For a user to view the propagation rules the same situation as viewing the effective permission applies, where the user is required to traverse through the several Windows interface to retrieve the required information as seen in Figure 4.

D. Accumulation

Accumulation is the possibility for the subject to receive the effective permission of multiple different policies. This fea-

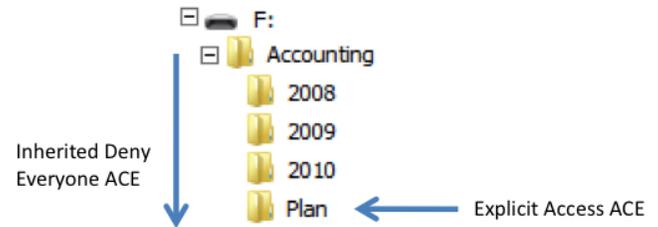


Fig. 3. Explicit before inherited demonstration

ture is prominent within the NTFS resulting in the possibility for a subject to receive permissions from multiple different ACEs within the same ACL. Furthermore, any subject that interacts with the NTFS can be assigned to any number of groups, which can be entered into the ACE. This means that the user does not have to be directly entered into the ACE, they could simply be a member of the group that is entered.

The policy combination is handled within the operating system by the Local Security Authority Subsystem Service (LSASS). This service combines the permissions together to effectively create the union of all the policies. There are few complexities within permission accumulation due to the structured way in which ACEs are processed. These are:

- 1) Explicit permissions take precedence over inherited permissions.
- 2) Explicit deny permissions always take precedence over apply permissions.
- 3) Permissions inherited from closer relatives take precedence over relatives. further away.

It might expect that deny permissions always take precedence over apply permissions to ensure that during the policy combination stage the user always operates as the least possible privilege elevation. However, the first point regarding explicit permissions taking precedence over inherited permissions can result in a situation where an inherited deny permission is never reached. Considering the folder structure in Figure 3, where the folder Accounting has an explicit deny permission for the Everyone group, which is set to propagate to all its children. This means that all the subfolders to the Accounting folder will receive an inherited deny Everyone ACE. If the case was to arise, like in this example, where a single user now requires access to the Plan folder, an explicit ACE to allow access could be entered. Now when the user visits the Plan folder, the LSASS would process the explicit allow permission first and allow for it to take precedence over any other permission. This goes against a fundamental aspect of policy combination to ensure that a deny permission is never ignored. If the case where a user is able to ignore a deny permission to receive access was to either intentionally or unintentionally arise, the system administrator needs to be made aware of this situation.

To summarise, the precedence hierarchy for policy accumulation is as follows:

- 1) Explicit deny.

- 2) Explicit allow.
- 3) Inherited deny.
- 4) Inherited allow.

In addition to the explicit permissions taking precedence over inherited permissions, inherited permissions that of closer distance to the invoked object will take precedence over more distant relatives. For example, a folder's inherited permissions will take precedence over those from their grandparent.

Accounting for permission accumulation has currently been made possible by using the standard Windows Explorer feature of displaying the effective permission. This feature allows for the user to enter a specified user or group and the effective permission that they hold on that specific directory will be displayed. Unfortunately, performing this evaluation on several folders soon becomes infeasible.

E. Group Membership

A fundamental aspect of access control within the NTFS is that of group membership. A subject (group, user or process) that interacts with the file system can be a member of any group. This means that permissions can be inherited from any of the associated groups if they are entered within any ACL. Subjects, in this case users, will often be grouped together by (separation of duty) to make management easier, and as Hanner, 1999 [4] identifies, understanding effective file permissions can become significantly more complex by group association. To correctly evaluate a user's effective permissions you would have to know which groups they are a member of. We should note that this is not directly related to the mechanism of how NTFS implements access control, it is an unavoidable component of how Microsoft allows for users, groups and processes to be managed by group association.

III. NOVEL SOLUTION

This section describes the design of a solution based on the NTFS's inner-workings which can cause the identified administrative complexities as seen in Section II.

A. Coarse- and Fine-Grained Permissions

As previously described, the NTFS allows for the standard set of coarse permissions, but also allows for the creation of special fine-grained permissions.

An alternative method of display, special permissions could be displayed by a character-to-attribute representation. This way a string can be constructed to display the full granularity of the permission by only using little space. For example, if a special permission was constructed to have the attributes enabled:

- 1) Read (R).
- 2) Write (W).
- 3) Delete subfolders and files (Dc).
- 4) Read permissions (Rp).
- 5) Change permissions (Cp).

Using the character-to-attribute would result in the production of the string 'R-W-Dc-Rp-Cp'. After some time the user would become accustomed to this relationship and the key would no longer be required.

B. Multiple Folders

Algorithm 1: Depth-first recursive directory search, analysing and filtering security permissions.

Input: Initial directory d
Input: Set of ACEs to be filtered out
 $F = (f_1, f_2, f_3, \dots, f_n)$
Output: Set of ordered directories and ACEs
 $P = (d_1, (p_1, p_2, p_3, \dots, p_n))$ where d_n is the directory and p_n are the permission entries for that directory.

```
1 Algorithm algo()
3    $P \leftarrow \text{proc}(d)$ 
5   return
6
1 Procedure proc(directory  $d$ )
2    $pACL \leftarrow d(ACL)$ 
3   foreach subdirectory  $c$  of  $d$  do
4      $cACL \leftarrow c(ACL)$ 
5     if  $cACL \neq pACL$  then
6       foreach ACE  $a$  in  $cACL$  do
7         if  $a \notin F$  then
8           if  $isSpecial(a)$  then
9              $p \leftarrow compress(p)$ 
10          else
11             $p \leftarrow a$ 
12          end
13           $P \leftarrow (c, p)$ 
14          proc( $c$ )
15        end
16      end
17    end
18
```

It has previously been identified that Windows Explorer allows for the examination of an objects' ACL, however, it is often the case that evaluating multiple ACLs is necessary. A useful way to view multiple ACLs would be to allow the examination of a whole directory structure simultaneously. This would provide the means to also examine how the propagation and inheritance aspects of the ACLs are interacting. Algorithm 1 describes the recursive depth-first examination search technique that has been implemented for analysing the permissions of multiple folders. This algorithm traverses the directory structure, analysing each directories permissions. In each analysis, the algorithm evaluates whether:

- 1) It is necessary to display the current ACL to the user based on whether it is different from the parent's ACL.
- 2) Each ACE in the ACL contains a special permission.
- 3) Report the ACE to the user, displaying the level of permission.

C. Compression

As seen on line 9 of Algorithm 1, a compress function is called if a special permission is identified. This compress function performs the character-to-attribute mapping as described in Section III-A. In this method, an enumerated type is used for changing the permission attributes to the associated character.

D. Filtering

Filtering of groups is easily performed as shown on line 7 of Algorithm 1 where a check is made to ensure that the current ACE a is not present in the set of groups to filter F . This provides the facility to filter for multiple user or group objects, therefore removing excess information.

E. Per User View

When performing a per user search of the file system, Algorithm 1 is used, however, line 7 is substituted with a condition to check that the ACE in question is the one that is being searched for ($a \in F$). This means that all groups and user objects are excluded if they are not represent in the filter list. When viewing per user, the filer list contains the user or group that the user wants to analyse.

F. Accumulation

Algorithm 1 identifies provides a search strategy that can report the file system permissions for an entire directory structure, whilst considering compression and filtering. Although the returned permission information is what is visible in the ACE, it might not be the user's effective permission as no consideration to permission accumulation as described in Section II-D is taken. Algorithm 2 provides an alternative method where the search concentrates on calculating the effective permission that the user and or group hold. Algorithm 2 shows an algorithm that can be used to store the explicit ex and inherited in permissions based on the inheritance and propagation. This algorithm considers both the inheritance and deny hierarchies. For speed purposes the algorithm can identify deny permissions and stop the algorithm from continuing the examine the ACL. Line 16 shows that once the explicit and inherited permissions have been identified a function is then called to calculate the effective permission. In this algorithm $calculatedEffective(explicit, inherited)$ represents a native Microsoft .NET command that is able to return the effective permission. Using this native method ensures that the correct effective permission is reported.

G. Group Membership

User and group membership is fundamental mechanism that allows users to inherit file system permissions from group objects. A simple recursive method can be used to examine a user or groups membership. There are two possible directions in which the group membership can be analysed. The first is to examine which groups an object is a member of. This is where a search is performed to recursively report which groups a user or group is a member of. The second method is the members of displaying a user or groups members. This is where a recursive search is performed to reporting on a groups members.

IV. DEVELOPED SOLUTION

The developed software-based tool is programmed in C# .NET 3.5 with the use of the Microsoft Management Console (MMC) System Development Kit (SDK) to produce a MMC SnapIn application. The motivation behind making the application run in the MMC was to bring consistency with other Microsoft management tool, therefore, making the software self-intuitive for the users.

Algorithm 2: Depth-first recursive directory search, returning the effective permission of a specified user or group.

Input: Initial directory d

Input: Initial group or user u

Output: Set of ordered directories and ACEs

$P = (d_1, (p_1, p_2, p_3, \dots, p_n))$ where d_n is the directory and p_n are the permission entries for that directory.

```
1  Algorithm algo()
3  |    $P \leftarrow \text{proc}(d)$ 
5  |   return
6
1 Procedure proc(directory  $d$ )
2  |    $pACL \leftarrow d(ACL)$ 
3  |   foreach subdirectory  $c$  of  $d$  do
4  |   |    $cACL \leftarrow c(ACL)$ 
5  |   |   if  $cACL \neq pACL$  then
6  |   |   |    $ex = \emptyset, in = \emptyset$ 
7  |   |   |   foreach ACE  $a$  in  $cACL$  do
8  |   |   |   |   if  $isExplicitDeny(a)$  then
9  |   |   |   |   |    $P \leftarrow (c, a)$ 
10  |   |   |   |   |   break
11  |   |   |   |   else
12  |   |   |   |   |   else if  $isExplicitAllow(a)$  then
13  |   |   |   |   |   |    $ex \leftarrow a$ 
14  |   |   |   |   |   |   else if  $isInherited(a)$  then
15  |   |   |   |   |   |   |    $in \leftarrow a$ 
16  |   |   |   |   |   |   |    $P \leftarrow (c, calculatedEffective(ex, in))$ 
17  |   |   |   |   |   end
18  |   |   |   |   end
19  |   |   |   end
20  |   end
```

The software runs under the credentials of the executing user, therefore, only receiving access to view file system permissions that they have been assigned to. The software runs in real-time, processing the desired ACLs upon request. This means that the software requires only a minimal amount of installation, and does not require an additional database to store permission entries. The overheads caused by the application on both the host machine and any interacting file servers are very small and do not affect normal performance at all.

In this remaining of this section, the provided functionality is discussed, using examples where possible.

A. Application Layout

As seen in Figure 4, the interface has three main sections. Firstly on the left is the control pane. The control pane is where the user can see all the physical and remote mounted NTFS volumes. The user is able to browse the folder structure of all local and remote drives in a Windows standard hierarchical tree view. In addition, any effective permission searches that the user performs will be listed here. The middle pane is where the associated results from the item selected within the control pane are displayed. On the right is the action pane. This pane contains functionality associated with each of the items selected within the control pane that can affect the contents of the results pane.

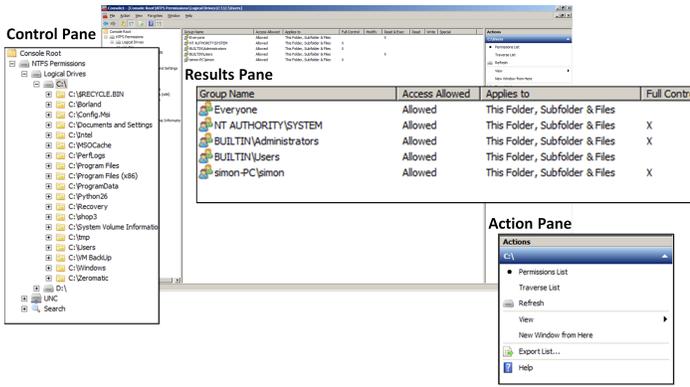


Fig. 4. Developed MMC Application

The results pane shows the ACL for the specified local or remote drive, providing that the executing user has permission to view the ACL. This pane contains the same ACL information as present in the Windows Explorer interface. The ACEs are classified into the standard NTFS sets although List Folders is not classed as a set because the permission is the same as Read & Execute, just the propagation is different, which is correctly displayed.

B. Coarse- and Fine-Grained Permissions

As described in the design, the application does have a different way of representing special permissions. To allow the user to easily and correctly see the fine-grained permissions the special permissions are displayed as a hyphen separated character string, where each character is associated with a different special permission attribute.

As shown in Figure 5 the group 'BUILTIN\Users' has a special permission entry that is displayed by the hyphenated character string. On further inspection of this permission it is possible to view the character-to-attribute relationship, which is also displayed in Figure 5. After using the application we might start to remember the character-to-attribute relationship, meaning that we do not need to inspect the special permission, therefore, further speeding up the process of reporting fine-grained special permissions. The results pane also shows information regarding whether each permission (ACE) is an allow or deny permission, and also the propagation level of each of the ACE entries.

C. Traversal View and Custom Filter

Another highlighted problem was difficulties within trying to view the ACL for multiple folders at any one time. The developed application avoids this issue by firstly allowing a user to simply traverse the file system in the control pane to view the ACL for a single folder, and secondly, allowing the user to view the ACLs for a whole directory in one traversal view. To reduce the quantity of displayed information and help display what is useful to the user, by default the traversal view will only show the ACL for a folder that is not the same as its parents'. A custom filter has also been implemented so that the user can select groups and users that they do not wish to include in the traversal view.

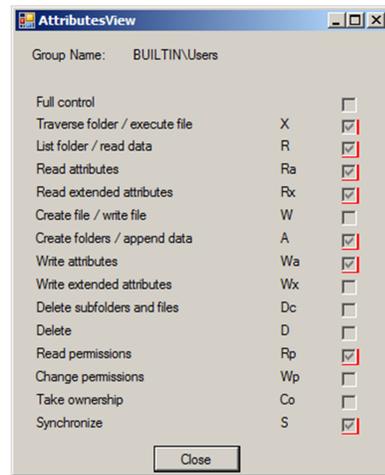


Fig. 5. Developed MMC Application

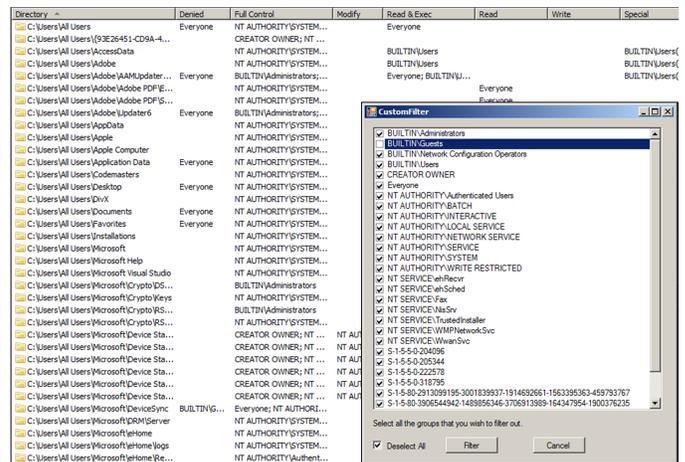


Fig. 6. Traversal view with custom filter

Figure 6 shows the results pane when the traversal function is applied to the local folder C:\Users. The illustration also shows the filter interface where the user can select groups that they wish to remove from view. The traversal view also displays both fine- and coarse-grained permissions in the same way as the individual view where the permissions are classified as the standard or special sets.

D. Permission Accumulation

Policy combination can be one of the most time consuming aspects of the NTFS when trying to evaluate the permission that a subject holds on any given location. As described earlier, accumulation of deny and access permissions, group membership as well as consideration to the ACE processing hierarchy results in several complication factors to the evaluation. The developed application has a built-in search feature to show the exact effective permissions for a given subject on the selected location. Figure 7 shows the interface after performing a custom search for the user 'simon-PC\simon' on the directory 'C:\User'. The same logic applies when performing a search where only permissions that differ from their parent object are displayed by default, and special permissions are displayed using the hyphenated character representation.

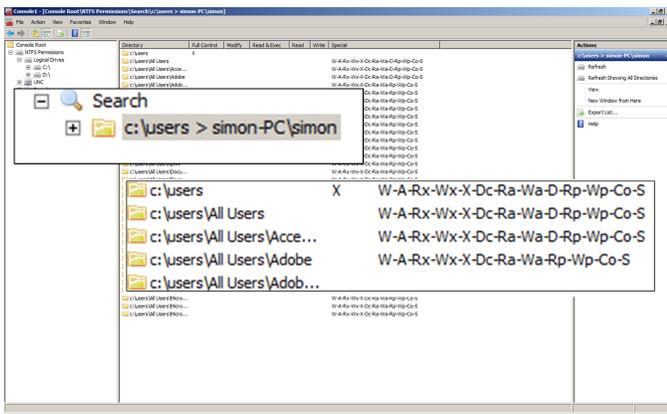


Fig. 7. Permissions accumulation search results

V. CONCLUSIONS

We began by examining in detail the workings of access control within the NTFS to highlight the potential causes of complexity, which could ultimately lead to unintended access. Next, we discussed the common usability problems that can be experienced when examining NTFS permissions. Following this, we developed a Microsoft Management Console SnapIn application to provide a new way of examining NTFS permissions that can help overcome the identified complexities. We believe that our study and software solution helps to improve file system security by providing an intuitive, efficient and thorough method for permission examination.

This paper provides a contribution to system administrators by aiding them with permission examination and allocation. The requirement to provide a software-based tool to overcome the identified complexities can be established from the in excess of four thousand downloads the tool has received since production. This shows that NTFS administrators are actively seeking support for their duties. In addition to the number of downloads, the tool has also received promotion through a rated software site [19] and a useful list of system administration tools [20]. This emphasises how requirement for such tool.

VI. FUTURE SCOPE

Future work involves allowing for the user to modify file system permissions once a problem has been identified. Another possibility is a software tool that can automatically identify configuration problems and suggest intelligent solutions.

VII. ACKNOWLEDGEMENT

The authors would like to express great thanks to Michele Puri of the European University Institution for passing on vast amounts of knowledge regarding the implementation and administration of NTFS permissions within a large organisation. Thanks should also be expressed to Alan Radley and Malcolm Merrington of the University of Huddersfield for providing additional insight to the problems and for testing the developed software.

REFERENCES

- [1] C. Russel, S. Crawford, and J. Gerend, *Microsoft windows server 2003 administrator's companion*. Microsoft Press, 2003.
- [2] D. A. Solomon, "Microsoft windows internals: Microsoft windows server 2003, windows xp, and windows 2000."
- [3] *Microsoft Windows Server 2003, Administrator's Companion*, 2nd ed. Microsoft Press, 2006.
- [4] K. Hanner and R. Hörmanseder, "Managing windows nt file system permissions—a security tool to master the complexity of microsoft windows nt file system permissions," *Journal of Network and Computer Applications*, vol. 22, no. 2, pp. 119 – 131, 1999. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1084804599900863>
- [5] K. Beznosov, P. Inglesant, J. Lobo, R. Reeder, and M. E. Zurko, "Usability meets access control: challenges and research opportunities," in *Proceedings of the 14th ACM symposium on Access control models and technologies*, ser. SACMAT '09. New York, NY, USA: ACM, 2009, pp. 73–74. [Online]. Available: <http://doi.acm.org/10.1145/1542207.1542220>
- [6] X. Cao and L. Iverson, "Intentional access management: making access control usable for end-users," in *Proceedings of the second symposium on Usable privacy and security*, ser. SOUPS '06. New York, NY, USA: ACM, 2006, pp. 20–31. [Online]. Available: <http://doi.acm.org/10.1145/1143120.1143124>
- [7] R. A. Maxion and R. W. Reeder, "Improving user-interface dependability through mitigation of human error," *International Journal of Human-Computer Studies*, vol. 63, no. 12, pp. 25 – 50, 2005, {ce:title}HCI research in privacy and security{/ce:title}. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1071581905000601>
- [8] S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati, "Access control: principles and solutions," *Software: Practice and Experience*, vol. 33, no. 5, pp. 397–421, 2003. [Online]. Available: <http://dx.doi.org/10.1002/spe.513>
- [9] Microsoft, "How to use xcalcs.vbs to modify ntfs permissions," 2006. [Online]. Available: <http://support.microsoft.com/kb/825751>
- [10] "Accesschk v5.01," 2010. [Online]. Available: <http://technet.microsoft.com/en-gb/sysinternals/bb664922>
- [11] Microsoft, "Accessenum v1.32," 2006. [Online]. Available: <http://technet.microsoft.com/en-us/sysinternals/bb897332>
- [12] "Security explorer v7.5.0.," 2010. [Online]. Available: <http://www.scriptlogic.com/products/security-explorer/>
- [13] B. Carrier, *File system forensic analysis*. Addison-Wesley Boston, 2005, vol. 3.
- [14] L.-y. WANG and J.-w. JU, "Analysis of ntfs file system structure," *Computer Engineering and Design*, vol. 3, p. 018, 2006.
- [15] L. J. Z. Yue, "The main data structure of ntfs file system," *Computer Engineering and Applications*, vol. 8, p. 038, 2003.
- [16] J. Crampton, G. Loizou, and G. O'Shea, "A logic of access control," *The Computer Journal*, vol. 44, no. 2, pp. 137–149, 2001.
- [17] R. W. Reeder, L. Bauer, L. F. Cranor, M. K. Reiter, K. Bacon, K. How, and H. Strong, "Expandable grids for visualizing and authoring computer security policies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 1473–1482. [Online]. Available: <http://doi.acm.org/10.1145/1357054.1357285>
- [18] O. Thomas, "Are ntfs and share permissions a bit too complicated," *Windows IT Pro*, vol. 16, p. 78, 2010.
- [19] SoftSea.com, "Ntfs permissions explorer snapin," <http://www.softsea.com/review/NTFS-Permissions-Explorer-SnapIn.html>, accessed: 2013-08-17.
- [20] C. Goggi, "101 free admin tools," <http://www.gfi.com/blog/101-free-admin-tools>, accessed: 2013-08-17.

Probabilistic Distributed Algorithm for Uniform Election in Triangular Grid Graphs

El Mehdi Stouti

FS–Abdelmalek Essaâdi University
P.O. Box. 2121 M’Hannech II
93030 Tetuan Marocco
Email: stouti@uae.ma

Ismail Hind

FS–Abdelmalek Essaâdi University
P.O. Box. 2121 M’Hannech II
93030 Tetuan Marocco
Email: ismailhind@gmail.com

Abdelaaziz El Hibaoui

FS–Abdelmalek Essaâdi University
P.O. Box. 2121 M’Hannech II
93030 Tetuan Marocco
Email: hibaoui@uae.ma

Abstract—Probabilistic algorithms are designed to handle problems that do not admit deterministic effective solutions. In the case of the election problem, many algorithms are available and applicable under appropriate assumptions, for example: the uniform election in trees, k -trees and polyominoids.

In this paper, first, we introduce a probabilistic algorithm for the uniform election in the triangular grid graphs, then, we expose the set of rules that generate the class of the triangular grid graphs. The main of this paper is devoted to the analysis of our algorithm. We show that our algorithm is totally fair in so far as it gives the same probability to any vertex of the given graph to be elected.

Keywords—Uniform Election, Distributed Algorithms, Probabilistic Election, Markov Process, Randomized Algorithm Analysis.

I. INTRODUCTION

Election in a network is to chose one and only one element of this network. The elected element may be used to manage such shared resources (printer, connection, etc.), or to centralize some network informations (size, diameter, etc.).

The election problem holds the attention of many researchers since it was first proposed by LE LANN [1]. Therefore, it has been studied under various assumptions: the proposed network could be oriented or not, synchronous or asynchronous (no shared global clock), anonymous or with identifiers (no unique identity is attributed to elements), size knowing or not, etc.

The solutions take also into account the network topology. Some solutions are deterministic while others are probabilistic. Another aspect is that we want also to study the uniform election. In this type of election, we attribute the same chance to all nodes and at any position in the network to be elected. The algorithms known in the literature are probabilistic and run on well defined topologies. We quote for trees [2][3], for k -trees [4] and for polyominoids [5]. The work presented here is a continuation of this researches. Thus, we introduce a probabilistic algorithm for uniform election in a network with the topology of triangulated grid graphs.

The triangular grid graph is, in graph theory, a finite sub-graph induced from the infinite graph associated with

the two-dimensional triangular grid [6]. It is a subclass of planar graphs [7]. However, networks discussed in this work have the topology of a triangular grid graphs. We assume that the network can be synchronous or asynchronous, and it is anonymous; no unique identity is attributed to its vertices.

The main objective behind this study is to suggest and analyse the uniform election of a probabilistic distributed algorithm in the triangular grid graphs. So and for a given graph G , each vertex $v \in G$ generates its lifetime duration depending to its weight w_v . The lifetime of a vertex v is an exponential random variable of parameter λ_v equals to its weight w_v . According to our algorithm, when the lifetime of a vertex expired, it removed with its incident edges, and its neighbour in the standard spanning tree recovers its weight.

The analysis of this algorithm proves that, whatever the vertex is situated in the studied graph, it has the same probability to be elected. We can consider our algorithm as a probabilistic variant of the distributed algorithm introduced in [8], where random delays are presented.

We consider local computations in the cells. At each step of computation, the vertices of a cell can change their status (or labels). Indeed, the new label of a vertex depends on its previous label (state) and the labels of its neighbours. In our approach we used a random delay for labelling; a vertex can not change its state if its associated lifetime duration is not over. These delays are exponential random variables, independently defined, for active vertices.

The parameter of a random variable associated with a vertex is equal to the weight assigned to this vertex. The weight is locally calculated in term of initial weight and the weights collected from the vanishing neighbours. The labelling process continues until no transformation is possible, that is to say, the last configuration is reached. In this configuration, there was only one vertex which has a different tag (label) from the others, this vertex is considered as the elected one (leader) [5].

For the analyse of this algorithm we model the elimination process with a continuous time of a Markov death Process.

This paper is organized as follows. In Section II, we give

some definitions required to understand the rest of the paper. In section III we give a set of rules generating the class of triangular grid graphs. The section IV devoted to the analysis of a probabilistic algorithm for uniform election in this family of graphs. In section V we presents the operation of the algorithm, providing some tools for its analysis (section 5).

II. PRELIMINARIES AND NOTATION

A Triangular Grid Graph (**TGG**) is a finite graph where its vertices are points in $\mathcal{Z} = \mathbb{Z} \times \mathbb{Z}$, where \mathbb{Z} denotes the set of decimal integers. They are linked by neighbourhood relationships [7].

The edges are the links between pairs of points. They are of the forms : $\{(x, y), (x, y + 1)\}$ or $\{(x, y), (x + 1, y)\}$ or else $\{(x, y), (x + 1, y - 1)\}$ for all $x \in \mathbb{Z}$ and for all $y \in \mathbb{Z}$ (see Fig 1).

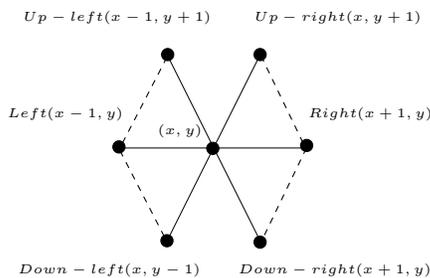


Figure 1. Vertex's neighbours in TGG.

Two vertices $v = (x, y)$ and $v' = (x', y')$ of \mathcal{Z} are neighbours if one of the following conditions are satisfied:

- $y = y'$ and $|x - x'| = 1$, or
- $x = x'$ and $|y - y'| = 1$, or
- $|x - x'| = 1$ and $|y - y'| = 1$.

Two neighbours are the ends of an edge.

For each vertex v of coordinates (x, y) , we use the usual terms such as 'up' to denote the neighbour of coordinates $(x, y + 1)$, 'down' for the neighbour $(x + 1, y - 1)$, 'right' for $(x + 1, y)$ and 'left' for $(x - 1, y)$ neighbour, see Fig 1.

Let \mathcal{S}_E be the set of all edges whose ends are neighbours and $\mathcal{I}_G = (\mathcal{Z}, \mathcal{S}_E)$ the infinite graph consisting of the set of vertices \mathcal{Z} and the set of edges \mathcal{S}_E . A cell is a sub-graph of \mathcal{S}_E , induced by a set of three pairwise neighbour vertices $\{(x, y), (x + 1, y), (x, y + 1)\}$ having the form Δ , called *up triangle cell*, or else $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$ for the vertices with form ∇ , called *down triangle cell*.

A path is a finite alternated sequence $\sigma = v_0, e_1, v_1, \dots, v_{k-1}, e_k, v_k$ of $k + 1$ vertices and k distinct edges ($k \geq 0$), such that v_{i-1} and v_i are the ends of the edge e_i , for $1 \leq i \leq k$.

We recall that the length of a path σ is the number of its edges k . It should be noted that a path may pass several times

through a vertex, but can not borrow an edge more than once. A cycle is a path of length $k \geq 3$ in which the first vertex v_0 and the last vertex v_k coincide. For a given cycle, we can easily and according to [9] define the vertices or edges inside this cycle.

Definition 2.1: A vertex (x, y) is inside the cycle $\gamma = (x_0, y_0), (x_1, y_1), \dots, (x_{k-1}, y_{k-1}), (x_0, y_0), ((x_k, y_k) = (x_0, y_0))$ if $(\text{card} \{i \mid y = y_i \text{ and } y \neq y_{i+1} \text{ and } x \leq x_i\})$ is odd. Therefore, the boundary vertices of γ are inside γ .

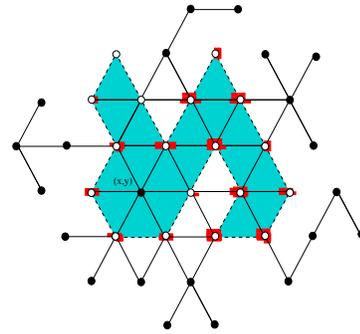


Figure 2. Example of vertex inside a cycle.

Proposition 2.1: For a given TGG and for a given vertex v inside a cycle γ in this graph, we have for each *linear path* (*LP*) including v , the number of vertices belonging to both *LP* and γ is even.

Example 2.1: The vertex (x, y) of the figure 2 is inside the cycle consisting by the white vertices and the edges in bold. For the linear path $LP = (x, y), (x + 1, y), (x + 2, y), (x + 3, y), \dots$, we have the cardinal of vertices set inside the said cycle and belonging to *LP* is even.

Definition 2.2: A triangular grid graph $G = (V, E)$ is a sub-graph of \mathcal{I}_G if the following conditions are satisfied:

- 1) V is finite,
- 2) G is connected and
- 3) G does not contain holes, i.e. for any cycle γ in G , the vertices inside γ are contained in V and if two neighbours are inside γ , so the edge connects these two vertices belongs to E .

We define the size of $G = (V, E)$ as the cardinal of V .

Definition 2.3: A triangular grid graph $G_s = (V_s, E_s)$ is called sub-triangular grid graph of the triangular grid graph $G = (V, E)$ if only if $V_s \subseteq V$ and $E_s \subseteq E$ such that $E_s = E \cap \{(u, v) \mid \exists e = (u, v), \text{ for all } (u, v) \in V_s^2\}$.

III. DISTRIBUTED CONSTRUCTION OF TRIANGULAR GRID GRAPH

Let \mathcal{S}_P be the set of partial sub-graphs of the infinite graph \mathcal{I}_G obtained by the following inductive rules:

- a) For all $(x, y) \in \mathcal{Z}$, the graph $G = (\{v = (x, y)\}, \phi)$ is in $\mathcal{S}_{\mathcal{P}}$. \implies Let $G = (V, E)$ be a TGG. We show by induction proof on the cardinal of the set V that G belongs to $\mathcal{S}_{\mathcal{P}}$.
- b) Let $G = (V, E) \in \mathcal{S}_{\mathcal{P}}$. Consider two neighbouring vertices v and v' such that $v \in V$ and $v' \notin V$, then $G' = (V \cup \{v'\}, E \cup \{\{v, v'\}\})$ is in $\mathcal{S}_{\mathcal{P}}$.
- c) Let $G = (V, E) \in \mathcal{S}_{\mathcal{P}}$. Suppose that V contains three neighbouring vertices $v_1 = (x, y)$, $v_2 = (x+1, y)$ and $v_3 = (x, y+1)$ or else $v_1 = (x, y)$, $v_2 = (x+1, y)$ and $v_3 = (x+1, y-1)$ located in a cell of $\mathcal{I}_{\mathcal{G}}$, such as two edges of this cell are in E and the third one, called e , is not then the graph $G' = (V, E \cup \{e\})$ is in $\mathcal{S}_{\mathcal{P}}$.
- o If V is of cardinality 1, then obviously $G \in \mathcal{S}_{\mathcal{P}}$.
 - o Now suppose that a graph G with size $n \geq 2$ is a TGG, then $G \in \mathcal{S}_{\mathcal{P}}$. We prove that it's true for a graph G' with size equals to $n+1$.

Let $G' = (V', E')$ be a TGG of size $n+1$. If G' has a vertex v of degree 1, then when we delete v and its incident edge, G' will transform to G . It is clear that G preserves the properties (1)–(3) of Definition 2.2 and therefore, by the recurrence assumption, it belongs to $\mathcal{S}_{\mathcal{P}}$. Indeed, an application of Rule (b) allows that G' is also in $\mathcal{S}_{\mathcal{P}}$.

The construction is totally distributed and applying rewrite rules, as seen in [10], requires only knowledge of neighbouring areas that are in a ball of radius 1. Therefore, the local construction can be expressed by considering transformations assigned to a vertex v and the set of all its neighbours. In this case, it is difficult to show that the set $\mathcal{S}_{\mathcal{P}}$ is the class of all the triangular grid graphs on $\mathcal{I}_{\mathcal{G}}$. The following proposition proves the equivalence of the two definitions.

Proposition 3.1: A partial sub-graph $G = (V, E)$ of $\mathcal{I}_{\mathcal{G}}$ is a triangular grid graph iff it belongs to $\mathcal{S}_{\mathcal{P}}$.

Proof:

\Leftarrow Let $G = (V, E) \in \mathcal{S}_{\mathcal{P}}$ and prove that G is a TGG. We just need to prove that the constructions given by the rules (b) and (c) preserving the structure of the triangular grid graphs. So, suppose that G is a TGG and prove that the sub-graph G'_b of $\mathcal{I}_{\mathcal{G}}$ obtained by (b) and the sub-graph G'_c of $\mathcal{I}_{\mathcal{G}}$ obtained by (c) are also triangular grid graphs. The properties of the connectivity and the finiteness are obvious. We will show that no hole is created during the application of the rule (b) or the rule (c).

- o Applying the rule (b), a new vertex v' is added to G . Since v' is of degree 1, there is no new cycle in G'_b and all the vertices inside a cycle in G remain inside the same cycle in G'_b . Obviously, the same fact is verified for every edge whose ends are in G'_b .
- o Let $G'_c = (V, E'_c)$ be an extension of the triangular grid graph $G = (V, E)$ obtained by applying the rule (c). Let v be a vertex inside the cycle γ included in G'_c . If all edges are in E , then v should be in V . Otherwise, we use an edge of a cell formed by the set of the vertices $\mathbf{S} = \{v_1, v_2, v_3\}$, say $\{v_1, v_2\}$, which does not belong to E . In addition $E'_c = E \cup \{v_1, v_2\}$. In this case, it is possible to transform γ into another cycle γ' included in E avoiding v_1 and borrowing other vertices of the set \mathbf{S} .

Suppose now that all vertices of a triangular grid graph $G = (V, E)$ are of degree greater than or equal to 2. We have $|E| - |V| \geq 0$, if not, G is a tree and admits a vertex of degree 1.

We use now a second recurrence on $|E| - |V|$. It is clear that G has at least one cycle. Let γ be a maximum cycle in G . It is easy to see that if we remove an edge from γ , the residual graph obtained, denoted \mathcal{R} , preserves properties (1)–(3) seen in Definition 2.2. Thus the induction assumption on $|E| - |V|$ gives $\mathcal{R} \in \mathcal{S}_{\mathcal{P}}$.

An application of the rule (c) on the triangular grid graph \mathcal{R} allows to reconstruct the graph G as a member of $\mathcal{S}_{\mathcal{P}}$. ■

IV. UNIFORM ELECTION IN TRIANGULAR GRID GRAPH

A. Model used

We represent a communication network by a graph, where a nodes (stations) are represented by a vertices and the edges represent communication links.

The election algorithm presented here is designed for anonymous networks which have the topology of a triangular grid graph. So, We assume that each vertex does not know the size of the graph neither its own coordinates in the plan. Its only knowledge is the directions of its incident edges.

We use the asynchronous system where no global clock is shared. This means that the transmitter sends the message but there is no information on when the receiver actually receives it. Hence, the processes execute the instructions with arbitrary speeds and the messages reach their destinations in a finite time but also arbitrary.

B. Distributed election

In this section, we describe our election algorithm by a graph rewriting system. The rewriting systems or

more generally the local computations in graph are a powerful models providing general tools to encode distributed algorithms, to understand their power, and to prove their validity [8][10][11][12].

Our distributed algorithm is based on the rewriting systems presented in [12]. Each vertex (resp. edge) has a label that represents its state. In fact, the labels attached to the vertices and the edges are locally modified.

So, initially, all vertices of the TGG have the same weight 1 according to the anonymity condition imposed on the graph. The election process behaves as a continuous-time Markov process.

Each vertex has a weight local knowledge and depending to the situation it changes its state by applying one of the set of R_i and R'_i rules described below.

The random delay associated to each removable vertex are independent and can locally be generated. These vertices may be removed in a random delay which is an exponentially distributed random variable with a parameter equal to the weight of the vertex. Whenever the lifetime (delay removal) of a vertex has expired, it is removed with all its incident edges. The weight of the removed vertex is collected by one of its neighbours according to the R'_i -rules.

The rewriting is performed step by step, then after a number of rewriting steps, we obtained an irreducible graph where no rule is applicable. In this graph there is a special label attached to exactly one vertex. This vertex will be considered as elected one (called *leader*).

The rewriting system applied here uses the forbidden contexts [11][13][14]. The idea is to prevent the application of a rewriting rule whenever the related occurrences are included in some special configurations, called *forbidden contexts*. Thus, a rule can be applied if the two conditions are satisfied:

- 1) the rule does not occur in a prohibit context already mentioned, and
- 2) its associated delay has expired.

Formally, let G_R be a connected graph and two marking functions of G_R : the initial labelling λ_R and the final labelling λ'_R .

The rewrite rule with forbidden contexts is a quadruple $\mathbf{R} = (G_R, \lambda_R, \lambda'_R, \mathcal{F}_R)$ such that $(G_R, \lambda_R, \lambda'_R)$ is a rewrite rule and \mathcal{F}_R is a finite set of forbidden contexts (G_R, λ_R) .

$$\mathbf{R} : \{ \mathcal{F}_R; (G_R, \lambda_R) \longrightarrow (G_R, \lambda'_R) \}.$$

For our case, let $G = (V, E)$ be a TGG and $S_L = \{N, A, B, L\}$ the set of labels. The label **N** encodes the neutral state, **A** encodes the active state, and **B** encodes the beat state, and **L** encodes the leader state (elected).

We denote by **X** any state except **B**, i.e., $\mathbf{X} \in S_L \setminus \{B\}$.

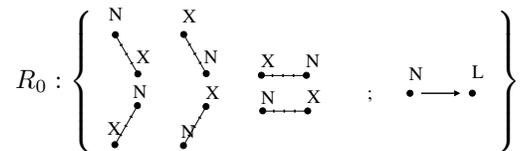
The election process on G runs in distributed manner as follows:
Initially, all vertices have the same weight $w = 1$ and each one is **N**-labelled.

Each **N**-labelled vertex v decides locally whether it is active or not according to the activation rules R_i below. So, if a vertex v becomes active, it generates its lifetime, which is an exponential random variable with parameter equal to its current weight. Once the lifetime of an active vertex is expired, its weight is transmitted to one of its neighbour. In the end, only one vertex is active. This surviving vertex is called the leader.

Activation rules:

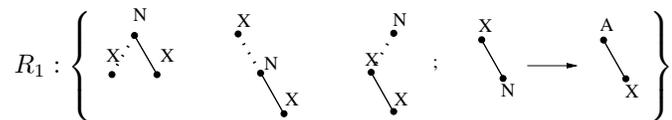
Each vertex in the graph can determine locally if it is active or not according to the rules R_i bellows.

- R_0 : If the degree of v is zero ($deg(v) = 0$), then the election is over, and v is the elected vertex. It is important to note that this vertex is considered as an active vertex.

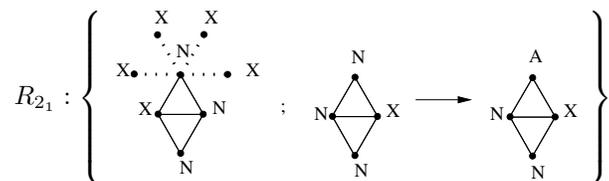


In this rule, the forbidden context shows that v shouldn't have a **X**-labelled neighbour.

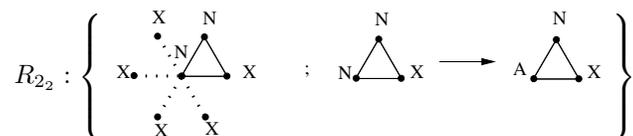
- R_1 : If the degree of v is 1, then v becomes active and generates its lifetime. Once the lifetime of the vertex v is expired, it disappears with the incident edge and its neighbour, noted u , recovers its weight.

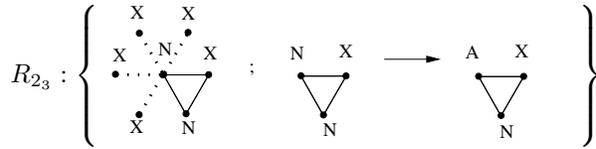


- R_2 : If the degree of v is 2 then depending on its position it could become active or not. We distinguish five sub-rules to ensure that v becomes active:

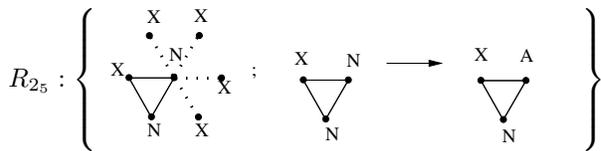
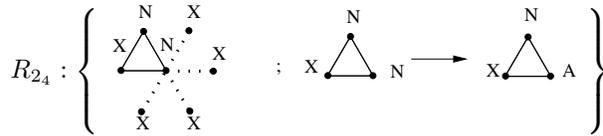


The sub-rule R_{2_1} expresses that if $v = (x, y + 1)$ is on the top of a up triangle cell $\{(x, y), (x + 1, y), (x, y + 1)\}$ with the existence of the down triangle cell $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$, then v becomes active.



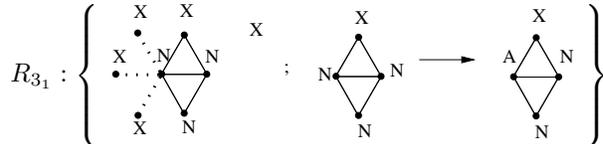


The sub-rule R_{2_2} (resp. R_{2_3}) expresses that if $v = (x, y)$ is on the left of a up (resp. down) triangle cell then v becomes active.

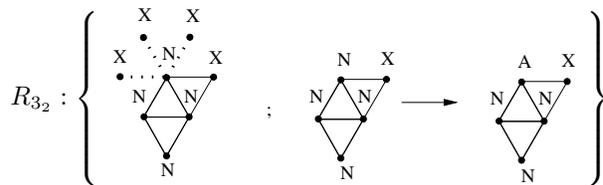


The sub-rule R_{2_4} (resp. R_{2_5}) expresses that if $v = (x + 1, y)$ is on the right of a up (resp. down) triangle cell then v becomes active.

- R_3 : If $deg(v) = 3$, then two cases arise:



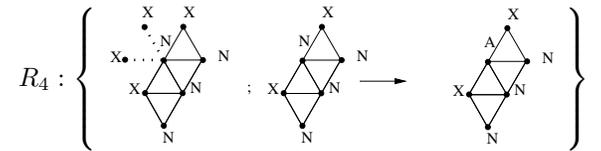
The sub-rule R_{3_1} describes that if v is either on the left of the up and the down triangle cells, then v becomes active.



The sub-rule R_{3_2} explains that if $v = (x, y + 1)$ is either on the top of the up triangle cell $\{(x, y), (x, y + 1), (x + 1, y)\}$ and on the left of the down triangle cell $\{(x, y + 1), (x + 1, y), (x + 1, y + 1)\}$ and also the the down triangle cell $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$ exists, then v becomes active.

- R_4 : If $deg(v) = 4$, then if $v = (x, y + 1)$ belongs to three cells $\{(x, y), (x, y + 1), (x + 1, y)\}$, $\{(x, y + 1), (x, y + 2), (x + 1, y + 1)\}$, and $\{(x, y + 1), (x + 1, y + 1), (x + 1, y)\}$ and the down triangle cell $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$

exists, then v becomes active.



Whenever a vertex v of weight w_v becomes active, it generates its lifetime $L_t(v)$ which is an exponentially distributed random variable (r.v.).

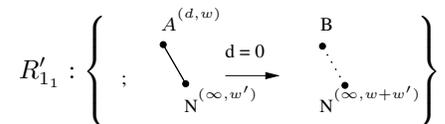
$$Pr(L_t(v) \geq t) = e^{-w_v(t)}.$$

This random variable has the expected value $1/w_v$.

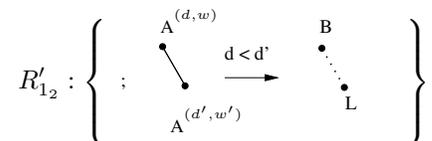
Weight transmission rules:

Once the lifetime has expired the vertex will be no longer A-labelled and the algorithm removes the vertex with all incident edges, giving its weight to the selected neighbours. The choice of the weight receiver neighbour is done according to rules R'_i . In those rules, d denotes the vertex lifetime and when a vertex is removed all incident edges are removed but we conserve the edges through which the weights are transmitted. Those edges are dotted.

- R'_0 : The election is terminated, the remaining vertex is considered as the leader.
- R'_1 : The neighbour vertex u of v recovers the weight of v , and it is either in the active state, or in the neutral state. In both cases, we have:
 - R'_{1_1} : If u is neutral then when it recovers the weight of v , it decides locally if it becomes active in the residual graph $G' = (V \setminus \{v\}, E \setminus \{v, u\}, u \in V)$.



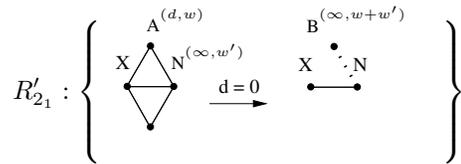
- R'_{1_2} : Otherwise, u is active before the time when v is removed. So, u becomes the elected one. We will be partially in the case of the rule R_0 .



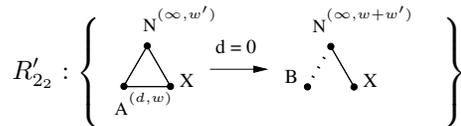
- R'_2 : If the lifetime of the vertex of degree 2 is expired, it is removed with its incident edges and the right

neighbour $(x + 1, y)$ recovers its weight (in a up triangle case), or else the down neighbour $(x + 1, y - 1)$ recovers its weight (in a down triangle case). We have:

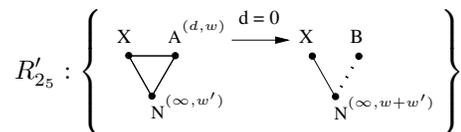
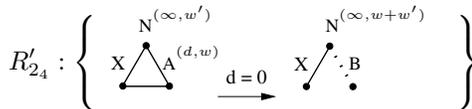
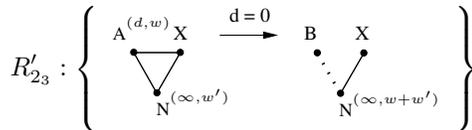
- R'_{21} : If the active vertex $v = (x, y + 1)$ is on the top of $T_u = \{(x, y), (x, y + 1), (x + 1, y)\}$ with condition of the existence of $T_d = \{(x, y), (x + 1, y), (x + 1, y - 1)\}$, then the neighbour $u = (x + 1, y)$ recovers its weight.



- R'_{22} : If the active vertex $v = (x, y)$ is the right-down vertex of the up triangle $T_u = \{(x, y), (x, y + 1), (x + 1, y)\}$, then its neighbour $u = (x, y - 1)$ recovers its weight when v is removed.

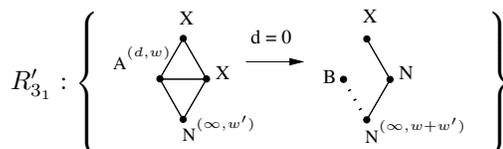


- R'_{23}, R'_{24} and R'_{25} : Like the rules R'_{21} and R'_{22} the transmission of the weight pass through the diagonal indecent edge of the removed vertex.

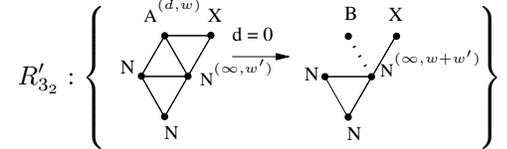


- R'_3 : If the $deg(v) = 3$ then v disappears and the right-down neighbour gets its weight. In this case we distinguish the three following sub-rules.

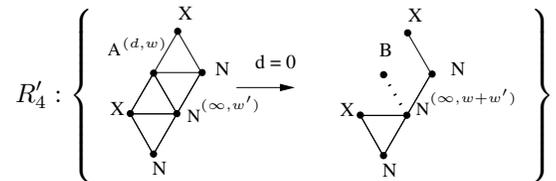
- R'_{31} : If the active vertex $v = (x, y)$ belongs to the two pairs of cells $\{(x, y), (x, y + 1), (x + 1, y)\}$ and $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$, then when its lifetime ends, its neighbour $u = (x + 1, y - 1)$ recovers its weight.



- R'_{32} : If the active vertex $v = (x, y)$ belongs to the two pairs of cells $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$ and $\{(x, y), (x + 1, y - 1), (x, y - 1)\}$, then its neighbour $u = (x + 1, y - 1)$ collects its weight when v is vanished.



- R'_4 : If the active vertex $v = (x, y)$ belongs to the tree cells $\{(x, y), (x, y + 1), (x + 1, y)\}$, $\{(x, y), (x + 1, y), (x + 1, y - 1)\}$, $\{(x, y), (x + 1, y - 1), (x, y - 1)\}$, then its neighbour $u = (x + 1, y - 1)$ recovers its weight at its disappearance.



C. Invariant proprieties

Our algorithm removes an active vertex once its lifetime expired. Thus, to ensure the continuity of the removal process, we must prove that the residual graph preserve the specific properties of TGG.

Proposition 4.1: Let $G = (V, E)$ be a TGG with size ≥ 2 , v an active vertex in G . The graph $G' = (V \setminus \{v\}, E \setminus \{\{v, u\}, \forall u \in V\})$ is a TGG.

Proof: Let $G = (V, E)$ be a TGG with size ≥ 2 , and v an active vertex in G and let the graph $G' = (V \setminus \{v\}, E \setminus \{\{v, u\}, \forall u \in V\})$. To prove the proposition, we must show that G' is a connected graph.

- If $deg(v) = 1$, then the removal of v and its incident edge in G doesn't introduce the disconnection of G' neither the creation of a hole in G' .
- If $deg(v) = 2, deg(v) = 3$ or $deg(v) = 4$, then let v, v_1, v_2 be a three vertices in the triangular grid graph G such as v is the active vertex whose lifetime has expired (in case of rules R_k and $R'_k, k = 2, 3, 4$). Consider the vertex $u \in V \setminus \{v, v_1, v_2\}$. then, if u is accessible to a node $v_i; 1 \leq i \leq 2$, via a path passing through v , then when v is deleted, u will still accessible to v_i in another way by taking the vertices $v_j \neq i; j = 1, 2$.

■

V. ANALYSIS OF THE ALGORITHM

A. Standard spanning tree

Let $G = (V, E)$ be a TGG and let F the set constituted by only the edges of E on which the weights of the vanishing vertices are transmitted. The set F can be built in advance with a distributed way as follows:

- If $e = \{(x, y), (x + 1, y - 1)\}$ is an edge in E then e belongs to F , i.e., any edge of the form $\{(x, y), (x + 1, y - 1)\}$ in E belongs to F .
- If $e = \{(x + 1, y - 1), (x + 1, y)\}$ belongs to E and to a single cycle γ of the form $\gamma = (x, y), (x, y + 1), (x + 1, y), (x + 1, y - 1)$. Then $e \in F$.

The graph $\mathbf{T} = (V, F)$ connects all vertices of G and it is acyclic. Then it is a spanning tree.

Proposition 5.1: The graph $\mathbf{T} = (V, F)$ as described above is a spanning tree of the triangular grid graph G .

Proof: We can prove this proposition by an inductive construction of \mathbf{T} on G :

- 1) If $G = (\{(x, y)\}, \emptyset)$, the triangular grid graph consists of only one vertex, then the proposition is asserted $\mathbf{T} = (\{(x, y)\}, \emptyset)$.
- 2) Let $G = (V, E)$ be a TGG and let $\mathbf{T} = (V, F \subseteq E)$ be the spanning tree of G obtained by the above rules. Consider two adjacent vertices v and v' such as $v \in V$ and $v' \notin V$. According to the inductive rules seen in Section IV-B , the graph $G' = (V \cup \{v'\}, E \cup \{\{v, v'\}\})$ is a TGG. So, it remains to prove that the tree $\mathbf{T}' = (V \cup \{v'\}, F \cup \{\{v, v'\}\})$ is the spanning tree of G' .

We can easily see that no cycle is created when the new edge $\{v, v'\}$ is added. Now let the tree $A = (V_A, F_A)$ where $V_A = \{v, v'\}$ and $F_A = \{\{v, v'\}\}$, then when we join the spanning tree \mathbf{T} with the tree A the residual graph is acyclic. So it is a spanning tree of the triangular grid graph G' .

- 3) Let $G = (V, E)$ be TGG and let $\mathbf{T} = (V, F \subseteq E)$ be its spanning tree. Suppose now that V contains three adjacent vertices $v_1 = (x, y)$ and $v_2 = (x + 1, y)$ and $v_3 = (x, y + 1)$ or else $v_1 = (x, y)$ and $v_2 = (x + 1, y)$ and $v_4 = (x + 1, y - 1)$ located in a cell such as two edges of the cell are in E and the third one, called e , is not. So according to inductive rules seen in section IV-B, the residual graph $G' = (V, E \cup \{e\})$, after the insertion of the new edge e , is a TGG.

However, it remains to prove that the weight transmission occurs through the spanning tree $\mathbf{T}' = (V, F')$ of G' .

Let $C_1 = \{v_1, v_2, v_3\}$ or $C_2 = \{v_1, v_2, v_4\}$ be a cell of $G' = (V, E \cup \{e\})$ and let $e_1 = \{v_1, v_2\}$, $e_2 = \{v_2, v_3\}$, $e_3 = \{v_2, v_4\}$, $e_4 = \{v_1, v_3\}$, and $e_5 = \{v_1, v_4\}$.

We have:

- If $e = e_1$ or $e = e_3$ or $e = e_4$ then $F' = F$. (In this case the spanning tree does not change.)
- If $e = e_2$ then $F' = F \setminus \{e_4\} \cup \{e_2\}$.
- If $e = e_5$ then $F' = F \setminus \{e_3\} \cup \{e_5\}$.

We can easily notice that the graph \mathbf{T}' is a connected graph and, moreover, no cycle is generated when e is added. Thus, \mathbf{T}' is a spanning tree of G . ■

Remark 5.1: The spanning tree constructed by these rules is unique.

Definition 5.1: The spanning tree $\mathbf{T} = (V, F)$ is called standard spanning tree of the triangular grid graph G .

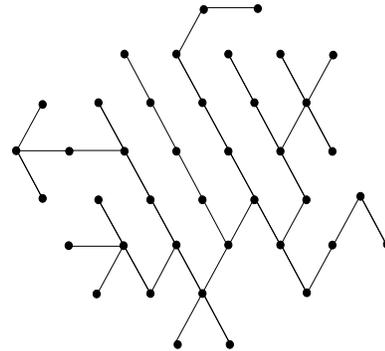


Figure 3. Standard spanning tree of the TGG given in Fig 2.

Proposition 5.2: Let $G = (V, E)$ be a TGG and $\mathbf{T} = (V, F)$ its standard spanning tree. The vertex $v \in V$ is an active vertex in G iff it is a leaf in \mathbf{T} .

Proof:

\Rightarrow Let the six vertices of a TGG $G = (V, E)$ defined as follows:

- $v_1 = (x, y)$
- $v_2 = (x, y + 1)$
- $v_3 = (x + 1, y)$
- $v_4 = (x + 1, y - 1)$
- $v_5 = (x + 1, y + 1)$
- $v_6 = (x, y + 2)$,

and let $C_1 = \{v_1, v_2, v_3\}$, $C_2 = \{v_1, v_3, v_4\}$, $C_3 = \{v_2, v_3, v_5\}$ and $C_4 = \{v_2, v_5, v_6\}$ four cells. If v one active vertex of those vertices then we will show that v is a leaf in the spanning tree \mathbf{T} of G .

- If $deg(v) = 1$, then certainly, v is a leaf in \mathbf{T} .

- If $deg(v) = 2$, then we enumerate the following cases:
 - (i) if v belongs to the cell C_1 and C_2 doesn't form a cell in G , then v is an extremity of the horizontal edge $\{v_1, v_3\}$, and since the weight transmission doesn't pass through this edge, thus, v is of degree 1 in \mathbf{T} (i.e. it is a leaf).
 - (ii) If v belongs to the cell C_2 , then v is an end of the horizontal edge $\{v_1, v_3\}$, and since the transmission of weight does not pass through this edge, thus, v is a leaf in \mathbf{T} .
 - (iii) If $v = v_2$ and the cells C_1 and C_2 exist in G then v is one end of the edge $\{v_2, v_3\}$. While the weight transmission doesn't pass through the edge $\{v_1, v_2\}$, it becomes a leaf in \mathbf{T} .
- If $deg(v) = 3$ and v belongs both to C_1 and C_2 , then we have the bellow cases:
 - (i) If $v = v_1$ then the weight transmission does not pass through the edge $\{v_1, v_2\}$ either $\{(v_1, v_3)\}$. So v becomes a leaf in \mathbf{T} .
 - (ii) If $v = v_2$ then the weight transmission does not pass neither through the edges $\{v_1, v_2\}$ and $\{v_2, v_3\}$. So v becomes a leaf in \mathbf{T} .
- If $deg(v) = 4$, and v belongs to the three cells C_1 , C_3 , and C_4 of G , then the weight transmission pass only through $\{v_2, v_3\}$. So v (equals v_2) becomes a leaf in \mathbf{T} .

⇐ Suppose now that v is a leaf in $\mathbf{T}=(V, F)$ and prove that v is an active vertex in G .

- If $deg(v) = 1$, then clearly v is an active vertex in G .
- If $deg(v) = 2$, then the two incident edges to v in G couldn't be in the same line, otherwise v is not a leaf in \mathbf{T} . In the case where those edges are in a different orientations, only the edge $\{v_1, v_4\}$ or else the edge $\{v_2, v_3\}$ is in F , in addition, v is in the context of the rules R_2 , then it is an active vertex.
- If $deg(v) = 3$, then with similar reasoning to the previous case, only one of the incident edges of v is in F . Thus, v is in the context of the rules R_3 . So it is active.
- If $deg(v) = 4$ and the cells C_1 , C_3 , and C_4 are in G , then v is in the context of the rule R_4 . Consequently, only the edge $\{v_2, v_3\}$ is in F , and according to the construction rules of F the vertex v is active.

■

B. Uniform election algorithm

Based on the results of the previous sections, we can summarize the distributed probabilistic election algorithm in triangular grid graph G as follows.

While G is not reduced to a single vertex **do**

- Each active vertex (rules R_0 - R_4) generates its lifetime according to its weight.
- Once the lifetime of an active vertex has expired, it is removed with its incident edges and its neighbour in the standard spanning tree collects its weight.

The election algorithm in a TGG is seen as an election algorithm in its standard spanning tree. The Proposition 5.2 shows that each active vertex in a triangular grid graph is a leaf in its standard spanning tree, and the weights of this vertex in the both configurations are equals.

Let $G = (V, E)$ be a TGG. Initially, all vertices have the same weight 1: $w(v) = 1, \forall v \in V$. According to the rules introduced in Section IV-B, when an active vertex disappears, its *successor* collects its weight and adds it to its current weight. At the time t when a vertex v becomes active, its weight is the number of the vanishing vertices of its sides in the standard spanning tree. The lifetime $L(v)$ of a vertex v is a exponential random variable of parameter $\lambda(v)$ such that:

$$\lambda(v) = w(v) : Pr(L(v) \geq t) = e^{-\lambda(v)t}, \forall t \geq 0.$$

This property is equivalent to say that the probability of the disappearance of v in the time interval $[t, t + h]$ is $\lambda(v)h + o(h)$, when $h \rightarrow 0$ at each instant t , and this is independently of what happening elsewhere and what happened in the past. The random process is a variant of pure death process which is, in its turn, a special example of the Markov process in continuous time.

C. Election process

Probabilistic election can be mathematically modelled by a Markov process in continuous time. The initial state of the process is $G = (V, E)$ (the entire TGG). Let \mathcal{S}_G be the set of all sub-triangular grid graphs of G and $G' \in \mathcal{S}_G$

We define \mathbf{R} by:

$\mathbf{R} = G' \cup (\{v\}, \{\{v, u\}\})$, u adjacent with v in \mathbf{T} , i.e. the remove of the vertex v and all its incident edges from \mathbf{R} leads to G' .

The transition probability from the triangular grid graph \mathbf{R} to the G' is:

$$P_{(\mathbf{R}, G')} = \frac{w(v)}{\sum_{u \text{ active in } \mathbf{R}} w(u)}$$

The following properties characterize the process of elimination in a TGG.

- The death rate of the triangular grid graph G is:

$$\lambda(G) = w(G) = \sum_{u \text{ active in } G} w(u)$$

- The lifetime of G is: $L(G) = \min_u \{L(u), u \text{ active in } G\}$ has the following distribution function:

$$Pr(L(G) \leq t) = 1 - Pr(L(G) \geq t) = 1 - e^{-\lambda(G)t},$$

$\forall t \in \mathbb{R}^+$

Proposition 5.3: Let G' be a TGG in \mathcal{S}_G , and let $P_{G'}(t)$ the probability that G' is the state of the election at time t . We have:

- (i) $\frac{dP_G(t)}{dt} = -w(G)P_G(t)$,
- (ii) for $G' \neq G$ of size ≥ 2 ,
 $\frac{dP_{G'}(t)}{dt} = -w(G')P_{G'}(t) + \sum_{v \text{ active in } R} w(v)P_R(t)$,
- (iii) $\frac{dP_{\{\{v\}, \emptyset\}}(t)}{dt} = \sum_{u \text{ adjacent to } v \text{ in } G} w(u)P_{\{\{u,v\}, \{\{v,u\}\}}(t)$
with the initial condition $P_G(0) = 1$.

Proof: Let G' is a sub-TGG of G and consider the evolution of the elimination process in the interval $[t, t+h]$. Let's calculate the probability of being in the state G' at time $t+h$.

- For $G' \neq G$ and G' is not reduced to a leaf, we have:

$$P_{G'}(t+h) = \sum_{v \text{ active in } R} P_R(t)\pi_{R,G'}(h) + P_{G'}(t)\pi_{G',G'}(h) + o(h),$$

where $R = G' \cup \{v\}$ and $\pi_{R,G'}(h)$ is the probability of a direct transition from R to G' in a time interval of length h ; the summation is performed for each vertex v adjacent to G' .

$$P_{G'}(t+h) = h \sum_v \lambda(v)P_R(t) + P_{G'}(t)[1 - \lambda(G')] + o(h).$$

Therefore,

$$\frac{P_{G'}(t+h) - P_{G'}(t)}{h} = -\lambda(G')P_{G'}(t) + \sum_v \lambda(v)P_R(t) + \frac{o(h)}{h}.$$

This proves (ii).

- To prove (i), we remark that in the case of $P_G(t+h)$, the sum $\sum_v (\dots)$ disappeared from the right side, and since G has no predecessor graph. This established (i).

- To prove (iii), we just need to remark that the singleton state $(\{v\}, \emptyset)$ is absorbing, and, thus, $\pi_{\{\{v\}, \emptyset\}}(h) = 1$. So, in $P_{\{\{v\}, \emptyset\}}(t+h)$, the negative term disappeared. A simple computing gives (iii). ■

Proposition 5.4: The strategy described above leads to a totally fair election: in a TGG, all vertices have the same probability of being elected.

Proof: In [3] the authors give the prove of the uniform election in trees. In our work we have showed that there is a similarity of the election process over a TGG and over its standard spanning tree \mathbf{T} . Using the similarity between the two structures, we can conclude, based on the results presented in [3], that for a triangular grid graph G of size n , the probability of being elected in G for any vertex $v \in G$ is $\frac{1}{n}$. ■

VI. CONCLUSION

In this paper, we proposed and analysed a probabilistic algorithm for uniform election in triangular grid graphs (TGG). We have introduced some rules to produce the family of those graphs.

Our algorithms use random delay associated to discovered vertices (active ones). These delays are an independent random variables and are locally generated when the vertices are discovered. To determine locally the active vertices we presented the activation rules. Also, we presented the weight transmissions rules for the successor of the vanishing vertex.

The election process is an elimination process that remove the active vertices of the TGG until the graph is reduced to only one vertex, called leader. Using a single pass and a local computations, the elimination process is modelled by a pure death Markov process in a continuous time.

Finally, we showed that our algorithm is totally fair, since it gives the same probability to each vertex to be elected.

Our further work will be focussed on the study of the uniform election in the chordal graphs.

REFERENCES

- [1] G. L. Lann, "Distributed systems – toward a formal approach," in *Proceedings of the IFIP Congress 77*, 1977, pp. 155–160.
- [2] Y. Métivier and N. Saheb, "Probabilistic analysis of an election algorithm in a tree," in *Colloquium on trees in algebra and programming*, ser. Lecture Notes in Comput. Sci., vol. 787. Springer-Verlag, 1994, pp. 234–246.
- [3] Y. Métivier, N. Saheb, and A. Zemmari, "A uniform randomized election in trees (extended abstract)," in *Proceedings of The 10th International Colloquium on Structural Information and Communication Complexity (SIROCCO 10)*. Carleton university press, 2003, pp. 259–274.
- [4] A. E. HIBAOU, N. SAHEB, and A. ZEMMARI, "A uniform probabilistic election algorithm in k -trees," *IMACS : 17th IMACS World Congress : Scientific Computation, Applied Mathematics and Simulation*, July 2005.

- [5] A. E. Hibaoui, J. M. Robson, N. Saheb-Djahromi, and A. Zemmari, "Uniform election in trees and polyominoes," *Discrete Appl. Math.*, vol. 158, no. 9, pp. 981–987, May 2010.
- [6] V. S. Gordon, Y. L. Orlovich, and F. Werner, "Hamiltonian properties of triangular grid graphs," *Discrete Mathematics*, vol. 308, pp. 6166–6188, 2008.
- [7] M. Benantar, U. Dogrusoz, J. Flaherty, and N. S. Krishnamoorthy, "Triangle graphs," *Applied Numerical Mathematics*, vol. 17, pp. 85–96, 1995.
- [8] I. Litovsky, Y. Métivier, and E. Sopena, "Different local controls for graph relabelling systems," *Math. Syst. Theory*, vol. 28, pp. 41–65, 1995.
- [9] R. Sedgewick, *Algorithms in C++*, 1st ed. Addison-Wesley Co., 1992.
- [10] I. Litovsky, Y. Métivier, and E. Sopena, "Graph relabelling systems and distributed algorithms," in *Handbook of graph grammars and computing by graph transformation*, H. Ehrig, H. Kreowski, U. Montanari, and G. Rozenberg, Eds. World Scientific, 1999, vol. 3, pp. 1–56.
- [11] A. Sellami, "Des calculs locaux aux algorithmes distribués," Ph.D. dissertation, Université Bordeaux I, 2004.
- [12] J. Chalopin, Y. Métivier, and W. Zielonka, "Election, naming and cellular edge local computations," in *Proc. of International conference on graph transformation*, vol. 3256, ICGT'04, LNCS, 2004, pp. 242–256.
- [13] I. Litovsky, Y. Métevier, and E. Sopena, "Definition and comparison of local computations on graphs and networks," in *MFCS'92*, ser. Lecture Notes in Comput. Sci., vol. 629, 1992, pp. 364–373.
- [14] E. Godard, "Réécritures de graphes et algorithmique distribuée," Ph.D. dissertation, Université Bordeaux I, 2002.

Correlated Topic Model for Web Services Ranking

Mustapha AZNAG*, Mohamed QUAFALOU* and Zahi JARIR**

* Aix-Marseille University, LSIS UMR 7296, France.

{mustapha.aznag,mohamed.quafalou}@univ-amu.fr

** University of Cadi Ayyad, LISI Laboratory, FSSM, Morocco.

jarir@uca.ma

Abstract—With the increasing number of published Web services providing similar functionalities, it's very tedious for a service consumer to make decision to select the appropriate one according to her/his needs. In this paper, we explore several probabilistic topic models: Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) to extract latent factors from web service descriptions. In our approach, topic models are used as efficient dimension reduction techniques, which are able to capture semantic relationships between word-topic and topic-service interpreted in terms of probability distributions. To address the limitation of keywords-based queries, we represent web service description as a vector space and we introduce a new approach for discovering and ranking web services using latent factors. In our experiment, we evaluated our Service Discovery and Ranking approach by calculating the precision (P@n) and normalized discounted cumulative gain (NDCGn).

Keywords—Web service, Data Representation, Discovery, Ranking, Machine Learning, Topic Models

I. INTRODUCTION

Web services¹ [25] are defined as a software systems designed to support interoperable machine-to-machine interaction over a network. They are loosely coupled reusable software components that encapsulate discrete functionality and are distributed and programmatically accessible over the Internet. They are self contain, modular business applications that have open, internet-oriented, standards based interfaces [2]. The Service Oriented Architecture (SOA) is a model currently used to provide services on the internet. The SOA follows the find-bind-execute paradigm in which service providers register their services in public or private registries, which clients use to locate web services. SOA services have self-describing interfaces in platform-independent XML documents. Web Services Description Language (WSDL) is the standard language used to describe services. Web services communicate with messages formally defined via XML Schema. Different tasks like matching, ranking, discovery and composition have been intensively studied to improve the general web services management process. Thus, the web services community has proposed different approaches and methods to deal with these tasks. Empirical evaluations are generally proposed considering different simulation scenarios. Nowadays, we are moving from web of data to web of services as the number of UDDI Business Registries (URBs) is increasing. Moreover, the number of hosts

that offer available web services is also increasing significantly. Consequently, discovering services which can match with the user query is becoming a challenging and an important task. The keyword-based discovery mechanism supported by the most existing services search engines suffers from some key problems:

- User finds difficulties to select a desired service which satisfies his requirements as the number of retrieved services is huge.
- Keywords are insufficient in expressing semantic concepts. This is due to the fact that the functional requirements (keywords) are often described by natural language.

To enrich web service description, several Semantic Web methods and tools are developed, for instance, the authors of [10], [23], [1] use ontology to annotate the elements in web services. Nevertheless, the creation and maintenance of ontologies may be difficult and involve a huge amount of human effort [3], [14].

With the increasing number of published Web services providing similar functionalities, it's very tedious for a service consumer to make decision to select the appropriate one according to her/his needs. Therefore mechanisms and techniques are required to help consumers to discover which one is better. In this case one of the major filters adopted to evaluate these services is using Quality of Service (QoS) as a criterion. Generally QoS can be defined as an aggregation of non-functional attribute that may influence the quality of the provided Web service [26], [21], [17]. Although, in various approaches [26], [17] the authors propose to calculate an overall score that combines the quality of service (availability, response time, ...) and use it to classify the web services.

To address the limitation of keywords-based queries, we represent web service description as a vector and introduce a new approach for discovering and ranking web services based on probabilistic topic models. The probabilistic topic models are a way to deal with large volumes of data by discovering their hidden thematic structure. Their added value is that they can treat the textual data that have not been manually categorized by humans. The probabilistic topic models use their hidden variables to discover the latent semantic structure in large textual data.

In this paper we investigate using probabilistic machine-learning methods to extract latent factors $z_f \in Z = \{z_1, z_2, \dots, z_k\}$ from service descriptions. We will explore several probabilistic topic models : PLSA (Probabilistic latent

¹<http://www.w3.org/standards/webofservices>

semantic analysis), LDA (Latent Dirichlet Allocation) and CTM (Correlated Topic Model) and use them to analyze search in repository of web services and define which achieves the best results. By describing the services in terms of latent factors, the dimensionality of the system is reduced considerably. The latent factors can then also be used to provide an efficient discovery and ranking system. In our experiments, we consider that web services are mixtures of hidden topics, where a topic defines a probability distribution over words.

The rest of this paper is organized as follows. In Section II we describe in detail our Service Discovery and Ranking approach. Section III describes the experimental evaluation. Section IV provides an overview of related work. Finally, the conclusion and future work can be found in Section V.

II. WEB SERVICE DISCOVERY AND RANKING APPROACH

In this section, we will first describe the necessary preprocessing of WSDL document to construct a web service representation. We then discuss the probabilistic machine-learning techniques used to generate the latent factors. Finally, we explain how these latent factors are used to provide an efficient discovery and ranking mechanism.

A. Web Service Representation

Generally, every web service has a WSDL (Web Service Description Language) document that contains the description of the service. The WSDL document is an XML-based language, designed according to standards specified by the W3C, that provides a model for describing web services. It describes one or more services as collections of network endpoints, or ports. It provides the specifications necessary to use the web service by describing the communication protocol, the message format required to communicate with the service, the operations that the client can invoke and the service location. Two versions of WSDL recommendation exist: the 1.1² version, which is used in almost all existing systems, and the 2.0³ version which is intended to replace 1.1. These two versions are functionally quite similar but have substantial differences in XML structure.

To manage efficiently web service descriptions, we extract all features that describe a web service from the WSDL document. We recognize both WSDL versions (1.1 and 2.0). During this process, we proceed in two steps. The first step consists of checking availability of web service and validating the content of WSDL document. The second step is to get the WSDL document and read it directly from the WSDL URI to extract all information of the document.

Before representing web services as TF-IDF (Text Frequency and Inverse Frequency) [22] vectors, we need some preprocessing. There are commonly several steps:

- *Features extraction* extracts all features that describe a web service from the WSDL document, such as service name and documentation, messages, types and operations.

- *Tokenization*: Some terms are composed by several words, which is a combination of simple terms (e.g., *get_ComedyFilm_MaxPrice_Quality*). We use therefore regular expression to extract these simple terms (e.g., *get, Comedy, Film, Max, Price, Quality*).
- *Tag and stop words removal*: This step removes all HTML tags, CSS components, symbols (punctuation, etc.) and stop words, such as 'a', 'what', etc. The Stanford POS Tagger⁴ is then used to eliminate all the tags and stop words and only words tagged as nouns, verbs and adjectives are retained. We also remove the WSDL specific stopwords, such as *host, url, http, ftp, soap, type, binding, endpoint, get, set, request, response*, etc.
- *Word stemming*: We need to stem the words to their origins, which means that we only consider the root form of words. In this step we use the Porter Stemmer [19] to remove words which have the same stem. Words with the same stem will usually have the same meaning. For example, 'computer', 'computing' and 'compute' have the stem 'comput'. The Stemming process is more effective to identify the correlation between web services by representing them using these common stems (root forms).
- *Service Matrix construction*: After identifying all the functional terms, we calculate the frequency of these terms for all web services. We use the Vector Space Model (VSM) technique to represent each web service as a vector of these terms. In fact, it converts service description to vector form in order to facilitate the computational analysis of data. In information retrieval, VSM is identified as the most widely used representation for documents and is a very useful method for analyzing service descriptions. The TF-IDF algorithm [22] is used to represent a dataset of WSDL documents and convert it to VSM form. We use this technique, to represent a services descriptions in the form of *Service Matrix*. In the service matrix, each row represents a WSDL service description, each column represents a word from the whole text corpus (vocabulary) and each entry represents the TF-IDF weight of a word appearing in a WSDL document. TF-IDF gives a weight w_{ij} to every term j in a service description i using the equation: $w_{ij} = tf_{ij} \cdot \log(\frac{n}{n_j})$. Where tf_{ij} is the frequency of term j in WSDL document i , n is the total number of WSDL documents in the dataset, and n_j is the number of services that contain term j .

B. A Probabilistic Topic Model Approach

Service Discovery and Selection aim to find web services with user required functionalities. While Service Discovery process assumes that services with similar functionalities should be discovered, Service Selection and Ranking aim to find a proper services with the best user desired quality of services. Thus, Service Ranking aims to give a value of

²<http://www.w3.org/TR/wsdl>

³<http://www.w3.org/TR/wsdl20/>

⁴<http://nlp.stanford.edu/software/tagger.shtml>

relevance to each service returned by the discovery process and proceeds to order the results in descending order starting from the most relevant ones. In our approach, we apply probabilistic machine-learning techniques; Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM); to extract latent factors (or topics) $z_f \in Z = \{z_1, z_2, \dots, z_k\}$ from web service descriptions (i.e., *Service Matrix*). In our work, topic models are used as efficient dimension reduction techniques, which are able to capture semantic relationships between *word-topic* and *topic-service* interpreted in terms of probability distributions. In our context, an observed event corresponds to occurrence of a word w in a service description s . We propose to use the learned latent factors as the base criteria for computing the similarity between a service description and a user query. The services can then be ranked based on the relevancy to the submitted query.

The Probabilistic Latent Semantic Analysis (PLSA) is a generative statistical model for analyzing co-occurrence of data. PLSA is based on the aspect model [11]. Considering observations in the form of co-occurrences (s_i, w_j) of words and services, PLSA models the joint probability of an observed pair $P(s_i, w_j)$ obtained from the probabilistic model is shown as follows [11]:

$$P(s_i, w_j) = \sum_{f=1}^k P(z_f)P(s_i|z_f)P(w_j|z_f) \quad (1)$$

We assume that service descriptions and words are conditionally independent given the latent factor. We have implemented the PLSA model using the PennAspect⁵ model which uses maximum likelihood to compute the parameters. The dataset was divided into two equal segments which are then transformed into the specific format required by the PennAspect. We use words extracted from service descriptions and create a PLSA model. Once the latent variables $z_f \in Z = \{z_1, z_2, \dots, z_k\}$ are identified, services can be described as a multinomial probability distribution $P(z_f|s_i)$ where s_i is the description of the service i . The representation of a service with these latent variables reflects the likelihood that the service belongs to certain concept groups [16]. To construct a PLSA model, we first consider the joint probability of an observed pair $P(s_i, w_j)$ (Equation 1). The parameters $P(z)$, $P(s|z)$ and $P(w|z)$ can be found using a model fitting technique such as the Expectation Maximization (EM) algorithm [11].

The Latent Dirichlet Allocation (LDA) is a probabilistic topic model, which uses a generative probabilistic model for collections of discrete data [4]. LDA is an attempt to improve the PLSA by introducing a Dirichlet prior on service-topic distribution. As a conjugate prior for multinomial distributions, Dirichlet prior simplifies the problem of statistical inference. The principle of LDA is the same as that of PLSA: mapping high-dimensional count vectors to a lower dimensional representation in latent semantic space. Each word w in a service description s is generated by sampling a topic z from topic distribution, and then sampling a word from topic-word

distribution. The probability of the i th word occurring in a given service is given by Equation 2:

$$P(w_i) = \sum_{f=1}^k P(w_i|z_i = f)P(z_i = f) \quad (2)$$

Where z_i is a latent factor (or topic) from which the i th word was drawn, $P(z_i = f)$ is the probability of topic f being the topic from which w_i was drawn, and $P(w_i|z_i = f)$ is the probability of having word w_i given the f th topic.

Let $\theta^{(s)} = P(z)$ refer to the multinomial distribution over topics in the service description s and $\phi^{(j)} = P(w|z = j)$ refer to the multinomial distribution over words for the topic j . There are various algorithms available for estimating parameters in the LDA: Variational EM [4] and Gibbs sampling [24]. In this paper, we adopt an approach using Variational EM. See [4] for further details on the calculations.

For the LDA training, we used Blei's implementation⁶, which is a C implementation of LDA using Variational EM for Parameter Estimation and Inference. The key objective is to find the best set of latent variables that can explain the observed data. This can be made by estimating $\phi^{(j)}$ which provides information about the important words in topics and $\theta^{(s)}$ which provides the weights of those topics in each web service.

The Correlated Topic Model (CTM) is another probabilistic topic model that enhances the basic LDA [4], by modeling of correlations between topics. One key difference between LDA and CTM is that in LDA, there is an independence assumption between topics due to the Dirichlet prior on the distribution of topics. In fact, under a Dirichlet prior, the components of the distribution are independent whereas the logistic normal used in CTM, models correlation between the components through the covariance matrix of the normal distribution. However, in CTM, a topic may be consistent with the presence of other topics. Assume we have S web services as a text collection, each web service s contains N_s word tokens, T topics and a vocabulary of size W . The Logistic normal is obtained by :

- For each service, draw a K -dimensional vector η_s from a multivariate Gaussian distribution with mean μ and covariance matrix Σ : $\eta_s \sim \mathcal{N}(\mu, \Sigma)$
- We consider the mapping between the mean parameterization and the natural parameterization: $\theta = f(\eta_i) = \frac{\exp \eta_i}{\sum_i \exp \eta_i}$
- Map η into a simplex so that it sums to 1.

The main problem is to compute the posterior distribution of the latent variables given a web service : $P(\eta, z_{1:N}, w_{1:N})$. Since this quantity is intractable, we use approximate techniques. In this case, we choose variational methods rather than gibbs sampling because of the non-conjugacy between logistic normal and multinomial. The problem is then to bound the log probability of a web service :

⁵http://cis.upenn.edu/ungar/Datamining/software_dist/PennAspect/

⁶<http://www.cs.princeton.edu/blei/lda-c/>

$$\begin{aligned} \log P(w_{1:N}|\mu, \Sigma, \beta) &\geq E_q[\log P(\eta|\mu, \Sigma)] \\ &+ \sum_{n=1}^N E_q[\log P(z_n|\eta)] \\ &+ \sum_{n=1}^N E_q[\log P(w_n|z_n, \beta)] \\ &+ H(q) \end{aligned} \quad (3)$$

The expectation is taken with respect to a variational distribution of the latent variables :

$$q(\eta, z|\lambda, \nu^2, \phi) = \prod_{i=1}^K q(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^N q(z_n|\phi_n) \quad (4)$$

and $H(q)$ denotes the entropy of that distribution (See [5] for more details).

Given a model parameters $\{\beta_{1:K}, \mu, \Sigma\}$ and a web service $w_{1:N}$, the variational inference algorithm optimizes the lower bound (Equation 3) with respect to the variational parameters using the variational EM algorithm. In the E-step, we maximize the bound with respect to the variational parameters by performing variational inference for each web service. In the M-step, we maximize the bound with respect to the model parameters. The E-step and M-step are repeated until convergence.

For the CTM training, we used the Blei's implementation⁷, which is a C implementation of Correlated Topic Model using Variational EM for Parameter Estimation and Inference. We estimate the *topic-service* distribution by computing: $\theta = \frac{\exp(\eta)}{\sum_i \exp(\eta_i)}$. Where $\exp(\eta_i) = \exp(\lambda_i + \frac{\nu_i^2}{2})$ and the variational parameters $\{\lambda_i, \nu_i^2\}$ are respectively the mean and the variance of the normal distribution. Then, we estimate the *topic-word* distribution ϕ by calculating the exponential of the log probabilities of words for each topic.

After training the three probabilistic topic model, a set of matched services can be returned by comparing the similarity between the query and services in the dataset. We propose to use the probabilistic topic model to discover and rank the web services that match with the user query. Let $Q = \{w_1, w_2, \dots, w_n\}$ be a user query that contains a set of words w_i produced by a user. In our approach, we use the generated probabilities θ and ϕ as the base criteria for computing the similarity between a service description and a user query. For this, we model information retrieval as a probabilistic query to the topic model. We note this as $P(Q|s_i)$ where Q is the set of words contained in the query. Thus, using the assumptions of the topic model, $P(Q|s_i)$ can be calculated by equation 5.

$$P(Q|s_i) = \prod_{w_k \in Q} P(w_k|s_i) = \prod_{w_k \in Q} \sum_{z=1}^T P(w_k|z_f)P(z_f|s_i) \quad (5)$$

The most relevant services are the ones that maximize the conditional probability of the query $P(Q|s_i)$. Consequently, relevant services are ranked in order of their similarity score to the query. Thus, we obtain automatically an efficient ranking of the services retrieved.

⁷<http://www.cs.princeton.edu/blei/ctm-c/index.html>

We propose also to use another approach based on the proximity measure called *Multidimensional Angle* (also known as *Cosine Similarity*); a measure which uses the cosine of the angle between two vectors [20], [7]. In the first time, we represent the user's query as a distribution over topics. Thus, for each topic z_f we calculate the relatedness between query Q and z_f based on *topic-word* distribution ϕ using Equation 6.

$$P(Q|z_f) = \prod_{w_i \in Q} P(w_i|z_f) \quad (6)$$

Then, we calculate the similarity between the user's query and a web service by computing the Cosine Similarity between a vector containing the query's distribution over topics q and a vector containing the service's distribution of topics p . The multidimensional angle between a vector p and a vector q can be calculated using Equation 7:

$$Cos(p, q) = \frac{p \cdot q}{\|p\| \cdot \|q\|} = \frac{\sum_{i=1}^t p_i q_i}{\sqrt{\sum_{i=1}^t p_i^2 \sum_{i=1}^t q_i^2}} \quad (7)$$

where t is the number of topics.

In our experiments, we will compare the results obtained for the two methods (i.e. Conditional Probability, Multidimensional Angle) for the three probabilistic topic models.

III. EVALUATION

A. Web Services Corpus

Our experiments are performed out based on real-world web services obtained from [27]. The WSDL corpus consists of over 1051 web services from 8 different application domains. Each web service belongs to one out of eight service domains named as: Communication, Education, Economy, Food, Travel, Medical and Military. Table I lists the number of services from each domain.

Before applying the proposed Web Service Discovery and Ranking, we deal the WSDL corpus. The objective of this pre-processing is to identify the functional terms of services, which describe the semantics of their functionalities. WSDL corpus processing consists of several steps: *Features extraction, Tokenization, Tag and stop words removal, Word stemming and Service Matrix construction* (See Section II-A).

#	Domains	Number of services
1	Communication	59
2	Economy	354
3	Education	264
4	Food	41
5	Geography	60
6	Medical	72
7	Travel	161
8	Military	40
Total		1051

TABLE I: Domains of Web services

We evaluated the effectiveness of our Web Service Discovery and Ranking for the three probabilistic topic models (labeled *PLSA, LDA* and *CTM*) using both methods Conditional

Probability (labeled *CP*) and Multidimensional Angle (labeled *MA*). The probabilistic methods are compared with a text-matching approach (labeled *Text-Search*). For this experiment, we use the services description collected from the WSDL corpus. As described previously, the services are divided into eight domains and some queries templates are provided together with a relevant response set for each query. The relevance sets for each query consists of a set of relevant service and each service s has a graded relevance value $relevance(s) \in \{1, 2, 3\}$ where 3 denotes *high relevance* to the query and 1 denotes a *low relevance*.

B. Evaluation Metrics

In order to evaluate the accuracy of our approach, we compute two standard measures used in *Information Retrieval*: *Precision at n* (*Precision@n*) and *Normalised Discounted Cumulative Gain* (*NDCG_n*). These evaluation techniques are used to measure the accuracy of a search and matchmaking mechanism.

1) *Precision@n*: In our context, *Precision@n* is a measure of the precision of the service discovery system taking into account the first n retrieved services. Therefore, *Precision@n* reflects the number of services which are relevant to the user query. The *precision@n* for a list of retrieved services is given by Equation 8:

$$Precision@n = \frac{|RelevantServices \cap RetrievedServices|}{|RetrievedServices|} \quad (8)$$

Where the list of relevant services to a given query is defined in the test collection. For this evaluation, we have considered only the services with a graded relevance value of 3 and 2.

2) *Normalised Discounted Cumulative Gain*: *NDCG_n* uses a graded relevance scale of each retrieved service from the result set to evaluate the gain, or usefulness, of a service based on its position in the result list. This measure is particularly useful in Information Retrieval for evaluating ranking results. The *NDCG_n* for n retrieved services is given by Equation 9.

$$NDCG_n = \frac{DCG_n}{IDCG_n} \quad (9)$$

Where *DCG_n* is the Discounted Cumulative Gain and *IDCG_n* is the Ideal Discounted Cumulative Gain. The *IDCG_n* is found by calculating the *DCG_n* of the first n returned services. The *DCG_n* is given by Equation 10

$$DCG_n = \sum_{i=1}^n \frac{2^{relevance(i)} - 1}{\log_2(1 + i)} \quad (10)$$

Where n is the number of services retrieved and $relevance(s)$ is the graded relevance of the service in the i th position in the ranked list. The *NDCG_n* values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. In our experiments, we consider only services with graded relevance values from 3 to 2 for this evaluation. *NDCG_n* values vary from 0 to 1.

C. Results and Discussion

We evaluated our Service Discovery and Ranking approach by calculating the *Precision@n* and *NDCG_n*. In this experiment, we have selected 8 queries - One query for each domain - (See Table II) from the test collection.

#	Domains	Query Name
1	Communication	Title Video Media
2	Economy	Shopping Mall Camera Price
3	Education	Researcher In Academia Address
4	Food	Grocery Store Food
5	Geography	Get Location Of City State
6	Medical	Hospital Investigating
7	Travel	City Country Hotel
8	Military	Government Missile Funding

TABLE II: Overview of the Queries used in our evaluation

The text description is retrieved from the query templates and used as the query string. We consider that the size of the services to be returned was set to 30.

Generally, the top most relevant services retrieved (i.e. the first 5 or 10) by a search engine are the main results that will be selected and used by the user. The *Precision@n* values and *NDCG_n* scores are obtained over all eight queries for the two probabilistic methods (i.e. CP: Conditional Probability, MA: Multidimensional Angle) based on the three probabilistic topic models (i.e. CTM, LDA, PLSA) and Text-Search.

The *Precision@5* and *Precision@10* values over eight queries are shown respectively in Table III and IV. The results show that the probabilistic method CP performs better than the MA for all the three probabilistic topic models. We remark that the CP based on CTM performs significantly than others methods. In fact, it gives a higher precision values (i.e. Average P@5 = 73% and Average P@10 = 68%) for all domains except Geography. We note also that the CP based on LDA performs better than MA based on LDA, CP/MA based PLSA and Text-Search. The methods based on PLSA and Text-Search were unable to find some of the relevant services that were not directly related to the queries. They give the lowest precision values.

The comparison of average *Precision@n* (See Figure 1) shows that the probabilistic method CP performs better than the MA for all the probabilistic topic models. The results show that the CTM and LDA perform better than Text-Search and PLSA. The probabilistic methods based on CTM and LDA used the information captured in the latent factors to match web services based on the conditional probability of the user query. Text-Search and PLSA were unable to find some of the relevant web services that were not directly related to the user's queries through CTM and LDA. The low precision results obtained by probabilistic method based on PLSA are due to limited number of concepts used for training the model. In this context, web service descriptions are similar to short documents. Therefore, the method based on PLSA model is not able to converge to a high precision using these limited concepts.

In Information retrieval, *NDCG_N* gives higher scores to systems which rank a search result list with higher relevance

Domain	Text-Search	PLSA		LDA		CTM	
		CP	MA	CP	MA	CP	MA
Communication	0.0	0.6	0.0	0.6	0.2	0.8	0.2
Economy	0.0	0.8	0.0	0.8	0.4	1.0	0.8
Education	0.0	0.8	0.4	0.6	0.2	0.4	0.0
Food	0.0	0.0	0.0	1.0	1.0	0.8	0.6
Geography	0.2	0.0	0.0	0.0	0.4	0.2	0.4
Medical	0.6	0.0	0.0	0.0	0.0	0.8	0.0
Travel	0.0	0.8	0.0	1.0	0.0	1.0	0.0
Military	0.0	0.0	0.0	0.6	0.6	0.8	0.6
Average	0.1	0.38	0.05	0.57	0.35	0.73	0.33

TABLE III: $Precision@5$ values for the eight queries. (CP: Conditional Probability, MA: Multidimensional Angle)

Domain	Text-Search	PLSA		LDA		CTM	
		CP	MA	CP	MA	CP	MA
Communication	0.4	0.6	0.0	0.5	0.5	0.6	0.4
Economy	0.0	0.8	0.0	0.7	0.7	0.7	0.9
Education	0.0	0.6	0.6	0.8	0.3	0.5	0.0
Food	0.1	0.0	0.0	0.9	0.9	0.9	0.8
Geography	0.1	0.0	0.0	0.0	0.2	0.2	0.2
Medical	0.5	0.0	0.0	0.2	0.2	0.8	0.1
Travel	0.0	0.8	0.0	0.6	0.0	0.9	0.0
Military	0.0	0.1	0.0	0.5	0.5	0.8	0.6
Average	0.14	0.36	0.08	0.53	0.41	0.68	0.38

TABLE IV: $Precision@10$ values for the eight queries. (CP: Conditional Probability, MA: Multidimensional Angle)

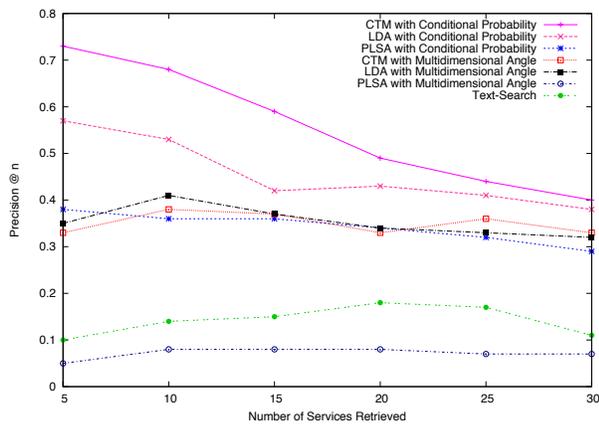


Fig. 1: Comparison of average $Precision@n$ values over 8 queries.

first and penalizes systems which return services with low relevance. The $NDCG_5$ and $NDCG_{10}$ values over eight queries are shown respectively in Table III and IV. The $NDCG_n$ values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. In our experiments, we consider services with graded relevance values from 3 to 2 for this evaluation. $NDCG_n$ values vary from 0 to 1. The results obtained for $NDCG_n$ show that the both CTM and LDA perform better than the other search methods. Thus, the probabilistic methods based on both CTM

Domain	Text-Search	PLSA		LDA		CTM	
		CP	MA	CP	MA	CP	MA
Communication	0.0	0.83	0.0	0.29	0.39	0.36	0.39
Economy	0.0	0.75	0.0	0.74	0.55	0.74	0.72
Education	0.0	0.57	0.27	0.44	0.17	0.64	0.0
Food	0.0	0.0	0.0	0.54	0.65	0.43	0.52
Geography	0.52	0.0	0.0	0.0	0.41	0.52	0.41
Medical	0.5	0.0	0.0	0.0	0.0	0.74	0.0
Travel	0.0	0.53	0.0	0.45	0.0	0.45	0.0
Military	0.0	0.0	0.0	0.83	0.69	0.52	0.5
Average	0.13	0.33	0.03	0.41	0.36	0.55	0.32

TABLE V: $NDCG_5$ values for the eight queries. (CP: Conditional Probability, MA: Multidimensional Angle)

Domain	Text-Search	PLSA		LDA		CTM	
		CP	MA	CP	MA	CP	MA
Communication	0.29	0.73	0.0	0.29	0.46	0.31	0.54
Economy	0.0	0.84	0.0	0.78	0.76	0.78	0.9
Education	0.0	0.49	0.4	0.68	0.33	0.58	0.0
Food	0.04	0.0	0.0	0.7	0.79	0.6	0.61
Geography	0.47	0.0	0.0	0.0	0.37	0.52	0.37
Medical	0.57	0.0	0.0	0.32	0.32	0.8	0.04
Travel	0.0	0.55	0.0	0.47	0.0	0.54	0.0
Military	0.0	0.1	0.0	0.64	0.61	0.48	0.52
Average	0.17	0.34	0.05	0.48	0.45	0.58	0.37

TABLE VI: $NDCG_{10}$ values for the eight queries. (CP: Conditional Probability, MA: Multidimensional Angle)

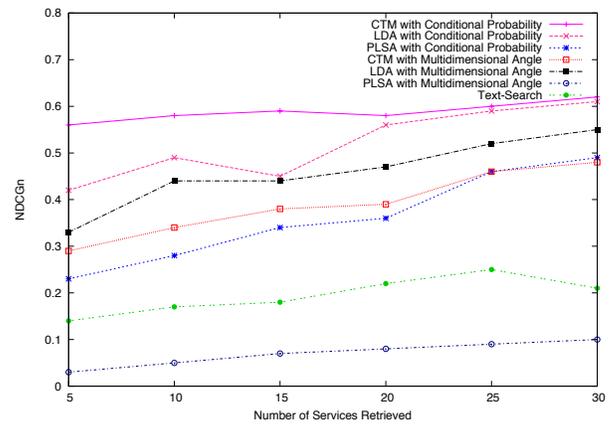


Fig. 2: Comparison of average $NDCG_n$ values over 8 queries.

and LDA give a higher $NDCG_n$ than all other methods for any number of web services retrieved (See Figure 2). This reflects the accuracy of the ranking mechanism used by our method. Text-Search and PLSA methods have a low $NDCG_n$ because, as shown in the $Precision@n$ results, both methods are unable to find some of the highly relevant services.

As can be seen from Figure 1 and 2, CTM based on the Conditional Probability performs significantly than others methods.

Finally, we evaluate the ranked lists obtained for both

ranking methods using the **Canberra distance**. In fact, the Canberra distance is used to measure the disarray for ranking lists, where rank differences in the top of the lists should be penalized more than those at the end of the lists [13]. Given two real-valued vectors $l, m \in \mathbb{R}^n$, their Canberra distance is defined as follows:

$$Ca(l, m) = \sum_{i=1}^N \frac{|l_i - m_i|}{|l_i| + |m_i|} \quad (11)$$

We consider only services with graded relevance values from 3 to 2 for this evaluation.

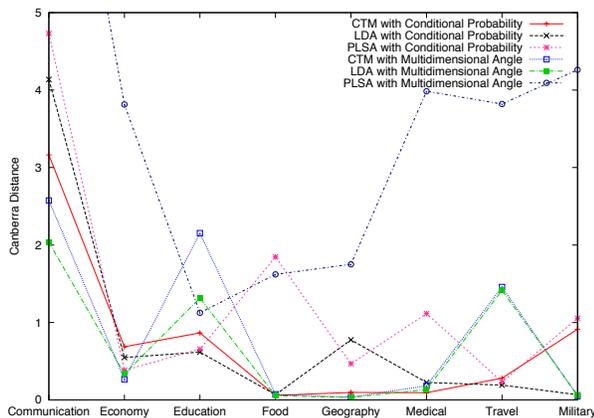


Fig. 3: Comparison of *CanberraDistance* values over 8 queries for the Ranking Methods.

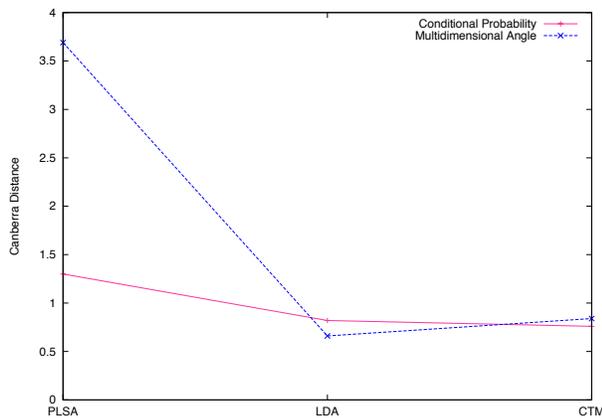


Fig. 4: Comparison of average *CanberraDistance* values for the Ranking Methods.

Figure 3 shows the Canberra Distance between the results obtained by both methods (CP and MA) based on the three probabilistic models and the relevant services for all eight queries. The comparison of average CanberraDistance values for the Ranking Methods is shown in Figure 4.

The results show that the *CTM with Conditional Probability* method based on the *Corelated Topic Model* gives the lowest CanberraDistance values. This reflects the accuracy of the ranking mechanism used by our method.

IV. RELATED WORK

In this section, we briefly discuss some of research works related to discovering Web services. In [1], the authors proposed an architecture for Web services filtering and clustering. The service filtering mechanism is based on user and application profiles that are described using OWL-S (Web Ontology Language for Services). The objectives of this matchmaking process are to save execution time and to improve the refinement of the stored data. Another similar approach [18] concentrates on Web service discovery with OWL-S and clustering technology. Nevertheless, the creation and maintenance of ontologies may be difficult and involve a huge amount of human effort [3], [14].

Generally, every web service associates with a WSDL document that contains the description of the service. A lot of research efforts have been devoted in utilizing WSDL documents [9], [3], [14], [15], [8], [16], [20]. Dong et al. [9] proposed the Web services search engine Woogle that is capable of providing Web services similarity search. However, their engine does not adequately consider data types, which usually reveal important information about the functionalities of Web services [12]. Liu and Wong [15] apply text mining techniques to extract features such as service content, context, host name, and service name, from Web service description files in order to cluster Web services. Elgazzar et al. [8] proposed a similar approach which clusters WSDL documents to improve the non-semantic web service discovery. They take the elements in WSDL documents as their feature, and cluster web services into functionality based clusters. The clustering results can be used to improve the quality of web service search results.

Some researchers use the proximity measures to calculate the similarity between services [18], [20]. Nayak et al. [18] proposed a method to improve the Web service discovery process using the Jaccard coefficient to calculate the similarity between Web services. Multidimensional Angle is an efficient measure of the proximity of two vectors. It is used in various clustering approaches [20]. This proximity measure applies cosine of the angle between two vectors. It reaches from the origin rather than the distance between the absolute position of the two points in vector space.

Ma et al. [16] proposed an approach similar to the previously discussed approaches [9], [1], [18] where the keywords are used first to retrieve Web services, and then to extract semantic concepts from the natural language descriptions in Web services. Ma et al. presented a service discovery mechanism called CPLSA which uses Probabilistic Latent Semantic

Analysis (PLSA) to extract latent factors from WSDL service descriptions after the search is narrowed down to a small cluster using a K-Means algorithm. The PLSA model represents a significant step towards probabilistic modelling of text, it is incomplete in that it provides no probabilistic model at the level of documents [4]. The Latent Dirichlet Allocation (LDA) [4] is an attempt to improve the PLSA by introducing a Dirichlet prior on document-topic distribution.

Cassar et al. [6], [7] investigated the use of probabilistic machine-learning techniques (PLSA and LDA) to extract latent factors from semantically enriched service descriptions. These latent factors provide a model which represents any type of service's descriptions in a vector form. In their approach, the authors assumed all service descriptions were written in the OWL-S. The results obtained from comparing the two methods (PLSA and LDA) showed that the LDA model provides a scalable and interoperable solution for automated service discovery in large service repositories. The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary. A limitation of LDA is the inability to model topic correlation [5]. This limitation stems from the use of the Dirichlet distribution to model the variability among the topic proportions.

The Correlated Topic Model (CTM) has been developed to address the limitation of LDA [5]. In CTM, topic proportions exhibit correlation via the logistic normal distribution. One key difference between LDA and CTM is the independence assumption between topics in LDA, due to the Dirichlet prior on the distribution of topics (under a Dirichlet prior, the components of the distribution are independent whereas the logistic normal models correlation between the components through the covariance matrix of the normal distribution). However, in the CTM model, a topic may be consistent with the presence of other topics. In this paper, we exploit the advantages of CTM to propose an approach for web service discovery and ranking. In our approach, we utilized CTM to capture the semantics hidden behind the words in a query, and the descriptions of the services. Then, we extracted latent factors from web service descriptions. The latent factors can then be used to provide an efficient discovery and ranking mechanism for web services.

V. CONCLUSION

In this paper, we have used several probabilistic topic models (i.e. PLSA, LDA and CTM) to extract latent factors from web service descriptions. The learned latent factors are then used to provide an efficient Service Discovery and Ranking. We evaluated our Service Discovery and Ranking approach by calculating the precision ($Precision@n$) and normalized discounted cumulative gain ($NDCG_n$). The comparison of $Precision@n$ and $NDCG_n$ show that the CTM performs better than the other search methods (i.e. LDA, PLSA and Text-Search). This reflects the accuracy of the ranking mechanism used by our method. The probabilistic methods based on CTM used the information captured in the latent factors to match web services based on the conditional probability of the user query.

Future work will focus on developing a new probabilistic topic model which will be able to tag web services automatically.

REFERENCES

- [1] Abramowicz, W., Haniewicz, K., Kaczmarek, M. and Zyskowski, D.: Architecture for Web services filtering and clustering. In ICIW'2007.
- [2] Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web Services - Concepts, Architectures and Applications. Springer Verlag, Berlin Heidelberg, 2004.
- [3] Atkinson, C., Bostan, P., Hummel O. and Stoll, D.: A Practical Approach to Web service Discovery and Retrieval. In ICWS'2007.
- [4] Blei, D., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation. J. Mach. Learn. Res., 3:993-1022, 2003.
- [5] Blei, D., and Lafferty, John D.: A Correlated Topic model of Science, In AAS 2007. pp. 17-35.
- [6] Cassar, G., Barnaghi, P. and Moessner, K.: Probabilistic methods for service clustering. In Proceeding of the 4th International Workshop on Semantic Web Service Matchmaking and Resource Retrieval, Organised in conjunction the ISWC'2010.
- [7] Cassar, G.; Barnaghi, P.; Moessner, K.: A Probabilistic Latent Factor approach to service ranking. In ICCP'2011, pp.103-109.
- [8] Elgazzar, K., Hassan A., Martin, P.: Clustering WSDL Documents to Bootstrap the Discovery of Web Services. In ICWS'2010, pp. 147-154.
- [9] Dong, X., Halevy, A., Madhavan, J., Nemes, E., Zhang, J.: Similarity Search for Web Services. In VLDB Conference, Toronto, Canada, pp. 372-383, 2004.
- [10] Hess, A. and Kushmerick, N.: Learning to Attach Semantic Metadata to Web services. In ISWC'2003, Sanibel Island, Florida, USA, 2003
- [11] Hofmann, T.: Probabilistic Latent Semantic Analysis. In UAI(1999), pp. 289-296.
- [12] Kokash, N.: A Comparison of Web Service Interface Similarity Measures. Frontiers in Artificial Intelligence and Applications, Vol. 142, pp.220-231, 2006.
- [13] Jurman, G., Riccadonna, S., Visintainer, R., Furlanello, C., Canberra Distance on Ranked Lists. In Proceedings of Advances in Ranking NIPS'2009 Workshop, 22-27.
- [14] Lausen, H. and Haselwanter, T.: Finding Web services. In European Semantic Technology Conference, Vienna, Austria, 2007
- [15] Liu, Wei., Wong, W.: Web service clustering using text mining techniques. In IJAOS'2009, Vol. 3, No. 1, pp. 6-26.
- [16] Ma, J., Zhang, Y. and He, J.: Efficiently finding web services using a clustering semantic approach. In CSSIA'08, pp 1-8. ACM, New York, NY, USA.
- [17] Maximilien, E.M., Singh, M.P.: Toward Autonomic Web Services Trust and Selection. In ICSOC'2004, pp. 212-221
- [18] Nayak, R. and Lee, B.: Web service Discovery with Additional Semantics and Clustering. In IEEE/WIC/ACM 2007
- [19] Porter, M. F.: An Algorithm for Suffix Stripping, In: Program 1980, Vol. 14, No. 3, pp. 130-137.
- [20] Platzer, C., Rosenberg F. and Dustdar, S.: Web service clustering using multidimensional angles as proximity measures. ACM Trans. Internet Technol. 9(3), pp. 1-26 (2009).
- [21] Rajendran, T., Balasubramanie, P.: An Optimal Broker-Based Architecture for Web Service Discovery with QoS Characteristics. In IWJSP'2010, Vol. 5, No. 1.
- [22] Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA (1989).
- [23] Sivashanmugam, K., Verma, A.P and Miller, J.A.: Adding Semantics to Web services Standards. In ICWS'2003, pp: 395-401.
- [24] Steyvers, M. and Griffiths, T.: Probabilistic topic models. In Latent Semantic Analysis: A Road to Meaning, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2007.

- [25] W3C (2004). Web services architecture. Technical report, W3C Working Group Note 11 February 2004.
- [26] Xu, Z., Martin, P., Powley, W. and Zulkernine, F.: Reputation Enhanced QoS-based Web services Discovery. In ICWS'2007.
- [27] Yu, Q.: Place Semantics into Context: Service Community Discovery from the WSDL Corpus. In ICDOC 2011, LNCS 7084, pp. 188-203.

Wideband Parameters Analysis and Validation for Indoor radio Channel at 60/70/80GHz for Gigabit Wireless Communication employing Isotropic, Horn and Omni directional Antenna

E. Affum¹ E.T. Tchao² K. Diawuo³ K. Agyekum⁴

Kwame Nkrumah Univ. of Science and Tech Kumasi, Ghana ^{1,2,3,4}.

eaffume@gmail.com¹, ettchao.coe@knust.edu.gh², kdiawuo.soe@knust.edu.gh³, kwame.agyekum⁴

Abstract—Recently, applications of millimeter (mm) waves for high-speed broadband wireless local area network communication systems in indoor environment are increasingly gaining recognition as it provides gigabit-speed wireless communications with carrier-class performances over distances of a mile or more due to spectrum availability and wider bandwidth requirements. Collectively referred to as E-Band, the millimeter wave wireless technology present the potential to offer bandwidth delivery comparable to that of fiber optic, but without the financial and logistic challenges of deploying fiber. This paper investigates the wideband parameters using the ray tracing technique for indoor propagation systems with rms delay spread for Omni-directional and Horn Antennas for Bent Tunnel at 80GHz. The results obtained were 2.03 and 1.95 respectively, besides, the normalized received power with 0.55×10^8 excess delay at 70GHz for Isotropic Antenna was at 0.97.

Index Terms—Indoor; Wideband; Isotropic; rms Delay; Power delay Profile; Excess delay.

I. INTRODUCTION

With end users ranging from corporate data centers to teenagers with iPhones demanding higher bandwidth, the demand for newer technologies to deliver this bandwidth is higher than ever before. A plethora of technologies exist for the delivery of bandwidth, with fiber optic cable considered to be the ultimate bandwidth delivery medium. However, the fiber optics are not unmatched [1] by any means, especially when all economic factors are considered. Millimeter wave wireless technology presents the potential to offer bandwidth delivery comparable to that of fiber optics, but without the financial and logistic challenges of deploying fiber. This paper is intended to analyze the wideband parameters of this new technology for different propagation indoor environment. Smulders studied wideband measurements of indoor radio channels operating in a 2 GHz frequency band [2] centered around 58GHz using a frequency step sounding technique. The results were presented for cell coverage and root mean square (RMS) delay spreads under both line-of-sight (LOS) and obstructed (OBS) situations. Again, various measurement campaigns and modeling activities were carried out [3] to obtain both the narrowband as well as the wideband characteristics of the 60 GHz channel for indoor and outdoor environments. A simple ray-tracing

was used to estimate the channel characteristics, for both narrow and wideband transmission systems in indoor as well as outdoor environments. Normalized received power, RMS delay profile and channel impulse response were simulated for indoor and outdoor radio channel. Ray tracing measurement of statistical parameters was comparatively studied and graphically represented by the group. Coherence bandwidth of wideband channel was estimated by Tlich and the group in [4]. They further provided sounding measurements in the 30 kHz100MHz band in several indoor environments. The coherence bandwidth and the RMS delay spread parameters were estimated from measurements of the complex transfer function and dispersion in the time domain, further, the variability of the coherence bandwidth and time-delay spread parameters with the channel class were presented based on the location of the receiver with respect to the transmitter and finally, related the RMS delay to the coherence bandwidth.

In April 2007, A 60GHz indoor propagation channel model based on the ray-tracing method was proposed by Chong et al, and in that study, they validated the proposed model with measurements conducted in indoor environment [5]. Moraitis and the group proposed the propagation models based on geometrical optics using ray-tracing theory for millimeter wave frequencies [6]. Expressions for Path loss, Received power, Power delay profile (PDP) and RMS delay spread were presented by Moraitis and the group in [7]. They performed propagation measurements at 60 GHz and determined the characteristics of indoor radio channels between fixed terminals that were illustrated. Path loss measurements were reported for line-of-sight (LOS), and non-line-of-sight (NLOS) cases, fading statistics in a physically stationary environment were extracted and effect due to the movement of the user on the temporal fading envelope were investigated. Path loss was predicted for models that provide excellent fitting with errors having dynamic range of fading in a quiescent environment.. In [8], a model for indoor radio propagation at 60GHz was used for predicting the performance of high speed wireless data networks, by using electromagnetic theory. Different models of the indoor environment were evaluated and corresponding received power, impulse response of the channel were pre-

dicted. Further, a statistical propagation model for the 60-GHz channel in a medium-sized room was presented by the Authors [9]. Extensive work was done in [10]. The paper they proposed an area prediction system, which was capable of accurately predicting indoor service areas using the ray-tracing method, when a base station antenna was installed indoors. Ray tracing techniques was discussed in various models and units. Received power and delay spread were estimated and simulated. Moreover, there was an in depth discussion on 60-GHz radio in different aspects. Propagation and antenna effects were studied in line-of-sight and non-line-of sight environments and Bit error rate were simulated for different noise level. The selections of wideband channel sounding measurements were performed as part of the AWACS (ATM Wireless Access Communications System). The results were obtained for two different indoor operating environments (mainly in line-of-sight conditions) at a carrier frequency of 19.37 GHz.. This paper analyzed the wideband parameters using a proposed indoor propagation environment with dimensions of $4.4m \times 2.5m$ representing a Straight Tunnel and a propagation environment using Bent Tunnel of angle of curvature of 45 with maximum height of 2m. Further, simulation results have been provided indicating normalized received power for different propagation environment. This paper is organized as follows: The propagation environments were first considered. Also the Narrowband and Wideband parameters were analyzed. This is followed by simulation results and discussion and finally conclusion.

II. PROPAGATION ENVIRONMENT

There are two general types of propagation modeling: site-specific and site general. Site-specific modeling requires detailed information on building layout, furniture, and transceiver locations [11]. It is performed using ray-tracing methods. For large-scale static environments, this approach may be viable. For most environments, however, the knowledge of the building layout and materials is limited and the environment itself can change by simply moving furniture or doors. Thus, the site-specific technique is not commonly employed. Site-general models provide gross statistical predictions of path loss for link design and are useful tools for performing initial design and layout of indoor wireless systems. In this work site-specific indoor environment is considered. Different types of environment were considered, namely: Plain Corridor, Corridor with equally spaced wooden door, glass door and lift, Straight Tunnel and Bent Tunnel

A. Plain Corridor

The propagation environment is a long plane corridor with dimensions $44 \times 2.20 \times 2.75m^3$ as shown in Figure 1. The left and right wall surfaces of the corridor are made of brick and plasterboard (relative permittivity $\epsilon_r = 4.44$ and $\epsilon_r = 5.0$). In order to simplify the simulation procedure, it was assumed that the surface is a uniform wall made of brick and plasterboard [9]. The floor is made of concrete covered with marble ($\epsilon_r =$

4.0) and furred ceiling is made of aluminum ($\epsilon_r = 1.0$) as shown in Figure 1

B. Corridor with Wooden Door, Glass Door & Lift

Figures 2 illustrate the corridor with wooden Door, Lift and Glass door. Between the distances of about 1-10m and 10.1-20m is a plane corridor, in between is a wooden door, similarly, in between 20-20.1m is a lift, and 30-30.1m glass door is present, from 30.1-40m is a plain corridor. Relative permittivity of wooden door is 3.3.

C. Straight Tunnel

The indoor propagation environment is a long tunnel with dimensions $44m \times 2.5m$ as shown in the Figure 3. The surfaces of the tunnel are made of concrete (relative permittivity (ϵ_r) = 5.0). The height of the transmitter is 2m and height of the receiver is 1.5m.

D. Bent Tunnel

In figure 4, the propagation environment is a bent tunnel with dimensions $44 \times 2.5m^2$. The angle of curvature of the Bent tunnel is 45o. The surfaces of the tunnel are made of concrete (relative permittivity (ϵ_r) = 5.0). The height of the transmitter is 2m and that of the receiver is 1.5m. The electromagnetic rays from the transmitter gets reflected, refracted, diffracted, scattered and absorbed by the propagation environment before reaching the receiver. The received signal is the combination of all the signals. At 60/70/80 GHz, the diffraction phenomenon is almost negligible and the diffracted power does not contribute to the total received power. So the Diffraction was not taken into account. The non-uniformities of the surface materials in indoor environments are such that the produced scattering is not a substantial contribution to the received power thus up to second order reflected rays were taken into consideration, since further reflected rays i.e., third, fourth and so on, have insignificant contribution to the total received power. Atmospheric propagation losses were not taken into account since in indoor environments the attenuation is very small (0.00116 dB/m). The beginning and the end of the corridor are open areas and were not taken into account in the simulations. The radio channel propagation modeling at millimeter wave frequencies can be realized based on ray-tracing theory. The ray-tracing method is among the available methods for the relatively accurate estimation of field strengths to deal with the type of complex layout that is often found in indoor environments. Ray-tracing allows fast computation of single and double reflection processes. In the 60/70/80 GHz region the diffraction phenomenon can be neglected and the sum of the direct ray and the reflected rays are enough to describe the behavior of the propagation channel with great accuracy.

III. NARROWBAND PARAMETERS

A. Received Power

The total received power (RR) of the multi-rays are calculated by [6] the summation of 'x single reflected and 'w double reflected rays given by

$$R_R = T_R \left(\frac{\lambda}{4\pi} \right)^2 \left| a_t a_r \left[\frac{e^{-jkd_1}}{d_1} + \sum_{i=1}^x R(\theta_0) \frac{e^{-jkd_2}}{d_2} + \sum_{j=1}^w R(\theta_1) R(\theta_2) \frac{e^{-jkd_3}}{d_3} \right] \right|^2 \quad (1)$$

where λ is the wave length; k is the wave number, d_1 is the distance of the direct path; d_2 is the distance of the single reflected path; d_3 is the distance of the double reflected path; a_t, a_r are the antenna functions; $R(\theta_0)$ is the reflection coefficient of the single reflected ray on the reflecting surface; $R(\theta_1)R(\theta_2)$ are the reflection coefficient of the double reflected rays on respective reflecting surfaces; and T_R is the transmitted power. For isotropic antennas ($a_t = a_r = 1$) the total received power [12] (R_R) is

$$R_R = T_R \left(\frac{\lambda}{4\pi} \right)^2 \left| \frac{e^{-jkd_1}}{d_1} + \sum_{i=1}^x R(\theta_0) \frac{e^{-jkd_2}}{d_2} + \sum_{j=1}^w R(\theta_1) R(\theta_2) \frac{e^{-jkd_3}}{d_3} \right|^2 \quad (2)$$

To examine the signal propagation in the indoor environment, we assumed three different transmission systems with different antenna characteristics and transmitted power. This was done so as to examine how the antenna radiation patterns affect the signal propagation in the indoor environment. The systems are: System 1: Isotropic antennas on both transmitter and receiver and 20 dBm output power. System 2: Transmitter power = 20 dBm, Transmitter Gain = Receiver Gain = 8.5 dBi for omnidirectional antenna System 3: Transmitter power = 10 dBm, Transmitter Gain = Receiver Gain = 20.8dBi also for horn antenna. Finally, the simulation was conducted with MATLAB script, using multi rays. The initial transmitter position is at the beginning of the propagation environment and the receiver is moving at almost constant speed. The total number of samples for the entire propagation environment (44 m) was 1024.

IV. WIDEBAND PARAMETERS

A. Power Delay Profile

The wideband multipath channel is often modeled as a time varying linear filter with complex impulse response [12]

$$(t, \tau) = \sum_{i=1}^N a_i(t, \tau) \exp(j\psi_i(t, \tau)) \delta(\tau - \tau_i) \quad (3)$$

For the indoor propagation environment where the time varying factors of the impulse response typically are human movement, it is appropriate to treat the channel as quasi-stationary. Assuming that the phase variations in the CIR have a uniform distribution we may consider only the amplitude and the delay components. The most significant parameter derived from the procedure is the received power as a function of the

time delay known also as the power delay profile (PDP). The power delay profile can be

$$P(\tau) = \sum_{i=1}^N P_r(d) \delta(\tau - \tau_i) \quad (4)$$

where, $P_r(d)$ is the received signal of the i^{th} ray and N is the total number of the rays used in the simulation procedure. The average received power in every bin is normalized to the maximum received power. The examination of the signal propagation in the indoor environment, using three different transmission systems with different antenna characteristics and transmitted power were considered. This was done so as to examine how the antenna

TABLE I
RMS DELAY SPREAD OF STRAIGHT TUNNEL

Antennas	60 GHz (ns)	70 GHz (ns)	80 GHz (ns)
Isotropic	1.92	1.93	1.94
Omni	1.84	1.85	1.86
Horn	1.77	1.78	1.79

TABLE II
RMS DELAY SPREAD OF BENT TUNNEL

Antennas	60 GHz (ns)	70 GHz (ns)	80 GHz (ns)
Isotropic	2.24	2.25	2.26
Omni	2.14	2.15	2.16
Horn	2.05	2.06	2.07

V. RESULTS AND DISCUSSIONS

Radiation patterns affect the signal propagation in the indoor environment. For Isotropic antennas both transmitter and receiver output power was 20 dBm with unity gain, for Omnidirectional antenna the transmitter power was also 20 dBm with transmitter and receiver gain as 8.5 dBi, besides, the transmitter power for horn antenna was 10 dBm with transmitter and receiver gain as 20.8 dBi. The height of the transmitter and receiver antennas in all propagation environments were 2m and 1.5m respectively. The distance between the transmitter and receiver was 44m and the total number of samples for the entire propagation environment (44 m) to be 1024. Figure 5, 6, 7 and 8 illustrate the total received power for different antenna system configurations under different propagation scenarios. It was observed that the received power of Horn antenna in Plain corridor at distance 10m was -30dBm, and Omni antenna were -45dBm and -60dBm for Isotropic antenna. Also, the received power for Plain Corridor at 70GHz at a distance 40m, was -30dBm, the received power for Horn antenna is -35dBm and for Omni antenna the received power was also observed as -50dBm and -60dBm approximately for Isotropic antenna. Again for Plain Corridor at 80GHz. for a distance 20m, the received power for Horn antenna -35dBm and the received power for Omni antenna is -50dBm and -65dBm for Isotropic antenna approximately. The received power for Corridor with wooden door, Lift and Glass door for 60GHz frequency at a distance of 10m, of Horn antenna was -30dBm, for Omni and

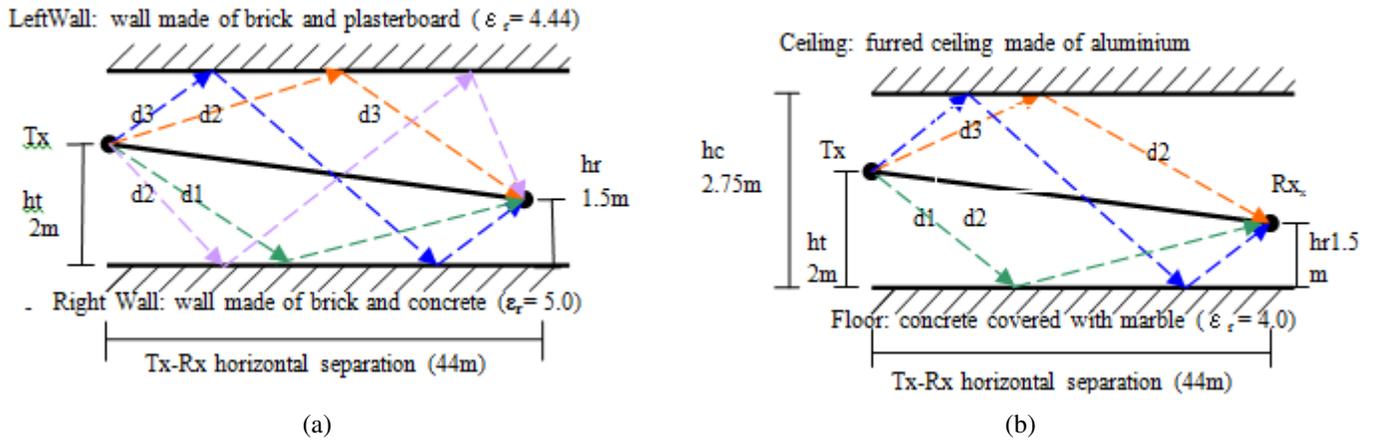


Fig. 1. Propagation Environment-Plain Corridor

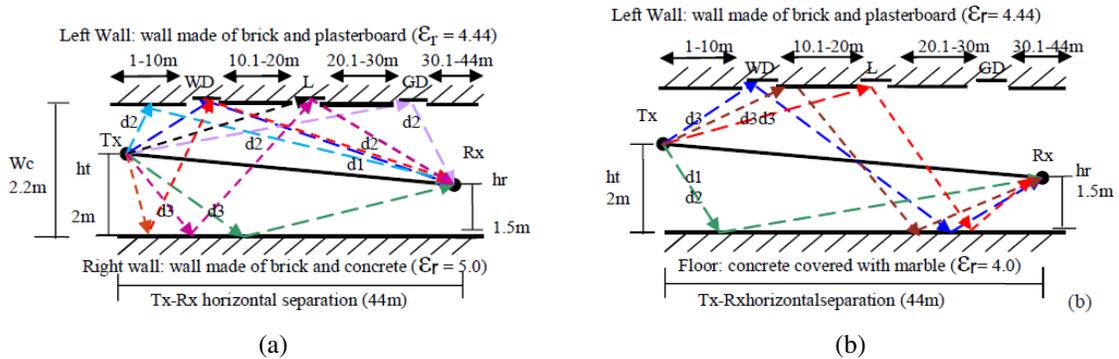


Fig. 2. Propagation Environment: Corridor with Wooden door, Glass and Lift [8]

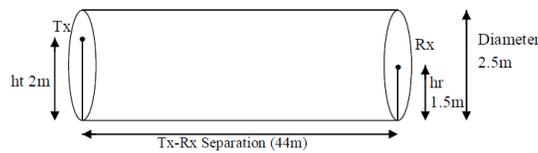


Fig. 3. Straight Tunnel made of concrete

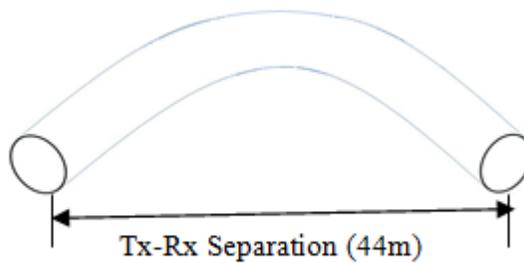


Fig. 4. Bent Tunnel made of concrete

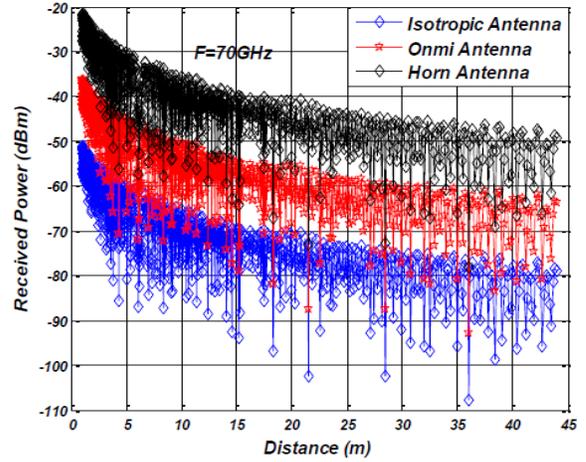
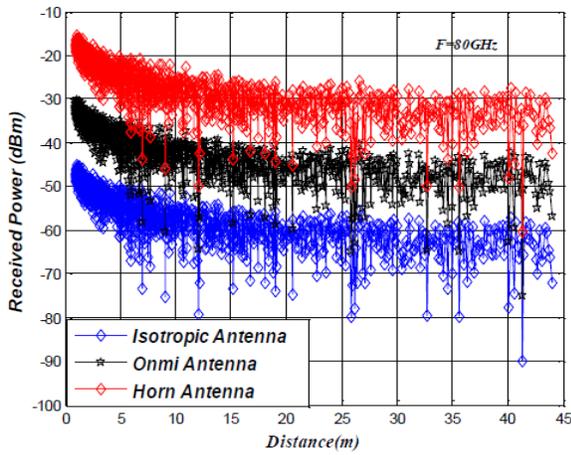


Fig. 5. Received power for different antenna systems at 80 GHz with Corridor with Wooden Doors, Glass Door and Lift

Fig. 6. Received power for different antenna system configurations at 70 GHz for Straight tunnel

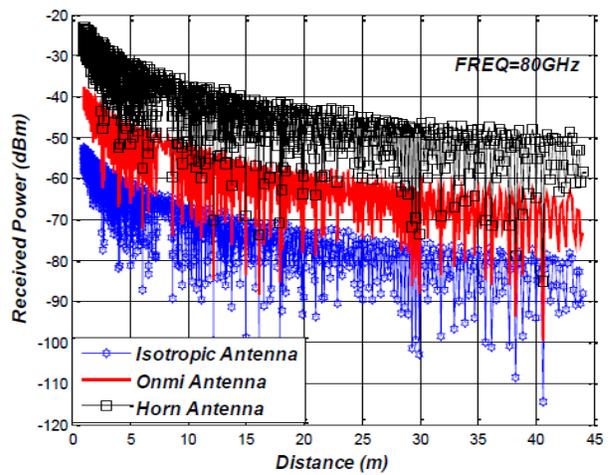
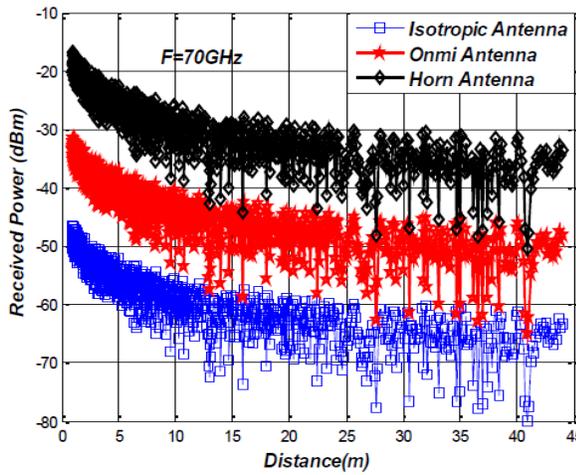


Fig. 7. Received power for different antenna system configurations at 70 GHz for Bent Tunnel

Fig. 8. Received power for different antenna system configurations at 80 GHz for Bent tunnel

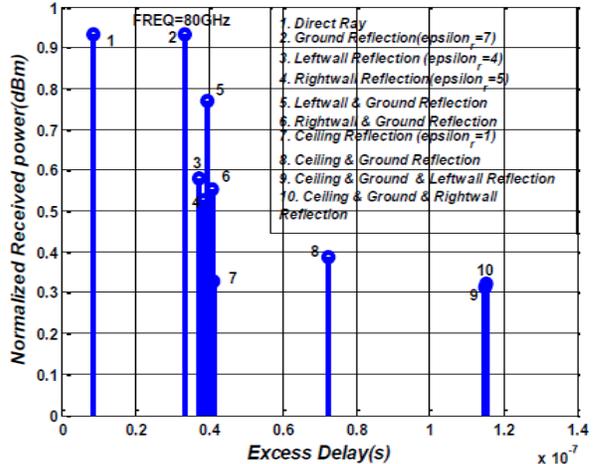
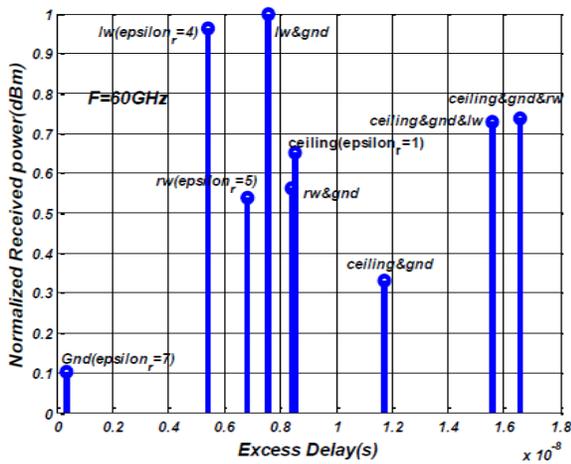


Fig. 9. Delay Spread of Straight Tunnel

Fig. 10. Delay Spread of Straight Tunnel

Isotropic antennas were -40dBm and -55dBm approximately. At Frequency 70GHz, the received power of Horn antenna at a distance of 20m was -30dBm, for Omni and Isotropic antenna the received powers -45dBm and 60dBm respectively. At 80GHz frequency, the received power of Horn antenna at a distance 30m, was -35dBm, for Omni antenna the received power is -50dBm. The received power in a Bent Tunnel at 60GHz was analyzed, at distance 10m, the power for Horn antenna was -35dBm, Omni antenna as 50dBm and -65dBm for Isotropic antenna. Also the for Bent Tunnel at 70GHz as shown in figure 7, for a distance of 20m, the received power for Horn antenna is -40dBm, and for Omni antenna the received power is -55dBm and -70dBm approximately for Isotropic antenna. The Power Delay Profile of Horn antenna at 60/70/80 GHz in Plain Corridor was also analyzed. The normalized received power of ceiling and ground reflected ray was 0.438 dBm with excess delay of $0.0724 \times 10^{-6}s$ and the normalized received power of ceiling, ground and left wall reflected ray was 0.364dBm with excess delay of $0.1152 \times 10^{-6}s$. Figure 10, illustrates the Power Delay Profile of Isotropic antenna at 70GHz. The Normalized received power of right wall and ground reflected ray was 0.527dBm with excess delay of $0.0386 \times 10^{-7}s$ and the normalized received power of ceiling reflected ray was 0.386dBm with excess delay of $0.0723 \times 10^{-7}s$ approximately. The normalized received power of direct ray at 80GHz is 0.931dBm with excess delay of $0.0333 \times 10^{-7}s$ and 0.818dBm was the normalized received power of ceiling and ground reflected ray with excess delay of $0.041 \times 10^{-7}s$.

A. Conclusion

This paper has presented the characteristics of the propagation channel at 60/70/80 GHz, using four different indoor environments namely: Plain Corridor, Corridor with wooden door, lift and glass door, Straight Tunnel and Bent Tunnel employing Horn Antenna, Isotropic Antenna and Omni Antenna, utilizing MATLAB. From the results analyzed so far the Horn antenna performance outweighs the Omni antenna and Isotropic antenna in all indoor environments at 60/70/80 GHz. The rms delay spread of propagation environments obtained also indicated that as the frequency increases, the rms delay spread of the propagation environment increased gradually

REFERENCES

- [1] P. Adhikari. Understanding millimeter wave wireless communication. In *San Diego*, 2008.
- [2] P. Smulders and A G Wagemans. Wideband indoor radio propagation measurements at 58 ghz. *Electronics Letters*, 28(13):1270–1272, 1992.
- [3] P Smulders and L Correia. Characterization of propagation in 60 ghz radio channel. *Electronics and Communication Engineering Journal*, pages 73–80, 1997.
- [4] M Tlich, G Avril, and A Zeddani. Home networking. *Springer Boston*, 256:129–142, 2008.
- [5] C C Chong, K Hamaguchi, P F M Smulders, and S K Yong. Millimeter-wave wireless communication systems: Theory and applications. *European Association for Signal Processing Journal*, 2007:1–2, 2007.
- [6] N Moraitis and P Constantinou. Propagation modeling at 60 ghz for indoor wireless lan applications. *IEEE Transactions on Wireless Communications*, 5:880–889, 2006.

- [7] N Moraitis. Indoor channel measurements and characterization at 60 ghz for wireless local area network applications. *IEEE Transactions on Antennas and Propagation*, 52(12):3180–3189, 2004.
- [8] P F Driessen. Development of propagation model in 20-60ghz band for indoor wireless communications. *IEEE Pacific Rim Conference on Communications, Computer and Signal Processing, Canada*, 1(4):59–62, 1991.
- [9] M Fryziel, C Loyez, L Clavier, N Rolland, and P A Rolland. Path-loss model of the 60-ghz indoor radio channel. *Microwave and Optical Technology Letters, France*, 34(3), 2002.
- [10] I Tetsuro, I Yuichiro, and F Teruya. Indoor microcell area prediction system using a ray-tracing method. *IEICE Transactions on Communications*, 83-B(11):1565–1576, 2003.
- [11] J Fisher, S Simpson, and T Welsh. An urban canyon multipath model for galileo, european navigation conference, copenhagen, england. 2002.
- [12] T S Rappaport. *Wireless communication*, upper saddle river, nj: Prentice hall. 1996.

Smart Grid Network Transmission Line RLC Modelling Using Random Power Line Synthesis Scheme

Ezennaya S.O¹, Udeze C. C², Okafor K .C³

^{2,3}R & D Department, Electronics Development Institute (FMST-NASENI), Awka, Nigeria.

Onyedikachi S.N⁴, Anierobi C.C⁵

^{1,4,5}Electronics Engineering Department, Nnamdi Azikiwe University, Awka, Nigeria.

Abstract—This work proposes Random Power line Synthesis (RPLS) as a quicker computational approach to solving RLC parameters of a modern smart grid transmission network. Since modern grid systems provide a holistic perspective of modern grid development, it is obvious that a transmission network that is ageing cannot serve the expanded load demand. The need to revolutionize the traditional transmission model while exploiting basic electrical theories and principles in Smart Grid (SG) architecture necessitated this paper. This work seeks to address the RLC parameter modelling for SG template to provision dynamic power in Nigerian context. Other schemes of transmission RLC modelling were studied as well as outlining their limitations. Consequently, we then proposed a fuzzy smart grid framework for RLC computation and developed a proposed SG overhead transmission line from its conductor characteristics and tower geometry considering the RLC parameters of the conductor while applying RPLS to generate the parameter metrics.

Keywords—RPLS; Smart Grid; Overhead; Conductor; RLC parameters.

I. INTRODUCTION

Generically, an electric grid consist of three main subsystems viz: The generation sources, delivery systems (transmission and distribution networks) and the end consumers. In Nigerian context, the ageing transmission power grids, operational challenges and high unreliability index characterises the grid network. This paper observes that the deregulation of the power sector will unleach unprecedented power (energy) trading across the regional power grids as such presenting power flow scenarios and complexities in vendor interfaces which the system may not be able to handle. Essentially, the transmission line transmits electrical power from one end of the line, sending end, to another, receiving end. A common method of analyzing this behavior is through parameterization and modeling of the transmission lines with passive components [1]. In a transmission model, the interconnection must incorporate distributed self and mutual inductance to accurately estimate time delay and crosstalk metrics in a multilevel network for large scale integration. Besides, the towers and conductors of a transmission line are familiar elements in any given scenario. However, on closer inspection, each transmission line has unique characteristics that have correspondingly unique implications for the environment.

The scope of this paper is majorly on the computational intelligence of RPLS based on fuzzy logic framework. Our proposed RPLS framework optimizes real time computation of required transmission line parameters while computing the vector metrics just in time. This paper defines RPLS computational intelligence in this context to be the maximum number of simultaneous, bidirectional transmission line parameter inputs into the fuzzy line inference system that can be supported in the SG architecture.

The paper is organized as follows: In section II, the related works covering the overview of transmission line specifications, the Smart grid architecture, etc was presented. The general system model and assumptions for RPLS for SG transmission line was presented in III. In IV, The RPLS fuzzy logic framework mechanism is presented alongside with the analytical models. Section V gives the the simulation results to support our propositions. The paper ends with the conclusions and future directions.

All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout the proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example.

II. RELATED WORKS

A. Methods of Transmission Line Computational Analysis

A representative sample of works on the methods of computing transmission line parameters have been studied in literature. The use of distributed-parameter state variable approach, State-space modelling of transposed lines using modal decomposition [2], lumped parameter approximation of line losses [2], distributed transmission line parameters [3], state equations [4], Compact Distributed RLC Interconnect Models [5] to compute line parameters as well as transients on transmission lines have been studied. Some observed limitations of these schemes include:

1) They involve complex mathematical models which could be lacking in precision and accuracy.

2) *Computational requirements for transmission line parameters is quiet large for scaled systems*

3) *These methods do not use any knowledge of the interior structure of the transmission model and in most cases allow only limited control of the closed-loop behaviour when feedback control is used in the parameter modelling. This work then proposed an intelligent approach to solving transmission line parameters. In the next sections, we shall introduce the line specification, smart grid architecture and the factors that limit maximum power transfer in a SG transmission line.*

B. Transmission Line Specifications

The work in [6], listed the design specifications (line characteristics) that are commonly required to define a transmission line. Many of these specifications could have their implications on the environmental as a whole. For the purpose of this work, a range of values is considered for these specifications, with the exception that a varying nominal voltage above 25 kV is assumed. The most basic descriptive specifications usually will include a line name or other identifier, nominal voltage, length of line, altitude range, and the design load district. They details as follows viz [6]:

1) Tower Specifications

The towers support the conductors and provide physical and electrical isolation for energized lines. The minimum set of specifications for towers are the material of construction, type or geometry, span between towers, weight, number of circuits, and circuit configuration. At 500 kV, the material of construction is generally steel, though aluminium and hybrid construction, which uses both steel and aluminium, could be used. The type of tower refers to basic tower geometry. The options are lattice, pole (or monopole), H-frame, guyed-V, or guyed-Y. The span is commonly expressed in the average number of towers per mile. This value ranges from four to six towers per mile. Also, the weight of the tower varies substantially with height, duty (straight run or corner, river crossing, etc.), material, number of circuits, and geometry.

2) Transmission Tower Minimum Clearances

The basic function of the tower is to isolate conductors from their surroundings, including other conductors and the tower structure in any deployment. In every design, clearances are specified for phase-to-tower, phase-to-ground, and phase-to-phase. For various line parameters, the Phase-to-tower clearances vary in distance ranges depending on specifications. These distances are maintained by insulator strings and must take into account possible swaying of the conductors. The typical phase-to-ground clearance is 30 to 40 feet. This clearance is maintained by setting the tower height, controlling the line temperature to limit sag, and controlling vegetation and structures in the ROW. Typical phase-to-phase separation is also 30 to 40 feet and is controlled by tower geometry and line motion suppression.

3) Transmission Line Insulators

Basically, insulator design varies according to tower function types. For suspension towers (line of conductors is straight), the insulator assembly is called a suspension string. For deviation towers (the conductors change direction), the

insulator assembly is called a strain string. For 500-kV lines, the insulator strings are built up from individual porcelain disks typically 5.75 inches thick and 10 inches in diameter. The full string is composed of 18 to 28 disks, providing a long path for stray currents to negotiate to reach ground. At this voltage, two to four insulator strings are commonly used at each conductor connection point, often in a V pattern to limit lateral sway.

4) Transmission Lightning Protection

Since the towers are tall with well-grounded metallic structures, lightning usually targets its structure creating unsolicited risk for end user electrical facilities/equipments. To control the effects of lightning, an extra set of wires is generally strung along the extreme top points of the towers. These wires are attached directly to the towers (no insulation), providing a path for the lightning directly to and through the towers to the ground straps at the base of the towers. The extra wires are called shield wires and are either steel or aluminium-clad steel with a diameter of approximately ½ inch.

5) Transmission Line Conductor Motion Suppression

Wind-induced conductor motion, aeolian vibration, can damage the conductors. A variety of devices have been employed to dampen these oscillatory motions. Dampers can prevent the formation of standing waves by absorbing vibration energy. Typically, a single damper is located in each span for each conductor.

Some of the transmission line components include:

1) *The Transmission towers which are the most visible component of the bulk power transmission system. Their function is to keep the high-voltage conductors separated from their surroundings and from each other. Higher voltage lines require greater separation. Some of the environmental implications of a transmission line result directly from these transmission tower design requirements. First, the physical dimensions of the towers, the resulting line arrangements and line spacing establish the necessary minimum dimensions, including clearances to natural and man-made structures. To create and maintain these clearances, it is often necessary to remove or trim vegetation during construction and operation. In addition, excavation, concrete pouring, and pile driving are required to establish foundations. All of these tasks require access roads and service facilities with dimensions and strength sufficient to handle large, heavy tower components, earthmoving equipment, and maintenance equipment.*

Figure 1a shows a lattice-type tower with a single-circuit 765-kV line with twelve conductors strung from insulators suspended on the crossbar as a single-circuit line. Figure 1b shows a shared corridor for a typical transmission line.

2) Conductors

In this regard, various materials like copper, aluminium material which has higher strength to weight ratio is preferable for deployments. Typical aluminium conductors are composed of multiple 1/8-inch-thick strands twisted together. There are about 50 varieties of multi-strand conductor cables, but a variety of conductor compositions and constructions are currently in use to meet a variety of specific requirements.

In designing and deploying a transmission infrastructure, the tower and its conductor geometry must be given utmost consideration in the capacity planning phase. Given the challenges of the traditional grid network for power provisioning, this thesis proposes Smart Grid power model for dynamic power stability in Nigeria. The two fundamental questions presented in this work (Q_n) are viz:

- 1) What is Smart Grid-SG?
- 2) How does the RLC parameter modelling leverage on SG to provision dynamic power in Nigeria context?

Firstly, the work proposes the modernization of the current electric power grid by leveraging on smart grid technology and consequently models the RLC parameters of a SG overhead transmission line from its conductor characteristics and tower geometry. The next section starts with the concept of Smart Grid.



Fig.1. a. Lattice (left) and Monopole (right) Towers (Source: Argonne Staff Photo)



Fig.1. b: Multiple lines in a power Corridor (Source: Argonne Staff Photo)

C. Smart Grid Architecture

The concept of SG is relatively new vis-à-vis transmission line designs. Explicit definitions on SG has been given in [7],[8],[9],[10],[11],[12], and [13]. In context, this paper defines SG as a highly scalable, robust and intelligent power network that provides reliable, high quality electric power in an environmentally friendly and sustainable way. According to

[14], the scope of SG extends all the interconnected electric power systems, from centralized bulk generation to distribution, from high voltage transmission systems to low voltage distribution systems, from utility control centers to end user home-area networks, from bulk power market to demand response service provider, and from traditional energy resources to distributed and renewable generation and storage. Table 1, shows the differences between the current power grid network and the SG network. The work in [14] presents five generic key technology areas (KTAs) as shown in figure 2. They includes:

- 1) *Integrated Communications which supports broadband networks, secure, low-latency channels connecting transmission stations and control centers.*
- 2) *Sensing and Measurements which handles phasor measurements and data streaming over high speed channels*
- 3) *Advance Components which includes the flexible AC transmission system (FACTS) devices, eg. Unified power flow controllers (UPFC), Static Var compensations (SVC), static synchronous compensators and High Voltage DC (HVDC).*
- 4) *Advance Control and protection methods which includes differential line relaying, adaptive settings, and system integrity systems that supports low latency communications.*
- 5) *Enhanced interfaces and decision supports which utilizes instantaneous measurements from phasor measurements units (PMUs). These KTAs are modelled into the SG networks at both the distribution and transmission domains [15].*

The report in [16] shows the chronological sequence of Smart Electricity Systems comprising the infrastructure of the past, present and the future

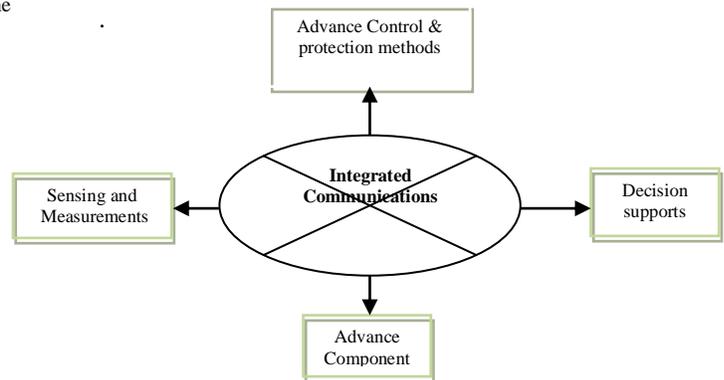


Fig.2. SG Key Technology Areas (KTAs)

Leveraging on Smart grid framework will not only guarantee stability in power provisioning but it will expedite efficiency in power generation, transmission and distribution. Also, Smart grids co-ordinate the needs and capabilities of all generators, grid operators, end-users and electricity market stakeholders to operate all parts of the system as efficiently as possible, minimizing costs and environmental impacts while maximizing system reliability, resilience and stability [16]. Figure 5 shows the generic Smart grid architecture modules with an additional SIM proposed in this work. Table 1

presents the differences between current power grid network and the SG network. The components of figure 5 are discussed below viz:

1) *Advanced Grid Components (AGC)* allow for a more efficient energy supply, better reliability and availability of power. Components includes viz: advanced conductors and superconductors, improved electric storage components, new materials, advanced power electronics as well as distributed energy generation.

2) *Advanced Control Systems (ACS)* monitors and control essential elements of the smart grid. It supports computer-based algorithms which allow efficient data collection and analysis, provide solutions to human operators while acting autonomously. With ACS, errors and faults can be detected much faster than in traditional grids and outage times can be reduced.

3) *Smart Devices and Metering (SDM)* include ipads, android tablets, wireless sensor networks (WSN) used at transformers, substations and meters at client residents. These facilitates remote monitoring as well as enabling demand-side management allowing for real-time determination and information storage of energy consumption and provide the possibility to read consumption both locally and remotely. The meters detect fluctuations and power outages, permit remote limitations on consumption by customers and allows the meters to be powered down. This work will present the AMS model that fits into the SG framework in context.

4) *Integrated Communication Technologies (ICT)* carries information provided by SDM to be transmitted via a communication backbone. This backbone is characterized by a high-speed and two-way flow of information. The different communication technologies form the communication backbone are LAN, WAN, Core Networking, Security, Power system operations, Network management.

5) *Decision Support and Human Interfaces (DSHI)*. This module will make data available to grid operators and managers in a user-friendly manner to support their decisions via software. It includes systems based on artificial intelligence and semi-autonomous agent software, visualisation technologies, alerting tools, advanced control and performance review applications as well as data simulation applications and geospatial information systems (GIS). The GIS provides geographic, spatial and location information and tailor this information to the specific requirements for decision support systems along the smart grid.

Smart Integrated Module (SIM) which is proposed to handle energy storage, and inverter technologies in the context of distributed generation. Figure 6 shows the proposed SG Power system architecture, however, some of the functionalities of the SIMs in figure 6 includes some functionalities in [14]: Static connection to the feeder, AC bus for AC loads, DC bus for DC loads and connections to energy storage and distribution generation, Voltage regulation in the steady state and in the transient mode, Fast real and reactive power compensation, Fault detection and fault current

limiting and isolation, Autonomous distributed intelligent control for shunt-time scale control, Coordinating an optimization for longer-time-scale control.

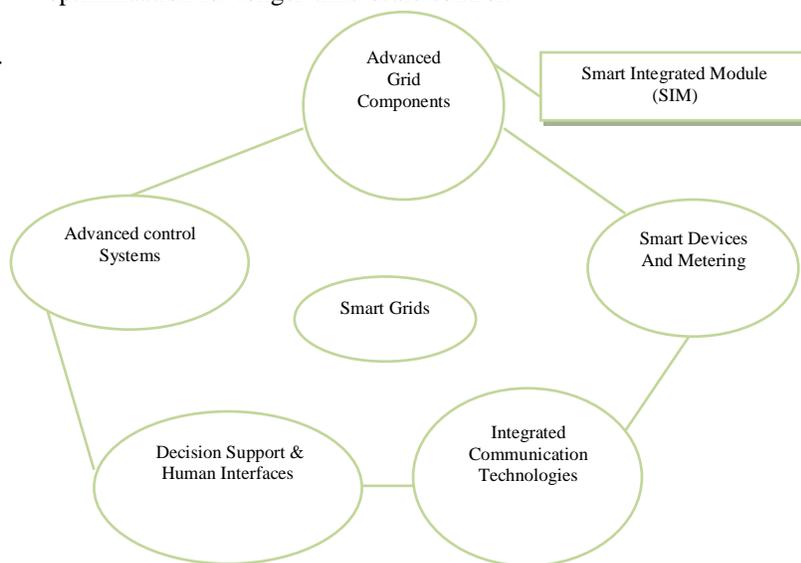


Fig.5. Smart Grid Architecture Modules

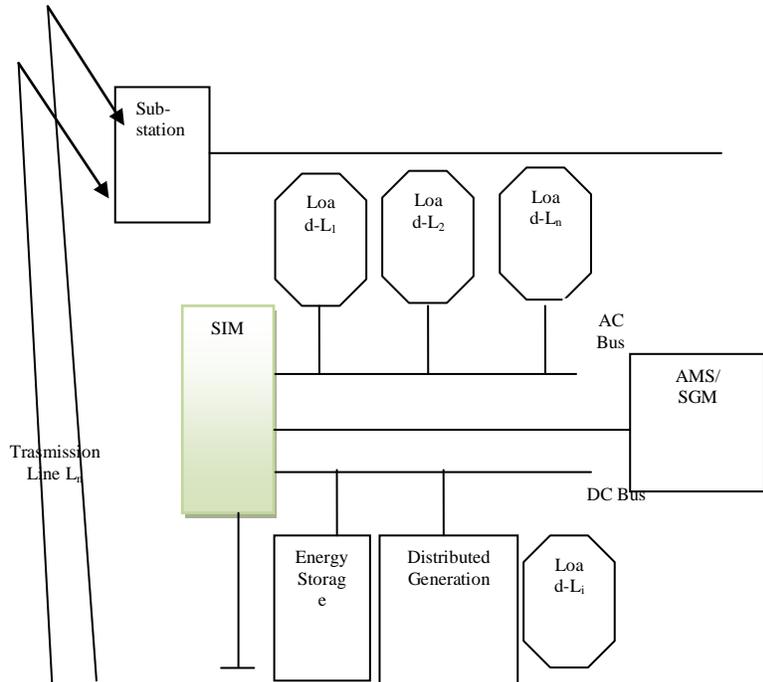


Fig6. A Conceptual Model of the SG Architecture For Dynamic Provisioning

In figure 6, the SIM contribute to energy loss reduction owing to:

- 1) Full Utilization of the distributed generation to reduce the real power flow on the grid
- 2) Provisioning of reactive power where it is consumed to reduce the reactance power flow on the grid. Following the role of power electronics in the voltage regulation functionality of SIMs, high-quality power

3) supply at every load connection point is assured. This is because it will maintain an optimized voltage levels while compensating for voltage drops.

Besides, the energy storage functionality will provide a short to-medium term power supply buffer capacity so that customer's service will not be interrupted in the event of short-term disruption on the distribution or transmission grid. This will create stability and relaxation in the transmission grid model. Future work will show the Advance metering sandbox (AMS) as a Smart grid meter which comprises of current and voltage sensors, partition dividers, signal conditioners, peripheral interface controller (PIC) with an embedded ADC and an LCD display. The AMS will leverage on the work in [17]. The design technique for digital meters is influenced by three major factors namely; desired device cost, efficiency and overall size, [17]. While the cost is influenced by users' general affordability, the efficiency and size must strictly comply with the SG standard. The Advanced metering Sandbox (AMS) will provide a wide range of functionalities viz:

- 1) Remote consumer price signals, which can provide time-of-use pricing information.
- 2) Ability to collect, store and report customer energy consumption data for any required time intervals or near real time.
- 3) Improved energy diagnostics from more detailed load profiles.
- 4) Ability to identify location and extent of outages remotely via a metering function that sends a signal when the meter goes out and when power is restored.
- 5) Remote connection and disconnection using mobile devices.
- 6) Losses and theft detection owing to the advance protection layer and alarm signalling.
- 7) Ability for a retail energy service provider to manage its revenues through more effective cash collection and debt management.

D SG Transmission Line Technical Limits to Power Transfers

1) Conductor resistance, Temperature rating, and line sag. As a transmission line receives power, resistance inherent in the line conductor material converts some of the electrical energy into thermal energy, thereby increasing the line temperature. Line temperature increases as the current flowing through the line increases. Power transfers above a predetermined safe operating transfer limit can cause excessive conductor temperature, which causes line conductors to expand in length. Also, excessive operating temperatures may weaken the conductor, reducing its expected life. For underground conductors, high operating temperatures can damage insulation. Because aboveground transmission lines are suspended on fixed-distance tower structures, an expanding conductor manifests itself as sagging that reduces conductor distance to ground at the midpoint

between towers. Because of line weakness at higher temperatures, this sagging can become permanent.

2) Voltage drop. The voltage drop increases as transmission line length increases. Similarly, the terminating voltage at the receiving end may vary above or below the recommended or nominal operating voltage, depending on the types of loads connected to the receiving end. Voltage constraints define the criteria needed to maintain receiving-end voltages within specified bounds (usually $\pm 5\%$ of the nominal voltage). Customer and utility equipment operates most efficiently when operated near the nominal voltage level.

3) Parallel flows. Because the electric power grid provides an interconnected set of transmission lines, the flows that one might expect to occur over the transmission line that directly connects Area A to Area B actually occur over all of the interconnected lines in varying amounts. It may be true that the direct line may transfer most (perhaps 60%) of the power from Area A to Area B, but lines that are parallel to the direct line will also carry some portion of the power between the areas. Because electric power does not flow between areas in a simple manner that follows the contract path, the presence of parallel flows can cause a violation of thermal constraints on other lines in the system.

4) Synchronization. When two or more generators operate using the same interconnected transmission system, the generators must be synchronized. In the United States, this frequency is very near 60 hertz. Assuring synchronization maximizes power transfers and minimizes utility and customer equipment damage. In addition, synchronization helps to avoid transient instability and small-signal instability.

TABLE I. CURRENT POWER GRID VERSUS SG

Current power grid network	New SG network.
Centralized generation	Generation is Everywhere
Power flows downhill	Power flows from Everywhere
Utility controls connections	Anyone can participate
It has a predictive behaviour	Random behaviour
Not scalable and Intelligent	Very scalable and Intelligent

III. METHODOLOGY

In this section, we present a modelling approach to the proposed SG transmission subsystem. The work computes RLC parameters of overhead transmission line from its conductor characteristics and tower geometry. This will form the baseline for modeling N-phase asymmetrical lines in the context of SG Transmission design and deployments. MATLAB Simulink 2009b [20] tool was used to develop our intelligent fuzzy framework for the RPLS Algorithm General System Model and assumptions for RPLS for SG transmission line.

1) In our approach, four line parameters were considered in context, while generating their corresponding RLC matrix

values. This includes: Case-1:Line_25kV_4wires.ie.(25-kV-three-phase distribution feeder with accessible neutral conductor.), Case-2:Line_315kV_2circuit.ie.(315-kV- three-phase, double-circuit line using bundles of two conductors), Case-3:Line_500kV_2circuit.ie.(500-kV, three-phase, double-circuit line using bundles of three conductors), Case-4:Line_735kV.ie.(735-kV-three-phase, line using bundles of four conductors).

- 2) The framework is based on fuzzy logic controller using mamdani FIS.
- 3) The SIM works concurrently with the Line parameters.
- 4) The Output switching is the defuzzified the RLC_matrix.

IV. RPLS FUZZY LOGIC FRAMEWORK FORMULATION FOR TRANSMISSION LINE.

A. RPLS Modeling For GNIS Fuzzy Logic Controller

In this framework, the considered computational variables/entities comprises of lines (case-1, Case-2, Case-3, Case-4, and Case-n+1), The SIM block and a defuzzified RLC_Matrix output. The priority computational scheme considers only fuzzy valid states for the input line parameters. All valid states are the fuzzy enable and as such solves the input matrix vectors.

B. Fuzzification Fuzzy Optimization

In the RPLS fuzzification process of the FLC design, let the four inputs Case-1, Case-2, Case-3, Case-4 be processed using four separate fuzzification blocks. In this process, there is need for memory elements M_e to store the results of the computations. For a scalable design, the fuzzification block can have $X_1 \dots X_{i+n}$ inputs and outputs with defined fuzzy value define in the inputs universe of discourse. However, in the design, the fuzzification process entails that for any single crisp value of the inputs Case-1, Case-2, Case-3, Case-4, only N adjacent values are significant (with non-zero membership values). We then configured the inputs in the FIS editor for the SG computational engine to generate the required matrix vectors of our line parameters. Figure 6 shows the block diagram of the RPLS fuzzy model. The middle block contains the rules which are formed using different combinations of the inputs supplied. The FIS editor displays the information about the fuzzy inference system.

Also from figure 6, the membership function editor of MATLAB environment was configured to process the inputs. The rule base in the MATLAB rule editor was configure for the inputs and now the defuzzified output is converted to surface diagrams and rule view for the computed RLC matrix vectors. The framework for RPLS using fuzzy logic framework is shown in figure 7.

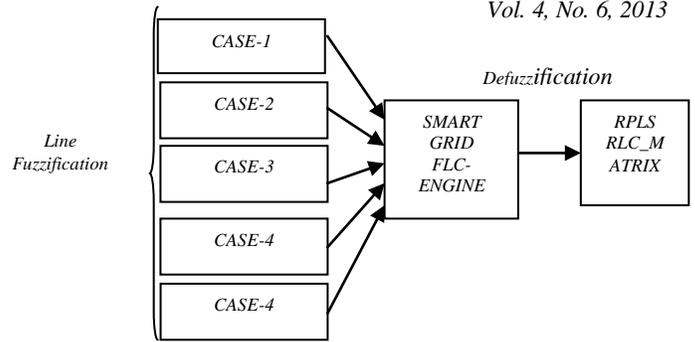


Figure 6: Block diagram of the RPLS Smart Grid fuzzy model

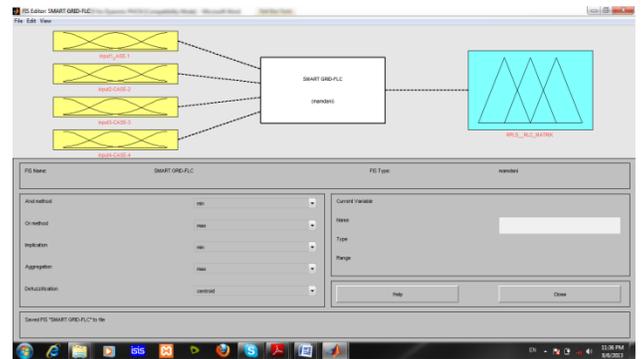


Fig.7. A Framework for RPLS using Mamdani fuzzy logic Structure.

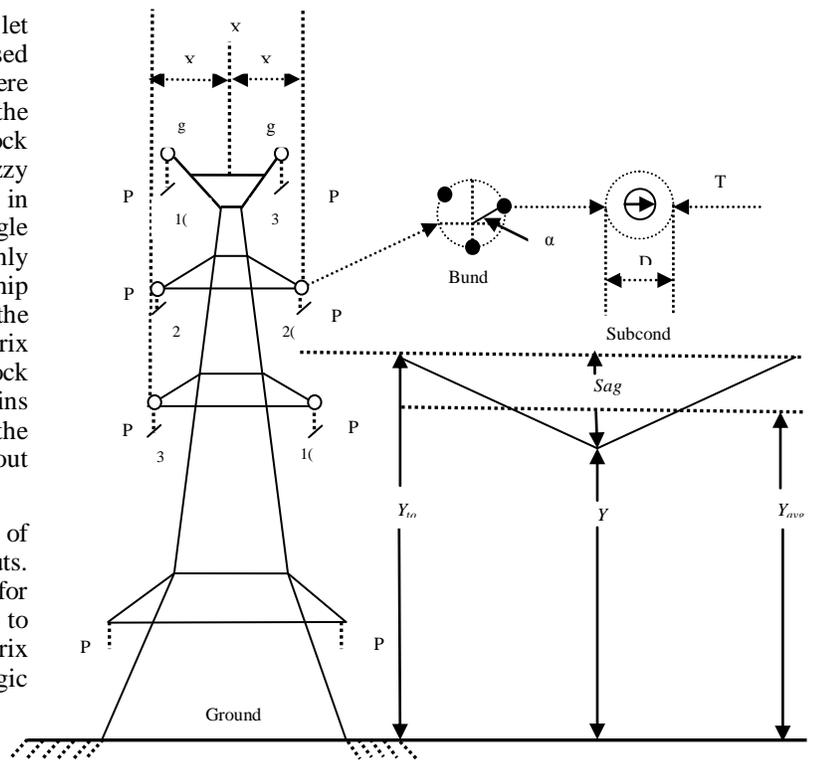


Fig.8. A Conceptual Model of the SG Transmission Line L (3-phase Double Circuit Line)

C. CSG Transmission Subsystem for RLC Parameters.

1) Modelling and Characterization

Following our framework design in figure 7, we then show the model of a SG transmission line with both line and tower geometry. The approach adopted in this paper follows the image conductor principle discussed in [19]. Using a proposed fuzzy RPLS scheme, any line transmission specification can be computed with ease in zero time.

By characterizing the input parameters both line and geometric tower components, the RPLS algorithm computes and displays the corresponding values. From figure 8, a conceptual model depicting only the transmission block with the tower and conductor geometry is presented. Recall that four line parameters were assumed for our matrix computation.

Let Case-1: represent a Line_25kV_4wires.ie.(25-kV-three-phase distribution feeder with accessible neutral conductor.), Case-2: represent a Line_315kV_2circuit.ie.(315-kV- three-phase, double-circuit line using bundles of two conductors), Case-3: represent a Line_500kV_2circuit.ie.(500-kV, three-phase, double-circuit line using bundles of three conductors), and Case-4: represent a Line_735kV.ie.(735-kV-three-phase, line using bundles of four conductors).

Considering each of the cases, we characterized the model to achieve our objectives. Now, let the Horizontal position of the conductor in meters = x_n and for symmetrical line, $x = 0$.

Frequency, Frq = 50hz; Ground resistivity, Rg =100Ω; Number of Conductor Nc = 2; Conductor internal inductance = T/D ratio = 0.5; Conductor skin effect = Enable

Number of Phase Conductor/bundles = 3; Number of ground wires = 2; For conductor bundles, let Phase numbers 1,2,3 = P₁,P₂,P₃; For a 3-Phase tripple circuit lines => Circuit 1= P₁,P₂,P₃; Circuit 2= P₄,P₅,P₆; Circuit 3= P₇,P₈,P₉

Ground wires = g₁,g₂,g₃.....g_n

Hence, P₁,P₂,P₃,P₄,P₅,P₆,P₇,P₈,P₉ = A,B,C,A,B,C,A,B,C

For a 3-Phase Double circuit lines => Circuit 1= P₁,P₂,P₃

Circuit 2= P₄,P₅,P₆

Figure 9 shows image conductor model for computing the resistance R, Inductance L, and Capacitance, C.

For a 3-phase double circuit line, the self and mutual resistance terms is given below:

$$R_{nn} = R_{int} + \Delta R_{nn} \quad \Omega/Km.....(1)$$

$$R_{nm} = \Delta R_{nm} \quad \Omega/Km.....(2)$$

The self and mutual Inductance terms is given below:

$$L_{nn} = L_{int} + \frac{\mu_0}{2\pi} \cdot \text{Log} \frac{2h_n}{r_n} \cdot \Delta L_{nn} \quad H/km.....(3)$$

$$L_{nm} = \frac{\mu_0}{2\pi} \cdot \text{Log} \frac{D_{nm}}{d_{nm}} + \Delta L_{nm} \quad H/km.....(4)$$

The self and mutual Potential coefficients terms is:

$$P_{nn} = \frac{1}{2\pi} \cdot \text{Log} \frac{2h_n}{r_n} \quad Km/F.....(5)$$

$$P_{nm} = \frac{1}{2\pi} \cdot \text{Log} \frac{D_{nm}}{d_{nm}} \quad Km/F.....(6)$$

$$[C] = [P]^{-1}.....(7)$$

Where, μ_0 = Permeability of free space = $4\pi \cdot 10^{-4}$ H/Km,

ϵ_0 = Permittivity of free space = $8.8542 \cdot 10^{-9}$ F/Km,

r_n = Radius of conductor n m in meters, d_{nm} = distance between conductor n in meters, D_{nm} = distance between conductor n and image of m in meters, L_n = Avg height of conductor n above ground in meters, R_{int} , L_{int} = Internal resistance and inductance of conductor, ΔR_{nn} , ΔR_{nm} : Carson R correction terms due to ground resistivity, ΔL_{nn} , ΔL_{nm} : Carson L correction terms due to ground resistivity.

The conductor self inductance is computed from the magnetic flux circulating inside and outside the conductor, and produced by the current flowing in the conductor itself. The part of flux circulating inside the conducting material contributes to internal inductance L_{int} , which is dependant on the conductor geometry. Assuming a hollow or solid conductor, the internal inductance in the model is computed from the T/D ratio where D is the conductor diameter and T is the thickness of the conducting material (As shown in figure 8).

Assuming the vertical position of the conductor (at the tower) wrt ground in meters = y_v , Vertical position of the conductor wrt ground at mid span in meters = y_{min}

Line parameters via Load = 750Kv, 500Kv, 450Kv, 315Kv and 25Kv. The Average Height of the conductor is given by :

$$Y_{avg} = Y_{min} + \text{Sag}/3 = [2Y_{min} + Y_{tower}]/3.....(8)$$

Where, Y_{tower} = Height of conductor at tower, Y_{min} = Height of conductor at mid span, and $\text{Sag} = Y_{tower} - Y_{min}$.

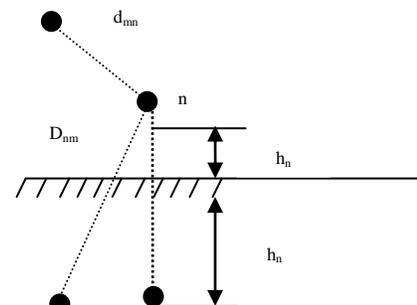


Fig.9. Image conductor model

V. IMPLEMENTATION RESULTS

The MATLAB 7.7.0 R2008b [20] was configured while loading the three case scenriors to compute the RLC line parameters matrix equivalents. First of all, we define the input line geometry and the conductor bundle characteristics (as shown in figure 10a and figure 10b) for the various cases in the MATLAB simulink environment and consequently runing the simulation and exporting the computed values to the command prompt for collection and data analysis. Below is

the RLC matrix results generated from the MATLAB Simulink.

$$R_matrix = \begin{bmatrix} 0.0890 & 0.0790 & 0.0773 \\ 0.0790 & 0.0915 & 0.0790 \\ 0.0773 & 0.0790 & 0.0890 \end{bmatrix}$$

$$L_matrix = \begin{bmatrix} 0.0016 & 0.0008 & 0.0006 \\ 0.0008 & 0.0016 & 0.0008 \\ 0.0006 & 0.0008 & 0.0016 \end{bmatrix}$$

$$C_matrix = 1.0e-007 * \begin{bmatrix} 0.1166 & -0.0213 & -0.0058 \\ -0.0213 & 0.1212 & -0.0213 \\ -0.0058 & -0.0213 & 0.1166 \end{bmatrix}$$

$$R1 = 0.0114 \quad 0.2466$$

$$L1 = 0.0009 \quad 0.0031$$

$$C1 = 1.0e-007 * 0.1343 \quad 0.0859$$

The positive-sequence and zero-sequence parameters of the transposed line are displayed in the Display Results window in the R1 [in Ω/km], L1 [in mH/km], C1[in nF/km] vectors.

VI. CONCLUSION AND FUTURE WORK

This paper have presented SG as a sustainable power model for the Nigeria environment while proposing a conceptual SG architecture that took cognizance of three fundamental research elements viz: the transmission line modelling, the SIM and the advanced metering sandbox. This paper outlined the current issues with the existing methods of solving transmission line parameters. In a SG design, a proposed RPLS offers an efficient method of computing the parametric matrices. Handling complex transmission line specifications can better be addressed with RPLS approach which results in less computational analysis by the system planner.

Furthermore, we argue that Smart Grid presents opportunities for utilities and consumers to benefit from efficient management of energy and advanced technology, equipment and devices under well designed transmission, and distribution infrastructures. Besides, it offers significant opportunities to intelligently manage the available energy options and resources by potentially eradicating monopoly, while integrating renewable and non-renewable generation sources into the electricity grid and enabling consumers to better manage their energy consumption. Its challenges have been outlined in [21]. Moreover, this work considered a method of modelling the transmission line parameters and quickly computing their values for system designers so as allow for futuristic prediction of the transmission grid

requirements. By using power line parameter computation in MATLAB Simulink, it was shown that it is possible to compute the RLC line parameters very conveniently.

In the future, we are going to investigate: i) the use of RPLS algorithm for arbitrary computations, ii) the optimal processing algorithm in order to provide the greatest correlation and accuracy, iii) complete the design and implementation of the AMS for the proposed Smart architecture and finally, various validation analysis will be presented to validate our proposal.

REFERENCES

- [1] Aaron St. Leger, "Transmission Line Modeling for the Purpose of Analog Power Flow Computation of Large Scale Power Systems", M.Sc, July, 2005.
- [2] Mehmet Salih MAM'IS, Asim KAYGUSUZ, Muhammet K'OKSAL, "State variable distributed-parameter representation of transmission line for transient simulations" Turk J Elec Eng & Comp Sci, Vol.18, No.1, 2010, doi:10.3906/elk-0905-2
- [3] A. R. Bergen and V. Vittal, *Power System Analysis*, 2nd ed: Prentice-Hall, 2000.
- [4] Mohazzab JAVED, Hussain AFTAB, Muhammad QASIM, Mohsin SATTAR, "RLC Circuit Response and Analysis (Using State Space Method)", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.4, April 2008.
- [5] Jeffrey A. Davis and James D. Meindl, "Compact Distributed RLC Interconnect Models—Part I: Single Line Transient, Time Delay, and Overshoot Expressions" IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 47, NO. 11, NOVEMBER 2000
- [6] J.C. Molburg, J.A. Kavicky, K.C Picel, " The Design, Construction, and Operation of long Distance High
- [7] Voltage Electricity transmission Technologies " Argonne National Laboratory, Nov, 2007.
- [8] Ravi Kaushal, " Challenges Of Implementing Smart Grids In India",2011
- [9] Whitepaper-Technology Roadmap, Smart Grids, www.iea.org/about/copyright.asp,2011
- [10] Whitepaper- Smart Sensor Networks: Technologies and Applications for Green Growth, December 2009
- [11] Climate Group, The and GeSI (2008), SMART 2020: Enabling the Low Carbon Economy in the Information Age, www.theclimategroup.org/assets/resources/publications/Smart2020Report.pdf.
- [12] Adam, R. and W. Wintersteller (2008), From Distribution to Contribution. Commercialising the Smart Grid, Booz & Company, Munich
- [13] Miller, J. (2008), "The Smart Grid – How Do We Get There?", Smart Grid News, June 26.
- [14] Electric Power Research Institute (EPRI, 2005), IntelliGridSM – Smart Power for the 21st century, www.epr-intelligrid.com/intelligrid/docs/Intelligrid_6_16_05.pdf.
- [15] Enrique. S. et al, Gettin SMART, IEEE power and Energy magazine,march 2010,pp.11-18.
- [16] Stanley I.H et al, The future of poer transmission, IEEE power and Energy magazine,march 2010,pp.4-10
- [17] Whitepaper-Technology Roadmap, Smart Grids, www.iea.org/about/copyright.asp,2011
- [18] M.C. Ndinechi, O.A. Ogungbenro, K.C. Okafor, "Digital metering system: a better alternative for
- [19] Electromechanical energy meter in Nigeria" International Journal Of Academic Research, Vol.3.No. 5. September, 2011, I Part.
- [20] EIA, 2002, *Upgrading Transmission Capacity for Wholesale Electric Power Trade*, Energy Information Administration, U.S. Department of Energy, Washington, D.C.

- [21] Available at http://www.eia.doe.gov/cneaf/pubs_html/feat_trans_capacity/w_sale.html. Accessed February 12, 2007.
- [22] Dommel, H., et al., Electromagnetic Transients Program Reference Manual (EMTP Theory Book), 1986
- [23] Mathworks. 2008. <http://www.mathworks.com>
- [24] Udeze Chidiebele .C, Prof. H.C. Inyama, Okafor Kennedy .C, Dr C.C. Okezie, Smart Grids: A New Framework for Efficient Power Management in Data Center Networks. International Journal of Computer Science and Application (IJACSA) Volume 3, Number 7 (2012), pp 59-66.